Inferring microbial species and strains directly from metagenome data

Dr Christopher Quince[1]

[1]Warwick Medical School, University of Warwick

In metagenome sequencing DNA from an entire microbial community is sequenced typically with short reads. The assemblies produced from these studies are usually highly fragmented comprising hundreds of thousands of partial assemblies or contigs. This is an inevitable consequence of intra- and inter-genomic repeats. Only from very simple communities can complete genomes be assembled. However, determining which contigs derive from which species or strain is almost as useful as a complete genome revealing gene complement. Metagenome analyses often comprise multiple samples from longitudinal analysis of the same community over time or horizontal sampling of multiple similar communities. We exploit this in a method, CONCOCT: Clustering cONtigs on COverage and ComposiTion, that combines sequence composition and coverage across multiple samples to automatically cluster contigs into species genomes. CONCOCT uses a dimensionality reduction coupled to a Gaussian mixture model, fit using a variational Bayesian algorithm, which automatically identifies the optimal number of clusters. We demonstrate high recall and precision rates on artificial as well as real human gut metagenome datasets. We then extend this principle, developing a probabilistic model of variant frequencies across samples on core genes within species clusters. These frequencies depend on the relative abundances of strains in each sample and their haplotype. Using a Gibbs sampling algorithm we can use this model to reconstruct the abundances of the strains and their genotypes on the core genes. These genotypes can then be used to determine the phylogenetic relationships between the strains present. Finally, we can apply this information to all the contigs associated with the species to reconstruct the accessory genomes of the different strains. This provides a methodology for de novo extraction of strain genome composition from metagenome analyses that does not rely on long read sequencing.