**Protocol for RNAseq Expression Analysis in Yeast**

Stefan Bohn\*

Author names will be published exactly as they appear on the title page of the manuscript. Please confirm the spelling of every author’s name.

Institute of Structural Biology, Helmholtz Zentrum München, Munich, Germany

\*For correspondence: [stefan.bohn@helmholtz-muenchen.de](mailto:stefan.bohn@helmholtz-muenchen.de)

Authors are required to carefully go through the writing style (i.e., flow, grammar, spelling, and verb tense) of the manuscript and correct any errors or make any change as needed.

To help our editing team view your modifications, please use "Track Changes" while editing.

To help readers follow the protocol more easily, we have made some minor modifications (in red). Please confirm them and make any change as needed.

**[Abstract]** Genome-wide sequencing of RNA (RNAseq) has become an inexpensive tool to gain key insights into mechanisms of cells and disease. Sample preparation and sequencing are streamlined and allow the acquisition of hundreds of gene expression profiles in a few days. However, in particular data processing, curation and analysis involve numerous steps and can be overwhelming to non-experts. Here, the sample preparation, sequencing and data processing workflow for RNAseq expression data in yeast is described. While this protocol covers only a small portion of the RNAseq landscape, the principle workflow common to most such experiments is described allowing the reader to adapt the protocol where necessary.

Graphic abstract:

电脑屏幕的照片上有字

中度可信度描述已自动生成

The graphic abstract we received recently (the bottom) is inconsistent with the previous one (the top) in your manuscript. Please confirm which one you would like to use as the final version.

**Keywords:** mRNA, Sequence analysis, Yeast, Relative expression levels, Next generation sequencing, Systems biology, Whole genome

**Comment:** **This is improved but the author needs to check the manuscript carefully. I found multiple mistakes.**

**[****Background]** Sequencing of RNA (RNAseq) has - with the emergence of next generation sequencing - become a powerful tool to measure the presence and quantity of RNA in a given cell population or even a single cell. Since its initial uses (Bainbridge *et al.*, 2006; Cheung *et al.*, 2006; Emrich *et al.*, 2007; Weber *et al.*, 2007; Nagalakshmi *et al.*, 2008) RNAseq has seen a large variety of applications: from gene expression analysis by quantifying relative amounts of RNA sequence reads to the discovery of novel transcripts or splice variants, ribosome profiling or detection of single nucleotide polymorphisms. Even though a manifold of specific RNAseq uses exist, the basic workflow remains the same: RNA molecules are extracted, amplified, sequenced, aligned to the genome of the host organism, and subsequently the data is analyzed. Sample requirements are relatively low, typically 1 μg down to 10 ng of input RNA are sufficient for downstream amplification and library generation. For single-cell RNAseq as little as 10 pg are required, where the low amount of input material is amplified before library generation (Haque *et al.*, 2017). RNAseq can be applied to any population of extracted RNA independent of the source organism. Due to the wide applicability of RNAseq, sample preparation kits varying in complexity are commercially available (from RNA extraction to whole RNAseq library generation including the computational analysis).

Integral part of RNAseq is the sequencing of the extracted RNA population. Most commonly, sequencing is performed by detection of fluorescently labelled nucleic acids bound to the surface of flowcells, *e.g.*, by platforms such as Illumina and PAC Biosystem. To this end, the RNA fragments are converted into a cDNA library, amplified and flowcell adapters are introduced. During each sequencing cycle DNA polymerases will attach fluorescently labelled nucleotides to the flowcell-bound library molecules which are then detected by the sequencer, typically generating read lengths of 150-300 bp to several kbp (for Illumina and PAC Biosystem, respectively). More recently emerging is sequencing by passage of nucleic acids through protein nanopores embedded in membranes (*e.g.*, by Nanoporetech) (Logsdon *et al.*, 2020), allowing for sequencing of much longer fragments (up to Mbp). At the time of writing the most commonly available sequencers (*e.g.*, from Illumina or PAC Biosystem) cost around $100 k for the machine alone, whereas table-top sequencers using the nanopore technology are considerably cheaper (~$10 k) promising wider applicability in the near future. Due to the considerable cost of the most commonly available sequencing systems from Illumina or PAC Biosystems resources are often shared among labs or institutes and managed by trained professionals that ensure the acquisition and integrity of high quality sequencing data.

