GENOME WATCH

The value of comparison

Nicholas Thomson, Mohammed Sebaihia, Ana Cerdeño-Tárraga, Stephen Bentley, Lisa Crossman and Julian Parkhill

With the number of published microbial genomes now in excess of 100, any new genome that is sequenced is likely to have a close relative available for comparison. Indeed, it is increasingly difficult to perform any genomic analysis that is not comparative. This should, however, not be seen as a drawback; it is often the case that a large amount of information can be drawn from these comparisons, especially between closely related organisms. Several genome sequences published recently indicate the value of comparisons at the genomic level.

Helicobacter hepaticus, an enterohepatic *Helicobacter* species, causes chronic hepatitis and liver cancer in mice. The genome of *H. hepaticus*[1] has recently been compared with the previously sequenced genomes of two *Helicobacter pylori* strains[2] and with *Campylobacter jejuni*[3]. *H. pylori* is a close relative of *H. hepaticus*, which colonizes the stomach and causes gastric inflammation and peptic ulcers, whilst *C. jejuni* shares the same habitat (the lower bowel) as *H. hepaticus* and causes diarrhoea in humans. Both *H. hepaticus* and *H. pylori* produce large amounts of urease. In *H. pylori*, urease production contributes to increasing the pH in the highly acidic environment of the stomach. Although the exact role of urease production in *H. hepaticus* pathogenesis is unknown, it has been shown that urease is essential for colonization. High levels of urease activity requires nickel, and although both *H. hepaticus* and *H. pylori* have similar urease gene clusters, they encode different nickel transport genes. As expected with differentially adapted organisms, most of the colonization and virulence factors of *H. pylori* are missing from the *H. hepaticus* genome, including the vacuolating cytotoxin gene, *vacA*. Instead, *H. hepaticus* has a cluster of genes (*cdtABC*) that are similar to those that encode the cytolethal distending toxin (CDT) of *C. jejuni*. The authors suggest that the genotoxic effects of CDT might contribute to the carcinogenic potential of *H. hepaticus*. *H. hepaticus* possesses only a small number of genes that encode transcriptional regulators, which comprise three sigma factors (σ^{70}, σ^{54}, σ^{28}), one flagellar anti-sigma factor (FlgM), and a flagellar transcriptional activator (FlgR). As in *H. pylori* and *C. jejuni*, this paucity of transcriptional regulators is compensated for by the high number of phase-variable genes owing to slipped-strand mispairing that occurs at homopolymeric tracts (FIG. 1). The authors detected 36 cases in the shotgun sequence where such tracts were seen to be variable, and a further 33 potential phase-variable genes. The genome carries a 71-kb region with unusual G+C content (33.2% compared with 35.9% elsewhere). This genomic island (HHGI1) contains 70 coding sequences (CDS), most of which code for hypothetical proteins; it also encodes three CDS that are similar to components of a type IV secretion system. Whole genome microarray analysis of 12 different *H. hepaticus* strains revealed that HHGI1 is either totally or partially deleted in different strains. These data indicate that HHGI1 was present in the common ancestor, but has subsequently been deleted. The availability of mice pathology records has established a direct role for HHGI1 in virulence; 5/6 mice infected with HHGI1-carrying strains had liver cancer, while 0/4 mice infected with HHGI1-negative strains developed liver disease.

A closer comparison is that between *Mycobacterium bovis*[4] and *Mycobacterium tuberculosis*. *M. bovis* is the agent of bovine tuberculosis, and the progenitor for the human vaccine strain bacillus CalmetteGuérin (BCG). Surprisingly, this genome does not have any unique genes, and is 99.95% identical to *M. tuberculosis* at the DNA level, lending further support to a previous study that indicated that *M. bovis* is a derivative of *M. tuberculosis* and not *vice versa*[5], and which led the authors to suggest that the reason for its host-specificity might lie in differential gene expression. The genome is 4.34 Mb, and has 11 regions of deletion with about 80 kb less DNA than *M. tuberculosis*. Apart from these deletions, there have been many single base-pair changes, which could also be partly responsible for the distinctive characteristics of *M. bovis*. The greatest variation, when compared with *M. tuberculosis*, was found in the genes that encode the cell wall and secreted proteins, such as the large, repetitive PE-PGRS/PPE (letters indicate amino acid repeats) families of proteins. More than 60% of these proteins differ,

