Output description for viral genome pipeline

Table of Contents

Pipeline overview

- FastQC v0.11.9 - read quality control.
- Cutadapt v3.1 - adapter and low quality trimming.
- miRDeep2 v2.0.1.3 - miRNA identification.
- DESeq2 v1.18.1 - Differential expression analysis and plots

Preprocessing

FastQC

FastQC gives general quality metrics about your reads. It provides information about the quality score distribution across your reads, the per base sequence content (%T/A/G/C). You get information about adapter contamination and other overrepresented sequences.

For further reading and documentation see the FastQC help.

Output directory: `01-fastQC`

- `{sample_id}/{sample_id}_R[12]_fastqc.html`
  - html report. This file can be opened in your favourite web browser (Firefox/chrome preferable) and it contains the different graphs that fastqc calculates for QC.
- `{sample_id}/{sample_id}_R[12]_fastqc`
  - older with fastqc output in plain text.
- `{sample_id}/{sample_id}_R[12]_fastqc.zip`
  - zip file containing the FastQC report, tab-delimited data file and plot images

Trimming

Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.

Parameters included for trimming are:

- Adapter: AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
- Minimum length > 16

Results directory: `02-preprocessing`

- Files:
  - `{sample_id}/{sample_id}_R[12]_trimmed.fastq`: trimmed reads for both forward and reverse in case.

  Note:To see how your reads look after trimming, look at the FastQC reports in the 03-preprocQC directory

  Note:From now on, all the steps will be host specific.

miRDeep2

miRDeep2 is a software package for identification of novel and known miRNAs in deep sequencing data. Furthermore, it can be used for miRNA expression profiling across samples. Last, a new module for preprocessing of raw Illumina sequencing data produces files for downstream analysis with the miRDeep2 or quantifier module.

Results:

- `04-mapping`: Results of mapping reads
  - `reads_collapsed.arf`: arf file with with mapped reads

    - `reads_collapsed.fa`: fasta file with processed reads
- `05-knownmiRNA`:expression of the corresponding miRNAs
    - `miRNAs_expressed_all_samples.csv`: A tab separated file with miRNA identifiers and its read count
    - `expression_output.html`: gives an overview of all miRNAs the input data
    - `pdfs_output`: contains for each miRNA a pdf file showing its signature and structure.
- `06-novelmiRNA`: microRNA detection
    - `miRNAs_expressed_all_samples.csv`: A tab separated file with miRNA identifiers and its read count for novel miRNAs
    - `expression_output.html`: gives an overview of all miRNAs and novel the input data
    - `pdfs_output`: contains for each miRNA and novel a pdf file showing its signature and structure.

Differential expression

DESeq2

Differential expression analysis with DESeq2. DESeq2 is a Bioconductor package for R used for RNA-seq data analysis. The script included in the pipeline uses DESeq2 to normalize read counts and create a heatmap / dendrogram showing pairwise euclidean distance (sample similarity). It also creates other plots to evaluate the sample dispersion. It also provides PCA plots to evaluate sample grouping.

Output directory `07-DESeq2`:

- DE_matrix.txt
    - Comparative table with the differential expression of two conditions.
- maPlot.pdf
    - MA plot of the DESeq analysis results for all the samples
- heatmap_sample_to_sample.pdf
    - Heatmap with the euclidean distance between samples.
- plotPCA.pdf
    - PCA plot of the samples for the rlog and the vsd.
        - rlog refers to the regularized log transformation, which transforms the count data to the log2 scale in a way which minimizes differences between samples for rows with small counts, and which normalizes with respect to library size.
        - vsd refers to variance stabilizing transformation (VST), which calculates a variance stabilizing transformation (VST) from the fitted dispersion-mean relation(s) and then transforms the count data (normalized by division by the size factors or normalization factors), yielding a matrix of values which are now approximately homoskedastic (having constant variance along the range of mean values). The transformation also normalizes with respect to library size.
- boxplot.pdf
    - PDF file with the box_plots
        - Box plot of the normalized Counts
        - Box plot of the counts cook distances to see if one sample is consistently higher than others.
- plotDispersions.pdf
    - PDF file with plots to analyze the dispersion of the samples
        - Dispersion calc is the per-gene dispersion estimate together with the fitted mean-dispersion relationship.
        - Histogram with the test of the differential expression pvalues
- hierarchical_clustering.pdf
    - PDF file with the hierarchical clustering of the samples. The input data comes from the normalization of the counts. For the normalization DESeq uses the normalization of the ratios where the counts are divided by sample-specific size factors determined by median ratio of gene counts relative ro geometric mean per gene.
- heatmapCount_top20.pdf

    - Heat map of the top 20 genes with the higher normalized mean count. The normalization is the same that the one of the hierarchical clustering.
- heatmapCount_all_genes.pdf
    - Heatmap of the normalized counts of all the genes.
- plotSD.pdf
    - Standard deviation of the transformed data, across samples, against the mean, using the shifted logarithm transformation.--->