Chapter 2

Whole-Transcriptome Sequencing for High-Resolution Transcriptomic Analysis in *Mycobacterium tuberculosis*

Andrej Benjak, Claudia Sala, and Ruben C. Hartkoorn

Abstract

RNA-seq uses next-generation sequencing technology to determine the transcription profile of an organism in a quantitative manner. With respect to microarrays, this methodology allows greater resolution, increased dynamic range, and identification of new features such as previously unannotated genes and noncoding RNAs. Here we describe how to extract RNA from mycobacterial cultures, how to prepare libraries for Illumina sequencing, and the bioinformatics analysis of the sequencing data to determine the transcription profile.

Key words Illumina, RNA-seq, Transcriptome, RNA extraction, Library preparation

1 Introduction

With the advent of next-generation sequencing (NGS) technology, RNA-seq (RNA sequencing) is now becoming a well-established approach for transcript quantification and gene expression studies. The advantage of RNA-seq compared to transcriptome analysis using microarrays is in lower background noise (reads can be unambiguously mapped to unique regions of the genome), a higher dynamic range of expression levels and the potential for discovering novel genes and other transcribed features. A broad overview and practical guidance for the RNA-seq work flow is given in ref. [1].

For the successful use of RNA-seq (as with microarrays) a basic requirement is that there exists a reliable reference genome sequence onto which reads can be mapped (Cross-reference to WGS chapter). In principle, the amount of transcription is then determined by sequencing the RNA from the test sample(s), mapping them onto the reference genome, and subsequently quantifying the number of times any particular feature (base/gene) is covered by the sequence reads from NGS technology.

Tanya Parish and David M. Roberts (eds.), *Mycobacteria Protocols*, Methods in Molecular Biology, vol. 1285,
DOI 10.1007/978-1-4939-2450-9_2, © Springer Science+Business Media New York 2015

Some aspects of RNA-seq experiments deserve more attention. In particular, the number of biological replicates is crucial for obtaining meaningful results [2]. With three or more biological replicates, the power to infer differentially expressed genes in pairwise comparisons greatly increases, while the difference in results among different statistical programs that can be used for the analysis becomes less prominent [3, 4]. The decision on the number of replicates is often affected by the sequencing costs and the available budget, but it should be pointed out that experiments without replicates are virtually the most expensive ones in respect to the amount of useful information gained over money invested. Importantly, to improve the detection power of RNA-seq experiments, it has been shown that increasing the number of biological replicate samples is significantly more beneficial than increased sequencing depth [3, 5]. Furthermore, the decrease of sequencing depth and the loss of genes covered by at least one fragment is not linear, i.e., a significant reduction in sequencing depth will be detrimental to only a small number of lowly expressed genes. To allow for the detection of numerous biological samples at the same time multiplexed sequencing is worth consideration.

To plan the experiment better, it is important to know the level of sequencing depth per sample needed to meet the requirements of the project's objectives. Haas and collaborators (2012) showed that a sequencing depth of 5-10 million non-ribosomal RNA (rRNA) fragments enables profiling the vast majority of transcriptional activity in diverse bacterial species grown under diverse culture conditions [5]. This brings us to the question concerning rRNA depletion. Commonly, more than 95 % of the total RNA-seq reads are rRNA, and therefore depletion or rRNA from the total RNA will allow for more coverage of the RNA of interest. This means that following rRNA depletion; several samples can be multiplexed on a single HiSeq lane (numerous samples per lane, decreasing sequencing cost), while sequencing total RNA requires a full single Illumina HiSeq lane (one sample per lane). The disadvantages of rRNA depletion relate to longer sample preparation time and the possible introduction of bias in the RNA population. However, in the case of a large number of samples, rRNA depletion will greatly reduce the sequencing costs, in which case it might be worth first testing and comparing the various commercially available kits on a subset of samples.

