BOOK CHAPTER

A Hitchhiker's Guide to Metatranscriptomics

Mariana Peimbert[1] and Luis D Alcaraz[2*]

*Correspondence:
lalcaraz@iecologia.unam.mx
[2]1. Departamento de Ciencias Naturales, Unidad Cuajimalpa, Universidad Autónoma Metropolitana. 2. Laboratorio Nacional de Ciencias de la Sostenibilidad (LANCIS). Instituto de Ecología, Universidad Nacional Autónoma de México, AP 70-275, Ciudad Universitaria, Coyoacán, 04510 Mexico city, MX
Full list of author information is available at the end of the article

Abstract

Recent technological development in high throughput DNA sequencing has opened many venues for biological systems analysis, where the data generation is no longer the bottleneck but its analysis. The sequencing technological advancements are being used in the study of a large and previously unknown microbe world; this unknown world has revealed itself due to the possibility to study taxonomic and functional diversity by means of culture-free metagenomic sequencing. The metagenomes provide an outlook about the coding potential or the probable metabolic functions of the studied microbial communities, thus just the potential outcome. To study community wide gene expression, under strictly determined conditions, metatranscriptomics by whole genome shotgun RNA sequencing is the preferred tool. The challenge of studying metatranscriptomes is quite complex, involving a correct experimental design, sequencing technology knowledge, wet laboratory and bioinformatic skills. Here, we present a guide with helpful hints, suggestions, tools, highlighting some of the complications down the road of metatranscriptomics to ease your journey in this winding road.

Citation: Peimbert, Mariana, and Luis David Alcaraz. 2016. "A Hitchhiker's Guide to Metatranscriptomics." In Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing, 313–42. Springer International Publishing. The final publication is available at Springer via http://dx.doi.org/doi:10.1007/978-3-319-31350-4_13

Keywords: metatrasncriptome; transcriptome; metagenomics

13.1 Transcriptomics, Metatranscriptomics, and Bacterial RNA Complications

Transcriptomics is defined as the complete set of RNA molecules produced in a cell (Güell et al. 2011). Metatranscriptomics is the assessment of environmental gene expression, be it in a population or a whole community. The rapid advance in sequencing technologies has allowed to rapidly increase the environmental genomics related works. At the beginning of the Next Generation Sequencing (NGS) about some 10 years ago from now, most of the works were only able to describe microbial taxonomic diversity by means of amplicon sequencing (16S/18S rRNA sequences), and then the introduction of 454 pyrosequencing led to multiple groups start working with Whole Genome Shotgun (WGS) metagenomics. Although the work was merely descriptive at the beginning of WGS metagenomics, it threw light on both taxonomic and functional diversity of the studied environments. Within the functional diversity, metagenomics is only describing the potential outcome, but to test the functional profile of a microbial community further methodologies for the expression (metatranscriptomics), and translation (metaproteomics) are required. The race for cheaper sequencing is still going on, and there is no such thing as a universal and

unique best solution platform in the market but there are several technologies leading the competition like is the case for Illumina®, and several newcomers are still on its way with promising technologies like nanopores, and solid state based solutions. The challenge of describing genome wide expression has been done historically by means of microarray chips, and they have the advantage of describing overall gene expression, but previous knowledge about the genomic sequence of the organism is mandatory. A previous NGS technique, to describe microbe's transcripts, is expressed sequence tags (ESTs); the current transcriptome sequencing strategies are just an up-scaling of ESTs. While microarrays have been proved as an effective tool for describing the expression profiles for model organisms, they are not still a major player in metatranscriptomics. The cause of the microarrays relegated role in metatranscriptomics is that for complex environments with high diversity there would be the need to sequence the metagenome, then select representative gene clusters, and print them into the microarray, making it expensive and laborious. Although it would be possible to design environmental microarrays looking for some particular genes (pathogenesis, virulence, etc.) or particular species, this would be limited when comparing to current RNA-seq approaches (Westermann et al. 2012). The main advantage of current NGS metatranscriptome is that is possible to associate gene expression patterns of even unknown genes, thus showing light that the unknown gene is transcribed under a particular condition. Hence, metatranscriptomics aids to identify novel genes related with environmental functions, with no necessary previous knowledge about any particular gene present in the sample (so no probe or primer design needed). The main drawback of environmental NGS metatranscriptomics is that most, sometimes >95%, of the environmental RNA isolated under any situation corresponds to ribosomal RNA (rRNA), and the prokaryotes do not have a polyA track in the 3 end of mRNA which is central for the transcriptome sequencing of eukaryotes, because it allows to start reverse transcription from the terminal polyA track and consequently the cDNA is almost exclu- sively formed by mRNAs (Sorek and Cossart 2010). Although rRNA is useful to determine community structure and having by PCR an unbiased picture of the active taxonomic diversity out there (by identifying, and annotating 16S rRNA fragments), when trying to define the community functional profile, getting rid of rRNA could be a challenge. However, with the current NGS technologies, it is feasible to think of having less than 5% of mRNAs in the total sample, and still have thousands of cDNAs to tell a story about, but nevertheless cleaning the rRNA is required. There has been an active development for technologies trying to enrich the amount of total mRNA and they could be divided in the following four main strategies: (1) Ribosomal RNA capture (rRNA hybridization), (2) 5–3 exonuclease degrading processed RNAs, (3) adding polyA to mRNAs by means of polyA polymerase (from Escherichia coli), and (4) antibody capture of mRNAs interacting with selected proteins (Sorek and Cossart 2010). The polyA and antibody capture methods are highly biased, thus not recommended. The cDNAs enrichment is a major issue when designing the overall strategy and experiments. A crucial factor in transcriptomics is whether you have a reference genome sequence to map the transcripts against or you will be performing *de novo* transcript assembly. It is the same situation with metatranscriptomics, if you have or not a reference metagenome obtained at the very same

time to map against. The major advantage of having a reference metagenome is that you can see if there is correspondence between raw gene abundance, and its expression levels. There are plenty of options to map NGS sequencing data against references like BWA, bowtie, and tophat (Langmead et al. 2009; Li and Durbin 2009), and at the end of the day you could build count tables with each transcript abundance, and mapping Single Nucleotide Polymorphisms (SNPs) for each of the transcripts. If you are just interesting to sequence the metatranscriptome without metagenomic reference you should assemble the reads first using some NGS assemblers like SOAPdenovo, Velvet, Celera, and then perform ORF prediction with some tool like Glimmer, or Metagenemark (see Table QG13.2). Up to date there are plenty of resources to address a metatranscriptome study. This work intention is to give an overall view of the metatranscriptomics process, experiments and analysis, and put the spotlight in the plenty of guides, tutorials and resources that have been systematically ordered for this purpose. Methodologically, the metatranscriptome uses the very same techniques and analytical tools as is sin- gle species precursor, the transcriptome.

