

### Genome Update: Protein secretion systems in 225 bacterial genomes

#### Genomes of the month

Seven new microbial genomes have been published since last month's Genome Update column was written. As usual, there is a heavy bias towards members of the *Proteobacteria* (which constitute about half of all the bacterial genomes sequenced so far; 107 out of 225).

The list of new genomes includes five *Proteobacteria*. Two are members of the  $\alpha$ -*Proteobacteria*, the rickettsia *Ehrlichia ruminantium*, and *Gluconobacter oxydans*, which is of interest to the food industry. Three are  $\gamma$ -*Proteobacteria*, the 'warfare germ' *Francisella tularensis*, the plant pathogen *Xanthomonas oryzae* and finally *Vibrio fischeri*. The other genomes are of the probiotic *Lactobacillus acidophilus* and the archaeon *Thermococcus kodakaraensis*. A brief overview of these genomes is given below and a summary is presented in Table 1.

*Ehrlichia ruminantium* is an obligate intracellular bacterium that causes heartwater, a tick-borne disease with high mortality in livestock. Although this infectious disease is currently restricted to Africa, it threatens to invade the Americas. It is hoped that the genome sequence of *Ehrlichia ruminantium* strain Welgevonden will facilitate the development of potential vaccines against heartwater. The genome is 1.5 Mb long, with an A+T content of 73 mol%, and encodes 888 proteins. There are 36 tRNA and 3 rRNA genes (Collins *et al.*, 2005). One remarkable property is the large number of tandemly repeated and duplicated sequences. Thirty-two pseudogenes were also found, most of them are truncated fragments of genes associated with repeats. It seems likely that the identified pseudogenes are products of ongoing sequence duplication events (Collins *et al.*, 2005). The pathogenicity genes are likely to be secreted by a Type IV secretion system (more on this below in Method of the month).

*Francisella tularensis* is a Gram-negative aerobic bacterium with two main serotypes: Jellison Type A and Type B. Type A is the more virulent form. It is one of the most infectious human pathogens and has long been considered as a potential biological weapon. Especially in the 1950s and 1960s, *F. tularensis* was examined by the US military and as recently as the early 1990s the Soviet Union displayed a special interest in tularaemia. The genome of *F. tularensis* strain SCHU S4 is 1.9 Mb long, with a high A+T content of 77 mol% (Larsson *et al.*, 2005). There are 38 predicted tRNAs and 3 rRNAs operons. The genome encodes a total of 1804 predicted genes, including 201 pseudogenes; 1281 of these predicted genes have homologues in one or more  $\gamma$ -proteobacterial genomes. Potential virulence or virulence-associated genes were identified as part of a putative pathogenicity island that is duplicated in the genome. The genome sequence should increase understanding of metabolic pathways in this organism and contribute to development of strategies against this potential biological warfare and bioterrorism agent (Larsson *et al.*, 2005). A type I secretion system could be identified, but complete gene clusters encoding type III, IV or V secretion systems were not found.

*Gluconobacter oxydans* belongs to the family *Acetobacteraceae*. These organisms have been used since ancient times in

biotechnological processes like the production of vinegar and are still used for industrial applications which take advantage of their ability to incompletely oxidize a great variety of carbohydrates, alcohols and related compounds. The genome of *Gluconobacter oxydans* strain 621H (Prust *et al.*, 2005) consists of one main 2.7 Mb circular chromosome and an additional five plasmids (pGOX1–pGOX5, ranging from 163 to 2.6 kb), comprising 232 ORFs. The genome has an A+T content of 39 mol% and contains 2432 ORFs (1877 of these have an assigned function). The unique metabolism of *Gluconobacter oxydans* makes it an ideal model organism to study microbial processing of food.

*Lactobacillus acidophilus* is a representative member of the lactic acid bacteria, which are used in food and feed fermentations, such as dairy and silage. Lactic acid bacteria include probiotic strains, several of which have been sequenced. The 2.0 Mb genome of *Lactobacillus acidophilus* NCFM (Altermann *et al.*, 2005) with 1864 predicted CDSs is relatively small and also has limited biosynthetic capabilities; most amino acids, cofactors and vitamins cannot be synthesized. On the other hand, the genome contains a relatively large fraction of genes involved in taking up various sugars and amino acids/peptides. This is consistent with its relatively nutrient-rich habitat of the gastrointestinal (GI) tract.