While it goes beyond the scope of this manuscript to describe all applications of RNAseq, this protocol aims to provide a workflow for RNAseq expression analysis that can be used as a reference backbone which the reader can adapt to their specific needs (*e.g.*, RNA extraction from different source, addition of splice-aware alignment steps for genomes of higher eukaryotes, *etc*.). RNAseq expression analysis is a powerful and commonly used tool to identify genes that are up- or downregulated in a stressed sample (*e.g.,* genomic mutations, UV light, drugs, chemical or nutrient stress) when compared to a relaxed sample (*e.g.,* wildtype cell population). A gene is “upregulated” or “downregulated” when more or less of its RNA is measured (*i.e.*, expressed in the cell) in the stressed condition when compared to wildtype, respectively.

Here, the workflow for RNAseq expression analysis in *S. cerevisiae* is described, from cell growth to RNA extraction, library generation, data processing and analysis. This protocol focuses on using commonly available lab resources wherever possible and utilizes open source and free-of-cost software packages provided by the bioinformatics community throughout. This workflow has proven robust and useful for the analysis of gene expression profiles in libraries of histone point mutants in yeast (Braberg *et al.*, 2020).

**Materials and Reagents**

1. 1.5 ml Eppendorf tube (*e.g.*, Eppendorf, catalog number: 0030120086)
2. Petri dishes, plastic, 10-cm diameter *(e.g.*, Falcon, catalog number: 353003)
3. Sterile pipette tips
4. Tooth picks (autoclaved)
5. Dry Ice
6. Agar (*e.g.*, Becton Dickinson, catalog number: 214030)
7. Bacto Peptone (*e.g.*, Becton Dickinson, catalog number: 211677)
8. CHCl3 (Acid Phenol, *e.g.*, ThermoFisher, catalog number: AM9720)
9. DEPC-ddH2O (Diethyl pyrocarbonate treated water, *e.g.*, catalog number: Invitrogen, 750024)
10. EDTA (Ethylenediaminetetraacetic acid disodium salt dihydrate, *e.g.*, Sigma-Aldrich, catalog number: E6635)
11. EtOH (Ethanol, *e.g.*, Sigma-Aldrich, catalog number: 459836)
12. Formamide (*e.g.*, Sigma-Aldrich, catalog number: 11814320001)
13. Glucose (*e.g.*, Molekula, catalog number: 13002238)
14. HCl (Hydrochloric Acid, *e.g.*, Sigma-Aldrich, catalog number: 320331)
15. NaAc (anhydrous Sodium acetate, *e.g.*, Sigma-Aldrich, catalog number: S2889)
16. NaOH (Sodium hydroxite pellets, *e.g.*, Sigma-Aldrich, catalog number: 1064980500)
17. SDS (Dodecyl sulfate sodium salt, *e.g.*, Merck, catalog number: 13760)
18. Tris (2-Amino-2-(hydroxymethyl)-1,3-propanediol, *e.g.*, Sigma-Aldrich, catalog number: T1503)
19. Yeast Extract (*e.g.*, Serva, catalog number: 24540)
20. Yeast Extract Peptone Dextrose (YEPD) media (see Recipes)
21. 1 M Tris-HCl solution, pH 7.5 (see Recipes)
22. 0.5 M EDTA solution, pH 8.0 (see Recipes)
23. 20% SDS solution (see Recipes)
24. Tris-EDTA-SDS (TES) solution (see Recipes)
25. 3 M NaAc solution, pH 5.2 (see Recipes)