13.2 Get to Know the Basics on Transcription Before Going Further

Previous work on systematizing the huge amount of information related to RNA-seq experiments in microorganisms has been done, and we strongly recommend to check out the biological and technical information before getting into the experimental design. A great starting point for understanding our current knowledge about bacteria transcription could be assessed in two excellent reviews the first by Sorek and Cossart 2010, and then read the review by Güell and collaborators (2011), both works on Nature Reviews Microbiology. Some previous protocols on metatranscriptomics are available as well (Gilbert and Hughes 2011), though thinking on a virtually retired technology (454 pyrosequencing), but all principles are still valid. The literature recommendations are based on first have a general outlook of what we know about bacteria gene regulation and how this is being enriched by transcriptomics. We also recommend to check some of the works to see the final publication output of metatranscriptomics, and how this is reported (Benítez-Páez et al. 2014; Frias-Lopez et al. 2008; Gilbert et al. 2008; Gosalbes et al. 2011; Hewson et al. 2009; Franzosa et al. 2014).

13.3 Experimental Design

If you want to try metatranscriptomics, the first thing would be choosing what kind of experimental approach is correct to your needs, and budget. Basically, there are two great first approaches to it, a qualitative or quantitative (Fig. 13.1).

For metatranscriptomics using RNAseq, the qualitative approach is highly valuable, because even the high amount of rRNA obtained, this could be used to describe community structure and describe the metabolically active members of it. The databases with 16S rRNA are still the best repositories for bacteria taxonomic diversity out there, and even though it may not be possible to perform the tasks done with PCR microbiome amplicons like multiple alignments, and diversity metrics derived from them (like Unifrac, Phylogenetic Distance methods, etc.), it is

possible to identify by homology each of the sequenced reads, and use some tools like the RDP classifier or Greengenes to classify the overall active bacteria diversity (Lozupone et al. 2011; Cole et al. 2009; Schloss 2010; DeSantis et al. 2006). The rRNA classification for a metatranscriptome has the additional advantages of not biasing the diversity due to primer election, and PCR amplification effects. Moreover, the expected 5 % of mRNA helps to identify expressed genes in the community, some of the genes are going to have known homologs in the databases and they will be annotated accordingly but for the orphan genes (with no homologs in DBs) we will have information about them being expressed under the tested situation, something not into reach with metagenomes and so the importance of knowing previously the tested variables and the metadata that will be available for future comparisons. The quantitative approach is the most used when doing transcriptomes on single organisms. This is because this approach allows us to detect significant differences between the overall gene expression in contrasting situations. Single organism transcriptomes in several contrasting experimental conditions had been proved to be a powerful tool when looking for Differential Gene Expression (DGEs). The success of getting DGEs depends on several factors like the number of conditions tested (biotic, abiotic), the number of biological replicas, sequencing coverage, read length. The sequencing coverage and read length could be easily planned if there is a reference genome. If there is no such thing like a reference genome one rule would be to dedicate equal sequence coverage for each of the replicas (i.e., if using Illumina HiSeq 2500® dedicate a single sequencing lane to each replica). If you are planning to conduct a metatranscriptome it would help a lot if you have some preliminary data on helping you to answer the basic how many sequences do I need? This could be the result of pilot studies on 16S rRNA amplicon diversity, a previous metagenome, or even diversity estimates from related systems of what you are currently studying. There are several tools aiding with the design and replica number in RNA-seq experiments, like EDDA (Experimental Design in Differential Abundance analysis) which is available like an R's Bioconductor package (EDDA), or as a web server (Luo, et al. 2014). Within EDDA you can upload some pilot data you might have and test about the experimental design. The key questions are: How many replicates should I use? How much sequencing depth? Is the experimental design helping out to capture biological variation? One rule of the thumb would be to use the same number of replicas for each condition tested and a minimum number of two replicates per condition to gain insight into the biological variance. Thus, considering one treatment and one control groups would be the simplest, and most widely used experimental design (See Table QG13.1). If you are trying a nested experimental design the number of replicates would increase dramatically but this is out of the reach of this chapter, please refer to experimental design guides, a good starting point for this was provided by Knight and collaborators (2012).

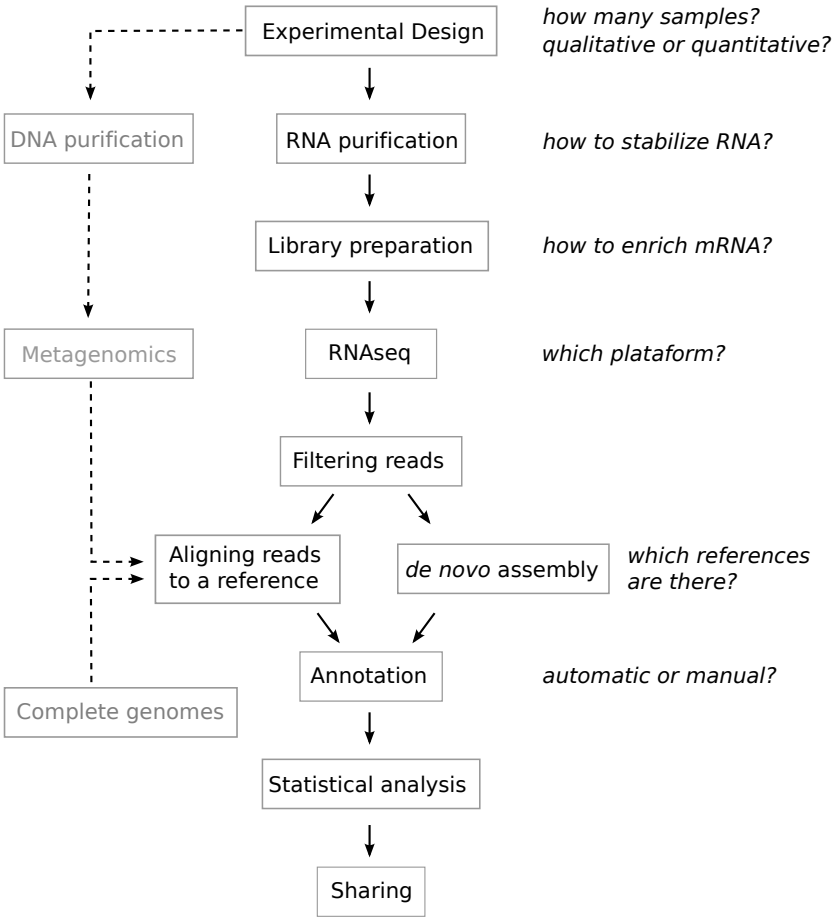

Fig. 13.1 Representation of the wet-lab procedure workflow.

