

Genome update: distribution of two-component transduction systems in 250 bacterial genomes

Genomes of the month

Twelve new microbial genomes have been published since the last 'Genome Update' column was written. The collection of this month's genomes, listed in Table 1, include five published bacterial genomes ('*Candidatus* Blochmannia pennsylvanicus', *Colwellia psychrerythraea*, *Mycoplasma hyopneumoniae*, *Mycoplasma synoviae* and *Pseudomonas syringae* pv. *syringae* B728a). An additional four bacterial genomes have been deposited in GenBank ('*Candidatus* Pelagibacter ubique', *Pseudomonas syringae* pv. *phaseolicola* 1448A, '*Psychrobacter arcticum*' and *Staphylococcus haemolyticus*). Furthermore, three protozoan genomes have been published (*Leishmania major*, *Trypanosoma cruzi* and *Trypanosoma brucei*).

Two new methodologies for fast sequencing of bacterial genomes have been reported (Pennisi, 2005). One method uses 'off-the-shelf' components to convert an epifluorescence microscope to a high-throughput sequencing machine, which was used to sequence an *Escherichia coli* K-12 genome with a phage insertion as a control (Shendure *et al.*, 2005). Using another method, a different group has sequenced the *Mycoplasma genitalium* genome on a desktop machine in 4 h (Margulies *et al.*, 2005). Although both of these technologies are still in development and neither of them has so far produced enough data for full assembly of the genomes into one contiguous piece, it is clear that this will soon be possible (for example, the *Mycoplasma* genome was assembled into 25 contigs). Add to this software improvements – for example, one can submit a DNA sequence for a bacterial genome over the web and get a fully annotated file within 24 h (Van Domselaar *et al.*, 2005) – it will soon be possible to come into the lab on a Monday, purify some

chromosomal DNA, add it to a desktop machine and have an EMBL or GenBank file by Wednesday morning! We (and many other groups) have scripts that will take that EMBL file and then compare it to all other bacterial genomes sequenced and produce organized lists and tables of similarities and differences between the new genome and what has already been published. So by Friday a draft of the paper could be ready. The thoughts of this are both pleasant and a bit of a nightmare in terms of the vast amounts of data that will become available. This month there are 12 new genomes; soon it could well double that, approaching one new bacterial genome published every day!

'*Candidatus* Blochmannia pennsylvanicus' is an endosymbiont of ants and is related to other insect mutualists (e.g. *Buchnera* of aphids and *Wigglesworthia* of tsetse flies). All of these are γ -*Proteobacteria* which have undergone genome reduction and have stable genome organization with low levels of rearrangement. Degnan *et al.* (2005) have sequenced the '*Candidatus* B. pennsylvanicus' BPEN genome and have taken advantage of this stability of endosymbiont genomes to compare the rate of gene evolution. Overall they found a 10- to 50-fold faster amino acid substitution rate for '*Candidatus* Blochmannia', compared to other bacteria. However, there is a strong conservation of gene order and strand orientation of the

genes of '*Candidatus* Blochmannia' species, so the underlying architecture of the chromosome is stable, even though the amino acid sequences are changing.

Colwellia psychrerythraea 34H is the type species of the genus *Colwellia* and is a good model organism for the study of life in permanently cold environments; although members of the γ -*Proteobacteria* group, *Colwellia* species are strictly psychrophilic and thus require temperatures of less than 20 °C. *C. psychrerythraea* prefers temperatures ranging from -1 °C to +10 °C, but can grow at lower temperatures in sugar solutions or under deep-sea pressures. A recent comparative genomic analysis of the 5.4 Mbp sequenced genome suggests that a collection of synergistic changes in the overall genome content and amino acid composition supports its psychrophilic lifestyle (Méthé *et al.*, 2005). An increase in polar residues (in particular serine), the favouring of aspartate over glutamate and a decrease in charged surface residues could all be likely to enhance architectural changes in enzymes, resulting in increased effectiveness at cold temperatures. In addition, of the 4937 predicted CDSs, many are likely to confer cryotolerance, e.g. polyhydroxyalkanoates (PHA) that may also aid in pressure adaptation, extracellular polysaccharides that can serve as cryoprotectants, genes involved in synthesis of branched membrane lipids that reduce membrane