**Equipment**

1. Autoclave
2. Centrifuge and table-top centrifuge
3. Vortexer
4. Flasks, autoclavable
5. Incubator
6. pH meter
7. Pipettes (1 ml, 200 μl, 20 μl, 2 μl)
8. Stir bar and Stir plate, magnetic
9. Thermocycler

**Software**

1. bbmap (BBMap – Bushnell, B.; Version 38.90; sourceforge.net/projects/bbmap)
2. Biocmanager (<https://cran.r-project.org/web/packages/BiocManager/vignettes/BiocManager.html>; Version 3.12; <https://bioconductor.org/install/>)
3. bioconda (doi:[10.1038/s41592-018-0046-7](https://doi.org/10.1038/s41592-018-0046-7); <http://bioconda.github.io/user/install.html>)
4. bowtie2 (Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359; Version 2.4.2; <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>;)
5. bwa (Burrows-Wheeler Aligner, Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60; Version 0.7.17; [http://bio-bwa.sourceforge.net](http://bio-bwa.sourceforge.net/)/)
6. DESeq2 (Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8; Version 1.30.1; <http://bioconductor.org/packages/release/bioc/html/DESeq2.html>)
7. fastqc (Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data; Version 0.11.9; <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)>
8. htseq-count (doi: [10.1101/002824](https://doi.org/10.1101/002824); Version 0.11.1; <https://htseq.readthedocs.io/en/release_0.11.1/install.html>)
9. Integrated Genomics Viewer (James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov.[Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)](http://www.nature.com/nbt/journal/v29/n1/abs/nbt.1754.html); Version 2.9.2; <https://software.broadinstitute.org/software/igv/download>)
10. R (R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [https://www.R-project.org/](https://www.r-project.org/) ; Version 4.0.4; <https://cran.r-project.org/bin/windows/base>)
11. samtools (Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup, The Sequence alignment/map (SAM) format and SAMtools, Bioinformatics (2009) 25(16) 2078-9; Version 1.11; <http://www.htslib.org/download/>)
12. tophat (Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25; Version 2.1.1; <https://ccb.jhu.edu/software/tophat/index.shtml>)
13. trimmomatic (Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170; Version 0.40; <http://www.usadellab.org/cms/?page=trimmomatic>)

**Procedure**

A) Sample preparation and RNA extraction

It is of utmost importance that when handling RNA all materials and reagents are RNAse free. Furthermore, it must be noted that the computation of relative expression values described in D) requires at least three biological replicas. Accordingly, if for example the expression levels of a mutant strain are to be compared to a wildtype strain, six RNA samples need to be prepared (three for each strain) which can then be used to create six RNAseq libraries.

Here, an efficient and reliable method to extract RNA despite yeast’s robust cell wall is described (Collart *et al.*, 2001).

1. With a sterile toothpick or pipette tip pick single colonies of *S. cerevisiae* and inoculate 2 ml YEPD liquid media for growth at 30 °C over night.
2. Inoculate 10 ml of liquid YEPD with 50 μl of overnight culture.
3. Harvest cells in mid-log phase (OD600 ≈ 1.0) by centrifugation and transfer to 1.5 ml Eppendorf tube. Resuspend in 300 μl DEPC-ddH2O, fast-spin in table-top centrifuge (up to 9,500x *g*), remove supernatant, flash-freeze on dry ice and store at -80 °C.
4. Resuspend cell pellet in 400 µl TES solution. Add 400 µl acid Phenol (CHCl3), cap the tube and vortex vigorously for 10 s (avoid leakage and handle careful!). Incubate for 60 min at 65 °C, with vortexing every 15 min. (Ref. 5).
5. Place on ice for 5 min. Spin in microfuge at 18,000x *g* for 10 min at 4 °C. Transfer aqueous top layer to a clean tube (avoiding the white protein phase). Add 400 µl CHCl3 and vortex vigorously 10 s. Spin in microfuge at 18,000x *g* for 10 min at 4 °C. Transfer aqueous top layer to a clean tube (pipette carefully and avoid the CHCl3-layer).
6. Add 1/10 volume of 3 M NaAc pH 5.2 and 2.5vol EtOH (-20 °C). Precipitate at –80 °C for at least 60 min. Spin in microfuge at 18,000x *g* for 10 min at 4 °C. Carefully remove supernatant and wash pellet by vortexing in 70% EtOH (-20 °C). Spin in microfuge at 18,000x *g* for 10 min at 4 °C.
7. Resuspend pellet in 100% formamide (from 4 °C). Try an equal volume of liquid to pellet first and move up from there. Most RNA should dissolve instantly. To aid solubilization allow to sit at room temperature for 15 min with pipetting every 5 min. If sample is to be very concentrated, store at 4 °C overnight.
8. Determine concentration by 1/100 in H2O and OD260/280 (OD260 1 ≈ 40 µg/ml for RNA). Remember to add formamide 1/100 to the blank.

