



ELSEVIER

Opening the pan-genomics box

Editorial Overview

Timothy D Read and David W Ussery

Current Opinion in Microbiology 2006, 9:496–498

Available online 1st September 2006

1369-5274/\$ – see front matter

Published by Elsevier Ltd.

DOI 10.1016/j.mib.2006.08.010

Timothy D Read

Biological Defense Research Directorate, Naval Medical Research Center, 503 Robert Grant Avenue, Silver Spring, MD 20852, USA
e-mail: readt@nmrc.navy.mil

Tim Read founded the Genomics group at the BDRD in 2003. The group performs high-throughput sequencing (using the 454 Inc GS20 sequencing technology) and microarray-based resequencing of biodefense pathogens and their near-neighbors. These data form the basis for functional genomics investigations, including vaccine target discovery. Previously, Dr Read was a faculty member at the Institute for Genomics Research (TIGR) in Rockville, Maryland.

David W Ussery

Center for Biological Sequence Analysis (CBS), BioCentrum-DTU, Technical University of Denmark, DK-2800 Lyngby, Denmark
e-mail: Dave@cbs.dtu.dk

Dave Ussery has been the Comparative Microbial Genomics group leader at CBS since 1998. His research focus is on developing high-throughput computational methods for comparison of bacterial genomes, and the role of DNA structures in chromosome architecture. His group also maintains the 'atlas web pages', which mainly focus on sequenced bacterial genomes (<http://www.cbs.dtu.dk/services/GenomeAtlas/>).

Overview

The six articles in this section center on and around the movement of prokaryotic genomics toward a more population-based science. In the early days of genome biology, when per base sequencing costs were relatively high, the emphasis was on the acquisition of a limited set of model organism sequences. The first bacterial project, *Escherichia coli*, was funded partially as a proof of concept for the human genome project. Since the application of the shotgun sequencing approach to whole bacterial genomes more than 10 years ago, the pace and scope of sequencing has increased, to the extent that more than a thousand diverse prokaryotic genomes have now been undertaken and can be searched against in public databases for a given protein or DNA sequence of interest. Now the next generation of ultra-low cost sequencing (ULCS) technologies [1], just starting to be employed, have opened the door to the sequencing of many genomes from the same species [2–4] in order to obtain a much better sampling of the gene population. This new concept has been called 'Pan-Genomics' [5] and is already starting to help us understand individual bacterial genomes in the context of their species [4,5].

Given this background, the first article by Field *et al.*¹ outlines some of the computational tools and approaches for large-scale microbial genome comparison. There are actually many tools available for comparative genome analysis; in our opinion, this article is a good place to start for those traditional microbiologists wanting to try and make sense of their newly sequenced genomes. Many microbiologists are likely to feel a bit overwhelmed with the explosion of genomic information, and many are either not aware of the available bioinformatics tools, or do not find them intuitive to use. There is a need for a concerted effort to build interfaces between comparative genomics data and the bench scientist. In addition to the tools used to compare genomes, it is now becoming evident that a complete genome sequence is not the same as a single gene entry in GenBank. In this regard, we need to learn more about the biology of sequenced organisms. These genome sequences should be considered as a valuable resource, and work is now being done to try and obtain a set of minimal information for each genome sequence in order to have the data for comparisons. With the possibility of thousands of genomes being sequenced, it makes sense for the people sequencing the genomes to report at least the conditions under which the respective organism was grown and the genomes isolated.

Of course, in order to compare hundreds of genomes, there should be some sort of consistent gene annotation, which is the subject for the article by

¹ As editors, we also wanted to have a similar discussion of methodology for environmental bacterial DNA sequencing ('metagenomics') as a companion for this article, but it was the opinion of the expert asked that there were currently more metagenomics review articles than original research papers; thus this topic is best saved for a later issue.

Stothard and Wishart. Most annotations submitted with sequencing papers use a combination of automated annotation with (more or less intensive) human curation. ‘Semi-manual methods’ for comparative analysis of many similar genomes are also available that can yield conserved genes and regulatory elements not readily found in a single genome [6]. **Field *et al.*** articulate the need for a standard, universally accepted mark-up language for bacterial annotation (Minimal Information about a Genome Sequence). One could imagine uploading a target list of genome sequences and have the computer to perform annotation using geographically dispersed web services that can communicate and output results through this shared language. What would be the role for human annotators in this process? Certainly, in collating high-level information about large groups of related genes and also perhaps for quality control of individual genome annotations. What is quite clear is that is that prokaryotic genome annotation in the future cannot be done in the same way that it has in the past, where it has taken several years to annotate a genomic DNA sequence that could now be generated in an afternoon using ULCS technologies.

