Manual

July 26, 2019

Provided files

excel files

At the moment, we provide two types of excel files. A summary of the results and information about the provided annotation (2 files). We generated the files in a general fashion, therefore some columns/sheets can be empty for the summary, as these information are not available for newly annotated ORFs. The provided excel files have a main sheet featuring all rows of the annotation (without regions) and multiple sheets for each feature.

summary.xlsx

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

- **Genome:** the genome accession
- **Source:** the source of the gene annotation (in this case merged from the results)
- **Feature:** the feature of the gene (in this case CDS)
- **Start:** the start position of the ORF
- **Stop:** the stop position of the ORF
- **Strand:** the strand of the ORF (+:forward, -:reverse)
- **Locus tag:** if the ORF is already annotated the locus tag is retrieved from the annotation (if available)
- **Name:** the name of the annotated gene (if available)
- **Length:** the length of the ORF
- **Codon count:** the number of codons
- **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.

- **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
- **Product:** empty
- **Note:** empty

annotation_total.xlsx

An excel sheet containing information about the annotated features (taken from the provided annotation). The RPKM values are calculated directly after mapping the reads, before removal of multi-mapping reads. This file contains:

- **Genome:** the genome accession
- **Source:** the source of the annotated feature
- **Feature:** the feature of the annotated feature (in this case CDS)
- **Start:** the start position of the annotated feature
- **Stop:** the stop position of the annotated feature
- **Strand:** the strand of the annotated feature (+:forward, -:reverse)
- **Locus tag:** the locus tag (if available)
- **Name:** the name of the annotated feature (if available)
- **Length:** the length of the annotated feature
- **Codon count:** the number of codons
- **RPKM:** reads per kilobase million measures for each method-condition-replicate triplet
- **Evidence:** empty
- **ORF type:** empty
- **Start codon:** the start codon for the annotated feature
- **Stop codon:** the stop codon for the annotated feature
- **Nucleotide sequence:** nucleotide sequence for the annotated feature
- **Amino acid sequence:** amino acid sequence for the annotated feature
- **Product:** the product annotated in the provided annotation
- **Note:** the notes from the provided annotation

annotation_unique.xlsx

An excel sheet containing information about the annotated features (taken from the provided annotation). The RPKM values are calculated directly after removal of multi-mapping reads, before removal of reads mapping to rRNA. This file contains:

- **Genome:** the genome accession
- **Source:** the source of the annotated feature
- **Feature:** the feature of the annotated feature (in this case CDS)
- **Start:** the start position of the annotated feature
- **Stop:** the stop position of the annotated feature
- **Strand:** the strand of the annotated feature (+:forward, -:reverse)
- **Locus tag:** the locus tag (if available)
- **Name:** the name of the annotated feature (if available)
- **Length:** the length of the annotated feature
- **Codon count:** the number of codons
- **RPKM:** reads per kilobase million measures for each method-condition-replicate triplet
- **Evidence:** empty
- **ORF type:** empty
- **Start codon:** the start codon for the annotated feature
- **Stop codon:** the stop codon for the annotated feature
- **Nucleotide sequence:** nucleotide sequence for the annotated feature
- **Amino acid sequence:** amino acid sequence for the annotated feature
- **Product:** the product annotated in the provided annotation
- **Note:** the notes from the provided annotation

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.

(To remove the featurecounts from the general statistics, use the **Configure Columns** and untick featurecounts. This might give you a better experience when comparing the different files.)

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.

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)

potentialStartCodons.gff

A genome browser track with possible start codons.

potentialStopCodons.gff

A genome browser track with possible stop codons.

potentialRibosomeBindingSite.gff

A genome browser track with possible ribosome binding sites.

potentialAlternativeStartCodons.gff

A genome browser track with alternative start codons.

combined.gff

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

combined_annotated.gff

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