

### Genome update: purine strand bias in 280 bacterial genomes

#### Genomes of the month

Eight new microbial genomes have been published since the last 'Genome Update' column was written. The collection of this month's genomes is shown in Table 1. The published genomes represent four bacterial phyla: one from the *Actinobacteria* (*Mycobacterium avium* subsp. *paratuberculosis*), one from the *Chlamydiae/Verrucomicrobia* (*Chlamydia trachomatis* A/HAR-13), two from the *Firmicutes* (*Staphylococcus saprophyticus* subsp. *saprophyticus* ATCC 15305 and *Streptococcus agalactiae* A909) and four from the *Proteobacteria* ('*Candidatus* Pelagibacter ubique' HTCC1062, *Pseudoalteromonas haloplanktis* TAC125, *Pseudomonas syringae* pv. *phaseolicola* 1448A and *Xanthomonas campestris* pv. *vesicatoria* 85-10).

*Mycobacterium avium* subsp. *paratuberculosis* (MAP) is a pathogenic bacterium thought to be involved in Crohn's disease in humans, although currently the evidence is inconclusive. In animals, MAP causes an intestinal illness known as Johne's disease, a chronic inflammatory disease affecting cattle, deer and sheep, and reported in livestock in many different countries. Li *et al.* (2005) sequenced and investigated the MAP strain K-10 which consists of a single circular chromosome of around 4.8 Mbp. The genome contains ~3000 genes with homologues to *Mycobacterium tuberculosis* and 161 unique genomic regions that encode 39 new *map* genes. Also noteworthy, a possible explanation for mycobactin dependence of MAP has been proposed, based on the genome sequence. (Mycobactin is a siderophore responsible for the binding or transport of iron into cells.) A truncation in a specific domain of a salicyl-AMP ligase, which is the first gene in the mycobactin biosynthesis gene cluster, has been identified.

*Streptococcus agalactiae* is a commensal bacterium colonizing the intestinal tract of a significant proportion of the human population. However, it is the major cause of invasive bacterial disease, including septicaemia, meningitis and pneumonia, in the neonatal period. Tettelin *et al.* (2005) sequenced six strains of *Streptococcus agalactiae* covering the five major disease-causing serotypes and compared these newly sequenced genomes with those of two already available strains. Analysis of these eight genomes shows that *S. agalactiae* contains a pan-genome consisting of a core genome shared by all investigated isolates, accounting for ~80% of any single genome and a dispensable genome composed of partially shared and strain-specific genes. Surprisingly, even after analysing eight genomes, unique genes were still detected and mathematical extrapolation predicts that more should be expected even after sequencing many more strains.

*Chlamydia trachomatis* is the causative agent of the disease chlamydia. Infection with *Chlamydia trachomatis* may result in urethritis, epididymitis, cervicitis, acute salpingitis or other syndromes when sexually transmitted. Isolates exist as 15 serovariants that are separated according to different pathobiotypes. Carlson *et al.* (2005) have sequenced the genome of an oculotropic trachoma isolate (A/HAR-13) and compared it to the already available genome of a genitotropic (D/UW-3) isolate. Analysis showed that the genomes share a remarkable sequence identity of 99.6% and single-nucleotide polymorphism investigations between the two strains confirmed the minimal genetic variation. An important outcome was the identification of a diagnostic marker that allows differentiation between genitotropic and oculotropic strains and between invasive and non-invasive serovars.

*Pseudoalteromonas haloplanktis* strain TAC125 is a psychrophilic gammaproteobacterium found in Antarctica (Médigue *et al.*, 2005). It is adapted to fast growth, suggesting that it

lives in areas rich in nutrients (i.e. plenty of plankton debris). Since *P. haloplanktis* lacks the necessary activities resulting in reactive oxygen species, it could prove useful for expression of foreign proteins in the cold and might be a useful tool in biotechnology. Médigue *et al.* (2005) provide an in-depth analysis of *P. haloplanktis* strain TAC125, reporting that it contains around 3.85 Mbp in two circular chromosomes. There is evidence that the smaller chromosome, chrII, was initially a plasmid that had been recruited to become a chromosome encoding essential genes and that it does not display a standard GC skew. Manual annotation identified 3488 protein coding genes (CDS). A relatively high number of tRNA genes, 106, is consistent with the findings in other *Gammaproteobacteria*.