B) RNAseq library generation and sequencing

Before sample preparation and submission to a sequencing facility it is strongly recommended to discuss the aims of the project with the trained personnel. For successful library generation the input RNA concentration is critical, commonly ranging from 1 μg down to 10 ng per sample. While it is possible to generate RNAseq libraries from scratch (*i.e.*, producing adaptors, buffers, polymerase, etc. using own materials) it is strongly recommended to use commercially available kits that require a minimum of common lab resources and are, most importantly, more reliable in the hands of researchers unfamiliar with RNAseq library formation.

During RNAseq library generation platform-specific adapters are attached to the extracted RNA molecules. Therefore, the library kit must be chosen according to the sequencing platform to be used. Here, the QuantSeq 3’ mRNASeq Library Prep Kit FWD for Illumina (Lexogen) was used for the generation of single-end (*i.e.*, fragments will be sequenced from one end only), 50 base pair reads sufficient for RNAseq expression analysis in yeast. For more complex eukaryotic genomes containing larger amounts of introns and when longer reads are required, consider paired-end library kits and sequencing (*i.e.*, fragments will be sequenced from both ends), after consultation with the sequence facility staff.

1. Generate cDNA libraries containing sequencer- and sample-specific adapters by carefully following the steps described in the manufacturer’s manual.
2. Check the quality of the generated libraries and measure the cDNA concentration.
3. Sequence the cDNA library. Here, an Illumina HiSeq 4000 sequencer was used.
4. Check the quality of the raw read data, typically supplied in fastq format (Figure 1). The most common checks involve the number of reads per sample (should be same order of magnitude for all sequenced samples which means files should be similar in size), the GC content (should match the overall GC content of the host organism) and the overall base quality. Several quality tools exist, here fastqc was used (also see Bérénice Batut, 2021). It is strongly recommended that quality control is performed after each processing step to ensure overall integrity of the data.

形状

中度可信度描述已自动生成

**Figure 1.** An example read sequenced on an Illumina platform in FASTQ format (line 1 contains basic read information, line 2 the actual sequence, and line 4 contains the quality score for each base in Phred33 or Phred64 encoding).

C.1) Preparation of data processing

The following steps describe the setup for the computational workflow described in C.2) as well as the data analysis described in D). This workflow is using open source programs available on Linux operating systems (and its derivatives). Whereas it is possible to process sequence files on Mac or Windows operating machines, the reader is strongly recommended to use Linux-based utilities, due to their wide applicability, timely updates and community-based trouble-shooting.

For most of the processing steps described in C.2) and D) multiple tools exist, in particular for the acquisition of genome assembly and gene annotations (15), creation of index files (16), adapter trimming (17), read alignment (18) and read filtering based on quality (19). While it goes beyond the scope of this manuscript to describe all tools in detail, alternatives to the programs used in this protocol are suggested.

Several of the tools used are available through so-called package managers such as Bioconda or Biocmanager, allowing for easy installation of software and dependencies of most recent versions. Hence it is recommended to follow the installation order described here.

