Metagenomics and beyond: new toolboxes for microbial systematics



Artem Men, Sue Forrest & Kirby Siemering

Australian Genome Research Facility
University of Queensland
Level 5, Gehrmann Laboratories
Research Road
St Lucia, Qld 4072
Tel (07) 3365 4448
Fax (07)3365 1823
Email artem.men@agrf.org.au

An extraordinary DNA sequencing revolution has taken place over the past decade, which has seen exciting, yet challenging times for microbial genomics and systematics. Numerous metagenomics and metatranscriptomics projects have provided us with an unprecedented glimpse at the vast biological diversity that exists in minute amounts of samples obtained from environments such as ocean water, soil or human distal gut. One of the key challenges is how we catalogue and classify this vast diversity of microbial life (much of which represents unculturable mixtures) discovered in the last few years alone. Of even greater challenge is the fact that biological mechanisms that rule bacterial plasticity and ecological fitness are far more complex than previously thought, resulting in new concepts of 'pan' or 'supra-genomes' that appear to be much larger than any individual bacterial genome.

For any new technology adopted in the field of biochemistry and genomics, microbes have always been 'proof-of-principle' organisms. The first pre-DNA techniques, like protein serotyping and detection of capsular polysaccharides were later strengthened by multilocus enzyme electrophoresis (MEE), a technique based on interrogating isoform patterns of over a dozen enzymes[1]. Microbial systematics later adopted DNA-based methods, such as restriction analysis and DNA-DNA hybridisation. The latter was further developed into fluorescent *in-situ* hybridisation (FISH) microbial tests and then into more specific applications, such as recognition of individual genes-FISH, or RING-FISH[2].

With the arrival of PCR coupled with Sanger sequencing, a robust bacterial genotyping platform was built on comparisons of conserved 16S rRNA loci[3]. Today, this methodology still serves as a major workhorse in microbial systematics, providing good phylogenetic resolution between, and often within, species. Currently about 99% identity between two 16S rRNA sequences is considered a threshold for a species (or phenotypic cluster). This level of resolution, however, is often insufficient, especially when distinguishing microbes within species, so it is often accompanied by other PCR/sequencing-based methods, like multilocus sequence typing (MLST). This technique interrogates ~450 bp amplicons derived from a set of (usually seven) conserved housekeeping genes[4].

MLST is far from perfect, as some pathogens, like *Bacillus anthracis* and *Salmonella,* appear to have conserved MLST sets making it impossible to subgroup them. If polymorphisms in virulence genes and pathogenicity islands do not correlate with MLST or 16S rRNA profiles, researchers have to develop a more detailed scheme based, for instance, on customised sets of single nucleotide polymorphisms (SNPs)[5]. Yet probably the biggest challenge for modern microbiology resides in the world of functional genomics, that is, in understanding the 'real biology' of a given microorganism, especially if it comes from a complex environmental sample and belongs to the 99% of microbes that are unculturable (or at least yet-to-be-cultured) in the laboratory. Even if 16S rRNA or MLST analysis could determine the phylogenic niche at an acceptable rate, the link between genomic information and biology or ecological function for many microbes remained elusive until entire genome sequencing became possible.

The genomics landscape changed dramatically when high-throughput sequencing of entire genomes became possible. In 1995, the first bacterial genome, 1.8 Mb *Haemophilus influenzae*, was published[6]. This tiny pathogen confirmed the power of concepts developed along the way, like shotgun cloning, paired-end Sanger sequencing and laboratory automation that would go on to revolutionise modern biology. Another success factor was the implementation of small-insert libraries to avoid dealing with fully functional genes of *H. influenzae* that could be unclonable

in *E. coli*. The genome of *H. influenzae* was produced with early fluorescent chemistry, a dozen slab gel AB 373 machines and the first version of TIGR Assembler that put 24,304 sequenced pieces together. Using a similar approach, hundreds of microbial genomes were sequenced in the following years.

