Meeting report

The first decade of microbial genomics: what have we learned and where are we going next?

David A Rasko*† and Emmanuel F Mongodin*

Address: *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. †Current address: Department of Microbiology, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75390-9048, USA.

Correspondence: David A Rasko. E-mail: david.rasko@utsouthwestern.edu

Published: 30 August 2005

Genome ***Biology*** 2005, **6:**341 (doi:10.1186/gb-2005-6-9-341)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/9/341

A report on the International Conference on Microbial Genomics, Halifax, Canada, 13-16 April 2005.

It is now a decade since the first microbial genome was sequenced. Although genomics is still in its infancy and the best is (hopefully!) still to come, amazing strides have been made since the completion in 1995 of the first genome sequence of a free-living organism, the bacterium *Haemophilus influenzae*. Just ten years later, 261 microbial genomes have been completed and an additional 669 are in progress. We have progressed from sequencing a single bacterial isolate, assuming that it was an adequate reference for that species, to metagenomics - sequencing an entire microbial community. We are just starting to discover the complexity and dynamic nature of the microbial world, which raises further questions. For example, what is a bacterial species? How many isolates need to be sequenced to capture the diversity of a single species? During the course of the recent International Conference on Microbial Genomics held in Canada, the question of "what is a bacterial species" was raised and discussed on many occasions. As pointed out by W. Ford Doolittle (Dalhousie University, Halifax, Canada), the notion of a bacterial species is classically defined as a "uniform and stable way for naming groups of similar bacteria". On the genetic level, it is well accepted that two isolates are part of the same species if their 16S rRNA genes share at least 98% identity. This definition is not, however, a good predictor of ecological and phenotypic differences. Furthermore, recombination and gene transfer among prokaryotes, as revealed by genomic, and more recently metagenomic, studies, create further difficulties in describing a microbial species. The concept of a bacterial species appears to take different forms depending on the scientific perspective. Genomic and clinical examinations of *Escherichia coli* and *Shigella* species clearly reveal significant differences, leading to subclassification based on gene content and disease presentation; comparison of the 16S rRNA sequences, however, clearly indicate that *E. coli* and *Shigella* are the same species.

In his talk, Doolittle discussed the species concept in relation to genomic data. He pointed out that while many people had felt that genomics would clarify the species concept in prokaryotes, it has actually done the exact opposite and made it harder to define. Large-scale genomic projects have identified an unexpected level of diversity among bacteria, which can often be linked to recombination and gene transfer between a variety of prokaryotic organisms. Thus, the use of reproductive barriers as a method of speciation in bacteria cannot be supported. Doolittle noted, however, that bacteria will fall into natural groups or clusters depending on the environment, the availability of other organisms with which to exchange DNA, and how readily each organism accepts the exchange of DNA. The concept of a 'species' was acknowledged to be necessary for comparative purposes; nevertheless, it probably does not have any reality at the level of the genome.

In her keynote presentation, Claire Fraser (The Institute for Genomic Research (TIGR), Rockville, USA) highlighted work at TIGR, starting from the genome of *H. influenzae* in 1995 to the current projects, one of which is to determine the number of genomes that need to be sequenced in order to assess the variability within any given species. It is clear that a species is not adequately represented by a single genome unless the species is evolutionarily young and relatively monomorphic. In the more diverse species, it seems as though each individual genome provides some unique information. The number of unique regions gets smaller with each genome sequenced, until a point of diminishing

returns is reached. This point appears to be unique to each species. According to James Tiedje (Michigan State University, East Lansing, USA), 13-15 genomes per species need to be explored to get 95% of the species gene pool, assuming that the strains chosen adequately represent the ecological diversity of the species. But there are exceptions, depending on the level of diversity (ecological niches, pathogen or non-pathogen, and so on) within a single species.

Metagenomic reconstruction has been taken to another level by Denis Le Paslier (Genoscope, Evry, France) using an iterative assembly process that uses cosmid sequencing data as a seed for building genome assemblies. This process has the advantage of being able to assemble larger and larger DNA fragments until a genome is complete or close to complete. He described how this approach led to the assembly of the genome of a virtual organism, suggested to be a free-living Gram-negative bacterium, with a 2.25 megabase (Mb) genome containing two rRNAs and 45 tRNAs. This method appears to be a promising way of assembling large genomic regions from organisms that cannot be cultured.

Eddy Rubin (US Department of Energy Joint Genome Institute (JGI), Walnut Creek, USA) described some of the metagenomic sequencing projects ongoing at JGI. One is a study comparing high- and low-nutrient environments: Wisconsin farm soil and Iron Mountain acid mine drainage, respectively. The results show that the high-nutrient environment (Wisconsin farm soil) contains many more species than the low-nutrient environment. This breadth of species diversity makes it difficult to assemble DNA shotgun fragments into large contiguous pieces, resulting in an inability to identify the dominant species. Rubin also described another JGI metagenomics project, which is studying deep-sea whale-fall regions, where whale carcasses have sunk to the sea floor. These environments are rich in lipid, and DNA encoding metabolic processes could be identified in samples that were geographically distinct but had similar nutrient content. In particular, two whale-fall regions separated by more than 8,000 miles contained similar functional genomic profiles when metagenomic data was analyzed using clusters of orthologous groups (COGs). As Rubin pointed out, identification of a functional process in a metagenomic project may lead to the recognition and study of a factor that was not previously examined in this environment. These functional identifications and sequence distributions could also be used as 'environmental genomic tags' (or EGTs, by analogy with ESTs, expressed sequence tags) that are representative of a particular environment.

Lindsay Eltis (University of British Columbia, Vancouver, Canada) highlighted further the functional genomic work that can take place once a genome has been sequenced. His work on *Rhodococcus* sp. RHA1, whose 9.7 Mb genome is composed of a linear chromosome (7.8 Mb) and three linear plasmids, raises the question of why this genome is so large, as there appears to be no obvious biological reason. The genome does not contain a large number of repeated elements, but does have genes for more than 25 non-ribosomal peptide synthetases and seven polyketide synthases, which tend to be large genes (more than 25 kb long). Interestingly, *Rhodococcus* RHA1 has never been shown to produce the products of these genes or the products of the enzymes' action, which are often biologically active compounds of pharmaceutical interest such as antibiotics and other drugs. In contrast, genes from *Streptomyces* have been shown to be expressed when introduced into *Rhodococcus* RHA1.

The tick-borne bacterial pathogen of cattle, *Anaplasma marginale*, can undergo significant antigenic variation. During an infection, bacteria expressing variants of a major surface antigen emerge. Guy Palmer (Washington State University, Pullman, USA), moving further down the path from sequence to function, discussed the unique method of variation employed by this pathogen. The small genome size (1.2 Mb) and the lack of plasmids or phage rule out antigenic variation by the recombination of complete pseudogenes from other genomic locations. This lack of extrachromosomal material suggests that the antigenic variation would have to come from within the existing genetic material. A number of short pseudogene segments were identified within the genome. It is these small segments that can recombine with the functional gene to create the antigenic variants. The accumulation of these recombination events over the course of an infection leads to increased antigenic presentation and the establishment of a low-level chronic disease.

The first decade of the genomics era has revolutionized our understanding of microbiology, and it is very likely that this process will accelerate, as new technologies are being developed that allow even more rapid generation of genomic data, which in turn will open more avenues of research. We are, however, currently only taking snapshots, not yet making movies. The challenge of the next decade will be to string all these pictures together, to really appreciate the complexity and the dynamic nature of the exchanges that are taking place in the microbial world and their functional implications.