

Genome update: sigma factors in 240 bacterial genomes

Genomes of the month

Ten new microbial genomes were published since the last 'Genome Update' column was written. The collection of this month's prokaryotic genomes, listed in Table 1, consists of one archaeon (*Sulfolobus acidocaldarius*) and five bacteria (*Corynebacterium jeikeium*, *Haemophilus influenzae*, *Pseudomonas fluorescens*, *Rickettsia felis* and *Xanthomonas campestris*). In addition, four microbial eukaryotic genomes have been published: *Dictyostelium discoideum* AX4 (Eichinger *et al.*, 2005), *Theileria annulata* (Gardner *et al.*, 2005), *Theileria parvarich* (Pain *et al.*, 2005) and *Toxoplasma gondii* (Khan *et al.*, 2005).

Corynebacterium jeikeium is an opportunistic pathogen and causes systemic infections, particularly in immunocompromised patients/hosts. Broad-spectrum resistance to antimicrobial agents is a common feature of *C. jeikeium* clinical isolates. Tauch *et al.* (2005) have published the genome sequence of the clinical isolate *C. jeikeium* K411. This strain contains a circular chromosome of ~2.5 Mbp and a ~15 kbp bacteriocin-producing plasmid (pKW4). About half the *C. jeikeium* genes (~52%) constitute a 'chromosomal backbone' of conserved genes found in all four *Corynebacteria* species sequenced to date (*C. glutamicum*, *C. efficiens*, *C. diphtheriae* and *C. jeikeium*).

Haemophilus influenzae strain Rd was the first bacterium to have its genome completely sequenced (Fleischmann *et al.*, 1995) and was also the first bacterial genome to be patented (O'Malley *et al.*, 2005), lending to this genome sequence a certain historical significance. In the decade since 1995, about 250 bacterial genomes have been sequenced and for many bacterial species multiple genome sequences have become available (for example, note that for all six of the genomes listed in Table 1, at

least one other genome has been sequenced from the same genus). A second *H. influenzae* isolate has now been sequenced (Harrison *et al.*, 2005). It is not generally appreciated that the originally sequenced *H. influenzae* strain was a rough form of a serotype not normally associated with disease. This was not stated as such in the publication (Fleischmann *et al.*, 1995) and in this light it is an improvement that the sequence of a pathogenic, non-typeable serotype of *H. influenzae* has now been completed.

The gene content of the newly sequenced *H. influenzae* strain 86-028NP was compared with the *H. influenzae* rough serotype d strain KW20 (Rd). In total, 280 ORFs were identified in strain 86-028NP that are absent in the previously sequenced Rd strain, and 169 of the genes found in the Rd strain were missing in strain 86-028NP. However, the Rd sequence had been annotated when genome sequences were largely *terra incognita*, and annotation by comparative genomic methods was not possible. Annotation techniques have improved substantially as hundreds of genomes have been sequenced and annotated. Several studies have indicated that bacterial genomes can be overannotated, and in the *H. influenzae* Rd genome there could be as many as 200 genes annotated that might not be real (Skovgaard *et al.*, 2001). In addition to the 'extra' genes, there could well be missing

genes, for example, any small non-coding RNAs that can play important regulatory roles. In our opinion, using the same set of 1709 genes annotated 10 years ago in the original *H. influenzae* GenBank file as 'gospel' is a missed chance to take advantage of the progress that has been made in 10 years of microbial genome annotation.

Of most interest are genes found in the new genome that are related to virulence. These include genes whose products are involved in adherence, of which 5 of 12 are present in strain Rd as well, and two others have contingency repeats (short stretches of simple base repeats, for example GGGGG, which can slip during replication, resulting in addition or deletion of a single base, changing the reading frame of the gene; Bayliss *et al.*, 2001). A total of 52 genes were identified for LPS biosynthesis, only four of which are unique to 86-028NP. At least eight LPS genes have contingency repeats. Iron acquisition and the oxidative stress response are important processes in determining virulence of *H. influenzae*. Strain 86-028NP has 21 genes involved in iron acquisition (of which 20 have homologues in Rd) and the genes thought to be involved in the oxidative stress response are relatively conserved between the two species. While these simple comparisons suggest that the difference in virulence may lie in differences in adherence, this needs to be confirmed with experimental data.

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*.

