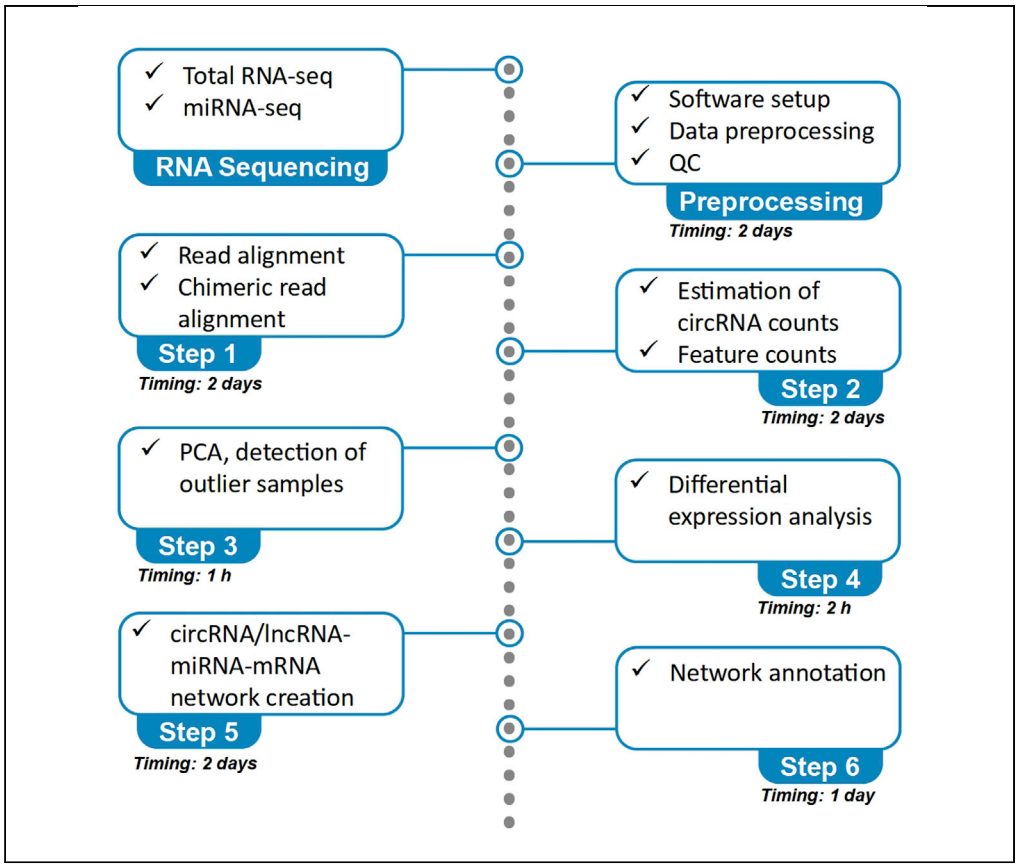
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Construction of transcriptional regulatory networks using total RNA-seq data



Total RNA sequencing allows capturing long non-coding and circular RNA along with mRNA. Additional sequencing of micro RNA (miRNA), using libraries with shorter fragments, provides the means to characterize miRNA-driven transcriptional regulation. Here, we present a protocol for processing total RNA and miRNA sequencing data to quantify circular RNA, long non-coding RNA, mRNA, and miRNA. Further, the protocol combines the quantification data with miRNA target annotation to construct likely transcriptional regulatory networks, which can be validated in the subsequent studies.

Philippe Chouvarine, Georg Hansmann

chouvarine.philippe@mh-hannover.de (P.C.)
georg.hansmann@gmail.com (G.H.)

Highlights

A network of likely transcriptional regulatory interactions based on total RNA/miRNA-seq

Circular RNA interactions can be analyzed in high-coverage total RNA-seq datasets

Functional labels can be added to the whole network or subnetworks of interest

Existing datasets can be reanalyzed for new transcriptomic insights

Chouvarine & Hansmann, STAR Protocols 2, 100769 September 17, 2021
https://doi.org/10.1016/j.xpro.2021.100769

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Construction of transcriptional regulatory networks using total RNA-seq data

Philippe Chouvarine[1,3,*] and Georg Hansmann[1,2,4,*]

[1]Department of Pediatric Cardiology and Critical Care, Hannover Medical School, Hannover 30625, Germany

[2]Competence Network for Congenital Heart Defects (CNCHD), Berlin, Germany

[3]Technical contact

[4]Lead contact

*Correspondence: chouvarine.philippe@mh-hannover.de (P.C.), georg.hansmann@gmail.com (G.H.)
https://doi.org/10.1016/j.xpro.2021.100769

SUMMARY

Total RNA sequencing allows capturing of long non-coding and circular RNA along with mRNA. Additional sequencing of micro RNA (miRNA), using libraries with shorter fragments, provides the means to characterize miRNA-driven transcriptional regulation. Here, we present a protocol for processing total RNA and miRNA sequencing data to quantify circular RNA, long non-coding RNA, mRNA, and miRNA. Further, the protocol combines the quantification data with miRNA target annotation to construct likely transcriptional regulatory networks, which can be validated in the subsequent studies.

For complete details on the use and execution of this protocol, please refer to Chouvarine et al. (2021).

BEFORE YOU BEGIN

RNA sequencing

Total RNA-Sequencing should be performed using a ribosomal depletion protocol. For sample preparation, we recommend using TruSeq transcriptome libraries following established protocols from Illumina. For sequencing, we recommend at least 100 million 100 bp paired-end reads per sample for total RNA libraries and 20 million 50bp single-end reads per sample for miRNA libraries. External datasets or older in-house datasets are recommended to be of comparable coverage. Lower total RNA coverage may result in poor detection of circular RNA (circRNA), however most long non-coding RNA (lncRNA) would still be adequately detected at lower coverage rates, e.g., 40 million reads per sample.

