# Appendix 3d: Unassembled and Assembled Reads

## Unassembled Reads

One approach to analysing metagenomics sequence read data is to assign each sequenced fragment ("sequence read") to a particular category (in this case a particular ARG in a reference database, or to no ARG if negative). This provides proportional frequencies of each category (which need to be treated with caution), relative to the set of input reads. The expectation is of course that the huge majority of reads will be ARG-negative.

Because our read data consists of paired-end (PE) reads (in this case 2 x 150 bp, separated usually by a gap of ~ 250 to 300 bp), ideally each pair of reads is analysed together (this is the case for RGI but not DeepARG - see below).

The main drawback of analysing the (pairs of) reads themselves ("unassembled reads") is that even when treating a PE pair together, the reads are quite short, reducing confidence of comparison with reference sequences (compared to say, whole gene sequences). Some short regions of reference sequences may be indistinguishable from each other (this is especially likely for different but related types of ARGs). These ambiguities may also be true between ARG and non-ARG sequences in some cases. The impact of mis-called bases is also higher for shorter sequences.

* **DeepARG** performs comparison of reads (the set of forward and reverse sequences are treated **independently**) versus references using a Blast-type approach (DIAMOND (Buchfink et al., 2015) which is a high-speed BlastX-analogue, i.e. compares translated DNA sequences versus protein references). The best match for each input sequence is then assigned a probability of correctness. We used the "DeepARG-SS" version of the software, which is optimised for short-read input.
* **RGI** (in "BWT" mode; referred to as "RGI BWT" throughout) uses a "mapping" approach which is widely used in high-throughput sequence analysis, due to speed and inefficiency issues regarding a traditional sequence similarity comparison like Blast. In mapping, reads are aligned to 0, 1 or more matching locations in the set of references. Reads which are ambiguous in their best location tend to have lower mapping quality scores. RGI uses one of two third-party tools to perform the mapping; we used RGI BWT in default mode (it uses BowTie2 (Langmead and Salzberg, 2012)). RGI then reports the numbers of mapped reads associated with each reference gene, along with the average mapping quality score.

We note that at the time we performed these analyses (beginning May 2020), **RGI BWT was formally still beta-release software**.

## Assembly of reads

Alternatively, a commonly applied approach to metagenomics and metatranscriptomics data is the assembly of the set of reads into longer contiguous sequences ("contigs"). This is akin to genome sequence assembly, by detecting overlaps of the short reads, but there is much more uncertainty; results depend on the total number of reads in the sample (the "depth") and the diversity and relative proportions of the biological sequences in the sample. Some fragments may have been sampled (by short reads) at high frequency, others much less; many of the short reads may not have any detectable overlaps with other reads, and so not contribute to any resulting assembled contigs.

However, if sequencing depth is sufficient, long contigs can result, which can in some cases be 100s of kbp or Mbp in length; in extreme cases, especially if bacterial diversity is quite low, even whole bacterial chromosome-length contigs may occur. The general principle is simply that longer, more complete DNA sequences from the samples enable more reliable matching to reference databases. This can of course be very dependent on the number, lengths and accuracy of the assembled sequences.

A major drawback of assembly in itself is that quantity information is lost, and so there are no proportional frequencies associated with each contig. In practice, this problem is addressed by "**back-mapping**", in which the reads which had been used as input to the assembly process are then aligned against each contig and the frequencies assessed - see normalisation. For various reasons, this is the standard method (rather than an attempt to count the contributions of the reads in the original assembly process), and it provides a consistency check. In NGS back-mapping exercises in general, it is common for the total proportion of reads to be mapped back to the contigs to be much less than 100%; for example we do not consider 50% untoward. This reflects the imperfections of the assembly process. However, for very high depth sequencing, consistently higher back-mapping proportions may occur.

Here, we refer to the sets of assembled contigs for each sample as "assemblies". These were analysed by the long-sequence specific version of DeepARG, "DeepARG-LS". The principal mode of RGI ("RGI MAIN") is for analysing assemblies such as genomes or assembled metagenomes.