Charles Dorman, 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 DDBJ.

| Name | Length | A+T (mol%) | No. of genes | tRNAs | rRNAs | $\sigma^{54}/\sigma^{70}/\text{ECF } \sigma$ | Accession no. |
|--|-----------|------------|--------------|-------|-------|--|---------------|
| <i>Corynebacterium jeikeium</i> K411 (Main) | 2 462 499 | 38.6 | 2137 | 50 | 3 | 0/2/7 | CR931997 |
| <i>Corynebacterium jeikeium</i> K411 (pKW4) | 14 323 | 46.2 | 16 | 0 | 0 | 0/0/0 | AF401314 |
| <i>Haemophilus influenzae</i> 86-028NP (Main) | 1 913 428 | 61.8 | 1792 | 58 | 6 | 0/2/2 | CP000057 |
| <i>Pseudomonas fluorescens</i> Pf-5 (Main) | 7 074 893 | 36.7 | 6137 | 71 | 5 | 1/4/28 | CP000076 |
| <i>Rickettsia felis</i> URRWXCal2 (Main) | 1 485 148 | 67.5 | 1400 | 33 | 1 | 0/2/0 | CP000053 |
| <i>Rickettsia felis</i> URRWXCal2 (pRF) | 62 829 | 66.4 | 68 | 0 | 0 | 0/0/0 | CP000054 |
| <i>Rickettsia felis</i> URRWXCal2 (pRFdelta) | 39 263 | 66.8 | 44 | 0 | 0 | 0/0/0 | CP000055 |
| <i>Sulfolobus acidocaldarius</i> DSM639 (Main) | 2 225 959 | 63.2 | 2223 | 48 | 1 | – | CP000077 |
| <i>Xanthomonas campestris</i> 8004 (Main) | 5 148 708 | 35.0 | 4273 | 55 | 2 | 2/3/10 | CP000050 |

Pseudomonas fluorescens is a commensal plant bacterium that can produce antimicrobial compounds to suppress plant pathogens and may even function as a growth promoter for plants. As a biofilm producer it is implicated in the fouling of dairy products in particular. *P. fluorescens* Pf-5 is the fourth publicly available genome of the pseudomonads; it is the largest of the four pseudomonads sequenced so far and it is composed of a 7 Mb circular chromosome, with 6144 annotated ORFs, 63 % of which have been assigned a function and 330 of which have no significant similarity to known proteins (Paulsen *et al.*, 2005). Other *Pseudomonas* genomes in public databases are from *P. aeruginosa* PAO1, *P. putida* KT2440 and *P. syringae* DC3000. Limited gene synteny exists between the genome of *P. fluorescens* Pf-5 and the other pseudomonads that have been sequenced. The authors suggest the existence of a core of over 4000 genes that are conserved between the four species. Features related to secondary metabolism have been localized to nine gene clusters. One such cluster encodes hydrogen cyanide production and is also found in *P. aeruginosa*. The chromosome contains one phage- and seven prophage-related genomic islands, constituting roughly 268 kbp in total (Paulsen *et al.*, 2005). *P. fluorescens* Pf5 is the largest genome shown in Table 1, and it also has by far the largest number of sigma factors – a total of 33 (1 σ^{54} , 4 σ^{70} and 28 ECF σ – see below for more details).

Rickettsia felis (Ogata *et al.*, 2005) is the largest of the seven currently finished *Rickettsia* genomes. Described in 1990 as a *Rickettsia*-like micro-organism, this

flea-borne bacterium was proposed in 1996 to be a distinct species, *Rickettsia felis*. Originally characterized as a typhus-like *Rickettsia*, phylogenetic analysis has reclassified *R. felis* into the spotted fever group of *Rickettsiae*. After its discovery in 1990 in fleas in the Americas, the first European human case of an *R. felis* infection was reported in August 2000 (Marquez *et al.*, 2002).

As *R. felis* is an obligate intracellular parasite, phenotypic characterization was difficult before post-genomic studies. There are several nice examples where predictions made from the *R. felis* genome sequence could be verified experimentally (Ogata *et al.*, 2005). For example, ORFs with close sequence similarity to bacterial pili-associated genes led to the electron microscopy discovery of both conjugation- and attachment-associated forms of pili. Perhaps the most important finding in *R. felis* is the presence of plasmids carrying conjugation-associated genes. As the gene content of the plasmids can be changed, this could provide a new tool for studying the more pathogenic species of *Rickettsia*.