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	AT (mol%)	No. of genes	tRNAs	rRNAs	Accession no.
' <i>Candidatus</i> Blochmannia pennsylvanicus' BPEN	791 654	70.4	610	39	1	CP000016
<i>Colwellia psychrerythraea</i> 34H	5 373 180	62.0	4 910	88	9	CP000083
<i>Mycoplasma hyopneumoniae</i> J	897 405	71.5	665	30	1	AE017243
<i>Mycoplasma synoviae</i> 53	799 476	71.5	672	34	2	AE017245
' <i>Candidatus</i> Pelagibacter ubique' HTCC1062	1 308 759	70.3	1 354	32	1	CP000084
<i>Pseudomonas syringae</i> pv. <i>syringae</i> B728a	6 093 698	40.8	5 137	64	5	CP000075
<i>Pseudomonas syringae</i> pv. <i>phaseolicola</i> 1448A	5 928 787	42.0	4 982	62	5	CP000058
<i>Pseudomonas syringae</i> pv. <i>phaseolicola</i> 1448A pLarge	131 950	45.9	127	0	0	CP000059
<i>Pseudomonas syringae</i> pv. <i>phaseolicola</i> 1448A pSmall	51 711	44.0	60	0	0	CP000060
' <i>Psychrobacter arcticum</i> ' 273-4	2 650 701	57.2	2 147	49	4	CP000082
<i>Staphylococcus haemolyticus</i> JCSC1435	2 685 015	67.2	2 678	60	5	AP006716
<i>Staphylococcus haemolyticus</i> pHsaeA	2 300	70.1	3	0	0	AP006717
<i>Staphylococcus haemolyticus</i> pHsaeB	2 366	68.9	2	0	0	AP006718
<i>Staphylococcus haemolyticus</i> pHsaeC	8 180	69.7	11	0	0	AP006719

viscosity at cold temperatures and cold-shock proteins (Methé *et al.*, 2005).

Mycoplasma hyopneumoniae is the aetiological agent of swine enzootic pneumonia. It is a major threat to swine health and is responsible for great economic damage every year in the swine industry. *Mycoplasma synoviae* is a poultry pathogen causing respiratory disease and synovitis. Vasconcelos *et al.* (2005) compared and analysed three complete *Mycoplasma* genomes, two *M. hyopneumoniae* strains (strain 7448, a pathogenic strain, and strain J, a non-pathogenic strain) and one strain of the avian pathogen *M. synoviae* (strain 53).

To examine different aspects of *Mycoplasma* evolution these genomes were also compared with eight other available *Mycoplasma* genome sequences. Genomic comparison revealed strain-specific regions, high rates of genomic rearrangements, alterations in adhesin sequences and possible horizontal gene transfer between *M. synoviae* and *Mycoplasma gallisepticum* (Vasconcelos *et al.*, 2005).

Pseudomonas syringae is a widespread bacterial pathogen of many plant species. This month, a second *P. syringae* genome has been published (*P. syringae* pv. *syringae* B728a; Feil *et al.*, 2005) and the sequence of

a third strain deposited in GenBank (*P. syringae* pv. *phaseolicola* 1448A). As can be seen in Table 1, the *P. syringae* pv. *syringae* B728a genome consists of only one circular chromosome of 6.1 Mb, with no plasmid, whilst the *P. syringae* pv. *phaseolicola* 1448A genome is a bit smaller (5.9 Mbp for the main chromosome) with an additional two plasmids. It is beyond the scope of this article to compare the three *P. syringae* genomes, but Table 2 lists the number of sigma factors (Kill *et al.*, 2005) and the number of two-component regulatory systems for all eight *Pseudomonas* genomes that have been sequenced so far. It is worth noting that,