*Pseudomonas syringae* pv. *phaseolicola* 1448A is a plant pathogen causing halo blight disease in the common bean (*Phaseolus vulgaris*), resulting in severe problems in developing countries. A comparative analysis between this new genome and the four other published *Pseudomonas* genomes identified a set of 3567 (67%) core *Pseudomonas* ORFs (Joardar *et al.*, 2005). Comparison with *P. syringae* pv. *tomato* DC3000 shows that most (81%) of the identified virulence factors of DC3000 are present in 1448A as well (Joardar *et al.*, 2005).

*Staphylococcus saprophyticus* subsp. *saprophyticus* strain ATCC 15305 is known to cause uncomplicated urinary tract infections (UTIs) in women (Kuroda *et al.*, 2005). The circular genome of *S. saprophyticus* is 2 156 575 bp long with an A + T content of 67% and has six rRNA operons and 60 tRNAs for all amino acids. A comparison of *S. saprophyticus* to *Staphylococcus aureus* and *Staphylococcus epidermidis*, the two other *Staphylococcus* species recognized as major human pathogens, revealed the absence of virulence factors in *S. saprophyticus*, such as extracellular matrix-binding proteins which are found in the other two species. This offers a reason as to why *S. saprophyticus*

**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	AT (mol%)	No. of genes	tRNAs	rRNAs	Frequency of YR tracts	Frequency of RR tracts	Accession no.
<i>Chlamydia trachomatis</i> AHAR13 Main	1 044 459	58.7	911	37	2	0.38	4.71	CP000051
<i>Chlamydia trachomatis</i> AHAR13 pCYA	7 510	63.7	8	0	0	0	3.06	CP000052
<i>Mycobacterium avium</i> k10 Main	4 829 781	30.7	4350	46	1	2.45	0.25	AE016958
<i>Pseudoalteromonas haloplanktis</i> TAC125 chr.1	3 214 944	59.8	2941	106	9	1.26	0.61	CR954246
<i>Pseudoalteromonas haloplanktis</i> TAC125 chr.2	635 328	60.6	546	0	0	1.27	0.57	CR954247
<i>Streptococcus agalactiae</i> A909 Main	2 127 839	64.4	1996	80	7	0.62	2.25	CP000114
<i>Staphylococcus saprophyticus</i> subsp. <i>saprophyticus</i> Main	2 516 575	66.8	2446	60	6	1.08	1.42	AP008934
<i>Staphylococcus saprophyticus</i> subsp. <i>saprophyticus</i> pSSP1	38 454	69.3	45	0	0	0.63	2.65	AP008935
<i>Staphylococcus saprophyticus</i> subsp. <i>saprophyticus</i> pSSP2	22 870	68.7	23	0	0	0.99	2.54	AP008936
<i>Pseudomonas syringae</i> pv. <i>phaseolicola</i> 1448A Main	5 928 787	42.0	4982	62	5	1.73	0.49	CP000058
<i>Pseudomonas syringae</i> pv. <i>phaseolicola</i> 1448A pLarge	131 950	45.9	127	0	0	0.94	0.76	CP000059
<i>Pseudomonas syringae</i> pv. <i>phaseolicola</i> 1448A pSmall	51 711	44.0	60	0	0	1.10	1.04	CP000060
' <i>Candidatus</i> Pelagibacter ubique' HTCC1062 Main	1 308 759	70.3	1354	32	1	0.42	2.89	CP000084
<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> 85-10 Main	5 178 466	35.3	4487	55	0	3.63	0.37	AM039952
<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> 85-10 pXCV2	1 852	43.4	2	0	0	1.40	1.24	AM039948
<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> 85-10 pXCV19	19 146	40.2	22	1	0	1.36	0.97	AM039949
<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> 85-10 pXCV38	38 116	39.3	43	0	0	2.24	0.63	AM039950
<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> 85-10 pXCV183	182 572	39.5	172	1	0	1.76	0.58	AM039951

does not cause a bacterial infection quite as severe as those that arise from the other virulent *Staphylococcus* species. Further studies also indicated the presence of a single unique ORF product in *S. saprophyticus* with the ability to mediate cell-wall-anchoring in adhesion to uroepithelial cells. This can be compared to *S. aureus* and *S. epidermidis*, where 20 and 11 such proteins are present, respectively. The complete genome sequence of *S. saprophyticus* confirms this bacterium as an important uropathogenic species in relation to UTIs.