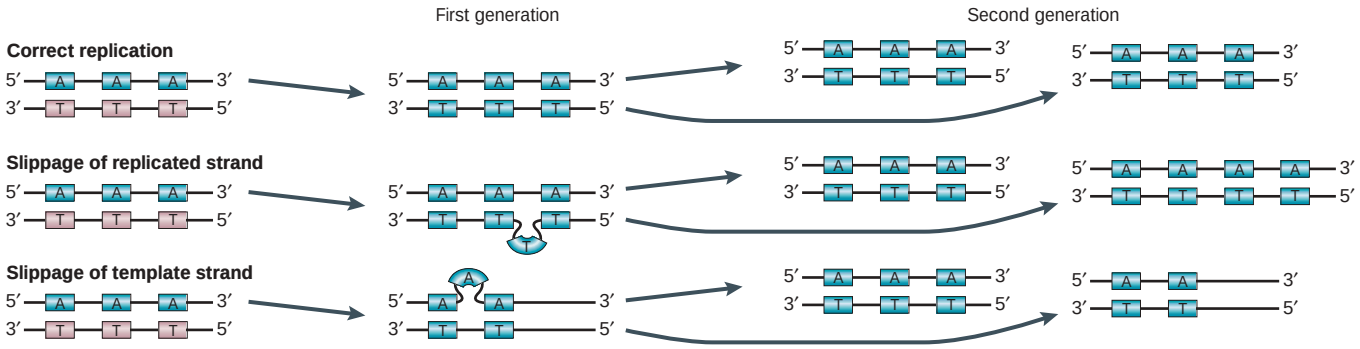

Figure 1 | **Slipped strand mispairing to generate phase variation.** In many prokaryotes, so-called slipped-strand nucleotide mispairing can generate variation in gene expression. Illegitimate base pairing in regions of repetitive DNA during replication, coupled with inadequate DNA mismatch repair systems, can produce deletions or insertions of repeat units. Bulging in the replicated and template strands, gives rise to larger and smaller numbers of repeat units, respectively. The figure shows a strand of DNA (blue) being carried through two rounds of replication.

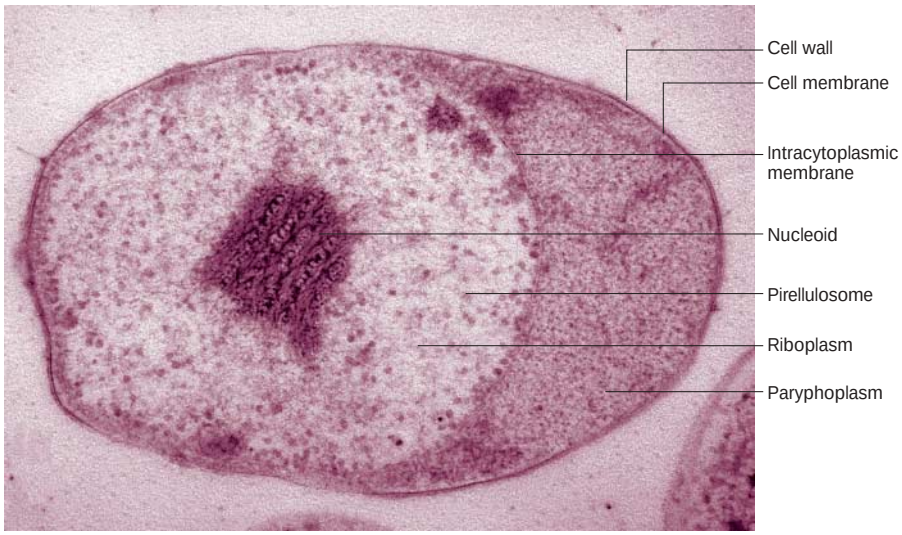

Figure 2 | **Compartmentalisation in *Pirellula marina*.** A cryosubstituted whole-cell thin section of *Pirellula marina* strain ACM 3344 (= ATCC 49060, DSM 3645) is shown false-coloured to improve contrast. Paryphoplasm (from the Greek paryphe, meaning 'border woven along a robe') is the peripheral cytoplasm surrounding the double-membrane-bounded pirellulosome, and lacks ribosomes. Riboplasm, inside the pirellulosome contains ribosomes[9]. We would like to thank John Fuerst for providing this image, which has been reproduced with permission from REF. 9 © (1997) Society for General Microbiology.

compared with 35% of all shared proteins, underlining the likely importance of these proteins in surface variation. There are significant differences between the growth requirements of these two mycobacteria *in vitro*: unlike *M. tuberculosis*, *M. bovis* cannot use glycerol as a source of carbon owing to mutations in the carbohydrate catabolic pathway. However this defect is not apparent in the BCG strain — it seems possible that this is owing to selection for reversion after 13 years of passage to create a BCG strain on glycerol-soaked potato slices.