Software setup

Timing: ~2 h

To facilitate automated installation of most Linux based software used in this protocol the following Miniconda script can be used. The download links for a few programs not included in this script are in the key resource table.

```
conda install -c bioconda fastqc

conda install -c bioconda multiqc

conda install -c bioconda bwa

conda install -c bioconda samtools
```



```
conda install -c bioconda picard

conda install -c bioconda prinseq

conda install -c bioconda ngs-bits # includes SeqPurge

conda install -c bioconda star

conda install -c bioconda rsem

conda install -c bioconda cytoscape # Can also be installed on Windows
```

Data preprocessing and quality control

Timing: 2 days

Note: The command line examples in this protocol include placeholders {…} that should be customized by the user. In all cases, please refer to the command line manuals of the corresponding tools.

Note: Some command line arguments are mandatory (universally applicable) in all situations for a particular step in the protocol. These mandatory arguments are shown in bold font and explained in the notes following the examples.

1. Initial quality control
   a. Analyze the quality of the sequenced reads for each sample using fastqc. (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
   ```
   fastqc {input fasta file} –o data
   ```

   Note: The –o flag specifies the output directory. Multiple input files can be analyzed, for example, by putting this command in a for loop within a script.
   b. To create a single quality report for all samples, merge the fastqc results using multiqc. (Ewels et al., 2016)

   ```
   multiqc data/
   ```

   Note: This step also applies to miRNA-seq (no further preprocessing is needed for miRNA data unless quality issues are detected, which then should be resolved by either contacting the sequencing center or filtration of poor quality reads, e.g., using prinseq-lite.pl.(Schmieder and Edwards, 2011)

2. Removal of ribosomal contamination sequences
   a. Download ribosomal cDNA FASTA sequences from Ensembl Biomart (https://www.ensembl.org/biomart/martview) by selecting "Ensembl Genes 104"; selecting your reference genome (e.g., GRCh38.p13); setting filters to Gene type: ribozyme, rRNA, rRNA_pseudogene; and setting Attributes to "Gene Stable ID" and "cDNA sequences" (this option appears after selecting the Sequences radio button).
   b. Using the resulting FASTA file, create a BWA (Li and Durbin, 2009) reference database.
   ```
   bwa index -a bwtsw {fasta file with reference sequences}
   ```
   c. Align reads from all samples to the database.
   ```
   bwa mem -t 12 -M -R "@RG\tID:{sample_id}\tSM: {sample_id}\tPL:illumina\tLB:
   {sample_id}\tPU: {sample_id}" {reference bwa DB} {input fasta_1.fq} {input
   fasta_2.fq} > {sample_id.rRNA.pe.sam}
   ```

Note: This step is useful to assess ribosomal depletion (the quality of the libraries prepared by the sequencing center). Moreover, once rRNA reads are removed after a quick BWA alignment, it should speed up and improve quality of the transcriptomic alignment using STAR.

Note: The arguments M (Mark shorter split hits as secondary) and R (read group header line) are necessary for compatibility with picard-tools (see Preprocessing step 2.d).

d. Extract the unaligned reads for further processing using samtools (Li et al., 2009) and picard-tools (http://broadinstitute.github.io/picard/)

```
samtools view -u -f 12 -F 256 -h -S {sample_id.rRNA.pe.sam} | java -jar picard-tools-1.77/SamToFastq.jar INPUT=/dev/stdin FASTQ={input fasta_1.fq} SECOND_END_FASTQ={input fasta_2.fq}
```

Note: the required flags –f 12 –F 256 used together result in output of a pair of PE reads, both of which are unmapped. The piped picard SamToFastq command converts the unaligned SAM file into a pair of Fastq files, which are free of ribosomal contaminants.

3. Remove adapter sequences using SeqPurge with default parameters.(Sturm et al., 2016)

```
SeqPurge -in1 {input file1}.fastq.gz -in2 {input file2 with reverse reads}.fastq.gz -out1 {outfile1}.fastq.gz -out2 {outfile2}.fastq.gz
```

4. Trim and filter reads using Prinseq.(Schmieder and Edwards, 2011)

For example, in our study we used:

```
prinseq-lite.pl -fastq {input file1} -fastq2 {input file2 with reverse reads} -out_good {sample name}.fil -out_bad null -log {sample name}.log -min_qual_score 10 -ns_max_n 2 -noniupac -trim_qual_left 20 -trim_qual_right 20 -min_len 30
```

Please refer to the manual for details on the command line arguments.

5. Repeat step 1 to verify whether additional preprocessing may be required. For example, adapter contamination may still be present, quality distribution may not meet the desired standard, any other sequencing artifacts reported by FastQC as problematic, may require a clarification from the sequencing facility.
6. Create STAR (Dobin et al., 2013) reference database for read alignment (using the FASTA and GTF files acquired from ENCODE: ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh38.primary_assembly.genome.fa.gz and ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.primary_assembly.annotation.gtf.gz)

```
STAR –runThreadN 8 –runMode genomeGenerate –genomeDir {directory for STAR reference DB} –genomeFastaFiles GRCh38.primary_assembly.genome.fa –sjdbGTFfile gencode.v38.primary_assembly.annotation.gtf –sjdbOverhang 99
```

Note: The –sjdbOverhang flag should be set to the read length of each paired end minus 1 (i.e., 99 for 100 bp PE reads). This parameter specifies how many bases can be concatenated from the donor and acceptor sides of a junction, i.e., up to 99 on one side and as few as one on the other side. Thus, the generated reference will be able to handle any split reads (split by a junction) of length up to 100 bases.