'*Candidatus* Pelagibacter ubique' strain HTCC1062 (originally named SAR11) is a marine bacterium that is possibly the most numerous bacterium in sea water and accounts for about 25 % of all microbial plankton cells. This bacterium plays a key role in the carbon cycle by oxidizing the dissolved organic carbon in the oceans. Giovannoni *et al.* (2005) sequenced the '*Candidatus* P. ubique' genome to discover an extremely compact and efficient genome, containing only 1.3 Mbp, which is considered to be the smallest among free-living organisms. Moreover, there are no pseudogenes, viral genes, transposons or junk DNA of any kind. Even the intergenic spacers are considered the shortest yet observed. Despite the small number of genes encoded (1354), it has complete biosynthetic pathways for all 20 amino acids and all but a few cofactors. The streamlined genome, minimizing the costs of cellular replication, is apparently a great advantage when replicating in hostile environments. '*Candidatus* P. ubique' is a unique member of this phylum since it is free-living and the loss of functional capabilities is minimal.

*Xanthomonas campestris* is a Gram-negative plant-pathogenic bacterium that some use as a model organism in the study of Gram-negative bacteria protein secretion systems. It is phytopathogenic to cruciferous plants and causes worldwide agricultural loss. It also produces an exopolysaccharide called xanthan which can be used in many areas such as spinning and paper making. *X. campestris* pv. *vesicatoria* strain 85-10 is the third *X. campestris* (and the fifth *Xanthomonas*) genome to be published. *X. campestris* pv. *vesicatoria* can cause bacterial spot disease

in pepper and tomato plants, but strain 85-10, described by Thieme *et al.* (2005), is pathogenic only for pepper plants. The genome of *X. campestris* pv. vesicatoria strain 85-10 consists of one circular chromosome of 5 178 466 bp and four plasmids. Like other *Xanthomonas* strains, its chromosome has a low AT content (37%), and this number varies between 43 and 40% among the plasmids. The whole genome contains 4726 CDS, including two rRNA operons (in the order 16S–23S–5S) and 54 tRNA genes. However, 697 out of the 4726 CDS are conserved hypothetical CDS and 949 CDS have unknown function. A comparison to three other *Xanthomonas* genomes revealed 548 CDS (12.2%) unique to *X. campestris* pv. vesicatoria. Controlled by two key regulatory proteins, HrpG (an OmpR family regulator) and HrpX (an AraC-type regulator controlled by HrpG), a type III protein secretion system (TTSS) is responsible for the pathogenicity of *X. campestris* pv. vesicatoria. Additionally, strain 85-10 has all other types of protein secretion systems known in Gram-negative bacteria. The largest plasmid encodes a putative type IV secretion system similar to the Icm/Dot system of human pathogens *Legionella pneumophila* and *Coxiella burnetii*. This is the first report of a putative Icm/Dot like type IV secretion system in a plant-pathogenic bacterium. Also, six novel type III effector proteins and several other virulence factors were predicted through comparisons with other completely sequenced plant pathogens.

### Method of the month – bias of purines and alternating pyrimidine/purine stretches

The highlighted method chosen this time is used for comparison of microbial genomes by looking for stretches of purines (R) and alternating pyrimidine/purines (YR) in sequenced chromosomes or genomes. A length of 10 bp or longer was chosen, as this represents roughly at least one complete turn of the DNA helix. The expected fraction of a chromosome with a sequence of 10 bp or longer of either purines (RR) or alternating YR stretches is about 1.07% (Ussery *et al.*, 2002). We have examined 287 totally sequenced chromosomes from bacterial genomes in 13 phyla. Links to the individual chromosome sequences and references can be found on our 'Genome

Atlas' web page (<http://www.cbs.dtu.dk/services/GenomeAtlas/>).

