# Appendix 3b: Characteristics of Sequence Data

1. The short-read data is in the form of pairs of reads, which are two fragments of around 150 nucleotides length each, representing two regions of the original biological DNA spaced apart by a variable amount (typically of the order of 250 nucleotides).The unassembled short-read data has advantages in that it is essentially quantitative. With some caveats, the proportion of reads which are fragments of a particular gene has a direct relationship to the relative abundance of that gene in the extracted DNA. In contrast, abundances are lost when overlapping short reads are collapsed into longer segments (quantitative information can still be estimated, however). The assembled data may also be of poor quality if the DNA is highly diverse and so few reads overlap, yielding few contigs. Also, the assembled contigs will not always be perfect, and some may contain mis-assemblies which do not represent any biological DNA completely accurately.
2. Conversely, the longer assembled sequences are much better for clearly identifying genes of interest (and taxonomic identification), instead of the unassembled reads which often contain only fragments of the genes. They also enable co-location analysis, which is extremely limited with unassembled reads.
3. The long-read data represents fragments of biological DNA, so is not affected by the mis-assembly issue. This means that long reads may provide definitive information on co-location. However, the accuracy of the ONT sequencing platform, while rapidly improving, is still rather less than that of the Illumina short-read technology. That is, the long reads have a higher frequency of erroneous nucleotide base calls, and erroneous insertions and deletions. This does prevent the accurate identification of some types of ARGs, while others are less affected. However, the long reads can also be used to "anchor" the short reads which correspond to the same pieces of DNA, providing correct positional information and thus providing the best aspects of both platforms.
4. The assembly of the long-reads potentially provides the longest, most complete sequences. A variety of assembly techniques are possible, including those utilising the short-read data.