13.4 What Sequencing Platform Is the Best for Metatranscriptomics?

This is the most frequent question for most of researchers entering into the metaomics world. There is no easy answer for this, as expected. The main trade-off would be between the overall cost, the read length, and the sequencing depth of each platform. The current major used platform is Illumina® due to its overall cost- benefit, though it has several possible confi gurations (MiSEQ, HiSEQ, etc.), the major platform used not that long ago was 454, and now is practically retired from metatranscriptomics, the message here is that the market is still far from being stable and new players are coming all days into it. The actual major players are: Illumina's HiSeq (X, 3000/4000, NexSeq, High-Output), and MiSeq, Life Technologies (PGM, Proton), Pacific Biosciences (RS), and the former 454. The sequence read length spans from 50 bp (Illumina) to 1.5 kb (PacBio), and the cost per Mb goes from USD$ 0.06 (Illumina) to USD$8.72 (454), and the output yield goes from 40 Mb (PacBio) to 300 Gb (Illumina). There are some recent works that show that in overall the gene expression profiles are similar across platforms and the main differences are the costs for detecting splice variants (Li et al. 2014). But keep in mind that the price is rather limiting but not the only variable to consider, please take into account the quality of the data, the support for the available technol-

ogy (aligners, assemblers) and compare the options offered by different providers, there are some places like http://allseq.com and https://genohub.com where you can quote multiple providers all at once. Also keep in mind that you can mix two strategies, i.e., Illumina's deep coverage mixed with PacBio long reads to aid in the assembly process. The main questions are: How many samples do you want to sequence? What is your desired read length? How many reads do you need per sample? How much money do you have?

13.5 Sequencing Depth or the Number of Aligned Reads Required for a Reliable Analysis

A bacterial genome is considered complete when it has an 8X coverage depth. For an average 5 Mb genome it would be necessary to sequence at least 40 Mb to have that amount of coverage. When talking about a metatranscriptome in the ideal scenario one would have previous data about the studied system, like the species abundance with 16S rRNA amplicon sequencing. Lets say that a given environment hosts 700 species and assuming a 5 Mb genome per species one would need at least 28 Gb of sequencing to have an 8X coverage depth. This is assuming some unrealistic situations like having equal abundances for each species and genes, and that they are all the same genome size. This is not an easy task, but with Illumina's deep sequencing it is expected to generate up to 300 Gb of sequencing that would be equivalent to a 85X coverage for each species of this hypothetical scenario, and considering that not every gene is always being expressed we can have an ultra-deep coverage of the metatranscriptome that can even be multiplexed. Most of the meta-omics analysis are highly biased to over-represented features, even of ultra deep sequencing we cannot be certain that we are not recovering rare species, or genes because of the sequencing effort. The rule of the thumb for sequencing depth is to be equitable for each condition and replicate tested.

13.6 General Considerations for Wet-Lab

When working with RNA is important to pay close attention to cleanliness of the bench working area, equipment and reagents. All living cells and all cell types produce intracellular and extracellular RNases. RNases are essential for the regulation of gene expression and are an important part of the immune system; that is the reason why there are several types of these enzymes, some of which are very resistant to inactivation treatments. Some RNases have several disulfide bridges so even after frozen or denatured they can be reactivated. RNase contamination main sources are the skin, saliva, hair, perspiration, clothing, fungi, bacteria, mites, plant, or any living cell (Sambrook and Russell 2012). This is why you should always take the following precautions: 1. Always wear gloves.

2. Change gloves frequently. Every time you touch the phone, the handle of the fridge, your face, skin, etc. you should change gloves.

3. Wear clean gown. The lab coat protects the experiment from dust on the clothes.

4. Use RNase-free tips and tubes. Providers indicate when their products meet this quality criterion. Bags and boxes must remain closed otherwise they are no longer RNase-free.

5. Work in a specific clean area with low traffic and free from air currents.

6. Use RNase-free reagents. We recommend using commercial kits, and reagents designed to work with RNA. Remember, tubes and bottles must be handled with gloves and must be closed as long as possible.
7. Clean every material to be used in a way that is free of RNase (see Sect. 13.6.1). Some labs still take extra precautions such as:
8. Use filter tips to avoid aerosols that could contaminate the sample.
9. Have a unique set of pipettes to work with RNA.
10. Aliquot reagents to reduce handling.
11. Use an RNase-free fumehood or cabinet.
12. Have a clean room equipped.

13.6.1 Treatments for RNase Cleaning

Contrary to common sense the autoclave does not inactivate all RNases. All the water in contact with the RNA must be free of RNase. The most commonly used protocol for this is treatment with DEPC (diethyl pyrocarbonate). DEPC covalently modifies the secondary amines inactivating RNases permanently. However, it also modifies RNA so it must be destroyed before use. For this treatment, a 0.1 % DEPC solution is prepared and incubated for 12 h at 37 C. Then, the solution is autoclaved for 15 min for DEPC degradation. Buffers and other reagents with amines (Tris, MOPS) should not be incubated with DEPC. To prepare these buffers water is first treated and then reagents are dissolved. All nondisposable material should be treated. Glassware should be washed and baked at 240 C for 4–16 h. Another protocol is to dip the glassware in water with 0.1 % DEPC for 12 h at 37 C and then autoclaved for 15 min to remove DEPC. It is important to wrap with foil glassware before putting it in the oven or autoclave. It is also recommended to have a clean area for all reagents and materials to be used. Electrophoresis tank must also be treated; it should be washed with detergent, rinsed with RNase-free water, and finally rinsed with ethanol. Some companies sell DEPC alternatives that do not require autoclave. RNase inhibitors are commercially available, inhibitors are high affinity proteins specific for RNase type A. RNase inhibitors are expensive, and it is recommended only to preserve the purified sample.

