# Answers to Deliverable VI

1. Explain what reads, contig and scaffold means. Draw the relationship between the 3

Reads are assembled into contigs. Contigs can be bridged into scaffold with additional information such as pared-end read information.



1. De Bruijn graphs are used for de novo assembly what is the main advantage compared to OLC or overlap-layout consensus assembly?

No alignment is used. K-mers are extracted from reads and mapped onto graph.

1. What is the optimal k-mer size for de Bruijn graph assembly of metagenomes?

There is no general optimal. The assembler Spades tries different k-mer sizes depending on input read length (default for reads with length 150bp is 21, 33, 55, 77).

1. What kind of sequence causes problems for de Bruijn graph assembly?

Repetitive & repeated sequences. One examples is horisontally transferred genes that will link the two genomes in the de Bruijn graphs

1. Why is de novo assembly harder with metagenomic samples than isolate DNA?

In metagenomes we not only have to deal with repetitive sequence within a genome but also between several genomes.

De novo assembly is always without a reference....it’s de novo!

1. Why would we do metagenomic binning?

Metagenomic binning provides context. We can link interesting genes to the encoding organism, which is especially important to understand if a antibiotic resistance or virulence gene is on a clinically relevant organism. It also allows us to sequence whole genomes of uncultivable organisms. We can also identify the replication rate of binned organisms using iRep.

1. Why would we not do metagenomic binning?

It demands deep and expensive sequencing. It is also expensive in computing power as all reads are mapped to all contigs etc. If our question does not require binning, don’t.

1. What kind of information does Metabat use for binning?

Tetranucleotide frequency & Co-abundance


1. CheckM is used for assesing the quality of metagenomic bins. It looks for single-copy genes, but this list of genes is not universal but based on?

CheckM uses sets of genes that are ubiquitous and single-copy within a phylogenetic lineage.

1. Why and when is bin dereplication used?

Since each samples is assembled individually there might be redundant contigs. Dereplication means identifying similar genomes from a larger set and picking the best ones