Metatranscriptomics: Pipeline Report

Eagle Genomics

Tuesday 8th August, 2023
10:38am

Contents

Go to page 4

1 Introduction

1.1 Publication-ready Methods

FastQC [1] was used to assess the quality of reads before and after preprocessing with fastp [2]. Fastp was used with default parameters and the option to detect adapter sequences in paired-end data. Bowtie2 [3], with the very-sensitive preset, was used to clean the reads from human and PhiX contamination. After quality control FastQC was run again to measure improvements in input data quality. MultiQC [4] was run on the outputs of FastQC to amalgamate the reports for all samples. Samples were additionally filtered to ensure that all read pairs fulfilled the following criteria a) the sequence length was higher than 70nt, b) the %GC was between 25 and 75, c) the minimum per base sequence mean quality was 20 or more and d) samples contained more than 500 reads. The quality control and decontaminated reads were used in downstream analyses to assemble contig sequences, generate taxonomic assignments, estimate microbial abundance and produce functional/secondary metabolite annotations.
Assembled contigs were produced from the quality controlled paired-end and single-end reads using RNA -Bloom [5] default parameters. To assess the quality of contig predictions, Bowtie2 was used to map the quality controlled reads to the contig assemblies.
To analyse KEGG pathways and modules, MinPath [6] was used to reconstruct/infer the presence of pathways and modules in each sample. The number of KEGG orthologs present in each sample was determined by Kofamscan [7] . For both pathways and modules, coverage was defined as the proportion of KEGG (K0) terms for each pathway/module identified in each sample. Pathway/module abundance was calculated as the harmonic mean of all KEGG term associated with that pathway/module using RSEM [8] .