The recent sequencing of the fourth group A *Streptococcus pyogenes* strain (GAS serotype M3) has advanced our understanding of phage evolution[6]. GAS strains are human pathogens that are responsible for several serious diseases, including necrotizing fasciitis, cellulitis, sepsis and post-streptococcal acute rheumatic fever. Many virulence factors are mainly carried on acquired phages. The sequencing of this strain has shown that five phages are located on one replichore (oppositely replicated halves of the genome) and one phage is located on the other. This might unbalance the genome and induce genetic rearrangements. In addition, the phages that are present on either side of the terminus of replication can exchange genes that encode virulence factors by direct recombination. One lysogenic phage in the previously sequenced M1 serotype has been chemically induced into a lytic cycle[7]. It is therefore possible that the GAS phages can exchange genetic information and then become lytic, allowing broad dissemination of virulence factors between GAS strains.

Corynebacterium efficiens is used for the industrial production of amino acids. Researchers are interested in this organism primarily owing to its ability to grow and produce amino acids at 40°C — 10°C higher than the closely related and commonly used industrial strain, *Corynebacterium glutamicum*. This feature offers potential savings in cooling costs for large-scale fermentation. Nishio *et al.*[8] have compared the complete genomes of these two bacteria to reveal the three most common amino acid substitutions that they conclude to be most likely to contribute to the increased protein thermostability in *C. efficiens*. Detailed analysis showed that although the increase in G+C content of the genome (63% compared with 54% for *C. glutamicum*) correlates with a shift in codon usage, it cannot account for all the thermostabilising amino acid changes.

Finally, this is biology, so there will always be an exception to any rule. *Pirellula* belong to a fascinating group of marine organisms that are members of the order *Planctomycetales*. Representatives of this order are unusual among bacteria in lacking peptidoglycan as the main structural component of their cell wall, a feature that is shared only with chlamydiae and mycoplasmas. Instead, the cell wall is a proteinaceous sacculus. Moreover, these organisms have compartments with an additional, single-layered intracytoplasmic membrane, which separates an outer, ribosome-free paryphoplasm from the inner riboplasm[9] (FIG. 2). Some members of the *Planctomycetales* also have a double-layered membrane enveloping the nucleoid. At 7.5 Mb, the genome of *Pirellula* spp. is the largest single circular genome published so far[10]. Notably, the genome contains only a single set of rRNA genes as well as another 7,322 predicted genes — 4,148 of which have no matches in the public databases. Of those that could be assigned a function, polyketide synthases, non-ribosomal peptide synthases, genes involved in carotenogenesis and heavy metal resistance were detected, as well as more than 100 sulphatases.

As a consequence of the degree of cell compartmentalisation it was expected that there would be a large number of genes involved in protein export. Consistent with this notion, *Pirellula* was found to encode an unprecedented number of proteins with motifs known to be important for protein targeting. Although the *Pirellula* cell wall lacks peptidoglycan, several genes involved in its synthesis are present within the genome. In contrast to previous theories, Glockner *et al.*[10] conclude that predecessors of *Pirellula* have developed from organisms with a Gram-negative-like cell envelope that has taken on its present form by the successive loss of genes involved in peptidoglycan synthesis. Consistent with this idea, several other vestiges of a Gram-negative-type cell wall have also been detected.

Nicholas Thomson, Mohammed Sebaihia, Ana Cerdeño-Tárraga, Stephen Bentley, Lisa Crossman and Julian Parkhill are at the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
e-mail: microbes@sanger.ac.uk
doi:10.1038/nrmicro734

1. Suerbaum, S. *et al.* The complete genome sequence of the carcinogenic bacterium *Helicobacter hepaticus*. *Proc. Natl Acad. Sci. USA* **100**, 7901–7906 (2003).
2. Alm, R. A. *et al.* Genomic sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180 (1999).
3. Parkhill, J. *et al.* The genome sequence of the food–borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668 (2000).
4. Garnier, T. *et al.* The complete genome sequence of *Mycobacterium bovis*. *Proc. Natl Acad. Sci. USA* **100**, 7877–7882 (2003).
5. Brosch, R. *et al.* A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl Acad. Sci. USA* **99**, 3684–3689 (2002).
6. Nakagawa, I. *et al.* Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res.* **13**, 1042–1055 (2003).
7. Ferretti, J. J. *et al.* Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl Acad. Sci. USA* **98**, 4658–4663 (2001).
8. Nishio, Y. *et al.* Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. *Genome Res.* **13**, 1572–1579 (2003).
9. Lindsay, M. R., Webb, R. I. & Fuerst, J. A. Pirellulosomes: a new type of membrane-bounded cell compartment in planctomycete bacteria of the genus *Pirellula*. *Microbiology UK* **143**, 739–748 (1997).
10. Glockner, F. O. *et al.* Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc. Natl Acad. Sci. USA* **100**, 8298–8303 (2003).