# Appendix 3e: Metagenome Assemblies

## Calculation of metrics describing distributions of contig lengths

In genomics, the N50 is the length in bp of the longest contig such that all contigs of that length or longer account for at least 50% of the base pairs (bp) of the genome. The L50 statistic is the smallest number of contigs whose summed lengths account for at least 50% of the bp of the genome. Whereas in genome sequencing, a perfect assembly would consist of a single contig containing the whole genome (thus N50 = size of genome and L50 = 1), in metagenomics there can be 1 (or more) complete-genome contigs and also a few (or very many) shorter (perhaps many much shorter) contigs.

Therefore N50 and L50 are less meaningful for a metagenome assembly, but do have some utility; a high N50 (and low L50) indicates that the metagenome meets all of these criteria: (i) dominated by a few genomes; (ii) those genomes have been assembled to a high level of completeness; (iii) *and* there are relatively few other genomes well represented (either because there are genuinely few in the microbiome, or because they are present in such low abundance as to be absent from the metagenome sequencing). In a metagenome assembly where the first two criteria are true but the third is not, i.e. there is a long "tail" of low abundance fragments of other genomes, there can be one (or more) complete and perfectly-assembled genomes, yet a low N50 (and high L50).

A related point is that for metagenomes with very uneven relative abundances, a poor overall read count can "truncate" the tail of low-abundance taxa (see (b) above). It may therefore be expected that low read-count, skewed-distribution samples will have quite high N50 and low L50 values.

With these limitations in mind, we have nonetheless determined N50 and L50 values for all samples' assemblies.

## Detection of ARG sequences in assembled metagenomes

The data for assemblies was obtained by using RGI MAIN (RGI in 'main' mode). RGI main in turn uses the Prodigal software (Hyatt et al., 2010) to predict ORFs, which are then compared to the CARD data. Very short ORFs (< 30 bp) are ignored. We ran RGI MAIN with the --low\_quality option ('use for short contigs to predict partial genes'). That is because depending on the outcome of the assembly, many of the contigs may not be long enough to capture entire ARG sequences. We note that the nucleotide sequences in the principal section of CARD range from 162 to 4,359 bp (median 861 bp; mean 946.6 bp). In the 'wildcard' section of CARD, the nucleotide sequences are between 60 bp and 6,237 bp (median 1,194; mean 1,431.9).

We note here that due to the nature of the metagenome sequence assemblies collectively, it was inappropriate to use these assemblies to generate production data of the ARG incidences. Nevertheless, we include the references here and in the other sections to RGI MAIN, because the ORF-prediction itself also provided invaluable information for assessing the qualities of those metagenome assemblies in the first instance.