Table 2. Number of histidine kinases (HisK) and response regulators (RRs) in this month's genomes and for eight *Pseudomonas* genomes

Organism	HisKA	RR	ECF	Sig70	Sig54
<i>Colwellia psychrerythraea</i> 34H	38	71	12	4	1
<i>Mycoplasma hyopneumoniae</i> J	0	0	0	1	0
<i>Mycoplasma synoviae</i> 53	0	0	0	1	0
' <i>Candidatus</i> Pelagibacter ubique' HTCC1062	4	5	0	2	0
' <i>Psychrobacter arcticum</i> ' 273-4	10	16	0	2	1
<i>Staphylococcus haemolyticus</i> JCSC1435	11	15	0	2	0
<i>Pseudomonas aeruginosa</i> PAO1	53	83	19	4	1
<i>Pseudomonas fluorescens</i> Pf-5	68	113	28	4	1
<i>Pseudomonas fluorescens</i> PfO-1	63	101	21	4	1
<i>Pseudomonas fluorescens</i> SBW25	64	102	25	4	1
<i>Pseudomonas putida</i> KT2440	58	89	19	4	1
<i>Pseudomonas syringae</i> DC3000	60	85	10	4	1
<i>Pseudomonas syringae</i> pv. <i>syringae</i> B728a	61	89	10	4	1
<i>Pseudomonas syringae</i> pv. <i>phaseolicola</i> 1448A	58	85	9	4	1

although the number of ECF sigma factors in all three *P. syringae* genomes is only about half that found in other *Pseudomonas* species (e.g. 10 vs about 20), the number of histidine kinases is about the same, roughly 60, and the number of response regulators (85–89) in the three *P. syringae* genomes is also close to that found in the other *Pseudomonas* genomes.

Baghdad boil, sleeping sickness and Chagas disease are three neglected parasitic diseases caused by *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*, respectively. Although responsible for the deaths of more than 150 000 people per year, there are still no vaccines and existing drugs often are toxic. Originally starting as several independent genome projects 15 years ago, the ‘TriTryp’ consortium has now published the genomes of these three parasites (El-Sayed *et al.*, 2005a). Perhaps one of the most interesting results is that despite differences in spreading, lifestyle and number of genes (ranging from 8000 to 12 000 genes), the ‘TriTryps’ share a core of over 6000 genes of which 2000 are unique to these three organisms and many of these species-specific genes are found near the end of the chromosomes where the genome tends to be more changeable. Hopefully, these gene sets will eventually provide targets for drugs that will affect the parasite, but not the host.

While *L. major* probably has genes permitting it to invade white blood cells (Ivens *et al.*, 2005), *T. brucei* can evade the immune system by creating a ‘smokescreen’ of millions of molecules on its surface (Berriman *et al.*, 2005); this is made by combining fragmented products of pseudogenes. Due to the huge genetic diversity, the sequencing of *T. cruzi* was especially difficult. Only with data from the other two genomes, combined with special build tools based on the defined nucleotide positions (DNP) method, was it possible for the *T. cruzi* genome to be assembled (El-Sayed *et al.*, 2005b).

Method of the month – distributions of two-component transduction systems in bacterial genomes

Two-component signal transduction systems (sometimes simply referred to as

two-component systems or here abbreviated to TCSs) are found in both prokaryotes and some eukaryotes. They provide an elegant system for signal transduction of extracellular signals. They are composed of a sensor histidine kinase (HK), which is typically membrane-spanning, and a response regulator (RR). When the HK receives a signal it autophosphorylates its HK domain and the phosphate group is then transferred to the receiver domain of the RR, thus activating the response regulator.