7. Create STAR reference database for miRNA alignment (using the same input files, but a different –sjdbOverhang value).

```
STAR –runThreadN 8 –runMode genomeGenerate –genomeDir {directory for miRNA STAR reference DB} –genomeFastaFiles GRCh38.primary_assembly.genome.fa –sjdbGTFfile gencode.v38.primary_assembly.annotation.gtf –sjdbOverhang 1
```

Note: for miRNA alignment reference database, splice junction references are effectively excluded using **–sjdbOverhang 1**, while still utilizing the GTF file for miRNA annotation.

8. Prepare RSEM (Li and Dewey, 2011) reference database using the same FASTA and GTF input files.

```
rsem-prepare-reference –gtf gencode.v38.primary_assembly.annotation.gtf GRCh38.primary_assembly.genome.fa RSEM/GRCh38_p13
```

9. Run the STARchip (https://starchip.readthedocs.io/en/latest/) setup script using the same FASTA and GTF input files. STARchip is the program designed to identify circular RNA (and fusions).

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human RV tissue (10–50 mg) was used for total RNA and miRNA extraction according to the TRIzol protocol (TRIzol, Life Technologies).	(Chouvarine et al., 2021)	N/A
Critical commercial assays		
RNAlater®-ICE Frozen Tissue Transition Solution	Ambion	Cat. No. AM7030
TRIzol™ Reagent	Invitrogen	Cat. No. 15596026
Deposited data		
Total RNA sequencing	(Chouvarine et al, 2021)	Controlled access via National Register for Congenital Heart Defects, Berlin, Germany (ubauer@kompetenznetz-ahf.de)
Total RNA sequencing (alternative dataset for testing)	(Ogoyama et al., 2021)	SRA: SRP298758
miRNA sequencing	(Chouvarine et al, 2021)	Controlled access via National Register for Congenital Heart Defects, Berlin, Germany (ubauer@kompetenznetz-ahf.de)
miRNA sequencing (alternative dataset for testing)	(Ogoyama et al., 2021)	SRA: SRP298758
Software and algorithms		
FastQC (V0.10.1)	Unpublished	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC (v1.10.1)	(Ewels et al., 2016)	https://multiqc.info/
BWA (0.7.12-r1039)	(Li and Durbin, 2009)	https://github.com/lh3/bwa
SAMtools (v. 1.7)	(Li et al., 2009)	http://www.htslib.org/download/
Picard-tools (v. 1.77)	Unpublished	https://broadinstitute.github.io/picard/
PRINSEQ (v. 0.20.4)	(Schmieder and Edwards, 2011)	https://sourceforge.net/projects/prinseq/files/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SeqPurge (v. 2020_03-159-g5c8b2e82)	(Sturm et al., 2016)	https://github.com/imgag/ngs-bits
STAR (v. 2.7.5a)	(Dobin et al., 2013)	https://github.com/alexdobin/STAR
RSEM (v1.3.0)	(Li and Dewey, 2011)	https://deweylab.github.io/RSEM/
STARChip (v. 1.3e)	(Akers et al., 2018)	https://starchip.readthedocs.io/en/latest/
Cytoscape (v. 3.7.2)	(Shannon et al., 2003)	https://cytoscape.org/
FactoMineR R package (v. 2.4)	(Lê et al., 2008)	http://factominer.free.fr/more/reference.html
DESeq2 R package (v. 1.30.1)	(Love et al., 2014)	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
EDASeq R package (v. 2.24.0)	(Risso et al., 2011)	https://bioconductor.org/packages/release/bioc/html/EDASeq.html
Gene filtering procedure to increase detection power of differential expression analysis	(Bourgon et al., 2010)	https://www.pnas.org/content/107/21/9546
Enrichr	(Chen et al., 2013)	https://amp.pharm.mssm.edu/Enrichr/
TargetScan 7.1	(Agarwal et al., 2015)	http://www.targetscan.org/vert_71/
miRDB	(Chen and Wang, 2020)	http://mirdb.org/index.html
Perl scripts for subtraction of circRNA counts from total RNA	(Chouvarine et al., 2021)	https://github.com/pch-code/circ-scripts

Note: It is not critical to use the specified software versions. Currently, the latest versions would provide the same functionality.

STEP-BY-STEP METHOD DETAILS

Read alignment

Timing: ~2 days (depending on the number of samples and computing capacity)

This major step results in three sets of read alignments applicable for three types of RNA: mRNA/lncRNA, circRNA, and miRNA.

1. Transcriptome read alignment using STAR

```
STAR –runThreadN 20 –runMode alignReads –twopassMode Basic –outSAMattrRGline ID:{sample name} LB:lib1 PL:ILLUMINA PU:unit1 SM:{sample name} –genomeDir Reference/Human/Genecode_GRCh38.p13/STAR_100bp –sjdbGTFfile Reference/Human/Genecode_GRCh38.p13/gencode.v34.primary_assembly.annotation.gtf –sjdbOverhang 99 –readFilesIn {input fasta_1.fq.gz} {input fasta_2.fq.gz} –outFileNamePrefix ./Alignments/$outfile. –quantMode TranscriptomeSAM –outSAMtype BAM SortedByCoordinate
```

Note: The alignment parameters for transcriptome alignment are according to the recommendations in the STAR manual.