An additional consideration to be made is whether a single-end (SE) library or paired-end (PE) library should be used, as well as the sequencing length to be used. For SE-libraries, each RNA fragment is sequenced from one side, while in PE-libraries each RNA fragment is sequenced from both sides, effectively doubling the number of reads per RNA. For a comparison of gene transcription between two samples, a SE-library of any size is sufficient (default size for HiSeq is currently 100 bases), as PE-libraries do not improve sensitivity, but increase the number of reads. In addition, sequencing of PE-libraries is more expensive than sequencing SE-libraries. Finally, strand-specific libraries are recommended for RNA-seq of mycobacteria because of high gene density and the presence of overlapping genes on opposite strands. Strand-specific reads can be assigned to their corresponding genes more accurately and can reveal potential antisense transcripts.

Here, we describe a detailed methodology for a comparative transcriptome study of mycobacteria using RNA-seq. Procedures described cover the extraction of total bacterial RNA, the preparation of strand-specific single-end library for Illumina sequencing, and detailed instruction on the basic bioinformatics methods used to map the reads to a reference genome and to count the number of reads per feature. Finally, we provide details on how to infer differentially expressed genes.

2 Materials

2.1 Extraction of RNA from M. tuberculosis (See Note 1)

1. *M. tuberculosis* (*see* **Note 2**).
2. RNA-free tubes, plasticware and glassware, and DEPC-treated water.
3. TRIzol (Life Technologies).
4. Bead beater (Biospec Products) or equivalent instrument.
5. 0.1 mm Zirconia beads (Biospec Products).
6. 1.5 mL MaXtract High Density Tubes (Qiagen).
7. Chloroform.
8. DEPC-treated 3 M sodium acetate pH 5.2.
9. Isopropanol.
10. 70 % ethanol.
11. DNase.
12. PCR or quantitative PCR reagents.
13. Agarose gel.
14. Spectrophotometer or NanoDrop.
15. Qubit (Life Technologies) (*see* **Note 3**).
16. Fragment Analyzer (Advanced Analytical).

2.2 Library Preparation for Illumina Sequencing

1. Molecular biology grade water.
2. LoBind tubes (Eppendorf).
3. Illumina TruSeq Stranded mRNA Kit for library preparation (Illumina).
4. Agencourt AMPure beads (Beckman Coulter).

5. Magnetic Particle Concentrator (Life Technologies).
6. Qubit (Life Technologies) (*see* **Note 3**).
7. Fragment Analyzer (Advanced Analytical).

2.3 Data Analysis

1. PC with at least 4 GB of RAM running under a 64-bit Unix-like operating system (*see* **Note 4**).

3 Methods

3.1 Extraction of RNA from M. tuberculosis

Good quality RNA is required for successfully performing transcriptomic analysis by RNA-seq (*see* **Note 5**). It is important to not allow RNA to be broken down, as this will impact the final expression profile. Various methods can be used for RNA preparation from *M. tuberculosis* cultures, including commercially available kits which involve column purification (*see* **Note 6**). Here we provide a protocol for RNA purification based on TRIzol reagent (*see* **Note 7**).

1. Grow the *M. tuberculosis* strain of interest to an OD_{600} of 0.3–0.4.
2. Pellet 40 mL of culture by centrifugation at 3,200×*g* for 10 min and discard the supernatant.
3. Snap-freeze the pellet in liquid nitrogen—at this point pellets can be stored at −80 °C.
4. Remove the bacterial pellet from the liquid nitrogen (or −80 °C freezer) and immediately resuspend it in 1 mL of TRIzol.
5. Transfer the bacterial suspension to a 2 mL screw-cap tubes containing 0.5 mL zirconia beads.
6. Place the 2 mL screw-cap tubes with sample into a bead-beater and bead-beat twice for 1 min with a 2-min interval on ice.
7. Incubate the sample at room temperature for 5 min, inverting periodically.
8. Centrifuge the sample for 30 s at 10,000×*g* and recover the top TRIzol layer.
9. Prepare a MaxTract tube by centrifugation for 30 s at 2,000×*g*.
10. Add the TRIzol layer to the gel in the MaxTract tube.
11. Add 200 μL of chloroform and shake vigorously (do not vortex) for 15 s.
12. Stand at room temperature for 10 min.
13. Centrifuge the MaxTract tube at 12,000×*g* for 10 min at room temperature.
14. Carefully collect the top aqueous phase into a new tube.