To recognize the RR we used a profile HMM (Hidden Markov model), downloaded from the Pfam database of protein families and HMMs (<http://pfam.wustl.edu/>), which targets the receiver domain. The HK domains are more diverse and four different HMMs are found in the Pfam database, which will recognize different classes. Table 2 details the HKs and RRs found in this month’s bacterial genomes. For the *Mycoplasma* genomes, there are no TCSs detected by our models. There are considerably more TCSs in the *Pseudomonas* genomes and, as can be seen from the table, this is generally true for all the *Pseudomonas* genomes sequenced so far. Compared to the phyla mean plotted in Fig. 1, the ‘*Candidatus Pelagibacter ubique*’, ‘*Psychrobacter arcticum*’ and *Staphylococcus haemolyticus* genomes have very few TCSs. Of course, at this stage we do not know really what to expect, since most of the bacteria sequenced are still from a fairly narrow taxonomic range.

Searching 250 complete bacterial genomes yielded the HKs and RRs shown in Fig. 1. In the *Actinobacteria* (top of the figure) we find mostly between 10 and 20 HKs, although we also see only two for the reduced genome of *Tropheryma whippelii*. In contrast, the large genomes from the *Streptomyces* species (*S. avermitilis* and *S. coelicolor*) have significantly more HKs (59 and 74 respectively). We see the same general trend with the RRs, although there are a few more RRs than HKs. The mean number of RRs for *Actinobacteria* is around 30, and we find 2, 69 and 81 in *T. whippelii*, *S. avermitilis* and *S. coelicolor*, respectively. In the next group of phyla in Fig. 1, we see that the *Bacteroides* genomes have significantly more HKs and RRs than the *Chlorobi*. *Bacteroides thetaiotaomicron* has

the most, with 67 RRs and 65 HKs. All the *Chlamydiae* seem to have one RR and one HK, except for *Parachlamydia*, which has two of each. The *Cyanobacteria* have the smallest median number of TCSs, whilst the two members of the genus *Anabaena* have the highest number of HKs and the second highest number of RRs (only surpassed by the proteobacterium ‘*Dechloromonas aromatica*’). In the *Deinococcus/Thermus* phylum we see that *Deinococcus radiodurans* has more RRs and HKs than the two *Thermus* species, with 25 RRs and 18 HKs versus around 13 RRs and 10 HKs for the *Thermus* species.

In the *Firmicutes* we find no HKs or RRs in the genera *Mycoplasma*, *Ureaplasma* and *Phytoplasma*; although it is not clear why this might be, it is worth noting that all of these organisms lack a cell wall. If the TCSs are responsible for sensing changes in the external environment, perhaps there is not a need for this in constant environments which change little. The *Firmicutes* with the most TCSs are members of the *Bacillus* and *Clostridium* genera. In the *Proteobacteria* (the phylum with the most genomes) we also find organisms that lack TCSs completely, for example in many of the reduced genomes of endosymbionts (i.e. there are no TCSs in ‘*Candidatus Blochmannia*’, *Buchnera* and *Ehrlichia canis*). *Wigglesworthia glossinidia* and ‘*Chlorochromatium aggregatum*’ both lack HKs, but have a single RR. On the other hand, there are some *Proteobacteria* (often environmental organisms) with a quite large number of TCSs. For example, ‘*Dechloromonas aromatica*’ and the *Pseudomonas* species, as well as genomes from *Bradyrhizobium*, *Geobacter*, *Burkholderia* and *Rhizobium* species. The *Spirochaetes* have a very large variance, with only the *Leptospira* in the high end; the rest (i.e. *Treponema* and *Borrelia*) have few TCSs.

