1.3 Workflow Steps

The following programs/steps have been used during this analysis

Quality Control

FastQC: Generates FastQC statistics for both forward and reverse sequences. FastQC aims to provide a simple way to perform quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis. FastQC is run on the raw data and the quality controlled data.

MultiQC: MultiQC is a reporting tool that parses summary statistics from results and log files generated by other bioinformatics tools.

Fastp: Fastp is used to trim low quality reads and adapter sequences.

Bowtie2: Bowtie2 is used during the quality control stage to remove human contaminant reads from the raw reads. Forward and reverse reads are mapped against indexed versions of the human genome (GRCh38). Reads matching either genome are discarded. During the contig assembly stage Bowtie2 is used to produce an index of the contig assemblies and then used to map quality controlled reads back to the contigs. A high rate of mapping here can be related to the quality of the contig assemblies.

FastQC Filtering Step: An internal script is used to process the FastQC results. The FastQC output file contains all the modules FastQC checks for the samples' quality. The internal script focuses on the Basic Statistics and the Per Base Sequence quality modules and checks whether a) the sequence length is higher than the default value (70nt); b) the %GC content is between 25 and 75 ; and c) the minimum per base sequence mean quality score is 20 or more; d) that more than a minimum (default 500) number of reads have passed the quality control. Samples that fulfil the above criteria are processed by the pipeline, otherwise they are discarded. Sequence length threshold depends on study criteria (type of sequence data, objectives, etc). In the literature sequence length between 50-100bp minimum is considered acceptable for paired end reads. Percentage GC content describes the guanine and cytosine content of a biological sequence and has historically been reported to range between 25% and 75% for bacterial genomes. A quality score of 20 represents an error rate of 1 in 100 (meaning every 100 bp sequencing read may contain an error), with a corresponding call accuracy of 99%. As quality scores decrease below 20 the confidence for accuracy is lost very rapidly (Phred-scaled quality scores).

Metatranscriptome Assembly and Quality Assessment

RNA-Bloom: This tool represents a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-seq data.

Mmseqs2: MMseqs2 (Many-against-Many sequence searching) is a software suite to search and cluster huge protein and nucleotide sequence sets.