The model *Crenarchaeota* organism, *Sulfolobus acidocaldarius* strain DSM639, was the first hyperthermoacidophile to be characterized from terrestrial solfataras and is the third *Sulfolobus* genome to be sequenced (Chen *et al.*, 2005). It has often been used in *in vivo* genetic studies of *Archaea* because of its ease of transformation and sensitivity to many ribosomal antibiotics. It grows optimally at 80 °C and pH 2 under aerobic conditions. The genome is A + T rich (63 mol%), 2.2 Mbp in length and carries 2292 predicted protein-encoding genes, of which

more than 50 % are specific to *Sulfolobus*. As is the case for *S. solfataricus*, there is evidence for three replication origins in *S. acidocaldarius*. Moreover, many single genes as well as the first genes in operons lack a well-defined Shine–Dalgarno motif. *S. acidocaldarius* differs from other known *Sulfolobus* species in that it accepts a more limited range of carbon sources for nutrition and contains genes for thermopsin, a UV damage excision repair system, an aromatic ring dioxygenase and a characteristic restriction modification system. It is worth noting that the integrated conjugative plasmid is likely to be involved in intercellular genetic exchange, and yet *S. acidocaldarius* appears to have a very stable genome organization.

Black rot is a disease of crucifers (cabbage family) caused by *Xanthomonas campestris* pv. *campestris* (Xcc), a species of Gram-negative, aerobic bacteria. Black rot symptoms initially appear as V-shaped areas along the outer leaf edges; in the latter stages of the disease leaf veins turn black, plants become stunted, wilt and usually die. The genome of *X. campestris* strain 8004 was sequenced, annotated (Qian *et al.*, 2005) and compared with that of *X. campestris* strain ATCC 33913 (da Silva *et al.*, 2002). There is a large degree of sequence conservation between the two genomes, both at the amino acid level and at the nucleic acid level. Analyses of the genome sequence have identified already known virulence factors, several additional metabolic pathways and regulatory systems, such as those for fatty acid degradation, type IV secretion and cell signalling. The *X. campestris* strain 8004 genome contains two σ^{54} genes, which normally exist in a

single copy in most proteobacterial genomes. This will be discussed in more detail below.

Method of the month – sigma factors in bacterial genomes

Sigma factors allow sequence-specific binding of RNA polymerase to bacterial promoters. We have constructed three profile hidden Markov models (HMMs), to identify the genes for σ^{54} , σ^{70} and ECF σ , based on experimentally verified sequences from the UNIPROT database (K. Kill, N. T. Hansen, L. J. Jensen, M. Skovgaard & D. W. Ussery, unpublished). Using these we have looked at the distribution of the three classes of sigma factors throughout the 240 bacterial genomes currently in our database. As a general rule, it appears that the number and diversity of sigma factor genes per genome relates to the environmental variation allowing growth for a given species.

Notice that in the first panel of Fig. 1, most genomes have either one or no σ^{54} genes, although a few of the *Proteobacteria* genomes can have two. So far, all sequenced *Xanthomonas* species have two σ^{54} genes, as does *Bordetella* (with the exception of *Bordetella pertussis*). The future will tell whether the apparent trend in the presence of σ^{54} genes according to phylum – either it is there in a single copy or it is absent – will continue to hold. Note that Fig. 1 illustrates the total number of chromosomes, not genomes. Thus, for example, in the

Spirochetes, the chromosomes with no σ^{54} are actually secondary chromosomes, which means that all spirochaete genomes do indeed have one σ^{54} .

The σ^{70} group is shown in the middle of Fig. 1; this group includes the household σ factor, as well as σ factors involved in the control of stationary phase, sporulation, flagella assembly and heat-shock response. There is a larger divergence in the numbers here than for σ^{54} , since each organism might only employ some of these mechanisms. For example, in the *Actinobacteria*, we see *Streptomyces avermitilis* and *Streptomyces coelicolor* with 13 and 14 σ^{70} genes, respectively, while the rest of the *Actinobacteria* have far less; but it is worth noting that the two *Streptomyces* genomes are about 9 Mbp, approximately three times the mean size of the other actinobacterial genomes. In the *Bacteroidetes/Chlorobi* group, we see that *Chlorobium tepidum* surprisingly lacks even an apparent household sigma factor. This is due to a frameshift in the gene, which according to the GenBank file is not a sequencing error. In the *Cyanobacteria* we find a quite high number of σ^{70} genes, which in this case control functions that are very different to those in the other sequenced bacterial genomes. The household sigma factor in *Deinococcus radiodurans* R1 is so divergent from anything else, that it is not picked up by our HMM. The *Firmicutes* are interesting, because they contain a quite large range in the number of σ^{70} genes. What is not seen in

Fig. 1 is that the *Bacillus* and *Clostridium* species have many σ^{70} genes (6–8), while the remaining *Firmicutes*, such as staphylococci and streptococci have only one or two.