2. Chimeric read alignment using STAR (for circRNA detection)

```
STAR –runThreadN 20 –genomeDir Reference/Human/Genecode_GRCh38.p13/STAR_100bp –readFilesIn {input fasta_1.fq.gz} {input fasta_2.fq.gz} –outFileNamePrefix Alignments_chim/{sample name}/ –outReadsUnmapped Fastx –quantMode GeneCounts –chimSegmentMin 15 –chimJunctionOverhangMin 15 –outSAMstrandField intronMotif –readFilesCommand zcat –outSAMtype BAM Unsorted
```

Note: the chimeric alignment parameters are according to the recommendations in the STARChip manual (https://starchip.readthedocs.io/en/latest/).

Note: chimeric alignments happen when a read or a pair of reads align to two distinct genomic segments (on different chromosomes, different strands, or far apart) indicative of a structural variation, e.g., a circle. The chimSegmentMin parameter indicates that either of the two segments in the chimeric alignment must be at least 15 bases. The chimJunctionOverhangMin parameter specifies that reads with less than 15 bases overhanging the chimeric junction should not be considered.

3. miRNA read alignment

```
STAR –outSAMattrRGline ID:{sample name} LB:lib1 PL:ILLUMINA PU:unit1 SM:{sample name} –readFilesIn {input fasta.fq.gz} –outFileNamePrefix ./Alignments/{sample name} –runThreadN 20 –genomeDir Reference/Human/Genecode_GRCh38.p13/STAR_sRNA_ENCODE –sjdbGTFfile Reference/Human/Genecode_GRCh38.p13/GENCODE_miRNA_subset.gtf –alignEndsType EndToEnd –outFilterMismatchNmax 1 –outFilterMultimapScoreRange 0 –quantMode TranscriptomeSAM GeneCounts –outReadsUnmapped Fastx –outSAMtype BAM SortedByCoordinate –outFilterMultimapNmax 10 –outSAMunmapped Within –outFilterScoreMinOverLread 0 –outFilterMatchNminOverLread 0 –outFilterMatchNmin 16 –alignSJDBoverhangMin 1000 –alignIntronMax 1 –outWigType wiggle –outWigStrand Stranded –outWigNorm RPM
```

Note: The miRNA alignment parameters are according the ENCODE protocol for miRNA-seq read alignment using STAR aligner (https://www.encodeproject.org/documents/b4ec4567-ac4e-4812-b2bd-e1d2df746966/@@download/attachment/ENCODE_miRNA-seq_STAR_parameters_v2.pdf).

Feature read counting

Timing: 2 days

This major step identifies total read counts for each feature. Since gene read counts will contain a mixture of mRNA and circRNA reads, they should be separated. This major step estimates the circRNA read counts for each detected circRNA and subtracts them from the total RNA counts to estimate mRNA read counts for the corresponding genes.

4. Run RSEM to perform initial feature read counting

```
rsem-calculate-expression –no-bam-output -p 12 –alignments –paired-end Alignments/{file name}.Aligned.toTranscriptome.out.bam {path to RSEM reference DB} {outfile}
```

Note: For miRNA alignments, running RSEM is not necessary. Instead the read counts should be extracted from the *.ReadsPerGene.out.tab files generated by STAR.

5. Estimate circRNA counts and calculate other feature counts
   a. Run STARchip with default parameters using the chimeric alignments from step 2.

Note: Below are the instructions from the manual for our in-house pipeline, also included in the Github repository (https://github.com/pch-code/circ-scripts/) to estimate the circRNA read counts and to correct the initial read counts by subtracting the circRNA read counts. This correction does not apply to the miRNA read counts.

b. Create a bed file of merged exonic coordinates and flanking introns based on the circRNA isoform coordinates (will use it to get exact GC and length for DESeq).
   i. MakeExonBedForMerging.pl -i {Main.gtf to create the *.exonsByGene.bed file}
   Using a rat genome example. Convert the main gtf file into a bed file.
   ```
   perl MakeExonBedForMerging.pl -i Rattus_norvegicus.Rnor_6.0.96.gtf -o Rattus_norvegicus.Rnor_6.0.96.exonsByGene.bed
   ```

The output looks like this:

1	396700	396905	ENSRNOG00000046319:AABR07000046.1
1	397780	397788	ENSRNOG00000046319:AABR07000046.1
1	399062	399070	ENSRNOG00000046319:AABR07000046.1
1	399557	399827	ENSRNOG00000046319:AABR07000046.1
1	400256	401059	ENSRNOG00000046319:AABR07000046.1
1	401912	402136	ENSRNOG00000046319:AABR07000046.1

   ii. Merge using bedops:
   ```
   cut -f4 Rattus_norvegicus.Rnor_6.0.96.exonsByGene.bed | sort | uniq | awk -F'_' '{ system("grep "$1" Rattus_norvegicus.Rnor_6.0.96.exonsByGene.bed | bedops –merge - "); print $0; }' > Rattus_norvegicus.Rnor_6.0.96.MergedExonsByGene.txt
   ```

The output looks like this (coordinates for each exon followed by a line with gene ID and gene symbol):

2	230660664	230660669
2	230660675	230662084
ENSRNOG00000000001:AABR07013255.1		
.....		

   iii. Make merged CircRNA.bed file from Rattus_norvegicus.Rnor_6.0.96.MergedExonsByGene.txt
   ```
   perl GetExonsAndFlankingIntronsRegionsPerAnnotatedCircRNA.pl -i Rattus_norvegicus.Rnor_6.0.96.MergedExonsByGene.txt -c circRNA.5reads.1ind.annotated –o circRNAExonicCoords.txt
   ```

Note: -i here is the output of step ii.

