Meeting report

Microbial metagenomics

Emmanuel F Mongodin, Joanne B Emerson and Karen E Nelson

Address: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.

Correspondence: Emmanuel F Mongodin. E-mail: emongodin@tigr.org

Published: 27 September 2005

Genome ***Biology*** 2005, **6:**347 (doi:10.1186/gb-2005-6-10-347)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/10/347

A report on the 105th Annual Meeting of the American Society for Microbiology, Atlanta, USA, 5-9 June 2005.

The 105th Annual General Meeting of the American Society for Microbiology held recently in Atlanta, Georgia, brought together the entire community of microbiologists to view and listen to the latest innovative developments in the field. The keynote lectures emphasized the contributions of genomics to microbiology, as well as the various applications to be derived from this continuously evolving field. As many speakers pointed out, the microbial world remains the last frontier in biology, as an estimated 99% of microbial species are still uncultured: a colossal reservoir of biodiversity and novel genes remains to be discovered. In her opening talk, Claire Fraser (The Institute for Genomic Research (TIGR), Rockville, USA) highlighted the major developments in microbial genomics since the completion of the genome sequence of *Haemophilus influenzae*, the first free-living organism to be sequenced, and conveyed a number of provocative ideas, including the revolution in the species concept, the forces that shape microbial genomes (leading, for example, to genome reduction in symbiotic species), and lateral gene transfer (LGT), now accepted to be far more widespread than originally appreciated.

Microbial genomics grows up through metagenomics

With an estimated 266 microbial genome sequences completed and an additional 730 in progress (Figure 1; see also GenomesOnline [http://www.GenomesOnline.org]), the new field of metagenomics, or the sequencing of an entire microbial community *en masse*, promises to reform and expedite gene discovery. In her acceptance speech for the 2005 Promega Biotechnology Research Award, Fraser described the human gastrointestinal (GI) tract metagenomics sequencing project, which is exposing the bias in our current understanding of different environmental niches because of our dependence on culture- and PCR-based techniques for exploring microbial diversity. The GI metagenomic study has revealed major differences among individuals, in addition to microheterogeneity within individual GI populations, which were not previously apparent. In previous culture-based studies of the human GI tract, Firmicutes and Bacteroides were thought to be the most abundant microbial groups present, but results from metagenomics approaches suggest that Actinobacteria and Archaea are among the most prolific.

Microbial diversity in the GI tract was also described by Jeffrey Gordon (Washington University School of Medicine, St Louis, USA). There are ten times more prokaryotic cells than human cells in the human body; in the intestines alone, there are more than 10 x 10^9 microorganisms. Our dependence on these microbial populations is highlighted by the fact that these symbionts have developed chemical strategies for regulation of host gene expression in ways that benefit both themselves and us. One direct relationship is that as much as 10% of our daily energy comes from short-chain fatty acids produced by our microbial inhabitants.

Metagenomics studies of microbial populations in an acid mine environment were described by Rachna Ram (University of California, Berkeley, USA). By undertaking a comprehensive genomic analysis complemented by an assessment of gene expression, she and colleagues from Jill Banfield's lab have revealed that the acid mine microbial community is relatively simple (three bacterial and three archaeal species), and that *Leptospirillum* is the dominant population. Shotgun sequencing enabled the reconstruction of two complete genomes and partial chromosomes of three other species but also raised some important questions about the roles of the hypothetical genes (more than 40% of all the predicted genes). Proteomic analyses could detect 17% of the

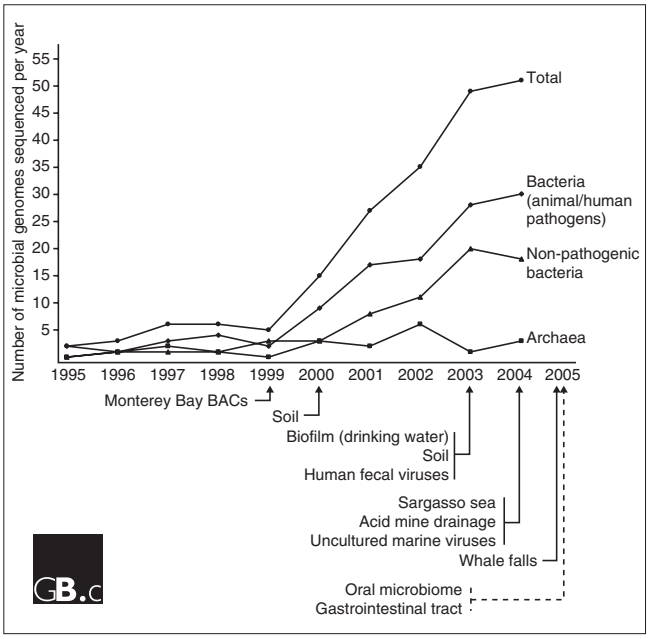

Figure 1
Progression of the number of sequenced microbial genomes since the completion of the genome of *Haemophilus influenzae* in 1995. The graph shows the number of microbial genomes that have been sequenced each year in various categories. The lower section of the figure highlights the acceleration in metagenomic projects, where all the microbial species in an environment can be sequenced without initial culturing; solid lines represent completed and published projects while the dashed line represents projects in progress.