15. Add 0.1 volume of 3 M sodium acetate pH 5.2 and 0.7 volumes of isopropanol.
16. Invert the tube several times to mix and then store at −20 °C for at least 2 h to allow the nucleic acid to precipitate (both RNA and DNA will precipitate).
17. Centrifuge the sample for 30 min at 16,000 ×*g* at 4 °C and remove the supernatant (pellet sometimes visible at the bottom of the tube).
18. Wash the nucleic acid pellet twice with 200 μL of ice cold 70 % ethanol (centrifuge, each time for 30 min at 16,000 ×*g* and 4 °C).
19. Dry the pellet under vacuum, or by leaving the tube open in a clean place (*see* **Note 8**).
20. Resuspend the pellet in 96 μL of DEPC-water and add 12 μL of 10× DNase buffer and 12 μL of 1 U/μL DNase.
21. Incubate the sample for 1 h at 37 °C.
22. Perform a phenol–chloroform extraction by adding an equal volume of phenol–chloroform–isoamyl alcohol to the sample, mix vigorously by hand, and let stand at room temperature for 5 min.
23. Centrifuge at 16,000 ×*g* at 4 °C for 10 min and recover the top aqueous layer.
24. Add 0.1 volume of 3 M sodium acetate (pH 5.2), followed by 0.7 volume of isopropanol.
25. Invert the sample until well mixed.
26. Incubate at −20 °C for at least 1 h.
27. To pellet the precipitated gDNA, centrifuge the sample (16,000×*g*, 4 °C for 30 min). Remove supernatant and wash the pellet (not always visible) once with 70 % ethanol. Centrifuge (16,000 ×*g*, 4 °C for 30 min), discard the supernatant, and air-dry the pellet.
28. Resuspend DNA in molecular biology grade water and store it at 4 °C.
29. Perform a PCR on a housekeeping gene, for example *sigA*, to confirm that no DNA is present (no amplification product). If there is still residual DNA present (this is quite common), perform a second DNase treatment.
30. Resuspend the final RNA pellet in DEPC treated water and store at −80 °C.
31. Determine the RNA concentration and purity using a spectrophotometer, NanoDrop or Qubit.
32. Check the RNA integrity on a 1 % agarose gel—the 23S, 16S rRNA should be clearly visible (a streak of RNA rather than clear bands would suggest RNA breakdown), or if a fragment analyzer is available, use this to check RNA quality.

3.2 Library Preparation for Illumina High-Throughput Sequencing Conditions (See Note 9)

1. Fragment the RNA by mixing 1 μL of RNA (100 ng/μL), with 19.5 μL Fragment, Prime, Finish Mix (FPF buffer) (*see* **Note 10**).
2. Vortex and incubate at 94 °C for 8 min.
3. Centrifuge briefly.
4. Carry out first-strand synthesis by adding 8 μL of thawed ice-cold First Strand Mastermix (FSM) to the fragmented RNA sample. Mix well and centrifuge to collect the sample at the bottom of the tube. Incubate in a PCR machine with the following program: 10 min at 25 °C/15 min at 42 °C/15 min at 70 °C/4 °C hold.
5. Carry out second-strand synthesis to form blunt-ended double-stranded cDNA. Bring the Agencourt AMPure XP magnetic beads, Resuspension Buffer (RSB buffer), and the Second strand Mastermix (SSM) to room temperature. Place the single-stranded cDNA from the first-strand synthesis into a heat thermal cycler set to 16 °C, and add 5 μL RSB, and 20 μL SSM. Mix the reaction mixture well and incubate at 16 °C for 1 h.
6. Purify double-stranded cDNA using the Agencourt AMPure XP magnetic beads as follows: mix 90 μL of Agencourt AMPure XP magnetic beads (vortex beads prior to pipetting to properly resuspend them), and 50 μL of the double-stranded cDNA mix. Mix thoroughly and incubate for 15 min at room temperature. Load the tube onto a magnetic rack for 5 min to allow the beads to separate from the solution. Remove the supernatant and wash the beads twice with 200 μL of 80 % ethanol (using the magnetic rack to separate the beads from the 80 % ethanol). For the final wash, remove all residual supernatant with a pipette and let the beads dry for 3 min at 37 °C.
7. Add 17.5 μL Resuspension Buffer (RSB) to the dried beads, mix well, incubate for 2 min at room temperature, and apply to magnetic rack.
8. Transfer cDNA containing supernatant (15 μL) to a fresh 0.2 mL tube.
9. Mix 2.5 μL of purified cDNA with 12.5 μL of A-Tailing Mix (ATL) (thawed on ice) (*see* **Note 11**).
10. Mix thoroughly and incubate in a thermal cycler at 37 °C for 30 min, then 70 °C for 5 min and 4 °C hold.
11. Ligate adaptors to the cDNA by adding 2.5 μL of adenylated with 2.5 μL of chosen adaptor and mixed thoroughly. Add 2.5 μL of Ligation Mix (LIG) and incubate at 30 °C for 10 min. Stop reaction using 5 μL of Stop Ligation Buffer (STL) (*see* **Note 12**).