The output looks like this:

chr10	101667886	101667980	ENSG00000107829:FBXW4:chr10:101667886:101676436
chr10	101672915	101673047	ENSG00000107829:FBXW4:chr10:101667886:101676436
chr10	101673488	101673673	ENSG00000107829:FBXW4:chr10:101667886:101676436

c. Create a file with circRNA counts
   ```
   bedtools nuc -fi ~/Reference/Rat/Rattus_norvegicus.Rnor_6.0.dna.toplevel.fa -bed circRNAExonicCoords.txt > MergedIntervalsGC.txt
   echo -e "Gene\tgc\tlength" > CircRNA_GC_Len.txt && awk '{print $4"\t"$6"\t"$13}' MergedIntervalsGC.txt | grep "ENS" >> CircRNA_GC_Len.txt
   perl AddExonLengths_WeightedAvgGC_Annot.pl -i CircRNA_GC_Len.txt -o CollapsedCircRNA_GC_Len.txt # this will also create
   ```

AnnotatedCircRNA_Counts.txt
The output contains circRNA coordinates in the first column and read counts for each sample in the subsequent columns.

d. Create a file with chromosomal coordinates for the identified circRNA
perl makeBed.pl -i AnnotatedCircRNA_Counts.txt -o circRNA.bed

The output contains circRNA coordinates:

10	16724062	16730140	Crebrf:ENSRNOG00000020769
10	51724607	51739115	Myocd:ENSRNOG00000003669
1	116642079	116660509	Ube3a:ENSRNOG00000015734
1	126573474	126583015	Tm2d3:ENSRNOG00000011290
...			

e. Run RSEM with the –output-genome-bam and –sampling-for-bam parameters. This will generate a genome-level BAM with each read mapped to a single location (${outfile}.genome.bam)

```
for file in Alignments_transcriptome/{sample name}.Aligned.toTranscriptome.out.bam
do
   outfile=${file%%.Aligned.toTranscriptome.out.bam}
   outfile=${outfile##Alignments_transcriptome/}
   rsem-calculate-expression –output-genome-bam –sampling-for-bam -p 12
   –alignments –paired-end –strandedness reverse
   Alignments_transcriptome/$outfile.Aligned.toTranscriptome.out.bam
   {rsem-reference} $outfile
done
```

f. Use the files created in steps d. and e. to generate circRNA coverage files

```
for file in *.genome.bam
do
   outfile=${file%%.genome.bam}
   bedtools coverage –abam $file –b circRNA.bed > $outfile.cov.txt
done
```

g. Calculate median insert size for every sample. Copy all output to a subfolder. It will be used to estimate percentage of circRNA in the next step.

Median insert size is will be passed to EstCircCounts_FragCountNormByExonicLen.pl in step h.

```
for file in *.genome.bam
do
   outfile=${file%%.genome.bam}
   java -jar picard.jar CollectInsertSizeMetrics I=$file
   O=$outfile.InsSize.txt H=$outfile.InsSize.pdf

done
```

h. Estimate counts and percentages of each circRNA isoform:

```
perl EstCircCounts_FragCountNormByExonicLen.pl -i circRNARegions_TotalRNACov -c AnnotatedCircRNA_Counts.txt -s circs5.1.investigate.consensus –m {average median insert size for PE reads across all samples} –r {read length of PE reads} -o EstCircRNAProportions.txt
```

-i is the directory with the Total coverages

-c created in step c. (AnnotatedCircRNA_Counts.txt)
-m median insert size is calculated in step g.

i. Subtract the estimated circRNA counts from the total RNA counts:

```
perl SubtractTheCircCounts.pl -t {TotalRNA counts file}.OrigNames.txt -c Est-
CircRNAProportions.txt.circCounts.txt -o mRNACounts.txt
```

Note: The {TotalRNA counts file}.OrigNames.txt, contains total RNA counts (gene level read counts generated by RSEM, in which the first line is a header and the subsequent lines contain Ensembl Gene ID in the first column and read counts for each sample in the following columns).

The file can be prepared using:

```
perl PrepareCountsForDESeq_FromRSEM.pl -i counts_file_list_in_Group1 -j
counts_file_list_in_Group2 -n Group1Name -m Group2Name -o out_for_DESeq.txt
```

Principal component analysis (PCA) for detection of outlier samples

Timing: 1 h

This major step outlines how to perform PCA to exclude outlier samples whose abnormal expression may indicate overlooked phenotypical features excluding such samples from further analysis.

Note: Read counts can be a) converted to variance stabilized input values for PCA (step 3) and b) used as input for differential expression analysis (step 4). In our study (Chouvarine et al., 2021) we used GC-bias correction as implemented in the EDASeq R package.(Risso et al., 2011) This tool can be integrated with DESeq2 as described in the EDASeq manual. In this case, the GC annotation for each gene is downloaded from Ensembl BioMart. An alternative simpler approach would be to use DESeq2 without the GC-bias correction.

6. Exclude features with the average read count across all samples $\leq$ 10.

```
filter <- apply(countsTable,1,function(x) mean(x)>10)

countsTable = countsTable[filter,]
```

7. Run the DESeq2/EDASeq R packages (as described in the EDASeq manual) applying the within-LaneNormalization function. (optional)

Briefly, the dataOffset object is created to pass the GC normalized counts to DESeq2:

```
dataOffset <- withinLaneNormalization(data,"GC",which="full",offset=TRUE)
```

Note: the data object was created using the newSeqExpressionSet function on the counts table that also includes a GC column, added by merging an annotation table containing Ensembl Gene ID, GC-content, and other columns of interest (can be obtained from BioMart).

```
dds <- DESeqDataSetFromMatrix(countData = counts(dataOffset),

                              colData = pData(dataOffset),

                              design = ~ conditions)
```

```
normFactors <- exp(-1 * offst(dataOffset))

normFactors <- normFactors / exp(rowMeans(log(normFactors)))

normalizationFactors(dds) <- normFactors
```

8. Acquire variance stabilized data.

```
vsd <- vst(dds, blind=FALSE)
```

9. Run PCA using the FactoMineR and factoextra R packages.(Lê et al., 2008) or any other preferred PCA package

```
pca_Pat = PCA(tvsd_df, scale.unit = F, quali.sup=1:5, graph = TRUE)
```

Note: tvsd_df contains transposed matrix with variance stabilized data t(vsd) and additional annotation, e.g., case-control group. In this example, Columns 1 through 5 (specified in quali.sup) contain annotation and are not used for dimensionality reduction.