1. Install the Conda package manager via github and R, bowtie2 and samtools using the specific channel Bioconda.  
     
    curl -O <https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh>  
    sh Miniconda3-latest-Linux-x86\_64.sh  
    conda install -c bioconda R  
    conda install -c bioconda bowtie2  
    conda install -c bioconda samtools
2. Install Biocmanager and the DESeq2-package from within R.  
      
    R  
    if (!requireNamespace(“BiocManager”, quietly = TRUE))  
    install.packages(“BiocManager”)  
    BiocManager::install()  
    BiocManager::install(c(“DESeq2”))

C.2) Processing of raw sequence read files

After passing quality checks the sequence reads will now undergo pre-processing and eventually alignment to the genome of the host organism. First, genome assembly, gene annotation and the genome index need to be prepared (C2)1 and C2)2). The sequence reads contain adapter contamination, random primer sequences and low quality tail reads which need to be removed (C2)3) before the alignment of filtered reads to the genome of the host organism (C2)4). Finally, reads are filtered based on their quality score (C2)5) and indexed for downstream analysis (C2)6).

1. Download *S. cerevisiae* genome assembly and gene annotation. Here, UCSC versions sacCer3.fa and sacCer3.ensGene.gtf were used, respectively (downloaded from <http://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/bigZips/>). Besides UCSC genome browser other platforms allow the download of genome assemblies and gene annotations (*e.g.*, NIH/NCBI or Ensembl). Since every platform maintains a slightly different naming convention, it is important that genome assembly as well as gene annotation are aquired from the same platform to avoid program errors during processing.
2. Create index files based on the genome assembly (“sacCer3.fa”) from Step C2)1. Here, the index output files are stored with the base filename “sacCer3”. This index will be used by the alignment program in Step C2)4. Hence, the indexing step must be adapted to the aligner of choice and the example for bwa is given below. Caution: bwa is not a splice-aware alignment tool. If splice events need to be considered during analysis, aligners like tophat need to be used for index creation as well as alignment (Step C2)4).  
     
    bwa index sacCer3.fa
3. Remove the random primer sequence, adapter contamination, and low quality tails. Here, bbmap was used for the library kit described in Step B1 according to the kit manufacturer’s recommended settings (see command below, from www.lexogen.com/quantseq-data-analysis). BBmap is a fast, splice-aware global alignment tool for RNA and DNA sequencing reads. The script “bbduk.sh” is used together with the polyA-tail sequence (polyA.fa.gz) and Illumina-specific adapter sequence information (truseq.fa.gz), located in the installation folder bbmap/resources/).  
   Adapter sequences and low quality reads can also be removed using different tools such as “trimmomatic” (see software section). Independent of the software being used, parameters need to be adjusted according to the library-kit and sequencing-platform used. As a result of trimming the size of the output file should be slightly smaller than the input file.   
     
    bbmap/bbduk.sh in=sample1.fastq out=sample1\_trimmed.fastq ref=bbmap/resources/polyA.fa.gz,bbmap/resources/truseq.fa.gz k=13 ktrim=r forcetrimleft=11 useshortkmers=t mink=5 qtrim=t trimq=10 minlength=20; done
4. Create alignments of the pre-processed sequence reads from Step C1)1 with an alignment tool such as tophat, bbmap, bowtie2 or bwa. Here, bwa was used. Depending on the sequence file size (*i.e.*, number of reads contained), genome size and CPU used this step can take several minutes to several hours. On an Intel® Zeon® CPU E5-2699 v3 @ 2.3GHz aligning took about 3 s per 100k reads.   
   The following command will align the input file reads (sample1\_trimmed.fastq) to the genome index from Step C2)2 (sacCer3) and write the results to “sample1\_trimmed\_aligned.sam”.  
     
    bwa mem sacCer3.fa sample1\_trimmed.fastq > sample1\_trimmed\_aligned.sam
5. Filter the data based on their quality using MAPQ filtering using samtools. Here, all reads with an average base read quality score of less than 50 (*i.e.* the probability of correct mapping is > 99.999%, Figure 1) are removed from the mapped read files generated in Step C2)4. As a result of the quality filtering the output filesize will be smaller than the input filesize. If no or very few reads are remaining, try filtering with a less stringent quality score (*e.g.*, 20). If this recovers the number of reads downstream analysis may still be possible albeit being less reliable.  
     
