Bacterial Genome Reconstruction

Matthew D’Iorio and Ken Dewar

As an alternative to *de novo* assembly, this procedure uses Illumina-style short read datasets and (i) identifies the genetically closest reference genome; (ii) calculates the estimated genome coverage in optimally quality trimmed reads; and (iii) aligns optimal reads to the closest reference to create a guided genome assembly that can be used to identify conserved and variant loci. It is primarily aimed to generate guided assemblies for individual bacterial genome sequencing projects and/or to assemble one or more genomes from whole genome microbiome sequencing projects. It is envisioned that the procedure will entail a series of modules, with each module providing reports allowing the user to decide whether to continue or not.

**Module 1: Data acquisition and identification of the genetically closest species.**

Step 1: After uploading a selected Illumina R1 and R2 set of fastq read files (or R1 alone, if paired-end sequencing was not performed), create a file of 1000 random reads from each data set. These files are made available to the user as 1000 read fasta files. The reads in each file are aligned (BLAST) to the NCBI RefSeq database and for each read, any best hit with a score >95% of a perfect score has the species and accession number retained.

Step 2: A summary table is prepared listing the species in order of most to least hits. Given the total number of reads available, and the genome size of identified species, another column is added to show the estimate of genome coverage within the full dataset. A second summary table is prepared listing the numbers of reads at different read lengths, as a result of automated Illumina quality trimming.

**Module 2: Data acquisition and identification of the genetically closest reference genome.**

Step 1: NCBI RefSeq contains a representative genome for every species, even though multiple genomes may have been completely assembled. Using the same set of 1000 random reads as in Module 1, the user will additionally enter the accession numbers of any or all complete genomes for the species (This information can be found in the NCBI Microbial Genomes Database <https://www.ncbi.nlm.nih.gov/genome/microbes/>). The 1000 reads are aligned (BLAST) to each genome and the alignment score is retained.

Step 2: The sum of all 1000 scores for each genome is tabulated, and a summary report is prepared listing genome accession numbers ordered by highest to lowest total score.

**Module 3: Calculating an optimal read length.**

Step 1: Illumina sequencers offer a variety of read lengths, where the sequencing error rate increases towards the end of the read. Empirically determining an optimal trim length and trimming all reads to a fixed length simplifies future analyses. Using the 1000 read set from Module 1 and the highest scoring genome from Module 2, reads are given a range of fixed lengths, as determined by the user according to the sequencing conditions employed. Given a set of read lengths, reads will be trimmed and aligned to the reference genome, and the number of perfectly matching reads will be recorded.

Step 2: For each specified read length a summary table reporting read length and number of perfectly matching reads will be generated. In addition, the product of the two values will be reported. The maximal product value will indicate the optimal length for the data set.

Step 3: This may be an iterative process, where a first analysis chooses read lengths at length intervals, then a second analysis fills in key intervals to select the best read length.