proteins predicted by shotgun sequencing, and the number of proteins recovered per organism paralleled the number of cellular organisms in the biofilm. Nancy Moran (University of Georgia, Athens, USA) discussed a shotgun approach to sequencing various ocean samples using bacterial artificial chromosome (BAC) and fosmid libraries. The initial results and comparisons to samples from Monterey Bay (from the group of Ed DeLong at the Massachusetts Institute of Technology, Cambridge, USA) showed that approximately 20% of the genes in her study belong to the *Roseobacter* clade (which has 30-40 genera). Using GenBank sequences from 118 genera (19 of which are marine), Moran's group identified 33 of 613 operons that are found more often in marine genomes than would be expected by chance. Of those 33, half encode transporters, mostly associated with sodium transport, consistent with what one might expect for interaction with the ocean environment.

Despite the increasing number of sequencing projects, from single organisms to entire bacterial communities, the number of hypothetical and conserved hypothetical proteins in each project seems to remain constant. As remarked by Frank Collart (Argonne National Laboratory, Argonne, USA) the validation of function for hypothetical proteins has not kept pace with genome sequencing. The '70% hurdle' - the fact that only 70%, at most, of the proteins predicted in a given sequencing project can be assigned a function - will definitely be the major challenge of the post-genomic era. As Michael Galperin (National Center for Biotechnology Information (NCBI), Bethesda, USA) pointed out, many of these hypothetical genes are expressed *in vivo*, and it is likely that they represent a reservoir of new functions yet to be discovered.

Towards a redefinition of the concept of species

The struggle to define the concept of a microbial species continues. Fraser introduced a new concept: define a microbial species by its 'pan-genome', that is, the set of both core and non-essential genes that includes all the genes present in the species gene pool, although not necessarily present in each individual strain. The number of genomes necessary to fully define a pan-genome, however, seems to vary depending on the species. To illustrate this work, Fraser presented the work of Hervé Tettelin (TIGR) and described how Group B *Streptococcus* (eight isolates sequenced) does not appear to have a well defined pan-genome, as each genome brings an average of 33 new, strain-specific genes to the group's pan-genome. Group A *Streptococcus* gives similar results, leading to the concept of an 'open pan-genome' for *Streptococcus* Groups A and B. In contrast, the five genomes available for *Bacillus anthracis* show a well defined 'closed pan-genome', which seems to remain consistent as more strains are sequenced. One possible explanation for this difference is that *Streptococcus* strains live in highly variable environments with selective pressure to adapt to different environments, whereas strains of *B. anthracis* do not experience as much selective pressure (or it may just be that strains from the same *B. anthracis* clade have been sequenced). Continuing the pan-genome theme, James Tiedje (Michigan State University, East Lansing, USA) reported that in *Burkholderia* the number of genomes necessary to capture 95% of the genes of that species is between 13 and 15. This number varies depending on the species studied and whether the strains chosen represent well the ecological diversity of the species (ecological niches, pathogenic and non-pathogenic strains of the same species, and so on).

W. Ford Doolittle (Dalhousie University, Halifax, Canada) pointed out that, while many people had thought the genomics era would clarify the microbial species concept, recombination and LGT events among prokaryotes, as revealed by genomics and more recent metagenomics studies, have actually made it harder to define. While the concept of species was previously considered a necessity, it may not be applicable at the level of the individual genome. Robert Charlebois (NeuroGadgets, Ottawa, Canada) elaborated on this point and showed that as more genomes are sequenced, fewer and fewer genes turn out to be possessed by all of them. Large-scale genomic projects have demonstrated an unexpected level of diversity among bacteria. This

issue was also addressed by studies presented by Camilla Nesbo (Dalhousie University) and Fraser, regarding differences within the genus *Thermotoga*, where genomic diversity seems to be driven by the ecotype and the substrates available for bacterial growth, and not so much by the species to which the bacterial strains belong. LGT has been shown to play an important role in shaping genomes in the Thermotogales, and therefore appears to be an extraordinary tool contributing to this genomic diversity and the acquisition of new genes and functions.

The field of microbial genomics continues to develop, and the explosion of sequence data (see Figure 1) continues to highlight how much diversity there is in the environment. This can only be expected to increase, as sequencing costs reduce, technologies improve, and more environments become the subject of metagenomic analyses.