    samtools view -bq 50 sample1\_trimmed\_aligned.sam > sample1\_trimmed\_aligned\_mapq50.bam
6. Sort the filtered, aligned reads from Step C2)5 and create index files using samtools. This will create an index file of the same name as the input file with the additional ending of “.bai”.  
   Aside from quality check of the output file using tools such as fastqc, mapped reads can be visualized using the Integrated Genomic Browser (IGV). In addition to ensuring that the overall mapping of reads is correct, tools like IGV allow to confirm the presence of intended genomic mutations or gene deletions (Figure 2).   
     
    samtools sort sample1\_trimmed\_aligned\_mapq50.bam -o sample1\_trimmed\_aligned\_mapq50\_sorted.bam  
    samtools index sample1\_trimmed\_aligned\_mapq50\_sorted.bam

应用程序

中度可信度描述已自动生成

**Figure 2. Snapshot of the IGV browser visualization.** Here, reads (grey bars) mapping to the genomic region of Set2 (blue, YJL168C) in yeast are compared between wildtype (upper lane) and the ∆Set2 mutant (lower lane).

D) Data analysis, calculation of expression values and visualization of results

The sorted and indexed files prepared in C) contain all reads that were successfully aligned to the host’s genome (here, *S. cerevisiae*), numbering typically from several hundreds of thousands to millions of reads per replica. This vast amount of information is a major hurdle for analysis by the researcher. Differential gene expression (DGE) analysis aims to determine which, if any, genes show a higher or lower amount of aligned reads across the tested conditions. To this end, reads belonging to a feature (*i.e.*, a gene) will be summed up for each replica and differential expression values are calculated across conditions taking into consideration the variance within a condition among replicas. Hence, it is critical for DGE that several replicas of the same condition are considered (typically, n=3). Gene expression values are usually reported as log2-fold changes in conjunction with adjusted p-Values describing the significance of the change (cut-offs vary, but typically p=Values < 0.05 are considered reliable).

The number of aligned reads can differ strongly between replicas due to technical reasons (*e.g.* fluctuations in amount of input RNA, variations of temperature in the thermocycler during library amplification, differences in binding capacity of the flowcell lanes, *etc*.). Hence, reads must be normalized across replicas and conditions. Several normalization methods for the calculation of DGE values exist, such as Reads Per Kilobase of transcript per Million mapped reads (FPKM), Fragments Per Kilobase of transcript per Million mapped reads (RPKM), Transcripts per Million reads (TPM) or counts per feature (*i.e.*, gene), *etc*. (Dillies *et al.*, 2013).

Here, the counts-based normalization by DESeq2 was used, based on the assumption that most genes are not differentially expressed across conditions. Therefore, counts per feature are extracted from each file generated in Step C2)6 using htseq (Step D1), combined and indexed (Steps D2-D3). Finally, counts are normalized and DGE values are calculated using DESeq2 (Steps D4-D5). Results are visually represented using MA plots, where log-fold changes are plotted against the mean expression values (Step D5).

1. Extract counts for each sample using htseq-count. Here, the aligned, filtered and sorted reads (e.g. sample1\_trimmed\_aligned\_mapq50\_sorted.bam) from Step C2)6 and the gene annotation file (sacCer3.ensGene.gtf) from Step C2)1 are used. This command will generate a txt-file containing the number of reads assigned to each gene annotated in the gtf-file.  
     
    htseq-count -f bam sample1\_trimmed\_aligned\_mapq50\_sorted.bam sacCer3.ensGene.gtf > sample1\_trimmed\_aligned\_mapq50\_sorted\_counts.txt
2. Counts-based expression values are calculated using R and Dseq2. This requires the count data to be assembled in a text document (here, “counts.txt”) as well as an index file (here, “table.txt”, Step D3).   
   Generate a “counts.txt”-file that contains the counts of each replica of a given sample (here, MUT\_X) as well as the reference sample (here, WT\_X) generated in Step D1 as columns in a tab-delimited txt document (Figure 3). As a quality check it is recommended to check several lines (*i.e.*, genes) for consistency (*i.e.*, similar read counts among replicas of a certain condition). Importantly, read counts are not yet normalized to the total number of read counts in each sample and respective variations are expected.  
     