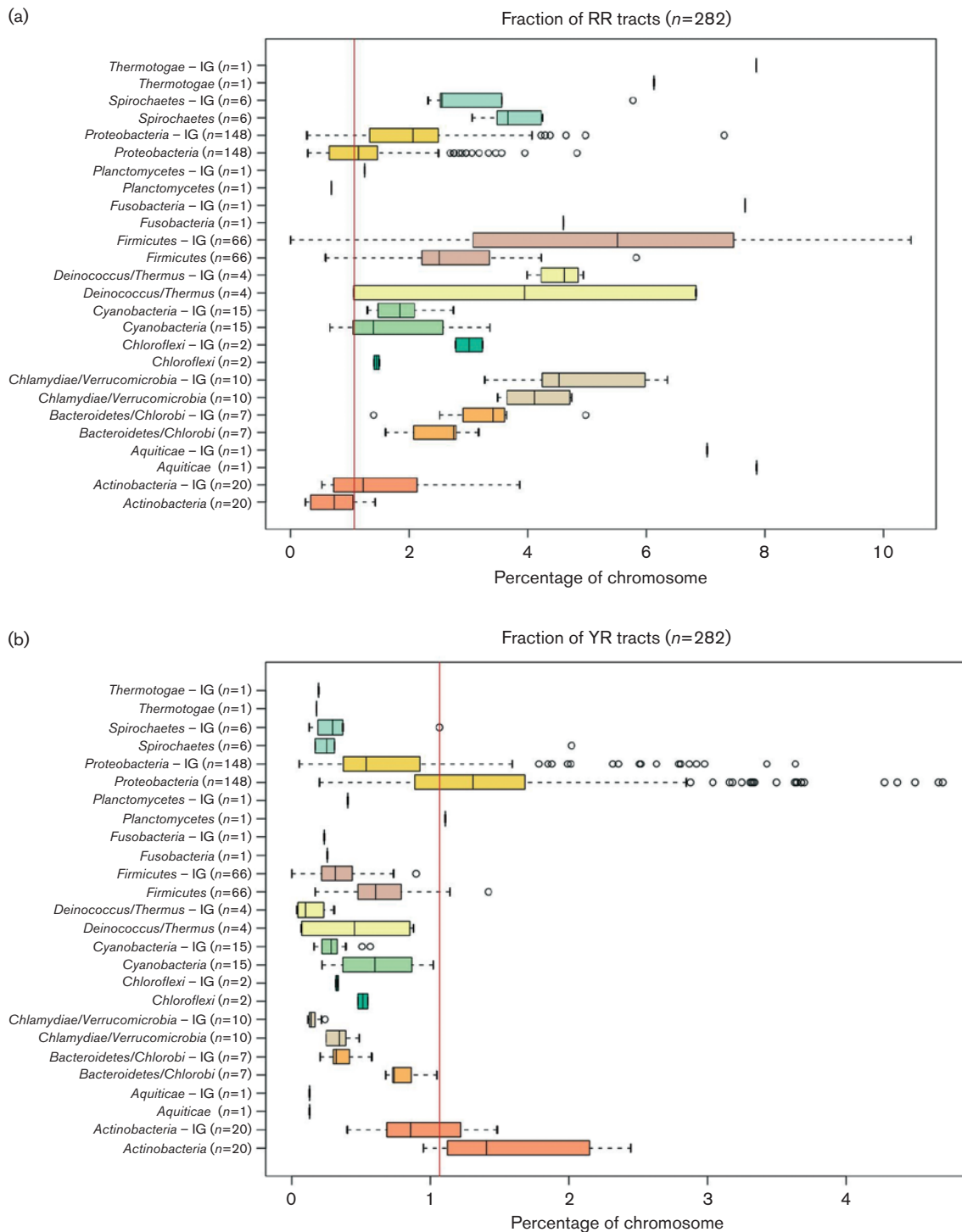
Fig. 1(a) shows that the observed fraction of purine tracts in the various phyla differs significantly. Indeed, each phylum seems to have its own distinct distribution. On average, nearly all phyla have more purine stretches than would be expected. Only the *Proteobacteria* and *Actinobacteria* have significant numbers of chromosomes with lower than expected numbers. Analysis of the proteobacterial and actinobacterial genomes with the largest bias show that many of these are GC-rich. The thermophiles in *Deinococcus/Thermus*, with only three organisms sequenced (or four chromosomes), show the greatest breadth of deviation from the mean. They are also GC-rich. However, *Spirochaetes*, with six organisms sequenced (or eight chromosomes), have a much tighter distribution and are AT-rich. A link to a web-based table showing the fraction for all sequenced bacterial chromosomes is provided in the supplemental web page (<http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp020/>) and can be sorted by the fraction of RR or YR stretches, as well as by the genus name, taxonomic group, length or percentage AT.

