NEWS & ANALYSIS

GENOME WATCH

Sequencing the species pan-genome

Stephen Bentley

This month's Genome Watch discusses the recent advances in sequencing technology that have allowed microbiologists to determine the genome sequences not just of individual species but also of entire genera and have allowed a view of the entire pan-bacterial genome.

Owing to the current acceleration in the rate of DNA sequence generation, microbiologists will be faced with increasing amounts of genomic data. The impact on microbiology will be enormous; ultra-high-throughput sequencing technology will make the sequencing of a single bacterial genome trivial and the sequencing of thousands realistic.

These advances will allow evolutionary hypotheses to be tested on a broader scale than before, and, as a result, the species concept for bacteria will become even more strained. The frequent gain and loss of genomic DNA will make it difficult to trace bacterial phylogenies, but the increasing amount of data should allow for clarification of patterns in the flux of genes. This will inevitably lead to new nomenclature, terminology and definitions to describe the newly acquired data: already we have the choice of the equivalent terms 'species genome', 'pan-genome' and 'supragenome'. Furthermore, the challenge of visualization of the data analysis is set to become a distinct field of bioinformatics research.

Analysis of multiple whole genome sequences from a single species is already possible in many cases. Efforts have also been made to study the pan-genome of an entire bacterial genus, and Lapierre and Gogarten[1] recently went a step further and used 573 genomes to determine the bacterial pan-genome, the set of all genes in eubacteria. Published analyses of pan-genomes clearly illustrate attempts to interpret the sequence data among widespread lateral gene flow.

Tettelin and colleagues[2] used the genomes of eight strains to characterize the pan-genome of *Streptococcus agalactiae* and concluded that it is 'open' (meaning that the pan-genome has an infinite size) and that each new added genome would provide, on average, 33 new genes. They first established orthologous groups of genes, which allowed the designation of a core genome (the genes present in all strains) of 1,806 genes (~80% of each individual genome) and showed that 95% of these core genes share greater than 90% identity. They then used regression analyses to extrapolate and predict the consequence of sequentially adding more genomes and concluded that the probability that the pan-genome is 'closed' (meaning the pan-genome has a definite size) is less than 6 x 10^{-4}. Importantly, they showed that whereas *Streptococcus pyogenes* also has an open pan-genome, the *Bacillus anthracis* pan-genome could be encompassed by as few as four genomes. However, they went on to suggest that *B. anthracis* is actually a clone of *Bacillus cereus* with a distinctive phenotype owing to its anthrax toxin plasmid, and that all true species genomes might be open. However, they also note that other species, such as maternally transmitted endosymbionts, may appear to have closed pan-genomes, reflecting their isolated niche with little opportunity for lateral acquisition of DNA.

> "The pan-genome sizes and rates of recombination are likely to reflect differences in niche and lifestyle of the different species..."

The core genome of *S. agalactiae* is composed of housekeeping, regulatory cell envelope and transport genes, whereas the strain-specific genes are dominated by genomic islands of atypical nucleotide composition, which is suggestive of horizontal acquisition. For both *S. agalactiae* and *S. pyogenes*, around 10% of strain-specific genes are phage-associated, highlighting the crucial role of phages in the generation of genomic diversity in these species.

Hiller and colleagues[3] took a slightly different approach to the analysis of the pan-genome (in this case referred to as the supragenome) of *Streptococcus pneumoniae*. They applied a clustering method to 17 genomes of *S. pneumoniae* to assign coding sequences into 'core', 'unique' or 'distributed' categories, in which distributed refers to genes present in some, but not all, strains. The distributed genes include the capsule biosynthesis loci, which add up to approximately 2 Mb and include approximately 2,000 genes, which is equivalent to a single pneumococcal genome[4]. Notably, Hiller and colleagues[3] set a cut-off of 70% identity over 70% of the peptide length for inclusion in an orthologous protein family or cluster, whereas Tettelin *et al.*[2] required "a minimum of 50% sequence conservation over 50% of the protein/gene length." Such differences in cut-off values will influence the output values of the analysis and may have consequences on subsequent interpretation. Crucially, this will affect the definition of the core genome, as genes that are shared by all genomes but that are inherently variable in sequence could easily be omitted from the core set. Examples of such genes include those that encode surface antigens and transporters. Variability in surface antigens is likely to be a function of host immune recognition, and the inclusion or exclusion of such surface antigens from the core genome would have serious