   形状

   中度可信度描述已自动生成

**Figure 3. Example of a counts.txt-file in tab-delimited format.** The first column designates the names of open reading frames (ORFs), the first row the names of the replicas of wildtype and mutant. The numerical matrix contains the number of reads mapped in each replica to the respective ORF.

1. Generate a “table.txt”-file for each sample indexing each column of data (Figure 4) in tab-delimited format.  
     
   形状

   中度可信度描述已自动生成  
   **Figure 4. Example of a table.txt-file in tab-delimited format.** Replica names as designated in counts.txt from Step D3 are indexed by their common condition (*e.g.*, wildtype or mutant).
2. In R, load the Dseq2 library, the combined counts-file from Step D3 and the table-file from Step D4.  
      
    library(DESeq2)  
    count\_table <- read.delim(‘counts.txt’,sep=’\t’,header=TRUE,row.names=’region\_name’)  
    sample\_table <- read.delim(‘table.txt’,sep=’\t’,header=TRUE,row.names=’sample\_name’)
3. Write RNAseq expression and p-Values to file using DESeq2. Generated txt-file (wt\_mutant\_p-values.txt) will contain log2-fold expression and p-values for the respective mutant in tab-delimited format and can now be used for further analysis or visualization. For data inspection a MA-plot is generated (Figure 5). In MA-plots generated by DESeq2 significant hits are colored in red – hence a first quality check is, how many data points are colored in black (*i.e.*, since most genes are not differentially expressed, most data points should be colored in black).  
     
   dds <- DESeqDataSetFromMatrix(countData = count\_table,colData = sample\_table,design = ~ condition)  
   dds <- DESeq(dds)  
   res <- results(dds)  
   resOrdered <- res[order(res$padj),]  
   plot <- plotMA(res, main = ‘mutant‘, ylim = c(-2,2), xlab = ‘mean count‘)  
   write.table(as.data.frame(resOrdered),sep=‘\t‘,quote=FALSE,file=‘wt\_mutant\_p-values.txt‘)

图表, 散点图

描述已自动生成  
**Figure 5. Example of MA-plot analysis as generated by DESeq2.** Genes that are statistically significantly up- or downregulated are marked in red above and below the x-axis, respectively, marked in red.

**Recipes**

1. Yeast Extract Peptone Dextrose (YEPD) media  
   For each liter of YEPD autoclave a mixture of 20 g Bacto Peptone, 10 g Yeast extract and 950 ml H2O. Add 50 ml of 40% (w/v) glucose, mix and cool down before use.  
   For YEPD plates add 24 g of Agar to the solution before autoclaving. Place the autoclaved solution on a magnetic stir plate, add stir bar and 50 ml of 40% (w/v) glucose and cool down solution while stirring. Pour warm media solution in Petri dishes, let it cool down until solid and store at 4 °C until use.
2. 1 M Tris-HCl solution, pH 7.5  
   Dissolve 121.14 g Tris in 800 ml H2O. Adjust pH to 7.5 with HCl. Bring final volume to 1 L with deionized water. Autoclave and store at room temperature.
3. 0.5 M EDTA solution, pH 8.0  
   Add 18.6 g of EDTA to 80 ml of H2O (use DEPC treated H2O). Stir on a magnetic stirrer until dissolved. Adjust the pH to 8.0 with NaOH (~2 g of NaOH pellets). Dispense into aliquots and sterilize by autoclaving.
4. 20% SDS solution  
   Dissolve 20 g of SDS in 90 ml of H2O (use DEPC treated H2O). Heat to 68 °C and stir with a magnetic until dissolved.
5. Tris-EDTA-SDS (TES) solution   
   10 mM Tris-HCl pH 7.5  
   10 mM EDTA pH 8.0  
   0.5% SDS
6. 3 M NaAc solution, pH 5.2  
   Add 24.6 g of sodium acetate to 80 ml of H2O. Stir on a magnetic stirrer until dissolved. Adjust the pH to 5.2 with glacial acetic acid. Fill up volume to 100 ml with H2O.

