Journal of
Microbial & Biochemical Technology

Altermann, J Microbial Biochem Technol 2012, 4:6
DOI: 10.4172/1948-5948.1000e111

Editorial Open Access

A Return to Microbial Genomes in the Metagenome Age

Eric Altermann*

Rumen Microbiology Team, AgResearch Limited, Grasslands Research Centre, Tennent Drive, Private Bag 11008, 4442 Palmerston North, New Zealand
Riddet Institute, Massey University, Palmerston North, New Zealand

Metagenomics has broadened significantly our understanding of microbial ecosystems, their underlying phylogenetic diversity and genetic complexity. In the course of only a few years microbial genomics has seen a dramatic rise from the 1.8 Megabase pair (Mbp) genome of the first free-living organism sequenced (*Haemophilus influenzae* Rd in 1995 [1]) to (meta) genome programmes now generating more than a Terabase pair of sequence data each. These advances have been made possible by increasingly more powerful sequencing technologies. Fluorescent slab-gel electrophoresis methods were replaced by capillary-based systems, which brought a significant increase in the level of throughput and automation. A step change came with the introduction of "sequencing by synthesis". This technology was commercialised as 'pyrosequencing', notably by 454 Life Sciences. While initially providing only shorter read lengths of 100-200 nucleotides (nt), and having a lower base call quality and problems with homopolymeric stretches of nucleotides, it also delivered a leap in sequencing capacity (up to 400 Mbp per run) from capillary Sanger-based sequencing technologies. Since then a number of other next-generation sequencing platforms have been commercialised (such as Illumina, SOLID, Ion Torrent), each increasing the amount of sequence information gained per run (Illumina HiSeq2500 currently delivers up to 600 Gbp per run). While single molecule real time (SMRT) sequencing is still in its infancy, it is likely to be the "next big thing" and prototypes (mainly from Pacific Biosciences) are currently being trialled.

All of the current next-generation sequencing platforms have the common drawback of short individual read lengths. Only the latest generation of pyrosequencing technology approaches the long, high-quality reads (in excess of 1000 nt) of Sanger based technologies. This inherent characteristic of high-throughput, next-generation sequencing technologies has shaped the field of metagenomics over the last 10 years, because it limited the degree of sequence assembly that was possible. Consequently, research concentrated on two main themes: community composition (who is there and in what numbers) and the overall genetic makeup of a microbial community. In 2006 Gill et al. [2] defined the 'superorganism' concept for humans and their individual intestinal microbiomes. The totality of intestinal microbes is seen to augment human metabolism by adding metabolic pathways, and each human and his or her microbes is perceived as one amalgamated unit. More recently, comparative metagenomic studies have been carried out, investigating geographical and temporal shifts in microbial communities [3,4]. Large programmes such as the European metaHIT or the US NIH-funded Human Microbiome Project (HMP) mainly aim to generate gene catalogues from metagenomes (supplemented by reference genomes) to investigate specific hypotheses. metaHIT, for example, focuses on two disorders of increasing importance in Europe, Inflammatory Bowel Disease (IBD) and obesity, whereas the HMP investigates the microbiomes of healthy humans from a range of body regions. While these and other programmes accumulate large amounts of metagenomic data (to date the HMP programme as sequenced over 3.5 Tbp), only a part of those reads can be aligned to reference genomes (the HMP programme aligned 57.6% of metagenomic data to 1,742 microbial genomes [5]).

A discrepancy exists between the amount of raw metagenomic data generated, the lack of coherent assembly, and the missing link to their respective microbial origin. Nelson et al. [6] observed a level of gene content similarity as low as 65% within sequenced reference *Lactobacillus* strains, while the evolutionary relatedness remained over 90%. The observed variability in a closely related group of microbes indicated early on that key genetic elements may be found outside the conserved core genome, and that those elements may subsequently fall outside the detection thresholds and therefore remain hidden in the data noise. A confounding problem might arise if specific or novel phenotypes of a microbial community are mediated by a numerically small subpopulation. In contrast, current metagenome programmes will struggle to identify such strain specific variation in the vast sea of short sequence reads. Huttenhower et al. [7] illustrated this conundrum by comparing the microbiomes of seven body sites for quantitative phylogenetic composition and metabolic modules. While the phylogenetic composition varied drastically across individual samples within a body site and across different body sites, metabolic profiles remained remarkably stable with very little variation across individual samples and sampling sites.

Such interstrain diversity is highly significant for biotechnological, medical, pharmaceutical and functional foods industries. Strain specific activity within the same ecosystem and across closely related or different ones is mediated by genetic changes such as point mutations in genetic regulators, and the uptake (by horizontal gene transfer) or loss of genetic information. These genetic variations can often only be identified and functionally assessed in their full genomic context. Traditionally, whole microbial genomes were attainable only from culturable microbes, which represent a miniscule fraction of the biological diversity found in a complex microbial community. Only very recently, a culture independent single-cell genomic sequencing approach has emerged [8] and been successfully applied to a number of microbial and archaeal genomes. Even so, the low throughput and the high technological challenges imply that this exciting new technology will likely take some time until it reaches mainstream science.