13.6.2 RNA Purification

Using commercial kits is recommended, mainly because they ensure that the solutions are RNase-free. Please pay attention to the amount of sample that is recommended by the supplier as excess can result in very low efficiencies. RNA purification is divided into the next steps: sampling, RNA stabilization, cell lysis, RNA isolation and treatment with DNAse I. Here we describe various protocols for each of these steps.

13.6.3 Sampling

The samples should be acquired quickly and aseptically. The sample should be processed immediately or snap-frozen. Generally, samples are frozen directly on the field in either liquid nitrogen, or dry ice/acetone to stop metabolism without damaging cell structures, however when samples are thawed RNases will be active. When planning your sampling you should anticipate how to stabilize RNA because usually this is done before freezing (see below).

13.6.4 Stabilization

As previously mentioned, all cells have intracellular RNase, the mRNA in bacteria generally have a few minutes life span so RNA can be degraded while purified. Moreover, transport and purification can induce the synthesis of new mRNA changing expression profiles. Several reagents may serve to inactivate endogenous RNase. The simplest is to add to the sample a 1:10 solution of 5 % phenol in ethanol. Another option is to start with the isolation process before freezing adding guanidinium thiocyanate–phenol–chloroform solution, commercially known as TRIzol®, Qiazol®, or TRi®. One of the most popular stabilizers is RNAlater® containing EDTA, sodium citrate, and ammonium sulfate, it is used for all cell types and has been tested in bacteria. RNAprotect® is a stabilizer designed for bacteria; this works for gram-negative and -positive bacteria.

13.6.5 Cell Wall Lysis

The three most popular methods to lyse the cell wall are: mechanical disruption (bead beater), enzymatic lysis (lysozyme or lysostaphin) and proteinase K digestion. In axenic cultures lysate efficiency is important for the total amount of RNA but when it comes to communities, lysis will also affect RNA distribution, as some bacteria are more sensitive to some treatments. If the aim is a qualitative study, it probably is best to mix all methods of lysis, to obtain as many as possible RNA, but if you want to make a quantitative study, you would better use a mechanical method that can lyse all bacterial types and is the most reproducible one. When working with soil communities is important to consider the contamination with humic acids, as they inhibit further PCR reactions. PowerSoil® kit is specially designed to deal with humic acids. If you do not have access to the kit, we recommend washing the cells several times with phosphate buffer and follow a purification protocol with CTAB.

13.6.6 DNase I Treatment

RNA samples often have trace contamination of genomic DNA, so the final step is to treat the samples with DNAse I, and its subsequent inactivation. DNAse I can interfere with the following steps, if not inactivated. Once again RNA can be purified by extraction and precipitation or by silica columns. The RNeasy® kit allow using DNAse when the RNA is bound to the column, which prevents the second purification. To prevent freezing and thawing we suggest to aliquot pure RNA samples. Store samples at 80 C before and after purification.

13.6.7 RNA Quality Determination

There are three factors to consider in determining the quality of a RNA sample: concentration, purity, and integrity. These three factors are important in deciding whether to continue the experiment or if repurification are necessary. We always advise to perform an UV absorbance spectrum (220–350 nm), NanoDrop® instrument allows to measure small volumes from 1 L; absorbance at 260 nm indicates the concentration of nucleic acids, absorbance at 280 nm allows to estimate the protein concentration; while 230 nm absorbance indicates the presence of humic acids salts or compounds that were used for purification. The disadvantage of this method is

that it cannot determine if the RNA is degraded and this not either distinguishes DNA contaminations. It is generally considered good quality samples when the 260/280 ratio is greater than 1.8 and the 260/230 ratio is greater than 1.7. If the sample is not pure, the concentration may be overestimate as contaminants also absorb at this wavelength (Fig. 13.2). Fluorescent dyes detect lower RNA concentrations, and these only emit in the presence of nucleic acids, so RNA concentration is more reliable. Fluorescent dyes, generally, do not discriminate between different nucleic acids and this technique cannot determine the purity and integrity of the sample. The agarose gel electrophoresis allows knowing RNA integrity; the criterion for determining that the RNA is intact is to observe 23S and 16S rRNA bands in a 1.8:1 ratio. The presence of genomic DNA can be identified in agarose gel because its size is much greater than 23S, but it do not allow us to estimate other kinds of contamination. One of its great advantages is that it is an inexpensive method that can be done in most laboratories; nevertheless, it is a qualitative method. The 2100-Bioanalyzer® is a quantitative method that uses cartridges ready to use for capillary electrophoresis. This equipment generates electropherograms and includes software that integrates the peaks to determine the RNA integrity number (RIN; Fig. 13.2). The big disadvantage of Bioanalyzer equipment and cartridges is their price, this method also allows to determine sample concentration.

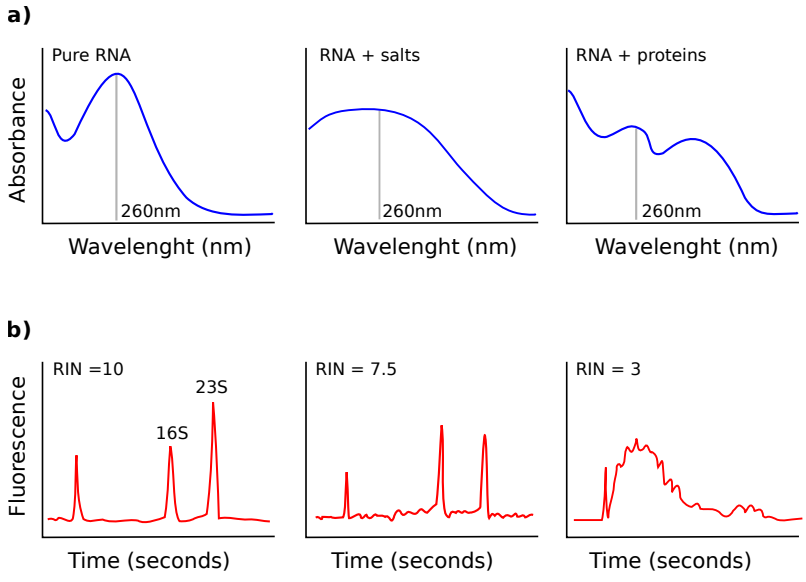

Fig. 13.2 Assessing RNA quality. (a) NanoDrop®'s absorbance UV spectrums, in the left plot an ideal sample with Pure RNA is shown, in the middle and right plots possible contaminations are shown. (b) Bioanalyzer® electropherogram profi les showing in the left plot the best case scenario with pure RNA, in the middle a plot of a partially degraded sample, and in the right a shred sample.