A decade ago, prototypes of 'next-generation' sequencing platforms no longer based on the traditional Sanger method were developed by companies like 454 and Solexa, and then became commercially available in 2005 (for review on second-generation sequencing and its applications to microbial genomics, see[7]). Immediately put to use for numerous genome sequencing projects, many of them microbial, these new platforms raised the field to unseen levels in terms of acquired sequencing data. However, they also brought new challenges to bioinformatics and annotation teams, particularly for those projects that dealt with complex metagenomics samples.

Lessons from early metagenomics experiments

In the late 1980s Jo Handelsman and several other groups commenced work in a field of study later termed as 'metagenomics' ('beyond the genome')[8]. Metagenomics first started by generating cloned libraries from environmental samples followed by 16S rRNA sequencing. At that time, the success was measured by the efficiency of cloning itself, that is, the ability to grow a reasonable number of clones produced in *E. coli*, by introducing environmental DNA into bacterial artificial chromosomes, cosmids or fosmids[9]. This approach led to comprehensive studies in the last decade, whereby numerous environmental samples were analysed by combining PCR and sequencing of conserved loci. Fascinating results were obtained on soil and marine populations, as well as human-related meta-microbiomes, such as distal gut, skin and different hospital environments[10-12].

Later, the J Craig Venter Institute applied the same principle used to sequence *H. influenzae*, namely computer assembly of shotgun Sanger reads, to a complex microbial sample collected from the Sargasso Sea[13]. In this ground-breaking experiment about a billion bases (giga base, or Gb) of non-redundant sequence was generated. The institute also sequenced samples from the Atlantic and Pacific oceans, placing almost eight million reads (6.3 Gb) into public databases[14]. This pioneering experiment not only produced extensive data, but showed just how little is actually known about abundance and diversity of microbial life in general. For example, in the Sargasso Sea sample alone, over 1,200 unknown genes were catalogued, about one unknown gene per each kilo base of sequenced DNA. This and subsequent experiments illustrated the need for new bioinformatics tools and easily accessible databases for assembly and annotation of microbial genomes, thus creating the ability to assign an ecological role to every piece of genomic information obtained from metagenomic samples.

Are we really in the post-genomics era yet?

The amount of novel sequencing data accumulated for microbes is second to none. Undoubtedly human genome sequencing still attracts major worldwide attention. However, much of the latter is dedicated to re-sequencing, with either a clinical or population flavour, and doesn't produce much novel genomic data *per se*. In contrast, every sequenced microbe reveals about 30% of novel sequence at the genus or species level, a quite remarkable influx of genomic information. This is being taken even further by next-generation sequencing capabilities where a single bacterial genome can be sequenced for a mere couple of hundred dollars (unless elaborative finishing processes are required). Today, over 200 Gb of sequence can be generated on each run of an Illumina HiSeq 2000TM machine. Together with sample bar coding, 192 single bacterial genomes the size of *E. coli* can be sequenced to 200x depth in just over a week!

Nevertheless, when it comes to complex microbiomes, even these sequencing capabilities are still painfully unsatisfactory. For example, one litre of marine sample contains on average 10^9 bacterial cells. For an average bacterial genome size of $5x10^6$ bases, it would need to produce $5x10^{15}$ bases just for 1x statistical coverage of that DNA. Based on the current coverage required for a human re-sequencing project (about 40 times), $2x10^{17}$ bases, or a million HiSeq 2000TM runs would be needed to sequence 10^9 individual bacterial cells to that depth. This is an enormous task even by current next-generation sequencing standards; however, the first steps in this direction have been made already. Recently an Earth Microbiome project has been initiated in order to sequence 10^{15} bases from 160,000 locations in the next three years[15].

Smart sampling the key

Today, genome sequencing can be performed quickly and efficiently, and many laboratories will now be able to afford a few giga bases, through grant funding or collaborations. On the back of next-generation sequencing technology, two questions do remain; first, why should we sequence one genome when we can sequence a hundred, and the second, is it really worthwhile to sequence a hundred if we can't make sense of the giant data set obtained? For many large metagenomics studies, careful planning starts with sample preparation. Not only thousands of genes of unknown function could be the result, but the analysis is further complicated by bacterial plasticity, environmental adaptation, and by the exchange of genomic content between species, such as horizontal gene transfer. Researchers now use concepts of 'core', 'dispensable' and 'pan-genomes', which describe the ability of microbes to exchange individual genes or even 'gene islands' among species isolates[16].