In Fig. 1(b) the fraction of YR tracts is plotted for various phyla. Although the expected values are the same (about 1% of the chromosome), here most organisms have fewer than expected, although the distributions are not quite as diverse as for the purine tracts. Again, as in Fig. 1(a), a significant fraction of proteobacterial and actinobacterial genomes show the opposite trend, with a large number of 'outliers'. These tend to be the same organisms which have less purine stretches, and are also often GC-rich.

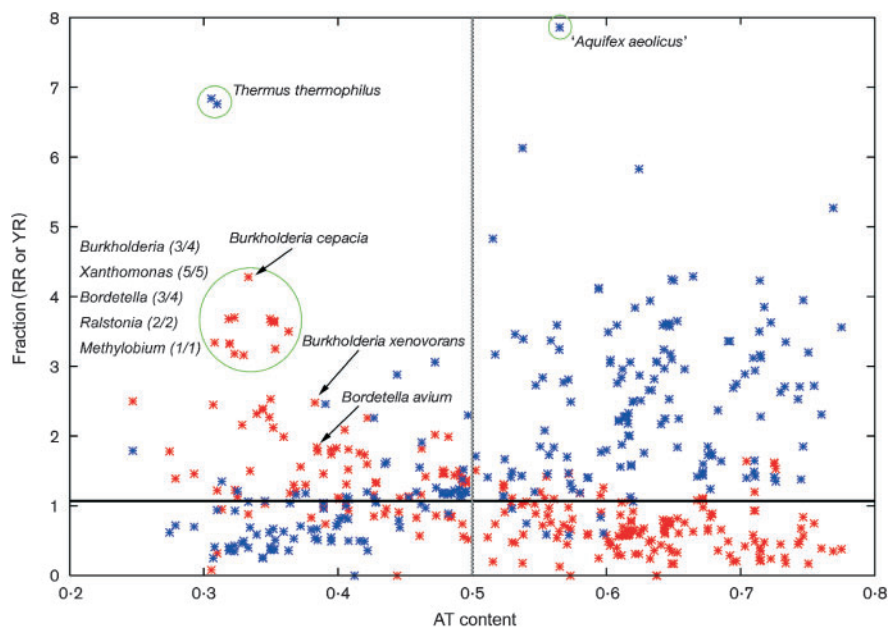
Why do we see such a trend? Although at this stage we cannot say for certain, there are several possible explanations, in terms of DNA structures and stability. In broad terms, the purine tracts tend to be more thermodynamically stable, whilst alternating pyrimidine/purine tracts require less energy to melt (with sequences like TATA melting quite readily, for example). Some purine tracts can stabilize an A-DNA type of structure, which is also the type of helix adopted by RNA-DNA hybrids; it is thought that 'normal' DNA in

cells consists of a mixture of A-DNA and B-DNA conformations. Furthermore, short runs of phased A-tracts can form stable DNA curvature, which could play an important biological role in DNA interaction with proteins. It is interesting to note that most of the genomes with the highest bias towards purine stretches tend to be AT-rich, whilst the genomes with less purine stretches tend to be GC-rich. The opposite trend seems to hold for YR stretches, which can be seen in Fig. 2. Alternating CG runs can form left-handed Z-DNA under the right conditions, and certain *Burkholderia* genomes (which are the most biased toward YR tracts) tend to have more CG dinucleotides than would be expected based on the mononucleotide composition. '*Aquifex aeolicus*' has been noted previously to be 'unique' in its overrepresentation or preference for polypurines in protein-coding regions (Raghavana *et al.*, 2000). In general, most soil bacteria tend to have considerably less purine tracts and more YR tracts than would be expected. These findings hint at the possibility of being able to find some general structures or maybe even sequence signatures that would be indicative of the environment in which an organism lives.