**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 DNA DataBase of Japan (DDBJ).

Name	Length (bp)	A+T (mol%)	No. of genes	tRNAs	rRNAs	Accession no.
<i>Ehrlichia ruminantium</i> Welgevonden-1	1 512 977	72.5	958	36	1	CR925678
<i>Francisella tularensis</i> SCHUS4	1 892 819	67.7	1804	38	3	AJ749949
<i>Gluconobacter oxydans</i> 621H	2 702 173	38.9	2432	55	4	CP000009
<i>Lactobacillus acidophilus</i> NCFM	1 993 564	65.3	1864	61	4	CP000033
<i>Thermococcus kodakaraensis</i> KOD1	2 088 737	48.0	2306	46	1	AP006878
<i>Xanthomonas oryzae</i> KACC10331	4 941 439	36.3	4637	54	2	AE013598
<i>Vibrio fischeri</i> ES114 (I)	2 906 179	61.0	2575	108	11	CP000020
<i>Vibrio fischeri</i> ES114 (II)	1 332 022	63.0	1172	11	1	CP000021
<i>Vibrio fischeri</i> pES100	45 849	61.6	55	0	0	CP000022

The complex sugar fructooligosaccharide (FOS), shown to promote growth of probiotic species in the GI tract, can be metabolized by *Lactobacillus acidophilus* NCFM.

As with other sequenced bacteria, a large fraction of the predicted genes have no known function. Subsets of genes, for example those involved in adhesion, may be of particular interest because probiotic species must interact in various ways with the epithelial cells of the GI tract to exert their various beneficial effects. Future research will be needed to identify if there are fundamental differences between adhesion and colonization properties of benign organisms compared to pathogens.

*Thermococcus kodakaraensis* is a hyperthermophilic archaeon which can reduce elemental sulfur during growth and lives in high-temperature ecosystems. *T. kodakaraensis* contains a single, circular chromosome of 2.1 Mb in which 2306 genes have been found, covering 92% of the genome, with a mean length of 833 bp. The genera of *Thermococcus* and *Pyrococcus* are closely related and both belong to the euryarchaeal order *Thermococcales*. A comparison of *T. kodakaraensis* to the *Pyrococcus* genomes of *Pyrococcus horikoshii*, *Pyrococcus furiosus* and *Pyrococcus abyssi* revealed 1204 shared proteins. The A+T content of *T. kodakaraensis* (48 mol%) is lower than the observed ~55–60 mol% for the three *Pyrococcus* genomes (Fukui *et al.*, 2005).

*Vibrio fischeri* is a species of bioluminescent bacteria that exists naturally in a free-living planktonic state, as a symbiont of certain

luminescent fishes or in squid. Genome sequences of several *Vibrio* species that cause human diseases (*Vibrio cholerae*, *Vibrio parahaemolyticus* and *Vibrio vulnificus*) have already been reported. The sequenced *V. fischeri* strain (ES114) is a representative of a non-pathogenic species (Ruby *et al.*, 2005). Its genome is 4.3 Mb and is divided into two chromosomes and a 45.8 kb plasmid (pES100). *V. fischeri* strain ES114 has an A+T content of 62 mol% and contains 3802 predicted genes. Most rRNAs and tRNAs (11 and 108, respectively) were found in chromosome I, although chromosome II encodes a single rRNA operon and 11 tRNA genes. Despite *V. fischeri* ES114 being considered non-pathogenic, by comparing and analysing the sequence, interesting parallels with *Vibrio cholerae* and other pathogens were found.

*Xanthomonas oryzae* is a  $\gamma$ -proteobacterium belonging to the pathovar *oryzae*, responsible for bacterial blight (BB) of rice (Lee *et al.*, 2005). This phytopathogenic strain causes endemic disease in tropical Asian countries. The *X. oryzae* strain KACC10331 genome is 4.94 Mb long with an A+T content of 36 mol% and 4637 predicted genes. Comparisons with two non-pathogenic *Xanthomonas* strains sequenced previously revealed 245 species-specific genes in pathogenic *Xanthomonas oryzae* strain KACC10331. This strain contains twice as many transposable elements as the two other sequenced *Xanthomonas* strains. Although a large number of genes have already been characterized, more work is needed to understand the many aspects of virulence

mechanisms of this important plant pathogen.