One way of dealing with metagenomics samples is to reduce its complexity prior to sequencing, through 'filtering' the

population through substrate-specific selection or by means of genome partitioning. A few groups came up with enrichment techniques for capturing cells that are biologically active in the presence of a particular chemical (substrate). For example, the use of methanol labelled with ^{13}C and rolling circle amplification allowed the separation of newly synthesised DNA containing ^{13}C, cloning it in fosmids and sequencing it to study microbes involved in methanol metabolism[17]. Another effective method is immunocapturing, where bacterial samples are fed with bromodeoxyuridine (BrdU) which actively incorporates into newly synthesised DNA. This DNA can be immunocaptured with anti-BrdU antibodies, thus providing a new sample of reduced complexity that corresponds to cells effectively dividing in the presence of the substrate[18].

If the study is focused on a particular set of genes which, for example, exhibit a unique function or a biochemical pathway, it is worthwhile looking at 'genome capture' technology widely used in human and mouse genomics for genome partitioning[19]. Capturing genomic regions of interest, like prokaryotic "fitness islands"[20], that is, sets of physically packed open reading frames (ORFs) related to pathogenicity or symbiosis, through oligonucleotides "baits" could be a very effective tool to analyse mixed samples. To the best of our knowledge, there is no literature yet describing the combination of microbial genome capture and next-generation sequencing, but it is potentially a very beneficial approach to study subsets of microbial genomes, for example virulence ORFs in closely related pathogens from complex clinical samples (Figure 1). Currently existing in 'on-chip' and 'in solution' forms, genome capture employs DNA or RNA oligonucleotides that are complementary to a genomic region of interest. Oligos can be designed against known or predicted gene clusters and then used to 'capture' sheared genomic DNA sample. The captured DNA is then amplified for library preparation and sequenced on a next-generation sequencing machine (Figure 1). Currently a single capture in humans and mice covers 1–30 Mb of sequence. As bacterial genome sizes vary between 0.15 Mb (*Candidatus Carsonella ruddii*) and 10 Mb (for *Solibacter usitatus*), a 30 Mb capture can be overkill for prokaryotes, but this can be compensated by more sophisticated probe design. By introducing highly polymorphic (degenerate) sets of capturing oligos based on multiple sequence alignments, maximal amounts of regions of interest can be captured from a complex microbial sample in a single experiment and sequenced to the depth of tens of thousands in search for rare variants.

Conclusion

Metagenomics is driven by many factors, one of which is commercial return. Many pharmaceutical companies are on the search for new bioactive molecules produced by microbes. The search for new genes and pathways that encode novel enzymes, hormones and other metabolites has potentially far-reaching