**Acknowledgments**

I would like to thank Dr. Pavel Sinitcyn, Dr. Assa Yeroslaviz and Dr. Rin Ho Kim from the Next-Generation Sequencing Core Facility at MPI Biochemistry for critical reading of the manuscript.

**References**

1. Bainbridge, M. N., Warren, R. L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., Mardis, E. R., Sadar, M. D., Siddiqui, A. S., Marra, M. A. and Jones, S. J. (2006). [Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach.](https://www.ncbi.nlm.nih.gov/pubmed/17010196) *BMC Genomics* 7: 246.
2. Bérénice Batut. (2021). [Quality Control (Galaxy Training Materials).](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html)
3. Braberg, H., Echeverria, I., Bohn, S., Cimermancic, P., Shiver, A., Alexander, R., Xu, J., Shales, M., Dronamraju, R., Jiang, S., Dwivedi, G., Bogdanoff, D., Chaung, K. K., Huttenhain, R., Wang, S., Mavor, D., Pellarin, R., Schneidman, D., Bader, J. S., Fraser, J. S., Morris, J., Haber, J. E., Strahl, B. D., Gross, C. A., Dai, J., Boeke, J. D., Sali, A. and Krogan, N. J. (2020). [Genetic interaction mapping informs integrative structure determination of protein complexes.](https://www.ncbi.nlm.nih.gov/pubmed/33303586) *Science* 370(6522).
4. Cheung, F., Haas, B. J., Goldberg, S. M., May, G. D., Xiao, Y. and Town, C. D. (2006). [Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology.](https://www.ncbi.nlm.nih.gov/pubmed/17062153) *BMC Genomics* 7: 272.
5. Collart, M. A. and Oliviero, S. (2001). [Preparation of yeast RNA.](https://www.ncbi.nlm.nih.gov/pubmed/18265096) *Curr Protoc Mol Biol* Chapter 13: Unit13 12.
6. Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., Jaffrezic, F. and French StatOmique, C. (2013). [A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.](https://www.ncbi.nlm.nih.gov/pubmed/22988256) *Brief Bioinform* 14(6): 671-683.
7. Emrich, S. J., Barbazuk, W. B., Li, L. and Schnable, P. S. (2007). [Gene discovery and annotation using LCM-454 transcriptome sequencing.](https://www.ncbi.nlm.nih.gov/pubmed/17095711) *Genome Res* 17(1): 69-73.
8. Haque, A., Engel, J., Teichmann, S. A. and Lonnberg, T. (2017). [A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications.](https://www.ncbi.nlm.nih.gov/pubmed/28821273) *Genome Med* 9(1): 75.
9. Logsdon, G. A., Vollger, M. R. and Eichler, E. E. (2020). [Long-read human genome sequencing and its applications.](https://www.ncbi.nlm.nih.gov/pubmed/32504078) *Nat Rev Genet* 21(10): 597-614.
10. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008). [The transcriptional landscape of the yeast genome defined by RNA sequencing.](https://www.ncbi.nlm.nih.gov/pubmed/18451266) *Science* 320(5881): 1344-1349.
11. Weber, A. P., Weber, K. L., Carr, K., Wilkerson, C. and Ohlrogge, J. B. (2007). [Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing.](https://www.ncbi.nlm.nih.gov/pubmed/17351049) *Plant Physiol* 144(1): 32-42.
12. [YPD media.](http://cshprotocols.cshlp.org/content/2010/9/pdb.rec12315.full) (2010). Cold Spring Harb Protoc.