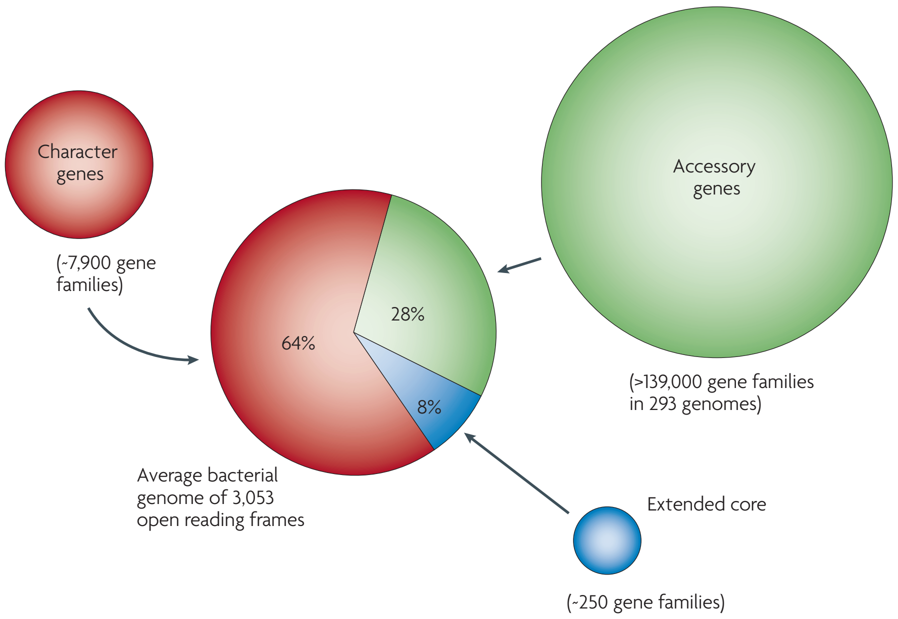

Figure 1 | **The bacterial pan-genome.** Each gene can be classified into one of three groups. First, the extended core genes, which include those that control translation, replication and energy homeostasis, are represented by approximately 250 gene families (blue). Second, the set of 'character' genes (red) is comprised of genes involved in adaptation to a particular environmental niche, such as those that control photosynthesis or endosymbiosis. There are approximately 7,900 character gene families. By contrast, the third group of genes, the accessory genes (green), is nearly limitless in size. These genes are often specific for a strain or serotype, and in many cases have no known function. Figure is reproduced, with permission, from REF. 1 © Elsevier Sciences.

implications for their potential for selection as vaccine candidates. Sequence variation of transporters encoded at equivalent genomic locations in different genomes could reflect a difference in substrate specificity, which raises the question of whether they should be considered orthologous.

As the analysis of pan-genomes is annotation dependent, a gene must be annotated to be included in the analysis. Both Tettelin and colleagues[2] and Hiller and colleagues[3] attempted to account for this problem by including steps to compare translations of gene-free regions against public protein databases to identify genes missed in the annotation. In the future, we are likely to see the use of annotation-independent methods in which the DNA sequences, rather than the coding sequences, define the core and pan-genome.

Lefébure and Stanhope[5] extended pan-genome analysis to the genus level by analysing 26 genomes of 6 streptococcal species. They used the OrthoMCL program to assign genes to orthologous groups, which allowed the determination of core and non-core gene sets, and concluded that the genus pan-genome contains slightly more than 6,000 genes. Furthermore, they showed that the pan-genome of *S. agalactiae* is larger than that of *S. pyogenes*, and that strains of particular species share around 75% of their genes whereas the streptococcal core genome comprises around 600 genes.

Using phylogenetic and base substitution methods, Lefébure and Stanhope[5] found evidence of recombination in the core genomes of all six species, although they detected higher levels of recombination in *S. pyogenes* than in *S. agalactiae*. The pan-genome sizes and rates of recombination are likely to reflect differences in niche and lifestyle of the different species, with greater niche diversity requiring larger pan-genomes and close host association promoting the generation of diversity through recombination with co-colonizing close relatives.

So what about the bacterial pan-genome? Lapierre and Gogarten[1] applied the pan-genome approach to 573 eubacterial genomes and concluded that the bacterial pan-genome is infinite and the size of the bacterial core genome is approximately 250 genes. To cut down the computational load for this larger-scale analysis, 15,000 genes were randomly sampled from 293 genomes and then searched against each of the genomes to determine their frequency distribution across the entire data set. This information was then used to extrapolate a sampling curve for the pan-genome, allowing estimates to be made of the size of the core. From the frequency distribution, the authors described the extended core genome (found in nearly all genomes), the accessory pool (found in few genomes) and the remainder, which the authors describe as character genes, as they could potentially be used to describe the character of a group of related organisms. The 7,900 character gene families make up most of the core genome and constitute 64% of an average bacterial genome (FIG. 1). The accessory gene families are shared by a small number of species, but constitute much of the infinite nature of the pan-genome and make up 28% of an average genome. The extended core category does not represent the bacterial minimal genome but rather a backbone from which an individual genome is built.

Finally, the authors discuss the possible roles of character and accessory genes in two distinct mechanisms of evolution. The character genes comprise families such as ATP-binding cassette transporters and metabolic enzymes, in which gene duplication, domain shuffling and sequence drift allow the generation of novel functions by generating variation in existing genes. The accessory genes are frequently small and associated with bacteriophages. Although a small proportion may represent misannotations, many may be fragments of genes that have been picked up by bacteriophages and now exist under low selective pressure, allowing for rapid evolution. Occasionally this may lead to the generation of a new protein fold, which will quickly be disseminated through the phage to the character gene pool if it provides a selective advantage.

Stephen Bentley is at the Sanger Institute,
Wellcome Trust Genome Campus, Hinxton,
Cambridge, CB10 1SA, UK.
e-mail: microbes@sanger.ac.uk

doi:10.1038/nrmicro2123

1. Lapierre, P. & Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends Genet.* 23 Jan 2009 (doi:10.1016/j.tig.2008.12.004).
2. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005).
3. Hiller, N. L. *et al.* Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J. Bacteriol.* **189**, 8186–8195 (2007).
4. Bentley, S. D. *et al.* Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet.* **2**, e31 (2006).
5. Lefébure, T. & Stanhope, M. J. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.* **8**, R71 (2007).

DATABASES
Entrez Genome Project: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
Bacillus anthracis | *Bacillus cereus* | *Streptococcus agalactiae* | *Streptococcus pneumoniae* | *Streptococcus pyogenes*

ALL LINKS ARE ACTIVE IN THE ONLINE PDF