`RiboExplorer` 1.1.0 - Result Guide

September 18, 2019

This manual describes the contents of the `RiboExplorer` results zip-archive. There are three Summary files placed directly in the archive, while specialized results can be found in dedicated subfolders*:

* click to jump directly to subsection

Summary

The top-folder contains the *sample.xlsx*, *annotation_total.xlsx* and *annotation_unique.xlsx* overview files. The sample sheet (*samples.xlsx*) shows which samples were used and how the corresponding results are named within the workflow. The excel sheets give detailed information about results for annotated genomic features, among others coordinates, locustag, RPKM, Translational Efficiency, amino acid sequence. The excel sheets contain multiple sheets for different genomic features like coding sequences, rRNA, tRNAs, pseudogenes and for all features together enabling the inspection of the individual groups and quality controll (rRNA depletion). The *annotation_total.xlsx* is computed for multi-mapping reads, while the *annotation_unique.xlsx* sheet is computed for unique mapping reads. Please note that the reads mapping to rRNA genes are removed and only uniquely mapping reads are used for downstream analysis.

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:

Header	**Description**
Genome	the genome accession identifier
Source	the source of the entry (e.g. Ensembl, NCBI, etc ...)
Feature	the feature of the entry (e.g. CDS, gene, rRNA, tRNA, etc ...)
Start	the start position of the entry
Stop	the stop position of the entry
Strand	the strand of the entry (+:forward, -:reverse)
Locus Tag	the locus tag (if available in the annotation)
Name	the name of the entry (if available in the annotation)
Length	the length of the entry
Codon Count	the number of codons
TE	the translational efficiency for each method-condition-replicate triplet
RPKM	reads per kilobase million measures for each method-condition-replicate triplet
Evidence	empty (only available for prediction result file)
ORF type	empty (only available for prediction result file)
Start Codon	the start codon for the entry (e.g. ATG, TTG, GTG)
Stop Codon	the stop codon for the entry (e.g. TAA, TAG, TGA)
Nucleotide sequence	nucleotide sequence for the entry
Amino acid sequence	amino acid sequence for the entry
Product	the product annotated in the user-provided annotation
Note	the notes from the user-provided annotation

The excel files are split into multiple sheets. The excel file contains one sheet for each feature. If the feature is not available, the sheet will be empty.

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:

Header	**Description**
Genome	the genome accession identifier
Source	the source of the entry (e.g. Ensembl, NCBI, etc ...)
Feature	the feature of the entry (e.g. CDS, gene, rRNA, tRNA, etc ...)
Start	the start position of the entry
Stop	the stop position of the entry
Strand	the strand of the entry (+:forward, -:reverse)
Locus Tag	the locus tag (if available in the annotation)
Name	the name of the entry (if available in the annotation)
Length	the length of the entry
Codon Count	the number of codons
TE	the translational efficiency for each method-condition-replicate triplet
RPKM	reads per kilobase million measures for each method-condition-replicate triplet
Evidence	empty (only available for prediction result file)
ORF type	empty (only available for prediction result file)
Start Codon	the start codon for the entry (e.g. ATG, TTG, GTG)
Stop Codon	the stop codon for the entry (e.g. TAA, TAG, TGA)
Nucleotide sequence	nucleotide sequence for the entry
Amino acid sequence	amino acid sequence for the entry
Product	the product annotated in the user-provided annotation
Note	the notes from the user-provided annotation

The excel files are split into multiple sheets. The excel file contains one sheet for each feature. If the feature is not available, the sheet will be empty.

genome-browser

This folder contains files for genome browser visualization, which we tested all files in both IGV and IGB genome browsers. The genome and the annotation provide the established state of research for the organism, while the additionally generated tracks for (alternative) start codon, stop codon and ribosome binding site (RBS) allow to judge the coding potential of a region. The coverage files show how many reads were mapping on the genomic region for RNA- and Ribo-seq experiments. Additionally coverage files with reads truncated to five prime, three prime or centered read region are available in the top-level supplementary folder. Finally additional ORF-prediction tracks, located in the top-level ORF-prediction directory, visualize the open reading frames detected using the read information from the experiments.

features

This folder contains further tracks for genome browser visualization. (pre-colored tracks will be available in the future)

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.

coverage

globaltracks

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

genome.fa

The user-provided genome file used for conducting the analysis.

annotation.gff

The user-provided annotation file used for conducting the analysis.

ORF-predictions

The ORF-predictions folder contains the resulting ORF-predictions created using the prediction tool *REPARATION*. We provide a file in *.gff3* format for genome-browser visualization and an excel-sheet which gives an overview over all predicted ORFs. The content of this file is explained in this section.

prediction_results.xlsx

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

Header	**Description**
Genome	the genome accession identifier
Source	the source of the ORF (**merged** result files)
Feature	the feature of the ORF (here 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	the locus tag (if available in the annotation)
Name	the name of the ORF (if available in the annotation)
Length	the length of the ORF
Codon Count	the number of codons
TE	the translational efficiency for each method-condition-replicate triplet
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 entry (e.g. ATG, TTG, GTG)
Stop Codon	the stop codon for the entry (e.g. TAA, TAG, TGA)
Nucleotide sequence	nucleotide sequence for the entry
Amino acid sequence	amino acid sequence for the entry
Product	empty (not available for the result file)
Note	empty (not available for the result file)

The excel files are split into multiple sheets. The excel file contains one sheet for each feature. If the feature is not available, the sheet will be empty.

prediction_results.gff

A genome browser track with the results of the ORF detection tools. Annotated using the existing annotation. If a locus_tag exists, then the detected ORF is also annotated in the user-provided annotation.

quality-control

This folder contains files that are useful to determine the overall quality of the input data and the progress after each key step of the workflow.

multiqc_report.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.

total_read_counts.xlsx

This file shows the overall read-counts for each feature annotated in the user-provided annotation, after mapping and before removal of multi-mapping reads.

unique_read_counts.xlsx

This file shows the overall read-counts for each feature annotated in the user-provided annotation, after mapping and after removal of multi-mapping reads.

heatmap_SpearmanCorr_readCounts.pdf

Spearman correlation coefficients of read counts. The dendrogram indicates which samples read counts are most similar to each other. Since there should be always a higher correlation between experiments with the same condition and experiment type (e.g. replicates) and not others, this is a rapid way to quality-control the labeling/consistency of input data.

supplementary

Contains additional tracks and files.

metagene

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.