The main difference seems to concern the presence or absence of sporulation factors. When looking at the *Proteobacteria* and the *Spirochetes*, we find only a low number of sigma factors. Although it seems that a number of these lack σ^{70} completely, this is just an artifact of multiple chromosomes – that is, all the *Proteobacteria* and *Spirochetes* have at least one σ^{70} per genome.

ECF σ factors are generally far more numerous than the other two classes, but since they are not essential, they are missing in many organisms. They are especially numerous in the *Bacteroides* and *Streptomyces* species, which have more than 40 ECF σ factors each. *Bacteroides thetaiotamicron* currently holds the record with 48 predicted ECF σ factors. In comparison, *Chlorobium tepidum* and *Porphyromonas gingivalis*, also of the *Bacteroidetes/Chlorobi* group, only have three and six, respectively. In the *Firmicutes* most genera do not have ECF σ factors at all. This seems to correlate with the presence of multiple σ^{70} factors, as it is only the *Bacillus*, *Clostridium* and *Listeria* species, plus *Lactococcus lactis*, that do have ECF σ factors. In the *Proteobacteria* we see many genomes without ECFs, and most have less than five, although some, especially *Xanthomonas* and *Pseudomonas* species, have more. Actually *Pseudomonas fluorescens* PF-5 holds the record amongst

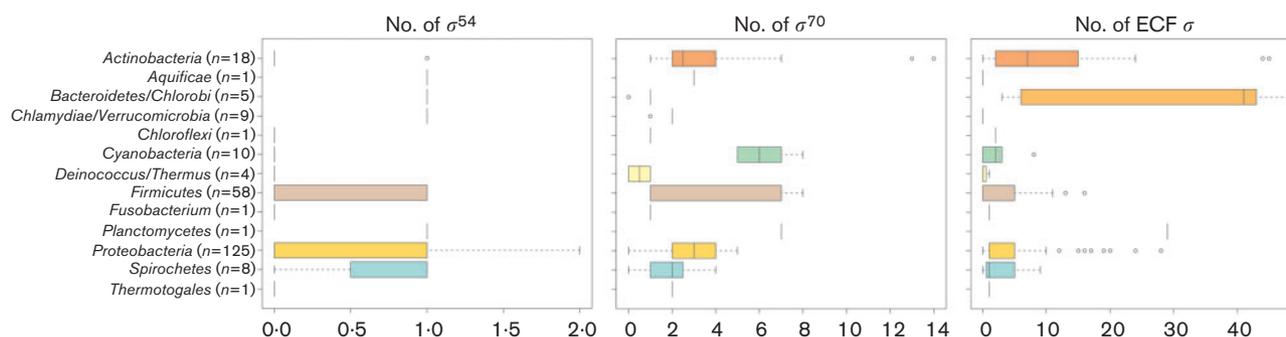


Fig. 1. Box and whisker plot of the number of sigma factor proteins in 13 different bacterial phyla. Note the difference in scales on the bottom axis. The colour scheme for the phyla is the same as found in the GenomeAtlas database (www.cbs.dtu.dk/services/GenomeAtlas/). The box represents the middle 50% of the data. The median of the data is shown by a vertical line. The 25th and 75th quartiles are shown on the left and right side of the median, respectively. The whiskers cannot extend any further than 1.5 times the length of the quartiles. Outlier data points outside the whiskers are shown by open circles. One single vertical line is shown where only one proteome is present.

the *Proteobacteria*, with 28 predicted ECF σ factors.

Supplemental web pages

Access to additional web pages containing supplemental material related to this article can be obtained via the following URL: <http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp018/>

Acknowledgements

This work was supported by a grant from the Danish Center for Scientific Computing.

Kristoffer Kill, Tim T. Binnewies, Thomas Sicheritz-Pontén, Hanni Willenbrock, Peter F. Hallin, Trudy M. Wassenaar and David W. Ussery

Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208, The Technical University of Denmark, DK-2800 Kgs Lyngby, Denmark

†Present address: Molecular Microbiology and Genomics Consultants, Zotzenheim, Germany.

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

Bayliss, C. D., Field, D. & Moxon, E. R. (2001). The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. *J Clin Invest* **107**, 657–662.

Chen, L., Brugger, K., Skovgaard, M. & 8 other authors (2005). The genome of *Sulfolobus acidocaldarius*, a model organism of the *Crenarchaeota*. *J Bacteriol* **187**, 4992–4999.

da Silva, A. C., Ferro, J. A., Reinach, F. C. & 59 other authors (2002). Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* **417**, 459–463.

