Manual

July 19, 2019

Provided files

summary.xlsx

A summary excel sheet containing the detected open reading frames (ORFs) together with some additional information. This file contains:

- **ORF ID:** composed of the genome accession, the start, the stop and the strand of the ORF.
- **start:** the start position of the ORF
- **stop:** the stop position of the ORF
- **strand:** the strand of the ORF (+:forward, -:reverse)
- **length:** the length of the ORF
- **RPKM:** reads per kilobase million measures for each method-condition-replicate triplet
- **evidence:** list of 'prediction tool'-'sample' combinations describing the tool by which the ORF was detected and the sample in which the ORF was detected.
- **annotated:** if the ORF is already annotated in the provided annotation file, the respective locus tag is shown in this column
- **name:** if the ORF is already annotated in the provided annotation file, the respective name (if available) is shown in this column
- **ORF type:** ORF type retrieved from reparation (if available)
- **start codon:** the start codon for the ORF
- **stop codon:** the stop codon for the ORF
- **nucleotide sequence:** nucleotide sequence for the ORF
- **amino acid sequence:** amino acid sequence for the ORF

multi_qc.html

The multiQC report collects information from different tools, including fastQC and subread featurecounts. The general statistics gives an overview over:

- the number of duplicates
- the GC content
- the average read lengths
- the number of reads (in millions)

These statistics are collected after each processing step of our pipeline.

- **raw:** the unprocessed data
- **trimmed:** the data after trimming the adapter sequences
- **mapped:** the data after mapping with Segemehl
- **unique:** the data after removing multi-mapping reads
- **norRNA:** the data after filtering the rRNA

Further, feature counts are provided for different features from the annotation file. (i.e. how many reads map to each feature) This includes, all(featurecount), rRNA, norRNA(after filtering), tRNA, ncRNA.

Following is a fastQC report including sequence counts, sequence quality histograms, per sequence quality scores, per base sequence content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, over-represented features and adapter content.

heatmap_SpearmanCorr_readCounts.pdf

Spearman correlation coefficients of read counts. The dendrogram indicates which samples read counts are most similar to each other.

genome.fa

The genome file used for conducting the analysis.

annotation.gff

The annotation file used for conducting the analysis.

annotation_rpkm.xlsx

An excel sheet containing information about the annotated features (taken from the provided annotation). This file contains:

- **Genome:** the genome accession
- **Source:** the source of the gene annotation
- **Feature:** the feature of the gene

- **Start:** the start position of the annotated gene
- **Stop:** the stop position of the annotated gene
- **Strand:** the strand of the annotated gene (+:forward, -:reverse)
- **Locus tag:** the locus tag of the annotated gene (if available)
- **Name:** the name of the annotated gene (if available)
- **Length:** the length of the annotated gene
- **Product:** the product information for the annotated gene (if available)
- **Note:** the notes for the annotated gene (if available)
- **RPKM:** reads per kilobase million measures for each method-condition-replicate triplet

unfilteredtracks

A folder containing single nucleotide mapping bigwig files for the data before removal of reads mapping to rRNA. Used for genome browser visualization.

globaltracks

A folder containing single nucleotide mapping bigwig files for the data after removal of reads mapping to rRNA. Used for genome browser visualization.

threeprimetracks

A folder containing single nucleotide mapping bigwig files for the data before removal of reads mapping to rRNA. For the region around the three prime end. Used for genome browser visualization.

fiveprimetracks

A folder containing single nucleotide mapping bigwig files for the data before removal of reads mapping to rRNA. For the region around the five prime end. Used for genome browser visualization.

centeredtracks

A folder containing single nucleotide mapping bigwig files for the data before removal of reads mapping to rRNA. For the region around the center. Used for genome browser visualization.

tracks

This folder contains further tracks for genome browser visualization. (colored tracks not yet completed)

0.0.1 potentialStartCodons.gff

A genome browser track with possible start codons.

0.0.2 potentialStopCodons.gff

A genome browser track with possible stop codons.

0.0.3 potentialRibosomeBindingSite.gff

A genome browser track with possible ribosome binding sites.

0.0.4 potentialAlternativeStartCodons.gff

A genome browser track with alternative start codons.

0.0.5 combined.gff

A genome browser track with the results of the ORF detection tools.

0.0.6 combined_annotated.gff

A genome browser track with the results of the ORF detection tools. Annotated using the existing annotation. (recommended for use)