Part of the annotation process involves determining the function of genes, and how they are regulated. Thus the subject of the article by **Luscombe and colleagues** is the transcriptional regulatory networks in bacteria, looking at the whole process from input signals to output responses. In the past decade, in tandem with sequencing of microbial genomes, there has been the development of network methodologies which enable the description of genes within a given organism in terms of interacting networks of proteins. For example, it was only at the end of the 1990’s that two key articles were published (each of which have been cited more than a thousand times!) establishing the foundations for current network analysis of biological systems [7,8]. From our perspective, one interesting point from this third article is how much we don’t know — even for *E. coli*, the most studied model bacteria, regulatory information in the current databases is lacking for roughly 50% of the transcription factors. Thus, even though the possible function has been predicted for ~96% of the genes in *E. coli* K-12 [9], we still have a long way to go in terms of gathering enough data about even *E. coli*. This is in context of the fact that there are now at least 20 different *E. coli* genomes that have been sequenced (see Table 1 in [10]). The first version of the *E. coli* pan-genome will include a core set of about 2000 genes, as well as another roughly 8000 or so different *E. coli* genes which are found in some but not all strains.

In addition to transcription factors, gene expression can be controlled by chromatin. The article by **Sandman and Reeve** focuses on archaeal histone proteins. Although likely to be equally as abundant as bacteria, archaeal genomes have so far been vastly under-represented in

terms of the genomes being sequenced: at the current count as of 27 June, 2006, 27 archaeal genomes against 327 bacterial genomes sequenced. Having said that, it is worth noting that the authors use metagenomic information from the Sargossa Sea to obtain information about Crenarchaeal histone sequences. Also, in keeping with the theme of pan-genomes, recently the pan-genome of the halophilic archaeon ‘*Haloquadratum walsbyi*’ has been published [11]. Furthermore, the sequence of a methane-producing archaea found in rice fields was recently obtained from metagenomic data [12]. It is our hope that these papers are indications of more archaeal genomes to come.

The article by **Siguier *et al.*** deals with insertion sequences (ISs) in prokaryotic genomes. ISs are a key feature of archaeal and bacterial genomes, and differences in the load and composition of these elements can differentiate very closely related strains. ISs, and related elements such as miniature inverted repeat transposable elements (MITEs), might play a major role in reductive evolution of bacterial strains and acquisition of genes through lateral transfer. However, as **Siguier *et al.*** point out, ISs are frequently mis-annotated or even missed completely if the genes contain frameshifts. In this regard, annotation seems to be lagging behind that of other key genes.

Coming back to the problem-solving aspect of microbiology, in the final article **Rino Rappuoli** outlines how reverse vaccinology (genome-based vaccine target filtering) has been updated by the new population genomics. It is clear that just having one genome sequence for a pathogen such as *Streptococcus agalactiae* was not enough to develop an effective vaccine. However, once several more *S. agalactiae* genomes were sequenced, and thus at least a start towards the pan-genome for these species, a universal vaccine could be developed [13].

One of the interesting results of this explosion of genomics is that microbiology is poised to rediscover its past but with much greater richness of information. Researchers are reconsidering top-down approaches to microbiological questions [14] after going through more than 50 successful years of reductionistic molecular biology. To do this we need to reconsider prokaryotic population biology, a discipline that has lagged behind human and vertebrate population biology. The irony here of course is that we would have come a full circle in genomics — using the human as a model organism to improve our understanding of prokaryotes.

Acknowledgements

TDR is funded by grants from the Defense Threat Reduction Agency, and DU is funded by a grant from the Danish Center for Scientific Computing.

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the US Government.

References

1. Shendure J, Mitra RD, Varma C, Church GM: **Advanced sequencing technologies: methods and goals.** *Nat Rev Genet* 2004, **5**:335-344.
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y, Chen Z *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**:376-380.
3. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science* 2005, **309**:1728-1732.
4. Tettelin H, Massignani V, Cieslewicz MJ: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'.** *Proc Natl Acad Sci USA* 2005, **102**:13950-13955.
5. Medini D, Donati C, Tettelin H, Massignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15**:589-594.
6. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241-254.
7. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393**:440-442.
8. Barabasi AL, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286**:509-512.
9. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T *et al.*: ***Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005.** *Nucleic Acids Res* 2006, **34**:1-9.
10. Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, La T, Hampson DJ, Bellgard M, Wassenaar TM, Ussey DW: **Ten years of bacterial genome sequencing: comparative-genomics-based discoveries.** *Funct Integr Genomics* 2006, **6**:165-185.
11. Legault BA, Lopez-Lopez A, Alba-Casado JC, Doolittle WF, Bolhuis H, Rodriguez-Valera F, Papke TR: **Environmental genomics of '*Haloquadratum walsbyi*' in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species.** *BMC Genomics* 2006, **7**:171.
12. Erkel C, Kube M, Reinhardt R, Liesack W: **Genome of Rice Cluster I archaea—the key methane producers in the rice rhizosphere.** *Science* 2006, **313**:370-372.
13. Maione D, Margarit I, Rinaudo CD: **Identification of a universal Group B streptococcus vaccine by multiple genome screen.** *Science* 2005, **309**:148-150.
14. Falush D, Bowden R: **Genome-wide association mapping in bacteria?** *Trends Microbiol* 2006, **14**:353-355.