### Method of the month – prediction of protein secretion systems in bacterial genomes

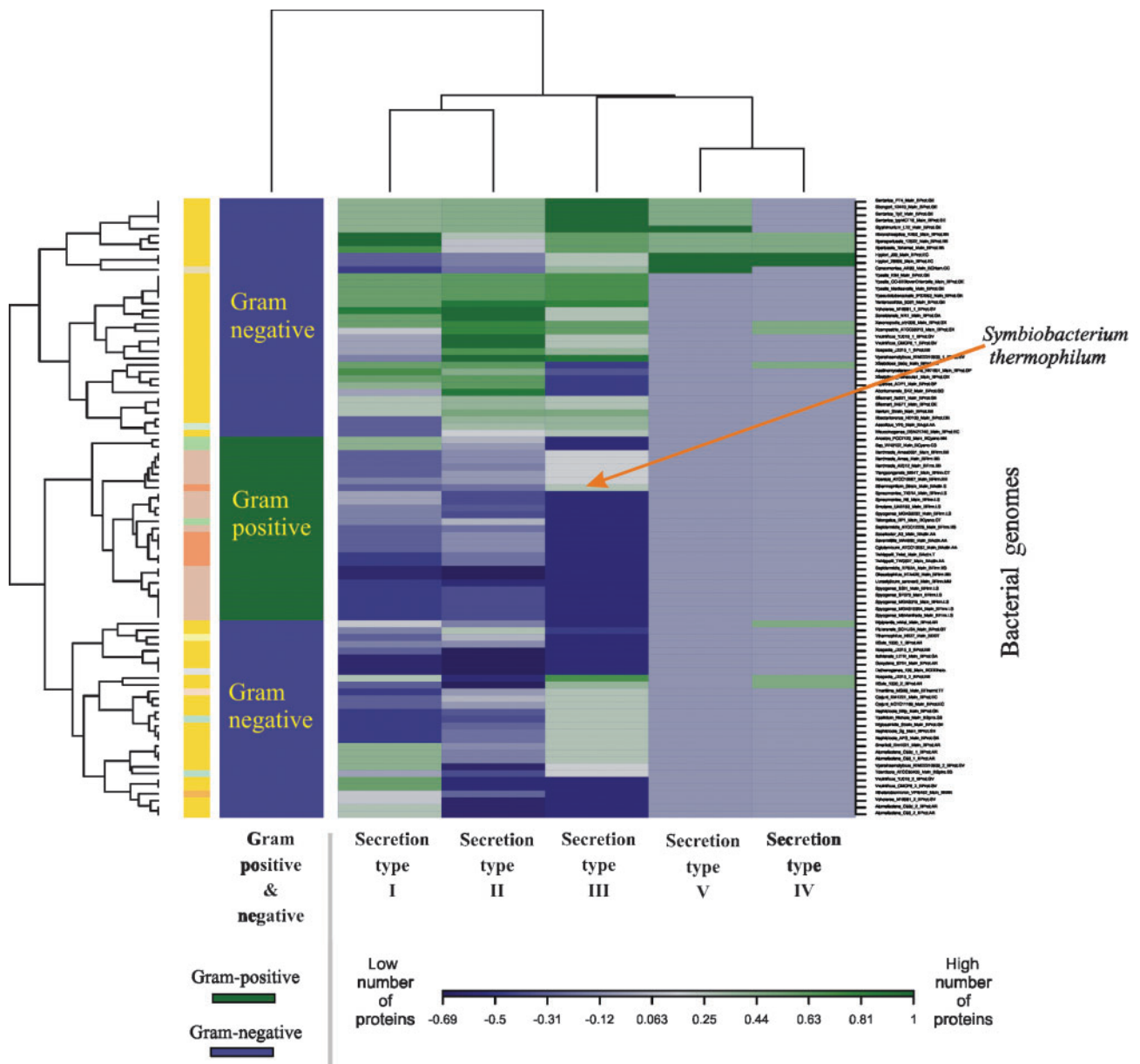
For most bacteria it is essential to secrete particular proteins, and there are several methods available for predicting which proteins will be secreted and how. This genome update deals with characterization of secretion systems in bacterial genomes. Next month, the prediction of secreted proteins will be described, and the following month we will look at prediction of membrane proteins and combining all this to characterize where all the proteins in a given bacterial proteome are likely to be localized – for Gram-negative bacteria, there are five possibilities: the cytosol ('inside'), embedded in the inner membrane, periplasm, outer membrane or secreted ('outside').

This month the focus is on the different bacterial secretion systems (types I–V). A database was constructed which includes information for Gram-positive and Gram-negative bacteria (J. D. Bendtsen, T. T. Binnewies, P. F. Hallin & D. W. Ussery, unpublished). We found most of the proteins for secretion systems type I–V manually by screening the UniProt database, since this database contains only proteins with experimentally verified function. It is important to emphasize that currently we do not cover all available proteins for each individual secretion system. We are still developing and expanding the secretion system database and we will continue work to update it.