Based on the results of the 250 genomes, it seems that the number of HKs and RRs are generally closely linked for most of the genomes, as can be seen in Fig. 1. In spite of the scale difference (larger scale for RRs), the ranges shown in the box and whisker plots look quite similar in both diagrams. Free-living bacteria have around 20 TCSs on average, although environmental bacteria can have considerably more. The

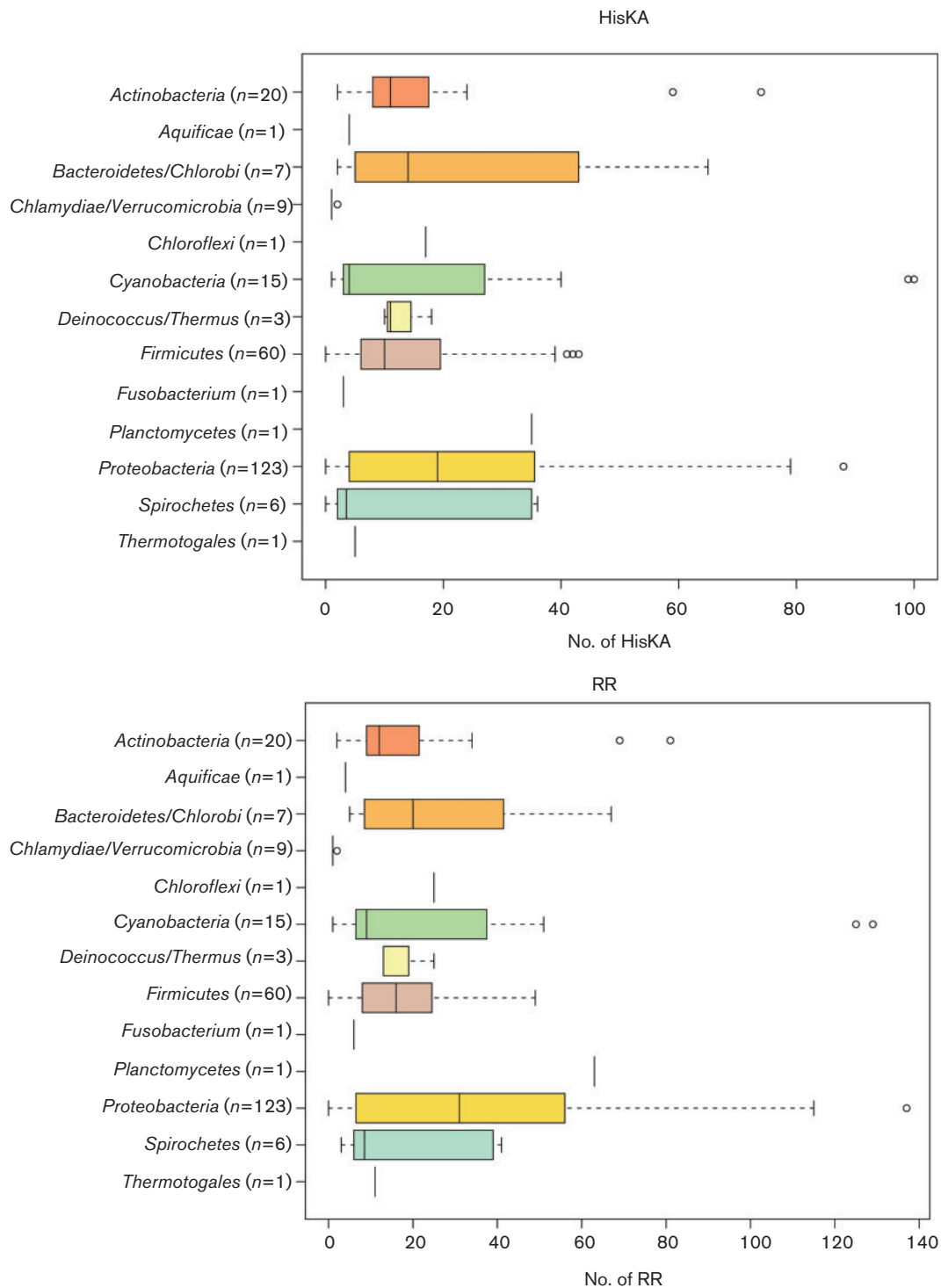


Fig. 1. Box and whisker plot of the number of two-component system proteins in 13 different bacterial phyla. Note the difference in scales on the bottom axis. The colour scheme for the phyla is the same as found in the GenomeAtlas database (www.cbs.dtu.dk/services/GenomeAtlas/). The box represents the middle 50% of the data. The median 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.