10. Display combinations of the first three components using the fviz_pca_ind function (assuming that the factoextra package is used for PCA visualization) to visualize potential outlier samples.

```
fviz_pca_ind(pca_Pat, col.ind = tvsd_df$Gr, label="ind", palette = c("#00AFBB",
"#FC4E07"), addEllipses = TRUE, legend.title = "Groups", repel = TRUE, title="In-
dividuals by disease - PC1 vs PC2")

fviz_pca_ind(pca_Pat, axes=c(1,3), col.ind = tvsd_df$Gr, label="ind", pal-
ette = c("#00AFBB", "#FC4E07"), addEllipses = TRUE, legend.title = "Groups",
repel = TRUE, title="Individuals by disease - PC1 vs PC3")
```

Differential expression analysis

Timing: 2 h

This major step should be separately applied to all types of RNA (mRNA, lncRNA, circRNA, and miRNA) to identify their differential expression for the groups in the study.

11. Run DESeq2 as described in the previous step, but this time generate differential expression output with FDR-adjusted P-values.

```
dds <- DESeq(dds)

res <- results(dds)
```

Note: By default, the DESeq function performs Wald test, which is appropriate in most cases (when two levels of a factor are compared, e.g., case vs. control). Another available method provided by this function is "LRT" (likelihood ratio test), which is appropriate for simultaneous comparison of multiple levels of a factor, e.g., multiple time points.

Note: The results function has an important parameter independentFiltering set to True by default. It applies the gene filtering procedure developed by Bourgon et al.(Bourgon et al., 2010) to improve the detection power. We recommend to not change this setting and perform

the gene filtering procedure. RNAs with FDR-adjusted P values < 0.05 after filtration will be considered significantly differentially expressed.

Construction of circRNA/lncRNA-miRNA-mRNA network

Timing: 2 days

This major step identifies targets of miRNAs and combines them in a regulatory network.

12. Identify gene targets for differentially expressed miRNA using miRDB(Chen and Wang, 2020) as the source of gene targets, considering all targets with the miRDB Target Score $\geq$ 50 as significant.

This step can be achieved either using the Target Mining tool provided by the miRDB website or by downloading the entire dataset from the website (http://mirdb.org/download/miRDB_v6.0_prediction_result.txt.gz) and setting up a database query.

13. Using TargetScan 7.1(Agarwal et al., 2015), identify lncRNA and circRNA targets based on the number of occurrences of matching 6-mers, 7-mers, and 8-mers. The affinity score is calculated as Score = 0.43* N_{8mer_1a} + 0.25* N_{7mer_m8} + 0.19* N_{7mer_1a} + 0.07* $N_{6mer.}$ Here, 8mer_1a is a site with an exact match to positions 2–8 of the mature miRNA (the seed) followed by an 'A', 7mer_m8 is a site with exact match to positions 2–8 of the mature miRNA (the seed + position 8), 7mer_1a is a site with an exact match to positions 2–7 of the mature miRNA (the seed) followed by an 'A', 6mer is a site with an exact match to positions 2–6 of the mature miRNA (the seed + position 6) followed by an 'A'. All matches with the score $\geq$ 1 will be considered significant. The scores can also be used to set the line thickness parameter in Cytoscape (see Step 15).

Calculating the scores as specified above can be automated using our script (https://github.com/pch-code/TargetScanScore/blob/main/scoreTargets.pl) that takes TargetScan output as input:

```
perl scoreTargets.pl -i {output from TargetScan} -o out
```

14. Select the differentially expressed mRNA, lncRNA, and circRNA identified in step 11 that are either on the gene target list (step 12) or the lncRNA/circRNA target list (step 13). The selected interactions will be the basis for competitive endogenous RNA interaction network focusing on differential expression regulation for a particular condition (compared to normal). Optionally, for a more focused view, we can select only those molecules whose differential expression is in the opposite direction compared to differential expression of the miRNAs targeting them.
15. Using the miRNA-target interactions identified in Step 14, create a network in Cytoscape.(Shannon et al., 2003) The interactions can be added to a *.sif file for creation of the network structure. Each line in a *.sif file represents a connection (edge) between two nodes (i.e., miRNA and its target). An exhaustive list of such interactions fully describes the likely competing endogenous RNA interaction network in a form of a graph. Log2(fold change) gene expression values can be used for coloring the nodes (passed to Cytoscape in a separate table). Line thickness can be used to indicate the strength of the interactions based on the corresponding target/affinity scores (passed to Cytoscape in a separate table).

Functional annotation of the network

Timing: 1 day

This major step outlines potential sources of annotation for the network created in the previous major step.

Note: Upon creation of the desired network (step 15), functional annotation labels can be added to visualize biological significance of individual nodes or a group of nodes. We recommend using a combination of automated annotation tools, such as Enrichr,(Chen et al., 2013) and manual annotation based on literature search and expert knowledge of the scientists involved in the project. In our study, automated annotation was primarily based on pathway and GO overrepresentation analyses of DEGs using the online resource Enrichr with the NCATS BioPlanet pathway annotation (https://ncats.nih.gov/pubs/features/bioplanet).