A more detailed analysis of the secretome (all proteins involved in secretion) will be given elsewhere (J. D. Bendtsen, T. T. Binnewies, P. F. Hallin & D. W. Ussery, unpublished) and will be only briefly sketched here. After collecting all homologues of the individual components of the five different secretion systems from UniProt, a Hidden Markov model (HMM) was built to specify conserved sequences for

each individual secretion system. The individual HMMs were used to search all available proteomes in our bacterial database. The number of 'hits' for a defined cut-off value for each secretion system was counted and stored in the database. A preliminary version of the secretion type database is available on our supplementary web page (<http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp015/>). Fig. 1

shows a difference in the distribution of the various secretion systems between Gram-positive and Gram-negative bacteria. For the Gram-positive organisms almost no type II or III secretion systems could be identified with the marked exception of *Symbiobacterium thermophilum*, which contains a type III secretion system (TTSS). Interestingly, according to Ueda *et al.* (2004) the phylogeny derived from 16S



**Fig. 1.** Two-dimensional clustering (Willenbrock *et al.*, 2005) of bacterial genome sequences versus secretion systems type I–V. Dark blue indicates that a low number of the selected proteins is present for the specific secretion type; dark green represents cases where we find that most of the proteins for a given secretion system are present. It should be noted that data within each column are normalized around the centre using minimum and maximum values.

rRNA suggests that this bacterium belongs to an unknown taxon in the Gram-positive bacterial cluster. In addition, a TTSS was predicted, perhaps assembled from Fli and FlhA/B proteins associated with flagellum assembly. These data are in agreement with the results from our secretion system database. This illustrates how genome sequences can raise, or answer, questions on taxonomic divisions of bacteria.

### Supplemental web pages

Additional web pages containing supplementary material related to this article can be accessed from <http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp015/>

### Acknowledgements

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

**Tim T. Binnewies,<sup>1</sup>  
Jannick D. Bendtsen,<sup>1</sup> Peter F. Hallin,<sup>1</sup>  
Natasja Nielsen,<sup>1</sup> Trudy M. Wassenaar,<sup>2</sup>  
Martin Bastian Pedersen,<sup>3</sup>  
Per Klemm<sup>4</sup> and David W. Ussery<sup>1</sup>**

<sup>1</sup>Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

<sup>2</sup>Molecular Microbiology and Genomics Consultants, Zotzenheim, Germany

<sup>3</sup>Genomics & Strain Development, Chr. Hansen A/S, Hørsholm, Denmark

<sup>4</sup>Microbial Adhesion Group, Center for Biomedical Microbiology, BioCentrum-DTU, Technical University of Denmark, Denmark

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

**Altermann, E., Russell, W. M., Azcarate-Peril, M. A. & 11 other authors (2005).** Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM. *Proc Natl Acad Sci U S A* (in press). doi:10.1073/pnas.0409188102

**Collins, N. E., Liebenberg, J., De Villiers, E. P. & 19 other authors (2005).** Genome of the heartwater agent *Ehrlichia ruminantium* contains multiple tandem repeats of actively variable copy number. *Proc Natl Acad Sci U S A* **102**, 838–843.

**Fukui, T., Atomi, H., Kanai, T. Matsumi R., Fujiwara, S. & Imanaka, T. (2005).** Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* Kod1 and comparison with *Pyrococcus* genomes. *Genome Res* **15**, 352–363.

**Larsson, P., Oyston, P. C., Chain, P. & 24 other authors (2005).** The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia. *Nat Genet* **37**, 153–159.

**Lee, B. M., Park, Y. J., Park, D. S. & 16 other authors (2005).** The genome sequence of *Xanthomonas oryzae* pathovar oryzae KACC10331, the bacterial blight pathogen of rice. *Nucleic Acids Res* **33**, 577–586.

**Prust, C., Hoffmeister, M., Liesegang, H., Wiezer, A., Fricke, W. F., Ehrenreich, A., Gottschalk, G. & Deppenmeier, U. (2005).** Complete genome sequence of the acetic acid bacterium *Gluconobacter oxydans*. *Nat Biotechnol* **23**, 195–200.

**Ruby, E. G., Urbanowski, M., Campbell, J. & 13 other authors (2005).** Genome sequence of *Vibrio fischeri*: a symbiotic bacterium with pathogenic congeners. *Proc Natl Acad Sci U S A* **102**, 3004–3009.

**Willenbrock, H., Binnewies, T. T., Hallin, P. F. & Ussery, D. W. (2005).** Genome Update: 2D clustering of bacterial genomes. *Microbiology* **151**, 333–336.

**Ueda, K., Yamashita, A., Ishikawa, J., Shimada, M., Watsuji, T. O., Morimura, K., Ikeda, H., Hattori, M. & Beppu, T. (2004).** Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism. *Nucleic Acids Res* **32**, 4937–4944.

DOI 10.1099/mic.0.27966-0