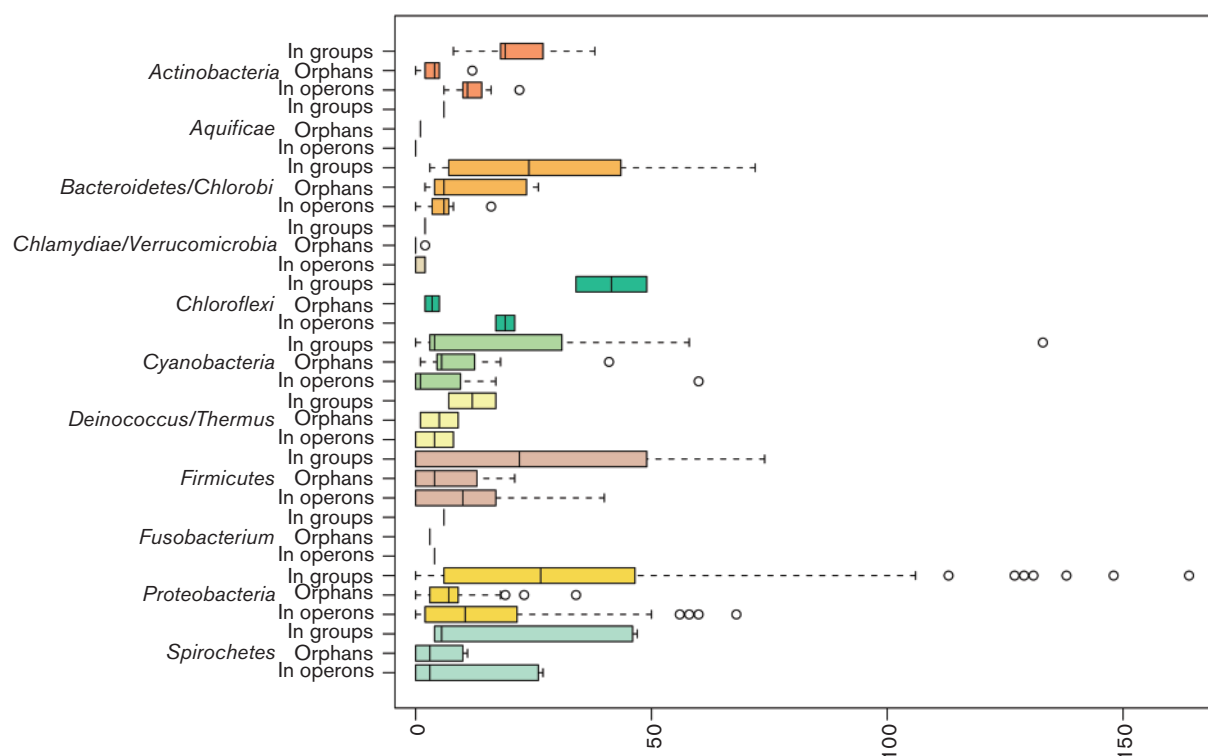


Fig. 2. Box and whisker plot of the distribution of the two-component system proteins, in terms of groups, 'orphans' and operons, in 13 different bacterial phyla. Note that this plot represents the total number of proteins (i.e. both RRs and HKs), so the numbers on the *x*-axis reflect a combination of the two distributions for RRs and HKs. The colour scheme is the same as for Fig. 1. A group is defined as occurring within 15 kbp of another RR or HK. 'Orphan' proteins are defined as those remaining RRs (and a few HKs) that are not found within a group. An operon refers to HKs and RRs occurring within 2 kbp of each other and oriented in the same direction.

Bacteroides have about 50 TCSs per genome, the largest mean number of TCSs for a phylum. It is hoped that the next 250 genomes sequenced will be more reflective of the biological diversity around us, and perhaps will help us to get a better grasp of what range of numbers to expect for a given phylum.