16. (Assuming Enrichr is used) generate a gene list as input. For a particular network, a gene list of selected nodes can be generated in Cytoscape.
17. Submit the gene list to the form on the Enrichr website.
18. Select the Pathways menu and the annotation source, e.g., BioPlanet. Download the results as a table.
19. Select the Ontologies menu and the annotation source, e.g., GO Biological Process. Download the results as a table.

Note: The overrepresentation analysis obtained in steps 18 and 19 is based on Fisher exact test and correction of the resulting p-values for multiple testing as implemented in Enrichr. Pathways with the adjusted p-value < 0.05 can be considered significantly overrepresented by DEGs.

20. Add the functional labels manually in Cytoscape, using the graphical user interface.

EXPECTED OUTCOMES

Standard command line processing tools used in this protocol produce output in the format described in the corresponding software manuals. The output formats of the online tools, i.e., Ensembl BioMart, miRDB, and Enrichr, are described in the corresponding online documentation.

Figure 1 illustrates the quality control step based on PCA output generated in step 3. Visual inspection of the original data (Figure 1A) suggested that Control.7 is a likely outlier. Inspection of the clinical data confirmed Control.7 as the outlier. Figure 1B shows the PCA output generated using the final cohort of the study, after the outlier sample was removed.

The expected output from steps 15–20 is presented in Figure 2. Depending on the size of the network, it may be desirable to output only a subset of the network. This can be done in Cytoscape by creating a subnetwork of selected nodes.

LIMITATIONS

Poor experimental design and/or sample quality can result in preparation of poor quality sequencing data, presence of outlier samples, and batch effects. In our protocol and the troubleshooting section below, we describe quality control steps and measures to mitigate these issues. However, if these problems are severe and the final quality control shows failure of the mitigation steps, then the protocol cannot be applied.

TROUBLESHOOTING

Problem 1

Read contamination, presence of poor quality sequencing reads, presence of adapter sequences (preprocessing step 1).

Potential solution

Step 2 of the data preprocessing and quality control section describes how to remove ribosomal sequences. The same procedure can be applied to remove any other contaminating sequences. Poor quality reads should be trimmed or filtered out (see step 3 in the preprocessing section). Adapter

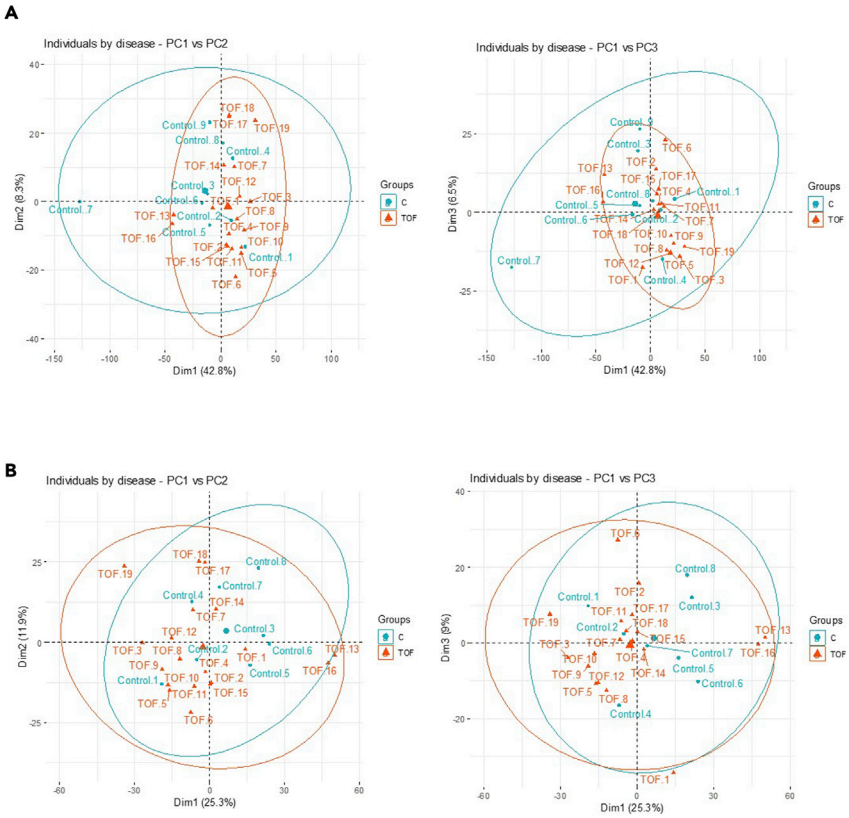


Figure 1. Principle component analysis (PCA) of individual samples allows finding potential outlier samples

(A) PCA graph of the first three components with the outlier (Control.7).

(B) PCA graph of the first three components after removal of the outlier.

sequences should normally be removed during preprocessing (preprocessing step 4). However, if their presence is still detected, they can be removed using alternative software (e.g., Cutadapt, https://cutadapt.readthedocs.io/en/stable/) in which the adapter sequences are specified as input parameters.

Problem 2

Failed ribosomal depletion during the library preparation (preprocessing Step 2).

Potential solution

Preprocessing step 2 may show that ribosomal sequences constitute the majority of the sequenced reads. In this case, after removal of the ribosomal sequences, the remaining read coverage may be insufficient for further processing (inability to detect lowly expressed genes/isoforms and circRNAs). Since this problem cannot be addressed programmatically, the affected library should be constructed and sequenced again.

Problem 3

Poor sample selection may result in presence of outlier samples (step 10).

Potential solution

As described in the expected outcomes section the PCA quality control step should detect any outlier samples that should be removed if the sample selection mistake is confirmed by clinical, demographic, or other data.