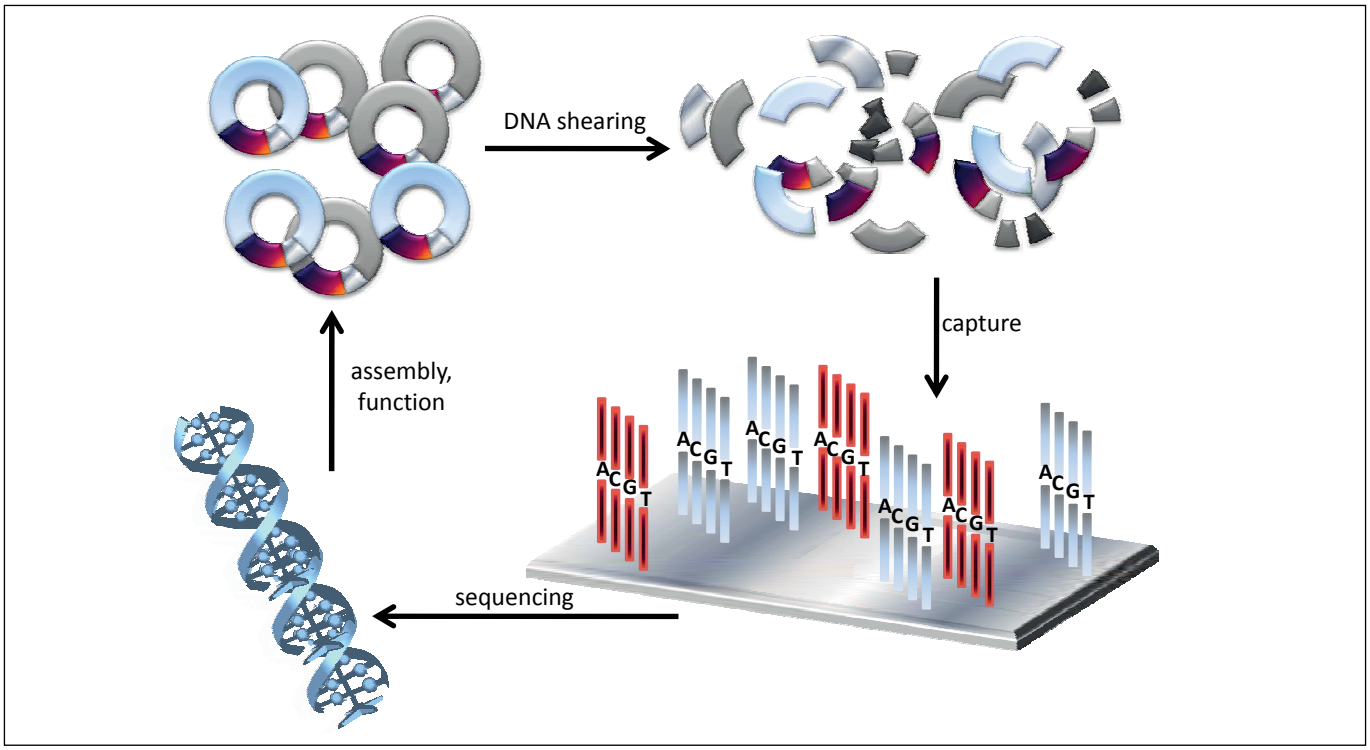

Figure 1. Potential application of 'genome capture' technology for analysis of microbial communities. Firstly, total DNA is isolated from a microbial sample, sheared and prepared for next-generation sequencing by polishing and ligating library-specific adaptors. Then the DNA sample is hybridised to single-stranded (DNA or RNA) oligonucleotide probes or 'baits' that capture region(s) of interest. Captured DNA is eluted, amplified and sequenced to appropriate depth. This technology could be particularly useful for studying genes which encompass a unique biological function or a pathway, or are located in compact genomic regions, such as 'fitness islands' (tightly linked groups of genes, usually 10–200 kb long[20], responsible for pathogenicity, secretion functions or symbiosis).

implications for the medical and research industries and the burgeoning area of synthetic biology[21].

The Australian Genome Research Facility (AGRF) is currently engaged in numerous projects dedicated to microbial genomics, and recognises the potential that microbial systematics brings to research opportunities. Some current platforms include individual 16S rRNA and MLST microbial typing pipelines, both based on PCR and Sanger sequencing. AGRF has been employing next-generation sequencing since 2006 for microbial genome sequencing and metagenomics, and currently runs a fleet of seven instruments capable of producing 70 Gb of sequence per day.

With rapidly increasing sequencing throughput and third-generation machines, such as Pacific Biosciences SMRT instrument due to hit Australian shores in the coming months, it is not unrealistic to imagine projects where terabase-scale single-pass sequencing experiments will be complemented by new methodologies equipped for finishing stages and genome assembly, such as optical mapping[22] and subassembly library construction[23]. In this aspect, the read length of a single sequencing run becomes increasingly important, especially for metagenome samples. For example, it was shown that up to 72% of BLASTX hits can be completely missed from short read data sets, for example when comparing classical Sanger reads with ones produced by FLX pyrosequencing, the longest reads for second-generation sequencing platforms[24]. An exciting development for projects where 'length matters' is the Pacific Biosciences SMRT machine, which has reported runs reaching beyond 2 kb of contiguous sequence, albeit according to early users the sequence accuracy (~81–83%) is still to be improved[25]. Additionally, coupling third-generation sequencing with methods which can reduce complexity of metagenomics samples (such as single genome amplification[26]) will undoubtedly help assemble full genomes of yet unculturable organisms. In any case, it is clear that new sequencing technologies that have already revolutionised genomics and are currently driving human genome costs to the $1,000 mark, combined with adequate bioinformatics resources will soon bring microbial systematics to previously unseen levels of understanding microbial ecosystems.