In addition to just counting the number of HKs and RRs in a given genome, the question remains as to whether all (or most) of the two-component systems are genes in the same operon, as classically described. In summary, we find that the majority of HKs and RRs are not found to be in the same operon (e.g. within 2000 bp of each other, in the same direction), although most of the time they are found to occur within clusters of 15 000 bp (see Fig. 2). In addition, there are some 'orphan' RRs, which are found in isolated places throughout the genome, outside the clusters containing HKs and RRs.

Supplemental web pages

Access to additional web pages containing supplemental material related to this article can be obtained via the following URL: www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp019/

Acknowledgements

This work was supported by a grant from the Danish Center for Scientific Computing. We also thank the Sanger Centre (www.sanger.ac.uk/Projects/) and the Joint Genomes Initiative (http://genome.jgi-psf.org/finished_microbes/) for making their genome sequences and preliminary annotations available to the public.

Kristoffer Kiil, Jean Baptiste Ferchaud, Christophe David, Tim T. Binnewies, Heng Wu, Thomas Sicheritz-Pontén, Hanni Willenbrock 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)

Berriman, M., Ghedin, E., Hertz-Fowler, C. & 99 other authors (2005). The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**, 416–422.

Degnan, P. H., Lazarus, A. B. & Wernegreen, J. J. (2005). Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res* **15**, 1023–1033.

El-Sayed, N. M., Myler, P. J., Blandin, G. & 42 other authors (2005a). Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309**, 404–409.

El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C. & 79 other authors (2005b). The genome sequence of *Trypanosoma cruzi*,

etiologic agent of Chagas disease. *Science* **309**, 409–415.

Feil, H., Feil, W. S., Chain, P. & 17 other authors (2005). Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. *tomato* DC3000. *Proc Natl Acad Sci U S A* **102**, 11064–11069.

Ivens, A. C., Peacock, C. S., Worthey, E. A. & 98 other authors (2005). The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**, 436–442.

Kill, K., Binnewies, T. T., Sicheritz-Pontén, T., Willenbrock, H., Hallin, P. F., Wassenaar, T. M. & Ussery, D. W. (2005). Genome update: sigma factors in 240 bacterial genomes. *Microbiology* **151**, 3147–3150.

Margulies, M., Egholm, M., Altman, W. E. & 53 other authors (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.

Méthé, B. A., Nelson, K. E., Deming, J. W. & 24 other authors (2005). The psychrophilic lifestyle as revealed by the genome sequence of *Colwellia psychrerythraea* 34H through genomic and proteomic analyses. *Proc Natl Acad Sci U S A* **102**, 10913–10918.

Pennisi, E. (2005). Biochemistry. Cut-rate genomes on the horizon? *Science* **309**, 862.

Shendure, J., Porreca, G. J., Reppas, N. B. & 7 other authors (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732.

Van Domselaar, G. H., Stothard, P., Shrivastava, S. & 7 other authors (2005). BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res* **33**, W455–W459.

Vasconcelos, A. T., Ferreira, H. B., Bizarro, C. V. & 83 other authors (2005). Swine and poultry pathogens: the complete genome sequences of two strains of *Mycoplasma hyopneumoniae*

and a strain of *Mycoplasma synoviae*. *J Bacteriol* **187**, 5568–5577.

DOI 10.1099/mic.0.28423-0

Clinical significance of seeding dispersal in biofilms

We read with interest the recent paper by Purevdorj-Gage *et al.* (2005) on seeding dispersal in *Pseudomonas aeruginosa* biofilms. The authors compared a mucoid cystic fibrosis (CF) *P. aeruginosa* strain (strain FRD1) with strain PAO1 and found it to be seeding-dispersal-negative. They concluded that seeding dispersal might not be utilized by mucoid variants of CF strains, but rather be a transmission mechanism utilized by environmental strains of *P. aeruginosa*. We have been investigating mucoid CF isolates in a flow-through biofilm model (Webb *et al.*, 2003) and have some observations that are relevant to this conclusion. Our studies have shown that CF isolates can exhibit biofilm developmental processes and seeding dispersal similar to strain PAO1 in this experimental model. We found that CF *P. aeruginosa* isolates ($n=6$) each exhibited a characteristic pattern of biofilm development and microcolony formation that was reproducible and ‘true to strain’ in replicated biofilm experiments. Some CF strains exhibited a developmental pattern similar to that reported for strain FRD1, without obvious seeding dispersal within