Eichinger, L., Pachebat, J. A., Glockner, G. & 94 other authors (2005). The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435**, 43–57.

Fleischmann, R. D., Adams, M. D., White, O. & 37 other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Gardner, M. J., Bishop, R., Shah, T. & 41 other authors (2005). Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* **309**, 134–137.

Harrison, A., Dyer, D. W., Gillaspay, A. & 10 other authors (2005). Genomic sequence of an

otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. *J Bacteriol* **187**, 4627–4636.

Khan, A., Taylor, S., Su, C. & 14 other authors (2005). Composite genome map and recombination parameters derived from three archetypal lineages of *Toxoplasma gondii*. *Nucleic Acids Res* **33**, 2980–2992.

Marquez, F. J., Muniain, M. A., Perez, J. M. & Pachon, J. (2002). Presence of *Rickettsia felis* in the cat flea from southwestern Europe. *Emerg Infect Dis* **8**, 89–91.

Ogata, H., Renesto, P., Audic, S., Robert, C., Blanc, G., Fournier, P. E., Parinello, H., Claverie, J. M. & Raoult, D. (2005). The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular parasite. *PLoS Biol* **3**, e248.

O'Malley, M. A., Bostanci, A. & Calvert, J. (2005). Whole-genome patenting. *Nat Rev Genet* **6**, 502–506.

Pain, A., Renaud, H., Berriman, M. & 47 other authors (2005). Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* **309**, 131–133.

Paulsen, I. T., Press, C. M., Ravel, J. & 26 other authors (2005). Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat Biotechnol* **23**, 873–878.

Qian, W., Jia, Y., Ren, S. X. & 25 other authors (2005). Comparative and functional genomic analyses of the pathogenicity of phytopathogen *Xanthomonas campestris* pv. *campestris*. *Genome Res* **15**, 757–767.

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet* **17**, 425–428.

Tauch, A., Kaiser, O., Hain, T. & 13 other authors (2005). Complete genome sequence and analysis of the multiresistant nosocomial pathogen *Corynebacterium jeikeium* K411, a lipid-requiring bacterium of the human skin flora. *J Bacteriol* **187**, 4671–4682.

DOI 10.1099/mic.0.28339-0

Intragenic position of UUA codons in streptomycetes

Streptomycetes have huge linear genomes (>8 Mbp), extreme GC content (around 70 mol%), and their life cycles involve vegetative growth, a phase with formation of aerial hyphae and a sporulation stage. In addition, they produce a range of secondary metabolites, including many of the antibiotics used in clinics today, and for this reason they have immense practical

importance. Production of antibiotics is typically closely linked to differentiation. One of the key genes involved in the differentiation switch is *bldA*, which encodes the tRNA recognizing the very rare UUA (leucine) codons (Lawlor *et al.*, 1987). It has been shown that the production of this tRNA is subject to temporal regulation as it becomes abundant in old cultures (Leskiw *et al.*, 1993). Conversely, genes containing UUA codons are often linked to antibiotic production or other aspects of differentiation (see reviews by Leskiw *et al.*, 1991a; Chater, 1993). As a consequence, heterologous expression of proteins, which often takes place in liquid cultures under vegetative growth, can be problematic if the foreign gene contains UUA codons (Leskiw *et al.*, 1991b; Ueda *et al.*, 1993). Codon usage in the streptomycetes is therefore an interesting phenomenon that deserves full attention.

Wright & Bibb (1992) investigated a limited number of streptomycete genes and concluded that codon usage largely reflected mutational bias. A full genomic analysis based on codon usage in *Streptomyces coelicolor* and *Streptomyces avermitilis* (both are fully sequenced and their genomes are available through GenBank) was recently published in *Microbiology* by Wu *et al.* (2005). The paper used a measure of synonymous codon usage bias called the codon adaptation index (Sharp & Li, 1987) to predict highly expressed genes. Technically, this involves identification of genes using codons that are particularly abundant in highly expressed genes such as ribosomal genes. Thus, UUA codons and other single codons had no particular focus in that study. In fact, UUA has to my knowledge not received particular focus in bioinformatic studies of the streptomycete genomes so far. The aim of this Comment therefore is to apply a genomic view on UUA, thereby supplementing the work by Wu *et al.* and adding to the general knowledge about codon usage in these important bacteria.

I hypothesize that since transcripts for genes containing UUA codons are produced throughout the life cycle, initiation of translation on such mRNAs is futile, and on this account the fitness among different mutants in a population could depend on the loss of energy associated with this futile