One possible explanation for the bias in purine tracts might have to do with strand bias of oligomers towards the leading or lagging DNA replication strands. In *Firmicutes*, there is a strong tendency for G and A residues to be on the leading strand; hence, one would expect a general bias towards purine tracts, as can be found in Fig. 1(a). In contrast, *Proteobacteria* tend to have G and T residues on the leading strand, resulting in less bias towards purine stretches, but more bias towards YR stretches, as can be seen in Fig. 1(b). The bias of A residues towards the leading strand is correlated with the presence of the *polC* gene; in the absence of *polC* there is a general tendency for the A residues to be on the lagging strand (Worning *et al.*, 2005). Finally, we have calculated the bias in the intergenic regions to test whether it is coming from the coding sequences (~90% of the DNA in many bacterial genomes). One could argue that most of the bias might come from choice of codon usage, for example. If this were true, in general one should see means closer to the expected value of around 1% for the intergenic



**Fig. 1.** Box and whisker plot of fraction of RR tracts (a) or YR tracts (b). The colour scheme for the phyla is the same as found in the GenomeAtlas database ([www.cbs.dtu.dk/services/GenomeAtlas/](http://www.cbs.dtu.dk/services/GenomeAtlas/)). IG is the value for the intergenic DNA. The box represents the middle 50% of the data. The median for each phylum 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 in open circles. One single vertical line is shown where only one proteome is present. The vertical red line in both plots represents the expected values.



**Fig. 2.** Observed frequencies of RR and YR stretches in bacterial chromosomes versus AT content. The chromosomal fraction containing at least 10 purines is shown in blue and the YR fraction is shown in red. Fifty per cent AT is marked with a vertical grey line; the expected fraction (independent of AT content) is shown by a horizontal black rule.

regions, which is the opposite of what is seen for most of the genomes in Fig. 1 (with the exception of *Spirochaetes*). Currently, we cannot offer a simple explanation for why this trend is seen.

### 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/GenUp020/>

### Acknowledgements

This work was supported by a grant from the Danish Center for Scientific Computing. We also thank the Sanger Centre (<http://www.sanger.ac.uk/Projects/>) and the Joint Genomes Initiative ([http://genome.jgi-psf.org/finished\\_microbes/](http://genome.jgi-psf.org/finished_microbes/)) for making their genome sequences and

preliminary annotations available to the public.

**P. Christoph Champ, Tim T. Binnewies, Natasja Nielsen, Guy Zinman, Kristoffer Kiil, Heng Wu, Jon Bohlin and David W. Ussery**

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

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

DOI 10.1099/mic.0.28637-0

**Carlson, J. H., Porcella, S. F., McClarty, G. & Caldwell, H. D. (2005).** Comparative genomic analysis of *Chlamydia trachomatis* oculotropic and genitotropic strains. *Infect Immun* **73**, 6407–6418.

**Giovannoni, S. J., Tripp, H. J., Givan, S. & 11 other authors (2005).** Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245.

**Joardar, V., Lindeberg, M., Jackson, R. W. & 29 other authors (2005).** Whole-genome sequence analysis of *Pseudomonas syringae* pv. phaseolicola 1448A reveals divergence among pathovars in genes involved in virulence and transposition. *J Bacteriol* **187**, 6488–6498.

**Kuroda, M., Yamashita, A., Hirakawa, H. & 10 other authors (2005).** Whole genome sequence of *Staphylococcus saprophyticus* reveals the pathogenesis of uncomplicated urinary tract infection. *Proc Natl Acad Sci U S A* **102**, 13272–13277.

**Li, L., Bannantine, J. P., Zhang, Q., Amonsin, A., May, B. J., Alt, D., Banerji, N., Kanjilal, S. & Kapur, V. (2005).** The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*. *Proc Natl Acad Sci U S A* **102**, 12344–12349.

**Médigue, C., Krin, E., Pascal, G. & 21 other authors (2005).** Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. *Genome Res* **15**, 1325–1335.

**Raghavana, S., Hariharanb, R. & Brahmachari, S. K. (2000).** Polypurine polypyrimidine sequences in complete bacterial genomes: preference for polypurines in protein-coding regions. *Gene* **242**, 275–283.

**Tettelin, H., Massignani, V., Cieslewicz, M. J. & 43 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.

**Thieme, F., Koebnik, R., Bekel, T. & 26 other authors (2005).** Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium *Xanthomonas campestris* pv. vesicatoria revealed by the complete genome sequence. *J Bacteriol* **187**, 7254–7266.

**Ussery, D., Soumpasis, D. M., Brunak, S., Staerfeldt, H. H., Worning, P. & Krogh, A. (2002).** Bias of purine stretches in sequenced chromosomes. *Comput Chem* **26**, 531–541.

**Worning, P., Jensen, L. J., Hallin, P. F., Staerfeldt, H.-H. & Ussery, D. W. (2005).** Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* (in press). doi:10.1111/j.1462-2920.2005.00917.x