Concurrently, however, sequencing technology has advanced sufficiently to begin assembling whole microbial genomes directly from microbial communities. Examples such as the draft assembly of bacterial species from enrichment cultures from the Tamar wallaby foregut [9], the reconstruction of a marine Euryarchaeota [10], or the genome sequencing of Segmented Filamentous Bacteria (SFBs) [11] prove that when using a focused approach, microbial genomes can be

***Corresponding author:** Eric Altermann, Rumen Microbiology Team, AgResearch Limited, Grasslands Research Centre, Tennent Drive, Private Bag 11008, 4442 Palmerston North, New Zealand, Tel: +64 6 351 8100; E-mail: eric.altermann@agresearch.co.nz

Received October 24, 2012; **Accepted** October 26, 2012; **Published** October 29, 2012

Citation: Altermann E (2012) A Return to Microbial Genomes in the Metagenome Age. J Microbial Biochem Technol 4: xiii-xiv. doi:10.4172/1948-5948.1000e111

recovered from metagenomic datasets. Yet, many hurdles remain, as illustrated in the assembly of the SFB genome, where a number of Single Nucleotide Polymorphisms (SNPs) and short inserts and deletions (indels) were observed in the sequenced SFB-'metagenome'. Even with the massive increase in sequencing capacity, still only the most frequent variants were used to build a consensus genome sequence, discarding the observed variability.

Clearly, the prospect of assembling individual genomes from a metagenome poses formidable challenges which can be divided into technical and biological aspects. Technical hurdles include insufficient sequence coverage of the metagenome, the occurrence of sequencing errors, and the inability of current sequence assemblers to optimally assemble metagenomic datasets. Recently some remarkable progress has been made to address these challenges, and in particular the assembly of metagenomic datasets is currently an active research area [12-15]. On the other hand, biological aspects such as repeat structures that may lead to ambiguity in assemblies, strain and species specific genetic features or relative strain/species frequencies in the microbial community provide challenges but at the same time also hold the most interesting promise of metagenomes. This biological variability must be captured to accurately reflect the genetic diversity and the resulting phenotypic differences in individual members of the microbial community.

Metagenomic datasets cannot trace back an individual sequence read to a specific microbial cell (yet). Any genome reconstruction will therefore not result into one genome, but into a genomic variability space where variants of the same species or strains are collated and form a conglomerated genome (con-genome). These variants will harbour all of the observed genetic variability commonly found in microbial genomes, such as SNPs, indels, insertions, deletions, small and large scale genomic rearrangements, gene duplications and movements of mobile genetic elements. Such con-genomes may be used to unify, simplify, and assess the genetic diversity found within microbial community and to compare equivalent conglomerated genomes from other samples or ecosystems. Con-genomes will therefore facilitate the identification of key genetic elements responsible for defined desirable or harmful phenotypes.

There is no format yet to describe con-genomes, but any format that emerges is likely to consider a range of questions. What should be the minimum frequency of any given genetic variation to be included? Which basic basic layout (use the most common variety versus the most comprehensive genotype as the baseline) is the optimal choice? In which way can the genetic variability be described and how can a set of variants be assigned to an individual 'layer' within the conglomerated genome structure? How can the underlying phylogeny be captured and to what extent does the species and strain concept still hold validity?

Con-genomes may represent the logical next step forward for metagenomics, maximising the amount of information retrievable from such datasets while moving away from science driven by data collection. Instead, comprehensively structured conglomerated genome systems can be employed to advance biomedical and biotechnological hypotheses and applications by understanding genes within a con-genome context.

References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269: 496-512.
2. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.
3. Vaishampayan PA, Kuehl JV, Froula JL, Morgan JL, Ochman H, et al. (2010) Comparative metagenomics and population dynamics of the gut microbiota in mother and infant.Genome Biol Evol 2: 53-66.
4. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, et al. (2012) Human gut microbiome viewed across age and geography. Nature 486: 222-227.
5. Human Microbiome Project Consortium (2012) A framework for human microbiome research. Nature 486: 215-221.
6. Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, Highlander SK, Worley KC, et al. (2010) A catalog of reference genomes from the human microbiome. Science 328: 994-999.
7. Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207-214.
8. Kalisky T, Blainey P, Quake SR (2011) Genomic analysis at the single-cell level. Annu Rev Genet 45: 431-445.
9. Pope PB, Smith W, Denman SE, Tringe SG, Barry K, et al. (2011) Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. Science 333: 646-648.
10. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, et al. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335: 587-590.
11. Kuwahara T, Ogura Y, Oshima K, Kurokawa K, Ooka T, et al. (2011) The lifestyle of the segmented filamentous bacterium: a non-culturable gut-associated immunostimulating microbe inferred by whole-genome sequencing. DNA Res 18: 291-303.
12. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30: 693-700.
13. Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, et al. (2012) De novo assembly of highly diverse viral populations. BMC Genomics 13: 475.
14. Pop M (2009) Genome assembly reborn: recent computational challenges. Brief Bioinform 10: 354-366.
15. Bashir A, Klammer AA, Robins WP, Chin CS, Webster D, et al. (2012) A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol 30: 701-707.