the time-frame of our experiments (7 days). However, other strains did form hollow structures with highly motile cells in the centre and seeding dispersal events, as described for strain PAO1, after 4–5 days of culture (Fig. 1a, b). Clearly, much still remains to be understood about the mechanisms underlying seeding-dispersal and ‘hollow-colony’ formation. Previous studies using *P. aeruginosa* strain PAO1 linked this behaviour with bacteriophage-mediated lysis of a subpopulation of cells inside microcolony structures. We also observed dispersal-associated death in our CF isolates. *BacLight* LIVE/DEAD staining (Molecular Probes) of CF biofilms after day 6 of culture showed that all six strains tested exhibited regions of cell death within microcolonies. For at least one strain the pattern of cell death was identical to that seen with strain PAO1 (e.g. Fig. 1c). Coincident with this microcolony death, bacteriophage titres reached levels of $>10^7$ p.f.u. ml⁻¹ in flow-cell effluents. Our evidence to date thus suggests that death-associated dispersal mechanisms, as have been described for strain PAO1 (Webb *et al.*, 2003, 2004), are also central to CF strain transmission.

S. M. Kirov,¹ J. S. Webb²
and S. Kjelleberg²

¹School of Medicine, University of Tasmania, Private Bag 29, Hobart, Tasmania 7001, Australia

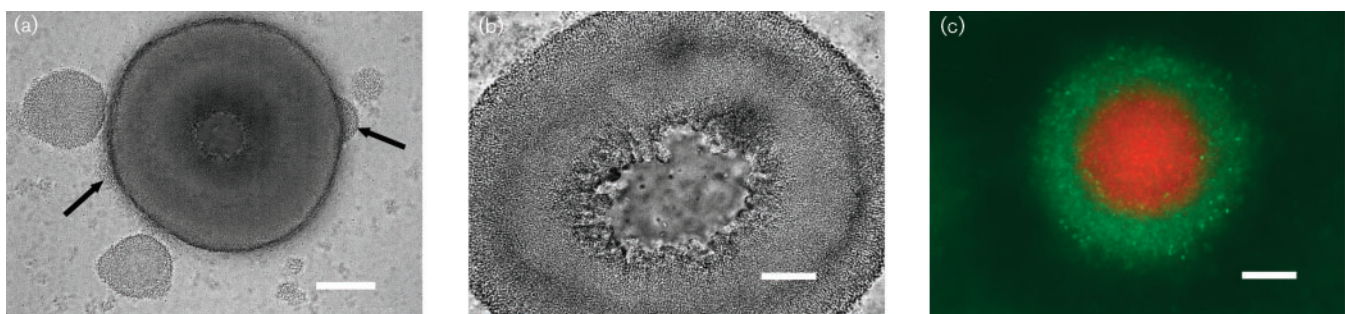


Fig. 1. Microcolonies of mucoid CF strains seen in a flow-cell model. (a) Sites of evacuation ‘blebs’ (arrows) from a hollow microcolony (strain U3A) after 4 days of culture. Bar, 46 μm . (b) A hollow microcolony (strain U3A) after 4 days of culture showing highly motile cells which appear blurred in the centre. Bar, 19 μm . (c) Cell death in the central region of a microcolony (strain 75) at ~ 7 days of culture (x - y plane view; *BacLight* LIVE/DEAD viability stain). Live cells fluoresce green; dead cells appear red in this confocal micrograph. Bar, 15 μm .