HRIBO 1.5.1 - Result Guide

July 27, 2021

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

- Genome-browser - Files for visualization
- ORF-predictions - Predicted Open reading frame files
- Quality control - MultiQC summary report for processing steps
- Differential Expression - Xtail and RiboRex results

* click to jump directly to subsection

Summary

The top-folder contains the *sample.xlsx* overview file. The sample sheet (*samples.xlsx*) shows which samples were used and how the corresponding results are named within the workflow. In addition, it contains this manual and an overview table.

overview.xlsx

An overview table containing all information gathered from the prediction tools and differential expression analysis. The contents of this table change depending on the workflow configuration. The overview table for the default table will contain reparation and differential expression output. Readcounts for this table are counted individually for each entry.

Header	**Description**
Identifier	identifier for use in genome-browsers (IGB/IGV/...)
Genome	the genome accession identifier
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 not novel)
Overlapping_genes	genes that overlap with the predicted ORF
Old_locus_tag	the old locus tag of a gene (if available in the annotation)
Name	the name of the ORF (if not novel)
Gene_name	The name of the ORFs associated gene feature. (if not novel)
Length	the length of the ORF
Codon Count	the number of codons in the ORF
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)
15nt_upstream	the 15nt upstream of the start codon
Nucleotide sequence	nucleotide sequence for the entry
Amino acid sequence	amino acid sequence for the entry
TE	the translational efficiency for each method-condition-replicate triplet
RPKM	reads per kilobase million measures for each method-condition-replicate triplet
Evidence_reparation	list of samples describing in which the ORF was detected
Reparation_probability	the probability with which a certain ORF is predicted the lower the value, the less likely it is an actual ORF (0.5-1)
Evidence_deepribo	list of samples describing in which the ORF was detected
Deepribo_rank	The deepribo rank for this ORF. (only available for deepribo predictions)
Deepribo_score	The score the deepribo rank is based on.
riborex_pvalue	The pvalue (determined by riborex)
riborex_pvalue_adjusted	The adjusted pvalue (determined by riborex)
riborex_log2FC	The log2FC (determined by riborex)
xtail_pvalue	The pvalue (determined by xtail)
xtail_pvalue_adjusted	The adjusted pvalue (determined by xtail)
xtail_log2FC	The log2FC (determined by xtail)

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 files

Normalizations

min: unnormalized entry $*$ $\frac{\text{minimum number of aligned reads}}{\text{number of aligned reads}}$
The min normalization is especially useful when comparing different libraries(samples) within one experiment. It is recommended to use these coverage files if you compare it with other coverage from this result folder exclusively.

mil: unnormalized entry $*$ $\frac{1000000}{\text{number of aligned reads}}$
The mil normalization is useful when comparing different libraries(samples) between different experiments.

Mappings

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.

genome.fa

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

annotation.gff

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

updated_annotation.gff

An updated annotation containing the reparation predictions as well as the original annotation provided by the user.

ORF-predictions

The ORF-predictions folder contains the resulting ORF-predictions created using the prediction tools `REPARATION` and `DeepRibo`. 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.

predictions_reparation.xlsx

An excel summary sheet containing the detected open reading frames (ORFs) together with some additional information. Readcounts for this table are counted individually for each entry. This file contains:

Header	**Description**
Identifier	identifier for use in genome-browsers (IGB/IGV/...)
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)
Pred_probability	the probability with which a certain ORF is predicted the lower the value, the less likely it is an actual ORF (0.5-1)
Locus Tag	the locus tag (if available in the annotation)
Old_locus_tag	the old locus tag of a gene (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)
15nt_upstream	the 15nt upstream of the start codon
Nucleotide sequence	nucleotide sequence for the entry
Amino acid sequence	amino acid sequence for the entry

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.

predictions_reparation.gff

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

predictions_deepribo.xlsx

An excel summary sheet containing the detected open reading frames (ORFs) together with some additional information. Readcounts for this table are counted individually for each entry. This file contains:

Header	**Description**
Identifier	identifier for use in genome-browsers (IGB/IGV/...)
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)
Pred_value	the prediction value DeepRibo attributes the given prediction
Pred_rank	the rank calculated from the prediction value (the best prediction has rank 1)
Novel_rank	a special ranking involving only novel ORFs that are not in the annotation
Locus Tag	the locus tag (if available in the annotation)
Old_locus_tag	the old locus tag of a gene (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)
15nt_upstream	the 15nt upstream of the start codon
Nucleotide sequence	nucleotide sequence for the entry
Amino acid sequence	amino acid sequence for the entry

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.

For DeepRibo, all results are available in the excel-sheet. The sheet is sorted regarding the best prediction rank. It is advised to start on the top of the list, which shows the best overall results. The last results are very unlikely which is mirrored by their predictive score.

predictions_deepribo.gff

A genome browser track with the results of the ORF detection tool `DeepRibo`. 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.

The *annotation_total.xlsx* and *annotation_unique.xlsx* 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. Readcounting is done by assigning fractional counts for overlapping entries of the same feature and fractional counts for multi-mapping reads. This file contains:

Header	**Description**
Identifier	identifier for use in genome-browsers (IGB/IGV/...)
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)
Old_locus_tag	the old locus tag of a gene (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
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)
15nt_upstream	the 15nt upstream of the start codon
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. Readcounting is done by assigning fractional counts for overlapping entries of the same feature. This file contains:

Header	**Description**
Identifier	identifier for use in genome-browsers (IGB/IGV/...)
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)
Old_locus_tag	the old locus tag of a gene (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
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)
15nt_upstream	the 15nt upstream of the start codon
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.

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.

differential-expression

riborex

contrast_sorted.csv

Table with riborex results.

contrast_significant.csv

Table containing only the significant results with adjusted p-value < 0.05.

xtail

contrast_sorted.csv

Table with xtail results.

contrast_significant.csv

Table containing only the significant results with adjusted p-value < 0.05.

fc_contrast.pdf

This figure shows the result of the differential expression at the two expression levels, where each gene is a dot whose position is determined by its log2 fold change (log2FC) of transcriptional level (mRNA_log2FC), represented on the x-axis, and the log2FC of translational level (RPF_log2FC), represented on the y-axis. The points will be color-coded with the pvalue_final obtained with xtail (more significant p values having darker color)

- blue: for genes whos mRNA_log2FC larger than 1 (transcriptional level).
- red: for genes whos RPF_log2FC larger than 1 (translational level).
- green: for genes changing homodirectionally at both level.
- yellow: for genes changing antidirectionally at two levels.

r_contrast.pdf

This figure shows the RPF-to-mRNA ratios in two conditions, where the position of each gene is determined by its RPF-to-mRNA ratio (log2R) in two conditions, represented on the x-axis and y-axis respectively. The points will be color-coded with the pvalue_final obtained with xtail (more significant p values having darker color)

- blue: for genes with log2R larger in first condition than second condition.
- red: for genes with log2R larger in second condition than the first condition.
- green: for genes with log2R changing homodirectionally in two condition.
- yellow: for genes with log2R changing antidirectionally in two condition.