13.6.8 Enrichment of mRNA

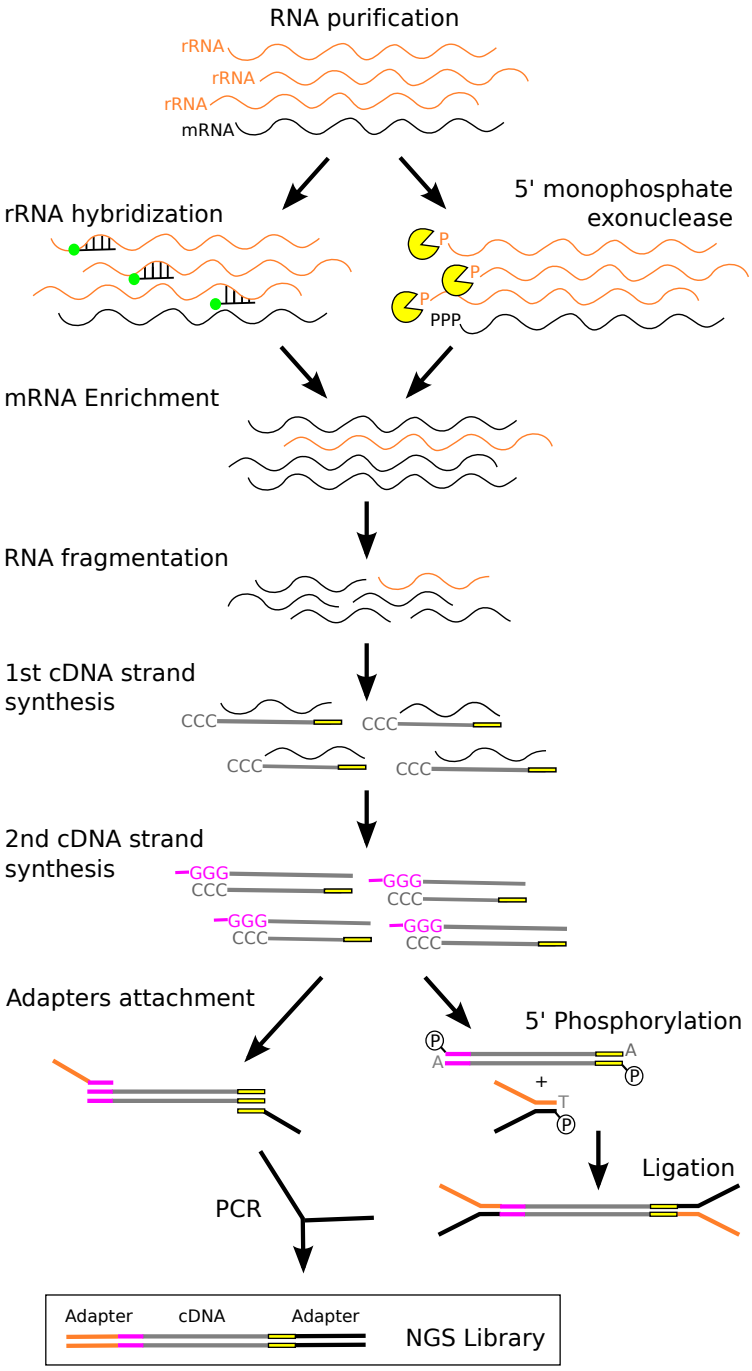

Fig. 13.3 The metatranscriptomics library preparation process. The main two strategies for mRNA enrichment are shown, first using rRNA separation by means of hybridization with 16S and 23S rRNA probes, and the second one is a depletion of rRNAs by means of a 5-exonuclease. Then, first strand of cDNA is synthesized by means of reverse transcriptase using random hexamers. Second strand of

cDNA is synthesized by a DNA polymerase. Finally, sequencing adapters need to be attached to the cDNA strands, and this could be done either by PCR or by ligation.

One of the most complicated steps in studying bacterial transcriptomes and metatranscriptomes is mRNA enrichment; in eukaryotes the problem is trivial by the presence of the polyA tail. The two most popular strategies to enrich the mRNA are rRNA hybridization, and degradation of processed RNA. rRNA hybridization is based on magnetic microbeads and oligo mixtures which hybridize with 16S and 23S (MICROBExpressTM, and Ribo-ZeroTM). The hybridization method is the most popular because RNA integrity is not required. This approach is sequence specific and does not eliminate all bacteria rRNA, for example those from high GC content. Another limitation is that oligos can also hybridize with some mRNA. Degradation of processed RNA requires a 5 monophosphate exonuclease for the removal of rRNA (mRNA-OnlyTM). Most mRNAs carry 5-end triphosphates therefore are not degraded. 5 monophosphate may be created by pyrophosphatase or endonuclease cuts. The advantage of this method is that sample diversity does not interfere; however, it requires very pure RNA as exonuclease is susceptible to inhibition by impurities; this also requires high RNA integrity (RIN >8) otherwise exonuclease degrades both rRNA and mRNA (Fig. 13.3). There are other strategies that enable deeper sequencing such as immunoprecipitations or duplex-specific nuclease digestion (DSN), these type of approaches only makes sense for specific experiments since strong bias is introduced. If your interest is to work with small RNA, these can be purified from an agarose gel. Specific biotinylated primers can be designed to eliminate other sequences, whether rRNA which are not recognized by hybridization kits or some other dominant messenger in the sample (Li et al. 2013).Transcriptomic analyses are based on cDNA synthesis so the polarity (5–3) information is lost. The polarity of the transcripts can give important information for antisense RNA and novel transcripts identification. If your interest is to know the polarity, there are protocols that incorporate dUTP in the synthesis of the second strand, allowing subsequent removal by uracil-DNA-Glycosylase (UDG) treatment (Parkhomchuk et al. 2009). The rapid development of sequencing technologies, and larger sequencing yields soon will make possible that rRNA would only need to be filtered in silico.