12. Purify cDNA with ligated adaptors using Agencourt AMPure XP magnetic beads. Add 42 μL of Agencourt AMPure XP magnetic beads and incubate for 15 min at room temperature.
13. Load the sample onto a magnetic rack for 5 min, and discard supernatant.
14. Wash the beads twice with 200 μL 80 % ethanol, and let the beads dry for 3 min at 37 °C.
15. To elute the cDNA resuspend the beads in 52.5 μL RSB buffer, incubate for 2 min at room temperature, place in magnetic rack, and recover the 50 μL of the supernatant that contains the cDNA.
16. Perform a second purification by adding 50 μL of Agencourt AMPure XP magnetic beads to the 50 μL cDNA.
17. Resuspend the dried beads in 22.5 μL of RSB buffer and recover 20 μL of the supernatant containing the cDNA.
18. Perform a 15 cycle PCR using specific primers that recognize the adaptors as follows: add 20 μL of the cDNA template, 5 μL of PCR Primer Cocktail (PPC), and 25 μL of PCR Master Mix (PMM), and cycled 15 times (98 °C for 10 s/60 °C for 30 s/72 °C for 30 s), with a final 5 min elongation at 72 °C and hold at 4 °C.
19. Purify the cDNA from the final PCR reaction with 50 μL Agencourt AMPure XP magnetic beads (as in **steps 12–15**), with two washes with 80 % ethanol.
20. Elute the purified DNA in 32.5 μL of RSB buffer.
21. Validate the library fragment size, purity, and concentration by Fragment Analyzer.
22. Submit the library to the sequencing facility.

3.3 Data Analysis

The data analysis workflow consists of mapping the reads against the reference genome, counting the number of reads that mapped to each gene and calculating the relative gene-to-gene expression levels between samples. This guide assumes that the reader has basic working knowledge of Unix systems and knows the basic principles of sequencing. We will describe a data analysis work flow for an *M. tuberculosis* RNA-seq experiment that includes biological replicates for two conditions (*see* **Note 13**). Note that the commands given below should be written in a single line for each step.

1. Download and install *Bowtie2* [6] (http://bowtie-bio.sourceforge.net/bowtie2) (*see* **Note 14**).
2. Download the *M. tuberculosis* H37Rv reference genome from NCBI (NC_000962.3) in FASTA format (*see* **Note 15**).
3. Build a Bowtie indexed reference:

```
bowtie2-build NC_000962.3.fasta H37Rv
```

4. Map the Illumina reads, for each sample separately. Example for sample “A”:

   ```
   bowtie2 -x /path/to/bowtie2_index/H37Rv -U /path/to/A_1.fastq.gz,/path/to/A_2.fastq.gz,/path/to/A_3.fastq.gz -S A_mapped-to-H37Rv.sam
   ```

   (*see* **Notes 16** and **17**).

