ITQB-Nova - Universidade Nova de Lisboa	EPP-SOP-ITQB03
	Version 1.0

EPP - Standard Operating Procedure

(only for selected experiments intended to transfer results from one lab to the other)

Title: Differential RNA-Seq data analysis

distribution list			
changes to prior version:			
	name	signature	date
experimenter 1	Vânia Pobre		14.03.2019

Instruction

Analysis of RNA-Seq data to determine differentially expressed genes

1 Introduction

This protocol describes the analysis of RNA-Seq data to identify differentially expressed genes between two samples. This protocol is simply the analysis of the data and does not include the sequencing protocol. All programs described in this protocol are free to use. A more detailed explanation for all steps can be found in the book chapter: Pobre, V. & Arraiano, C. M. Characterizing the Role of Exoribonucleases in the Control of Microbial Gene Expression: Differential RNA-Seq. Methods Enzymol 612, 1-24, doi:10.1016/bs.mie.2018.08.010 (2018).

2 Equipment, software and data

2.1Equipment

a. Computer with Linux operating system.

b. Data storage external drive (this type of data can have several terabytes and should be stored in a secured data storage independent of the computer to prevent loss of data).

2.2Software

a. FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

b. Cutadapt (Martin 2011)

c. Bowtie2 (Langmead and Salzberg 2012)

d. Samtools (Li, Handsaker et al. 2009)

e. FeatureCounts (Liao, Smyth et al. 2014)

f. RStudio (https://www.rstudio.com)

g. edgeR package from R Bioconductor (Robinson, McCarthy et al. 2010)

2.3Data

Illumina RNA-Seq paired-end data, 20M Reads, 150bp – two fastq files per sample (R1 and R2) corresponding to the reverse and forward strand sequencing. Biological triplicates per sample.

3 Procedures

3.1 Quality control and filtering of the data

a. Open the fastq files with FastQC.

b. Check the quality reports to identify low quality reads or contamination with adaptors and/or barcodes.

c. Use cutadapt to remove adaptor sequences contaminant and to trim low-quality nucleotides and reads from the data. The command line for cutadapt should be written in accordance with the FastQC results. An example of a command line to remove adapters and low quality nucleotides from the reads is: `cutadapt -u 20 -U 20 –-minimum-length 50 -a ADAPTER_FWD -A ADAPTER_REV -o out.1.fastq -p out.2.fastq reads.1.fastq reads.2.fastq`

d. Open the trimmed fastq files with FastQC to confirm that all samples now have good quality and no adapter and barcodes contamination.

3.2 Mapping

a. Download the *Pseudomona putida* genome (.fa) and annotation (.gff) from NCBI database.

b. To create an index of the genome run Bowtie2 command: `bowtie2-build genome.fa genome_ind`

c. Assuming that the libraries for the RNA-Seq were done with Truseq kit from Illumina the R1 files correspond to the reverse strand sequence and the R2 files correspond to the forward strand sequence. In this case the command to map the reads with the genome the command should be: `bowtie2 genome_ind -1 sample_R1.fastq -2 sample_R2.fastq --un sample_UN -S sample.sam`

d. Check that the alignment rate is higher than 90% and that most of the reads aligned exactly one time, if not recheck the quality of the data to identify problems that can be interfering with the mapping.

3.3 Reads Quantification

a. To be able to quantify the reads the mapping files need to be converted to BAM files and sorted. For these the following commands from Samtools should be used:

   i. `samtools view -b -o sample.bam sample.sam`

   ii. `samtools sort -O bam -T tem_bam sample.bam > sample_sort.bam`

b. To quantify the number of reads per transcript use the following command line: `featureCounts -s 1 -p -a genome.gtf -o featureCounts_sample sample.bam`

c. The number of reads can be normalized using the following formula:

i. Count up the total reads in a sample and divide that number by 1,000,000 – this is our “per million” scaling factor.

ii. Divide the read counts by the “per million” scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)

iii. Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

3.4 Differential expression analysis

a. Using RStudio is the easiest way to use the R package necessary for the differential analysis of RNA-Seq data. Initializing RStudio as administrator is essential so that the packages can be downloaded and installed as well as creating images and tables.

b. Create a TableOfCounts.txt that should be saved as a tab delimited file. The table to be prepared must have in the first column the identification of the transcripts, the second, third and fourth columns is the number of reads for each transcript for the sample 1 biological replicates and the fifth, sixth and seventh columns is the number of reads for sample 2 biological replicates.

c. The script for edgeR that calculates the differential expressed transcripts and creates a MA scatterplot and a table with the results (`edgeR_results.csv`) should be the following:

```
> x <- read.delim("TableOfCounts.txt",row.names="Transcript")

> group <- factor(c(1,1,1,2,2,2))

> y <- DGEList(counts=x,group=group)

> y <- calcNormFactors(y)

> design <- model.matrix(~group)

> y <- estimateDisp(y,design)

> et <- exactTest(y)

> topTags(et)

>edgeR_results <- topTags(et, n=6000)

>write.csv2(edgeR_results, "edgeR_results.csv")

>summary(de <- decideTestsDGE(et, p=0.05, adjust="BH"))

>detags <- rownames(et)[as.logical(de)]

>plotSmear(et, de.tags=detags)

>abline(h = c(-2, 2), col = "blue")

>abline(v = c(3), col = "blue")
```

4 Biosafety

There are no biosafety issues associated with this protocol.

5 Notes

In the edgeR script the line `edgeR_results <- topTags(et, n=6000),` **n** should be adjusted to the total number of transcripts that are present in the TableOfCounts.

6 References

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Liao, Y., G. K. Smyth and W. Shi (2014). "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." Bioinformatics **30**(7): 923-930.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." 2011 **17**(1): 3.

Pobre, V. & Arraiano, C. M. Characterizing the Role of Exoribonucleases in the Control of Microbial Gene Expression: Differential RNA-Seq. Methods Enzymol 612, 1-24, doi:10.1016/bs.mie.2018.08.010 (2018).

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

7 Acknowledgements

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 635536.