13.6.9 Library Preparation

Regardless of sequencing platform that will be used, the general idea is the same: to produce cDNA of a certain size (50–400 bp) that is flanked by adapters. So library preparation requires fragmenting the RNA, first strand synthesis, second strand synthesis, coupling adapters, and validating the library. Sequence service providers can perform the library preparation. cDNA should be of a certain size to optimize sequencing, depending on the platform is the size fragments must be. Fragmentation can be done with enzymes, metals, heat or sonication. Incubation times for fragmentation must be optimized for each case, as the integrity of each sample is usually different. The synthesis of the first cDNA strand is performed by a reverse transcriptase and generally random hexamer primers are used. The synthesis of the second strand of DNA is done with a DNA polymerase. In this case, primers with guanines at 3 are generally used since reverse transcriptase leaves a polyC

overhang (Fig. 13.3). Sequencing adapters include a region for binding to platform support and a region for primer hybridization. Additionally, they can include a barcode that serves to identify the sample if several samples are mixed in the same run (multiplexing). Depending on the used adapter kit is how many samples can be multiplexed. Illumina allows sequencing of the complementary strand, which allows for longer reads (pair-end). The adapters can be attached by a PCR or ligation reaction (Fig. 13.3). Currently the most widely used platform is Illumina for which there are kits like TruSeq® and SMARTer®. The superiority of the former is that it allows multiplexing up to 96 samples while SMARTer® allows only 16 samples. The advantage of the latter is that you can start with 1 ng of enriched RNA whereas TruSeq® requires at least 100 ng (Alberti et al. 2014). The last step is to validate the library. DNA concentration and size can be determined by the 2100-Bioanalyzer® coupled to a DNA chip like Agilent DNA 1000. We recommend contacting your sequencing provider, they have proven experience doing NGS on a daily basis, and they can assist you in fine-tuning the details about your samples. Sometimes your providers would even suggest some new sequencing platforms you have not noticed yet with higher yields at lower costs.

13.7 Bioinformatic Analyses

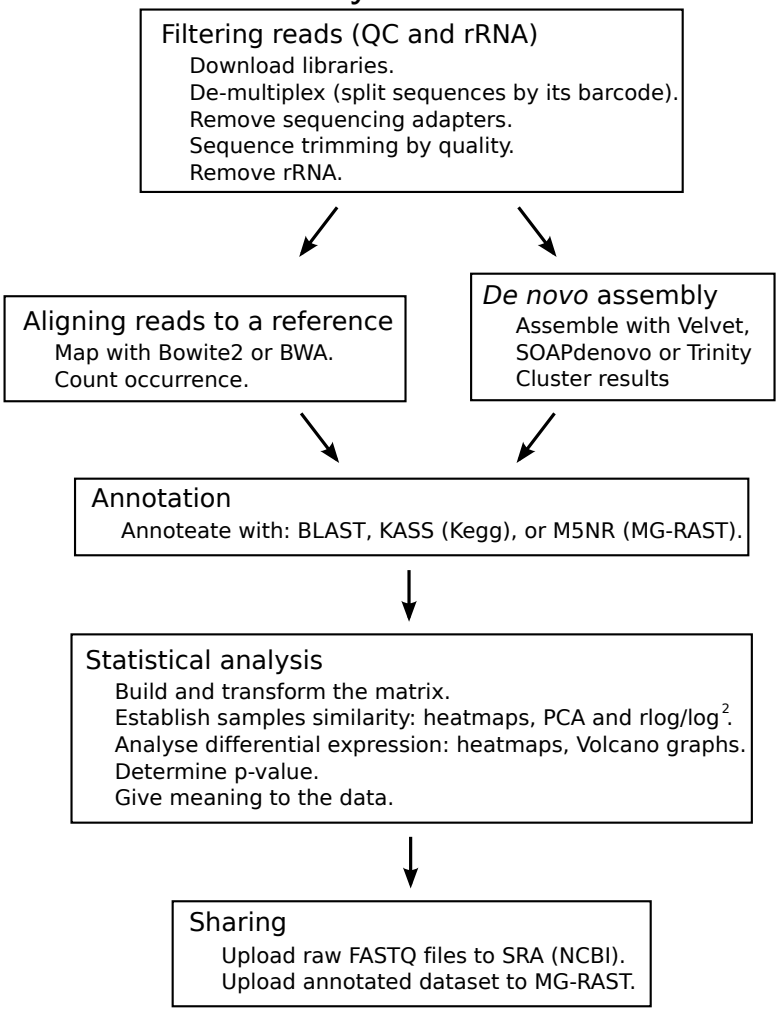

Fig. 13.4 The metatranscriptomics bioinformatic overall process. The main steps are: Filtering reads, choosing between aligning to reference sequences and performing *de novo* assembly, annotation, statistical analysis, and uploading the raw, assembled, and annotated data sets to the appropriate repositories.

The metatranscriptome analysis involves a conceptual and technical challenge when dealing with huge amounts of multivariate information. There is an intermediate level of computing knowledge required to be able to deal with this data, and we want to provide some basic steps previous to metatranscriptome analysis that should be fulfilled if you do not have bioinformatics experience, the overall bioinformatic analysis process is summarized in Fig. 13.4.

First basic steps:

1. Use the terminal (Linux, UNIX, Mac OS) or if you are running on Windows, immediately switch to Linux and learn how to use it. Use Ubuntu as it is the most supported Linux out there. And then, look for a Linux command line interface tutorial. Completing this exercise is highly recommended (see Table QG13.2).
2. Download your brand new transcriptome files from your provider FTP or provided URL. The file is normally a FASTQ, which is a text that contains both the sequence read and the base calling, encoded in ASCII characters (non directly human readable). Use any web browser, the web browser negotiates different transfer protocols (FTP, HTTP) in a Graphic User Interface (GUI), or you could automate this with Linux/UNIX's commands like wget, rsync, curl, and ftp.
3. Unzip and manipulate the files only on the terminal, this means in the Command Line Interface (CLI, also known as terminal). If you are using your mouse and clicking the files to open/unzip them, you will be out of your computer resources pretty soon.
4. Install the compilers (transforms source code to an executable), this is mandatory to install software from source, on Ubuntu's terminal type:
sudo apt-get install build-essential
For Mac OSX google: "Install the Command Line C Compilers in OS X" and follow the instructions.
5. Download your first program (fastx_toolkit, see Table QG13.2), and follow the install instructions.
6. Install R (see Table QG13.2).
7. Install Bioconductor (see Table QG13.2).
8. If you manage to do all the above tasks you are ready to install, and run almost any existent tools on Linux/UNIX.