Convert SAM files to coordinate-sorted BAM files as follows:

5. Download and install *samtools* [7].
6. Convert SAM to a sorted BAM:

   ```
   samtools view -Su A_mapped-to-H37Rv.sam | samtools sort - A_mapped-to-H37Rv_sorted
   ```

   (*see* **Note 18**).
7. Index the BAM file (*see* **Note 19**):

   ```
   samtools index A_mapped-to-H37Rv_sorted.bam
   ```

Counting the reads over genes as follows:

8. Download and install *featureCounts* [8] (*see* **Note 20**).
9. Download the gff3 file for the corresponding reference from NCBI. For H37Rv it is NC_000962.3.gff.
10. Convert the GFF file to SAF. Simplified annotation format (SAF) is a tab delimited file that contains five columns: feature identifier, reference name, start position, end position, and strand. SAF can be generated from a GFF file in a spreadsheet program (*see* **Note 21**). An example is shown below:

GeneID	Chr	Start	End	Strand
Rv0001	gi|448814763|ref|NC_000962.3|	1	1524	+
Rv0002	gi|448814763|ref|NC_000962.3|	2052	3260	+

11. Count the reads:

    ```
    featureCounts -b -F SAF -O -a /path/to/NC_000962.3.gff -o outpuname *.bam
    ```

    (*see* **Note 22**). At this step we have the raw expression levels for each gene (number of reads per gene). To look for differentially expressed genes, a statistical method must be applied that accounts for differences in the sequencing depths between samples, considers the variations of the expression levels among biological replicates and compares the expression levels of each gene between the two groups of samples.
12. Install *R* and the *DESeq* package (*see* **Note 23**).
13. Prepare the count table. The table generated with *featureCounts* should be edited to look like the tab delimited table below:

gene_id	A	B	C	D	E	F
Rv0001	123	111	222	321	456	789
Rv0002	10	12	30	88	99	50

14. Run *DESeq* (*see* **Note 24**). Below is an example set of commands that could be used for a dataset as given in the example count table above, where each column represents a biological replicate, samples A, B, and C are controls, and samples D, E, and F come from an experimental condition (comments are preceded by the hash character "#"):

```
# Run R
 R

# Load the DESeq package:
library("DESeq")

# Load the count data (in this case called countable.txt):
countTable <- read.table("countable.txt", header=TRUE, row.names=1 )

# Define conditions for the samples (any names can be given; here is "ctrl" for the control samples and "treated" for the condition samples. Note that the order corresponds to the order of samples in the count table:
condition = factor( c("ctrl ", " ctrl ", " ctrl ", " treated ", " treated ", " treated " ) )

# Define the CountDataSet, the central data structure in the DESeq package:
cds = newCountDataSet( countTable, condition )

# Estimate the effective library size:
cds = estimateSizeFactors( cds )

# Estimate the dispersions:
cds = estimateDispersions( cds )

# Look for differentially expressed genes between the two conditions:
res = nbinomTest( cds, " ctrl ", " treated " )

# Save the output to a file:
write.csv( res, file="DESeq.csv" )

# Recommended; plot some useful graphs that will help assessing the quality of the dataset or possible problems as well as to inspect the results visually (refer to the vignette for details).
# Load the necessary packages:
library("RColorBrewer")
library("gplots")
# Generate the heatmap of the sample-to-sample distances. First perform the variance stabilizing transformation:
cdsBlind = newCountDataSet( countTable, condition )
cdsBlind = estimateSizeFactors( cdsBlind )
cdsBlind= estimateDispersions( cds, method = "blind" )
vsd = varianceStabilizingTransformation( cdsBlind )
# Calculate the distances:
dists = dist( t( exprs(vsd) ) )
```

```
# Generate the heatmap and save it to a file:
jpeg('Heatmap.jpg')
hmcol = colorRampPalette(brewer.pal(9, "GnBu"))(100)
mat = as.matrix( dists )
rownames(mat) = colnames(mat) = with(mat)
heatmap.2(mat, trace="none", col = rev(hmcol), margin=c(13, 13))
dev.off()

# Plot the log2 fold changes against the mean normalized counts, and save to a file:
jpeg('plotMA.jpg')
plotMA(res)
dev.off()

# Plot the per-gene estimates against the mean normalized counts per gene and overlay the fitted curve, and save to a file:
plotDispEsts( cds )
jpeg('plotDispEsts.jpg')
plotDispEsts( cds )
dev.off()

# Plot the histogram of p-values, and save to a file:
jpeg('histogram_p-values.jpg')
hist(res$pval, breaks=100, col="skyblue", border="slateblue", main="")
dev.off()
```