References

1. Caugant, D.A. *et al.* (1988) Clonal diversity of *Neisseria meningitidis* from a population of asymptomatic carriers. *Infect. Immun.* 56, 2060–2068.
2. Zwirglmaier, K. *et al.* (2004) Recognition of individual genes in a single bacterial cell by fluorescence in situ hybridization – RING-FISH. *Mol. Microbiol.* 51, 89–96.
3. Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev.* 51, 221–271.
4. Spratt, B.G. (1999) Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the Internet. *Curr. Opin. Microbiol.* 2, 312–316.
5. Read, T.D. *et al.* (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* 296, 2028–2033.
6. Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512.
7. MacLean, D. *et al.* (2009) Application of 'next-generation' sequencing technologies to microbial genetics *Nat. Rev. Microbiol.* 7, 287–296.
8. Handelsman, J. *et al.* (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, 245–249.
9. Schmidt, T.M. *et al.* (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* 173, 4371–4378.
10. Gill, S.R. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359.
11. Coque, T.M. *et al.* (2002) Genes encoding TEM-4, SHV-2, and CTX-M-10 extended- spectrum beta-lactamases are carried by multiple *Klebsiella pneumonia* clones in a single hospital (Madrid, 1989 to 2000). *Antimicrob. Agents. Chemother.* 46, 500–510.
12. Dethlefsen, L. *et al.* (2007) An ecological and evolutionary perspective on human– microbe mutualism and disease. *Nature* 449, 811–818.
13. Venter, J.C. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74.
14. Rusch, D.B. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 5, e77.
15. Gilbert, J.A. and Dupont, C.L. (2011) Microbial Metagenomics: Beyond the Genome *Annu. Rev. Mar. Sci.* 3, 347–371.
16. Tettelin, H. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl Acad. Sci. USA* 102, 13950–13955.
17. Neufeld, J.D. *et al.* (2008) Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environ. Microbiol.* 10, 1526–1535.
18. Mou, X. *et al.* (2008) Bacterial carbon processing by generalist species in the coastal ocean. *Nature* 451, 708–713.
19. Kiialainen, A. *et al.* (2011) Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery. *PLoS ONE* 6, e16486.
20. Langille, M.G.I. *et al.* (2010) Detecting genomic islands using bioinformatics approaches *Nat. Rev. Microbiol.* 8, 372–382.
21. Antunes, L.C.M. *et al.* (2011) Mining bacterial small molecules. *Scientist* 25, 26–30.
22. Aston, C. *et al.* (1999) Optical mapping and its potential for large-scale sequencing projects. *Trends Biotechnol.* 17, 297–302.
23. Hiatt, J.B. *et al.* (2010) Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* 7, 119–122.
24. Wommack, K. E. *et al.* (2008) Metagenomics: read length matters. *Appl. Environ. Microbiol.* 74, 1453–1463.
25. Chin, C.-S. *et al.* (2011) The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* 364, 33–42.
26. Woyke, T. *et al.* (2009) Assembling the marine metagenome, one cell at a time. *PLoS ONE* 4, e5299.

Biography

Dr Artem Men completed his PhD in 1995 in Saint Petersburg, Russia. His thesis was on the molecular genetics of the symbiotic relationship between legumes and Rhizobia. He then moved to the University of Tennessee, Knoxville, and in 2000 to the University of Queensland, Brisbane, where he continued his work on mapping and cloning plant symbiosis genes. He joined AGRF in 2003 as a Senior Scientist and currently specialises in DNA sequencing, microbe typing, SNP analysis and laboratory automation. He has published more than 20 research articles and book chapters.