If you do not want to improve your CLI skills, there are Graphical User Interfaces (GUIs) designed to cope with most of the sequence files processing like the Galaxy Server (see Table QG13.2). If you manage to do a local installation, you are doing it right. This is for the basic processing of the data, QC filtering, and trimming. Also, this is manageable by most of modern personal computers. The overall process could be divided in the following stages: (1) Quality Control (QC), (2) Mapping against reference sequences, (3) *de novo* assembly, (4) annotation, (5) statistical analysis, (6) sharing your results. Each stage is described with useful hints at every step:

13.7.1 Sequences Quality Control

1. Split the libraries into individual files, this is also known as de-multiplexing, if you are using barcodes to mix several samples in a single run. Here the samples are split based on its barcode sequence.
2. Remove sequencing adapters. Removing this sequences that were used as templates for the sequencing is important and could help to further steps of mapping or assembly.
3. Quality Control, sequence trimming (and grooming). Each sequenced base has its own quality value, which is known as Phred score. Phred score serves as a proxy probability calculator, a Phred value of 30 accounts for 1 error every 1000 bases, or a 99.9 % of accuracy. This is a good standard to make a cut-off. Visualize the overall quality of your sequences via boxplots.
4. Filter rRNA. A quick way to do this step can be done with an rRNA DB and MegaBLAST (Altschul et al. 1997). There are other strategies using Interpolated Markov Models like Infernal and SSU-align and will help at this stage (Nawrocki et al. 2009; Nawrocki 2009).
The fast_toolx is a relatively easy way to perform the QC steps, plot qualities, and manipulate FASTA/FASTQ files. If command line is not an option, you should try the Galaxy servers to perform de-multiplexing, trimming adapters, and quality control (NGS QC and manipulation). The trade-off between working on the cloud or locally is the speed and fine tweaking of the pipelines, which are better controlled in our own computers. There are plenty of tutorials helping beginners to become familiar with Galaxy (see Table QG13.2; Kosakovsky et al. 2009).

13.7.2 Mapping Against Reference Sequences

1. Mapping against the reference metagenome/genomes. Use short read aligners. If there is no reference sequence(s), go to Sect. 13.7.3.

Here the standard options for short read mapping are Bowtie2 (Langmead et al. 2009), and BWA (Li and Durbin 2009). All of the mentioned programs are freely available online to be installed in CLI. There is also a cloud option provided by Galaxy under NGS Mapping. You should provide reference sequences, index the references if you are running this locally, and your metatranscriptome fastq files. After the alignment, you need to take the SAM/BAM resulting file and count the occurrence of each gene model (if available). The counting of each gene could be accomplished with R. R is a computer language intended for statistical computing and graphics, and the main recommended tool for downstream analysis (R Development Core Team 2004). For this purpose, use the libraries Rsamtools, summarizeOverlaps, and featureCounts of BioConductor (Huber et al. 2015).

13.7.3 De Novo Assembly

1. *De novo* assembly of metatranscriptome. This step applies if you do not have a reference, or you can do this step with the reads that were not aligned to it. You can perform *de novo* assembly if you do not have reference sequences, keep in mind that the most frequent limiting factor during assembly is the amount of RAM memory of

your computer. The amount of time required for assembly could last from minutes to days depending on the amount of sequences, and its complexity (repeats, SNPs, transcript forms, etc.). The most frequent choices are Velvet (Zerbino and Birney 2008), SOAPdenovo (Li et al. 2009), and Trinity (Grabherr et al. 2011). There is no clear better option when talking about assembly, you can try each one of them and can cluster the overall results at the end (with CD-HIT-est; Huang et al. 2010). All the mentioned programs are freely available online ready to be installed in your CLI. Trinity, has a cloud Galaxy based service that you could give a try (see Table QG13.2), this is recommended if you do not have enough computational resources locally.

13.7.4 Annotation

1. Annotate each transcript. If you have a metagenomic/genomic dataset already annotated, the coordinates could help you. Otherwise search by homology must be done. If there are no homolog sequences, you can try to use some RNA structural tools.

For the annotation, a hierarchical schema is suggested. If you know the species you are comparing and there are available annotated genome sequences for them, you could perform BLAST searches directly to them (Altschul et al. 1997). Then, for the sequences without homologs, go up to the next hierarchy a bacterial DB (see Table QG13.2). If there are still not homologs, try the largest DB, the NCBI's NR (see Table QG13.2). This could be tricky if you do not have the computational resources or the skills to perform it. Don't panic, there are some other cloud-based solutions like the KAAS, which is the KEGG's Automatic Annotation Server, where you can upload your assembled transcripts and annotate them, this is the most fast annotation tool that we are aware of (Moriya et al. 2007). The other main web-server solution is MG-RAST, which has the most elegant DB design which is named M5NR (Wilke et al. 2012). M5NR merges information from plenty of Databases in a nonredundant way like the annotation ontologies COG, SEED, eggNOG, KEGG, UniProt, IMG, Patric, RefSeq, SwissProt, TrEMBL, GO, and the NCBI's NR (Tatusov et al. 2000; Overbeek et al. 2014; Powell et al. 2014; Kanehisa and Goto 2000; UniProt Consortium 2008; Markowitz et al. 2008; Wattam et al. 2014; Pruitt et al. 2005; The Gene Ontology Consortium 2014), all this information is accessible through the metagenomics analysis server (MG-RAST; Glass and Meyer 2012). This is the source to have the most cost-effective annotation pipeline for a regular wet-lab, though you will not learn any bioinformatic skill with this. The MG-RAST accepts uploads of FASTQ or regular FASTA files but be aware that you will need to upload some experiment metadata, the data remains private until you ask the MG-RAST system to release it to the public, so it also serves as a sequence repository. If no homolog is present in your DB, you could use some tools like tRNA-SCAN (Lowe and Eddy 1997), and RNAFold (Denman 1993) to find out if there is a chance to classify your sequences by its secondary structure (i.e., hairpins, loops). The structural look at your data is demanding in computational and human resources to inspect the results. This approach could be useful if you are looking for particular class or regulatory elements (sRNAs, riboswitches). An excellent overview on annotation that should be reviewed, to understand the complexity of using multiple evidences to annotate, was done by Yandell and Ence (2012).