15. Interpret the results (*see* **Note 25**).

4 Notes

1. A clean work environment is required for performing experiments involving RNA. All glassware and plasticware must be RNase-free, wearing gloves is necessary at all times, DEPC-treated and autoclaved solutions should be used.
2. Manipulation of *Mycobacterium tuberculosis* cultures must be performed under Biosafety Level 3 (BSL3) containment. Adherence to local guidelines for BSL3 work is strictly required.
3. Alternative methods for DNA quantification can be used, such as the Quantus Fluorometer manufactured by Promega or the PicoGreen assay [9].
4. Alternatively one can run many bioinformatics programs on external servers, like Galaxy, which is a widely used and freely available platform (http://galaxyproject.org/).
5. Approximately 100 ng of total RNA is required for library preparation without ribosomal RNA depletion following the Illumina methods (http://www.illumina.com/), but this

amount is likely to decrease as new protocols are optimized. We therefore suggest careful consideration of all the available options and consultation with the sequencing facility where the library will be sequenced. If depletion will be included in the procedure, a few micrograms of total RNA will be necessary (usually between 1 and 5 μg).

6. Column purification may lead to removal of short transcripts and small RNAs. The user should therefore choose the most appropriate methodology according to the aim of his experiment.

7. In our experience the TRIzol (Life Technologies)-based protocol yields pure, intact total RNA suitable for the subsequent procedures. This method retains the small RNAs and is therefore recommended when the user is interested in obtaining a comprehensive expression profile including the small transcripts.

8. The nucleic acid pellet now contains both RNA and DNA. For RNA-seq, it is important that there is no DNA contamination of the pellet.

9. In order to prepare the purified RNA for Illumina sequencing, a cDNA library must be prepared, where the RNA needs to be fragmented, reverse-transcribed, adenylated, fitted with adaptors, and purified. Here we describe the methods used to prepare total RNA for sequencing, however, it may be of interest if experiments involve a lot of samples, to deplete ribosomal RNA from the total RNA, allowing for multiplexing. Protocols for removal of ribosomal RNA based on affinity purification are not described here, but have been developed and reported by different suppliers (ScriptSeq Complete kit by Epicentre is an example).

10. As Illumina sequencing allows for the sequencing of relatively short fragments of DNA (currently up to 250 bases), RNA needs to be fragmented to similar sized pieces.

11. To generate a 3′ overhang on the blunt-ended double-stranded cDNA (needed for ligation of adaptor in the next step), the 3′ ends need to be adenylated.

12. Adaptors act as "barcodes" that can be used to identify the origin of the cDNA, and therefore different adaptors can be used for different biological samples when sequencing them in a single lane by Illumina (multiplexing). Adaptors are also needed for the next step of enrichment.

13. While the protocol described here offers more options and flexibility for advanced usage, a new tool called Rockhopper was recently published [10] that automates the work flow described here and uses similar algorithms for each process. The usage of Rockhopper merely consists in loading the fastq

files and choosing a reference sequence. The final result is a table of gene expressions. Rockhopper also has the option for visualizing the results in the IGV browser.

14. Other mapping programs can be used [11]. In RNA-seq transcription levels are inferred by counting the number of reads that correspond to each gene. To do so, reads must first be aligned onto the annotated reference genome sequence (in this case *M. tuberculosis* strain H37Rv).

15. ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Mycobacterium_tuberculosis_H37Rv_uid57777/.

16. If multiple CPU cores are available use the "-p" option (like "bowtie2 -p 4 -x…").

17. FASTQ files obtained from the Illumina sequencer are by default gunzipped. There is no need to uncompress these files since Bowtie2 can read them. Also, Illumina reads for each sample usually come in a number of individual fastq files to avoid problems of handling one large file (therefore the three fastq files in the example, separated by commas without spaces around).

