Final Report on "Mathematical, Statistical and Computational Aspects of the New Science of Metagenomics", held at the Isaac Newton Institute: 24 March-17 April 2014

Organisers: Wally Gilks, Daniel Huson, Elisa Loza, Simon Tavaré, Gabriel Valiente,Tandy Warnow
Scientific Advisory Committee: Vincent Moulton, Mihai Pop

Vast numbers of microbes live in close association with eukaryotic organisms. These complex microbial communities, named microbiomes, support and sustain the life of their hosts. Humans depend on microbiomes for nutrition, metabolism and health. Metagenomics is a new discipline, exploiting modern DNA-sequencing technology to study the structure and function of microbiomes. Metagenomic experiments offer unprecedented opportunities for science and industry but generate huge amounts of data and present formidable analytical and computational challenges.

This programme brought together mathematicians, statisticians, computer scientists, bioinformaticians, biologists, biomedical scientists, ecologists and agronomists with research interests in metagenomics. The opening workshop and subsequent discussion meetings were designed to forge new multidisciplinary research teams around problems and techniques. An *Open for Business* networking event, with presentations from business and academia, illustrated an astonishing variety of industrial applications of metagenomics. Several research strands emerged.

Strand 1: *Taxonomic profiling* aims to identify the microbial taxa present in a metagenomic sample and to estimate their relative abundances. This involves mapping *reads* (short pieces of sequenced DNA) from the sample to previously sequenced genomes in reference databases. Initial discussions contrasted the problems of *binning* to identifying the source of each read, and *profiling* to identify the overall microbial composition. A promising new profiling method was developed, inspired by the taxonomic-binning method QUIKR of David Koslicki and a chromosome-painting technique of Daniel Falush. This method was shown to be capable of estimating relative abundances in a metagenomic sample at finer taxonomic levels than previously published methods.

Strand 2: *Reference-free methods.* Of the reads in a metagenomic sample, typically only a small percentage map unambiguously to fully sequenced genomes in reference databases. Analyses that do not involve referencing can give a more complete picture than taxonomic profiling. Approaches based on *k-mers* (subsequences of length k) observed in the sample of reads were studied. The complexity function (the number of unique k-mers as a function of k) gives insight into the diversity of a given set of reads. Additional information is contained within a *De Bruijn graph* of overlapping k-mers (see Figure). Metrics on a multiple-sample De Bruijn graph facilitate cross-sample comparisons, avoiding the difficult task of metagenomic sequence assembly. Connections were drawn between the shape of the complexity function and salient features of the De Bruijn graph.

Strand 3: *Ecological Modelling.* Ecological theory is divided between *niche models* which assume that species are constrained by their environmental niche, and *neutral models*, which assume that fluctuations in community structure are purely stochastically driven. The relevance of niche and neutral effects in microbial communities was debated and ideas for models incorporating both were developed. Unlike standard ecological datasets, metagenomic datasets also contain *phylogenetic* information on the evolutionary relatedness of taxa within the community. Statistical approaches relating microbial community structure and phylogeny were considered, as were statistical mixture models for disaggregating microbial data. Models were developed and applied to two specific metagenomic datasets: one showing how microbial community structure was influenced by fertilizer treatments in a long-term agricultural experiment; the other showing in a longitudinal medical study that the vaginal microbiome is more stable, and less likely to transition to a morbid state, in pregnant women than in non-pregnant women.

Strand 4: *Statistical design*. In addition to experimental-design issues common in biological studies, the metagenomic context presents issues of pooling of samples and the trade-off between sample-size and DNA-sequencing depth. Applications of the principles of statistical experimental design are at an early stage in metagenomic studies, but there are useful parallels with the more established area of RNA-seq experiments. Discussions emphasised the utility of paired and balanced block designs, which are easy to implement in metagenomic studies using DNA-barcoding and multiplexing, and the value of ascertaining prior information on the likely ranges of model parameters.

Strand 5: *Seeking the fourth domain of life.* Animals, plants and fungi all belong to the *Eukaryota* domain of the tree of life. *Bacteria* comprise another domain. A third domain, *Archaea*, was discovered only in 1977; these microbes exist abundantly in soils, the oceans, in the human gut where they aid digestion, and in many other environments. Eddy Rubin's inspirational lecture in the workshop asked: "Is there a fourth domain of life on earth?" If so, traces of it might be found in the huge volumes of DNA-sequence data generated by metagenomic experiments. Thus the search for a fourth domain of life might begin with a search for patterns in metagenomic sequence data that do not fit with existing knowledge. Statistical-bioinformatic approaches to this open problem were discussed, including methods to detect altered genetic codes, abnormally variant gene sequences within the cell's protein-production machinery, and recently discovered large DNA viruses.

Strand 6: *Critical assessment of metagenome interpretation (CAMI).* The interpretation of metagenomic data relies on sophisticated and computationally intensive approaches such as short-read assembly (to reconstruct microbial genome sequences from sampled reads), binning and taxonomic classification. All downstream analyses depend on the accuracy of these initial data-processing steps. Despite tremendous methodological progress in recent years, existing approaches each embody simplifying assumptions, leading to severe limitations and potential inaccuracies in their use. Assessment of these computational methods has so far been *ad-hoc*. Users are thus not well-informed about the limitations of specific methods, and method-developers must expend considerable time and computational resources to identify areas for improvement. To tackle this problem, a new event, CAMI, was proposed to evaluate methods independently, comprehensively and without bias. Discussions centered on suitable performance metrics and requirements for sufficiently realistic simulated benchmark datasets for this event. CAMI will take place as a follow-on to this meeting.

Follow-on meeting: Led by Alex Sczyrba and Alice McHardy, CAMI (see above and http://blogs.nature.com/methagora/2014/06/the-critical-assessment-of-metagenome-interpretation-cami-competition.html) will be held at the Isaac Newton Institute in September 2014, in the format of a one-week hackathon involving multiple developers of metagenome analysis software.

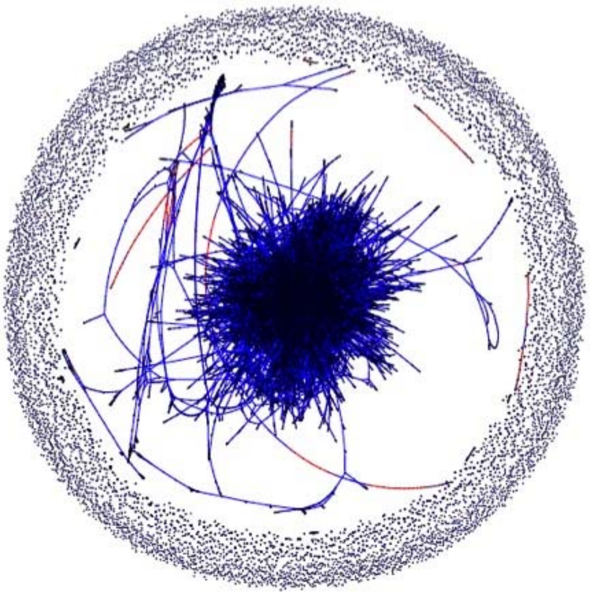

Figure: A De Bruijn graph of two metagenomes, coloured red and blue. Nodes represent distinct 20-mers, and edges are drawn connecting 20-mers that overlap by 19 basepairs (courtesy D. Koslicki).