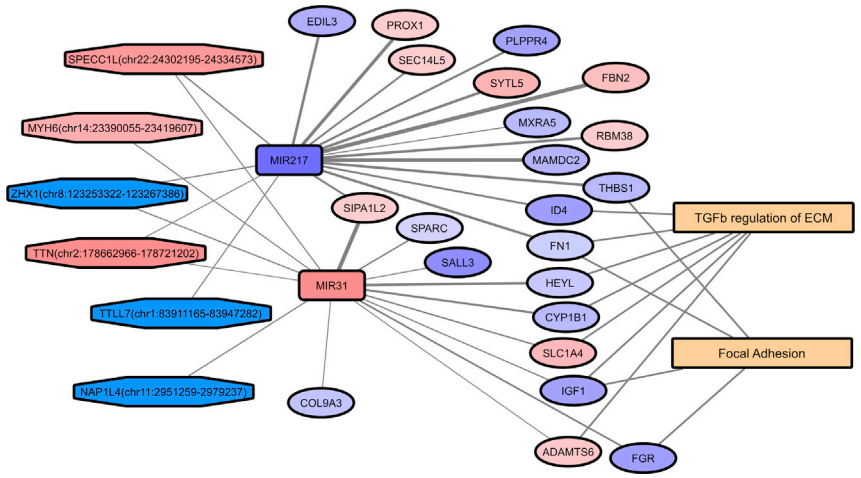


Figure 2. An example of competing endogenous RNA network with functional annotation labels

The network appearance is customizable in Cytoscape using the data obtained by following the steps described in this protocol.

Problem 4

Presence of batch effects (step 10).

Potential solution

We recommend planning the experiments, sequencing, and data processing as uniformly as possible to avoid any potential batch effects. However, if interpretation of the PCA plots points at potential batch effects, which can be traced to a particular factor (source of batch effects), e.g., multiple sample preparation dates, multiple technicians, sequencing flow cell or lane specificity, etc., then we recommend to apply the removeBatchEffect function from the limma R package:

```
assay(vsd) <- limma::removeBatchEffect(assay(vsd), vsd$batch, design=~Con
dition)
```

Redrawing the PCA plots after removal of the batch effect should indicate whether the problem has been addressed.

Problem 5

Due to various levels of ribosomal contamination or other technical artifacts, the depth of sequencing may vary between the samples. This difference in coverage will result in different effect sizes that would still yield the desired statistical power, e.g., $\geq$0.8. In the worst-case scenario, such difference can be specific to the groups (case and control).

Potential solution

To avoid interpretation of unreliable differential expression results not reaching the desired statistical power, the effect size (fold change) threshold can be set to filter out genes with the fold change below the calculated threshold. This effect size calculation based on predetermined coverage, sample size, coefficient of variation, false positive rate, and power can be performed using the rnapower function form the RNASeqPower R package. For example, rnapower(depth=15, n=10, cv=0.1, alpha=0.05, power=0.8) returns fold the change cutoff of 1.41. Please refer to the manual for description of the variables.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Georg Hansmann (georg.hansmann@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This study did not generate any software, except for Perl/bash scripts for subtracting circRNA read counts from the total RNA read count data, which are deposited to https://github.com/pch-code/circ-scripts. Sequencing data are not publicly available due to consent restrictions. Upon request, the data can be made available via controlled access (National Register for Congenital Heart Defects, Berlin, Germany). Alternatively, a publically available dataset (NCBI SRA accession SRP298758) can be used for testing.

ACKNOWLEDGMENTS

This study was supported by the German Research Foundation (DFG; HA4348/6-2 KFO311 to G.H.). G.H. receives additional funding from the German Research Foundation (DFG; HA4348/2-2), the Federal Ministry of Education and Research (BMBF ViP+ program 03VP08053; BMBF 01KC2001B), and the European Pediatric Pulmonary Vascular Disease Network (www.pvdnetwork.org). The Competence Network for Congenital Heart Defects and the National Register for Congenital Heart Defects have received funding from the Federal Ministry of Education and Research, grant number 01GI0601 (until 2014), and the DZHK (German Center for Cardiovascular Research; as of 2015).

AUTHOR CONTRIBUTIONS

P.C. performed the data analysis and wrote the manuscript. G.H. generated the hypotheses, developed the experimental design and concept of the study, performed RNA extraction and bioanalysis, and obtained funding. Both authors critically read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Agarwal, V., Bell, G.W., Nam, J.W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife *4*, e05005.

Akers, N.K., Schadt, E.E., and Losic, B. (2018). STAR chimeric post for rapid detection of circular RNA and fusion transcripts. Bioinformatics *34*, 2364–2370.

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. Proc. Natl. Acad. Sci. U S A *107*, 9546–9551.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128.

Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. *48*, D127–D131.

Chouvarine, P., Photiadis, J., Cesnjevar, R., Scheewe, J., Bauer, U.M.M., Pickardt, T., Kramer, H.H., Dittrich, S., Berger, F., and Hansmann, G. (2021). RNA expression profiles and regulatory networks in human right ventricular hypertrophy due to high pressure load. iScience *24*, 102232.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Ewels, P., Magnusson, M., Lundin, S., and Kaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An R package for multivariate analysis. J. Stat. Softw. *25*, 1–18.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Ogoyama, M., Ohkuchi, A., Takahashi, H., Zhao, D., Matsubara, S., and Takizawa, T. (2021). LncRNA H19-derived miR-675-5p accelerates the invasion of extravillous trophoblast cells by inhibiting GATA2 and subsequently activating matrix metalloproteinases. Int. J. Mol. Sci. *22*, 1237.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. BMC Bioinformatics *12*, 480.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Sturm, M., Schroeder, C., and Bauer, P. (2016). SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. BMC Bioinformatics *17*, 208.