18. The dash "-" in *samtools sort* defines the standard input. The vertical bar character "|", called a "pipe" is used to pass the standard output of one program to the standard input of the following one (in this case from *samtools view* to *samtools sort*). We could execute the two commands separately and have the *samtools view* write an unsorted BAM file to the disk which could be loaded to *samtools sort*. Pipelines are recommended to avoid IO bottlenecks and excessive hard disk usage. In fact, since we do not need the SAM file anymore after it is converted to BAM, we can pipe the *bowtie2*'s output (the SAM file) directly into *samtools view* without the need of writing it onto the hard disk, in a single command:

    ```
    bowtie2 -x /path/to/bowtie2_index/H37Rv -U /path/to/A_1.fastq.gz,/path/to/A_2.fastq.gz,/path/to/A_3.fastq.gz | samtools view -Su - | samtools sort - A_mapped-to-H37Rv_sorted
    ```

19. Many programs for downstream analyses or visualization of BAM files require the corresponding BAM's index file. It is good practice to index BAM files upon their generation.

20. Other feature counting programs can be used, like *htseq-count* (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html), *BEDtools* [12] etc.

21. In some cases it might be of interest to also check for transcripts deriving from intergenic regions in order to spot potential novel genes. In that case the SAF file should be modified to include the intergenic regions as additional features. It might be a good idea to omit very short intergenic regions.

22. Multiple BAM files can be input to the program, in which case the output table will contain individual counts columns for each BAM file. If all the BAM files are present in a single directory, using the wildcard character "*" will make *featureCounts* load them all. The option "-O" means that reads will be allowed to be assigned to more than one matched feature. For example, if a read spans over gene A and gene B, both genes will be counted. For bacteria this option makes sense because many genes are packed in operons. Reads spanning multiple genes derive from single transcripts.

    Note that *featureCounts* can handle strand specific reads with the "-s" option. For the library protocol provided here, the strand specific reads are in the reverse orientation. In that case the option should be "-s 2". Since strand specific library protocols often change, one can quickly check the orientation of the strand specific reads by running *featureCounts* twice using the "-s 1" and "-s 2" options, and then manually compare the resulting counts for a few genes, taking into account their strand.

23. Refer to Bioconductor for the installation instructions (http://www.bioconductor.org/install/). Other programs can be used for differential gene expression analysis [3, 4], including the recent *DESeq2* package.

24. DESeq is an advanced program that has a number of options. For proper usage and better understanding of the program please refer to its vignette and the user manual.

25. The main output of *DESeq* is the table of genes with expression values and relative expression changes between the compared groups of samples. There are no strict rules that define statistically significant or biologically significant differentially regulated genes. As a general guidance, one could consider only genes with the *padj* value smaller than 0.05 (*padj* indicates the false discover rate) to assure statistical accuracy, and genes that have at least a twofold difference in expression values for biological significance.

References

1. Wolf JBW (2013) Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol Ecol Resour 13:559–572. doi:10.1111/1755-0998.12109
2. Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. Genetics 185:405–416. doi:10.1534/genetics.110.114983
3. Rapaport F, Khanin R, Liang Y et al (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol 14:R95. doi:10.1186/gb-2013-14-9-r95
4. Kvam VM, Liu P, Si Y (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot 99:248–256. doi:10.3732/ajb.1100340
5. Haas BJ, Chin M, Nusbaum C et al (2012) How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics 13:734. doi:10.1186/1471-2164-13-734
6. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi:10.1038/nmeth.1923

7. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352

8. Liao Y, Smyth GK, Shi W (2013) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–30. doi:10.1093/bioinformatics/btt656

9. Ahn SJ, Costa J, Emanuel JR (1996) PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. Nucleic Acids Res 24(13):2623–2625

10. McClure R, Balasubramanian D, Sun Y et al (2013) Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res 41:e140. doi:10.1093/nar/gkt444

11. Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV (2013) Benchmarking short sequence mapping tools. BMC Bioinformatics 14:184. doi:10.1186/1471-2105-14-184

12. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi:10.1093/bioinformatics/btq033