13.7.5 Statistical Analysis

1. Build a count matrix. This could be done by counting the mapped reads or to cluster the sequences of all the experimental conditions by its identity and count the number of occurrences in each sample/experiment. This step is required for parsing the annotation data to the Data Analysis pipeline. If you have processed your datasets on MG-RAST there is an option to export the whole dataset in BIOM format (http://biom-format.org/). The BIOM format is an acceptable input to R. There are ways to switch from BIOM to plain tabulator separated file with biom-convert tools. If you do not feel like using BIOM matrix, you could build a "table" where each row represents each individual gene and each column accounts for each sample/replica, save the file in plain text would work fi ne for R's input. In R, be sure to read the data as matrix.

2. Transform your matrix. There are several methods to accomplish this, one is the regularized-logarithm transformation (rlog), when measuring distances and sample similarities, and other normalizations like DESeq, which uses a negative binomial distribution, are preferred for differential expression. The log 2 and regularized logarithm transformation, also known as r-log, are the usual choice. This works to normalize your data between experiments, samples, and replicas, diminishing the importance and dependence of mean values. To perform this we recommend to use the R's Bioconductor package DESeq2 and its function RNAseqGene (Love et al. 2014).

3. Assess sample/treatment similarity, using heatmaps, Principal Component Analysis and calculating the distance on the r log/log 2 transformed data. With the transformed matrix, we can now describe the dissimilarity between samples/ replicates/experiments by means of clustering analysis. The preferred option is to use heatmaps and Principal Component Analysis (or whatever ordination method you feel comfortable with). For this purpose we recommend to use the packages heatmap.2 and the function plotPCA, part of DESeq2 package.

4. Perform the differential expression analysis. In this point, you need to calculate the log 2 fold changes between your treatments (control vs. experiment). Here you will have to calculate the mean, log 2 fold change, its standard error, and test the null hypothesis that there is no change between treatments on each gene and, thus, reported as a p-value. For this step of the process, you could employ plenty of available tools some of the most used ones are: edgeR, DESeq, baySeq, NOISeq, and Cuffdiff (Trapnell et al. 2013; Tarazona et al. 2011; Hardcastle and Kelly 2010; Anders and Huber 2010; Robinson et al. 2010). The differences between the tools are based on what tests and assumptions they are based upon: Fisher's, negative binomial, parametric or nonparametric methods.

5. The p-value of RNA-seq is not what you are used to. You need to perform multiple testing correction, to calculate the amount of false discovery rate (FDR), or in other words the amount of false positives, and then assess the significance of the adjusted p-value. Remember that this is to answer how much of false positives could be accepted. There are multiple tools to calculate FDR and corrected p-values like metagenomeSeq which is available as part of Bioconductor and a standalone webserver (metastats), thus just working for pairwise control and experiment comparisons. This can also be done with DESeq2 package and its p-adjusted (p-adj)

values.
6. Visualize the amount of significant differentially expressed genes. You can do this by means of Volcano plots, and heatmaps. If you are running a pairwise comparison, one way to accomplish this is by means of Volcano plots (log 2 fold changes versus significance), or an MA plot (M = log ratios, A = average). This is done also by R's Bioconductor.
7. Connect the most abundant features with its annotation. To this purpose is extremely helpful to use an ontology. An ontology is a controlled dictionary about gene functions, organized in hierarchical way like: SEED, COG, GO, KO. After determining the overall significant differentially expressed genes, usually they are coded with an identifier to reduce the amount of data loaded into R. A new table with the DE-genes and its annotations is extremely useful. To build that table the use or relational databases (MySQL, PostgreSQL) makes this an easy task.
8. Make sense of the known and annotated genes to direct new working hypothesis about their gene expression under the tested circumstances. The whole dataset of significant genes derived from the previous steps could be divided into two main groups: genes with known functions, and genes with unknown functions. Most of the functional analysis will focus on the known annotated genes, and it is the easier part of the dataset to explain but most probably a large amount of the data from your metatranscriptome will be transcripts of unknown function and thus are suitable candidates to design further experiments to discover their function (mutants, heterologous expression, etc.). An expressed gene is better than a total hypothetical predicted gene. For the genes with a known function, a process of data mining will be necessary to get the most about the functions and processes involving their participation. There are several starting points for gene function data mining like the Protein Data Bank, UniProt, Pfam, InterPro, EcoCyc, STRING, and KEGG (Berman 2000; Finn et al. 2008; Karp et al. 2002; Szklarczyk et al. 2011; Kanehisa and Goto 2000; Hunter et al. 2012). The main advantages of using those starting points is to gain insights about the current knowledge of the proteins and access to the overall information like if there are any available crystal structures, the phylogenetic distribution, known and predicted interactions. The main resource to integrate the information would be spending hours searching PubMed for related literature, and connecting it on new associations something that not machine, for the moment, could not do better than our brains.

13.7.6 Sharing Your Results

1. Upload your RNA-seq experiments to appropriate databases and repositories. To upload your datasets the main repository is NCBI's Short Read Archive (SRA) where you need to register your project and then upload your raw FASTQ files to it. To upload your assemblies there is the Transcriptome Shotgun Assembly Sequence DB (see Table QG13.2). The suggested way to share the annotated dataset is through the MG-RAST server, this also assures you to have up-to-date annotations, and it becomes available to be compared with other publicly available datasets.

13.8 Final Remarks

Metatranscriptomics as its relative metagenomics is attracting newcomers from multiple disciplines. The potential outcome to study both the environmental genome

and its expression under certain conditions is a promising tool to describe the taxonomic and functional diversity out there. There is a hype about the meta-omics everywhere now, and everyone is trying to sequence; this is great and opens new opportunities to learn from a myriad of scientific perspectives. We just want to recommend to be cautious before getting into the omics fashion trend, and be aware that you need some prerequisites before getting into the adventure: a well-established and -equipped molecular biology laboratory, some computational hardware, and the most valuable asset of trained people on both experimental and analytical aspects. Take your time to plan the experimental design before getting started; do not be part of a growing disappointed crowd that ventures without any experimental design/ controls and thus not able to get trustworthy biological meaningful data, or people with large experimental background but lacking the required analytical skills to tackle millions of multivariate data. Recognize your strengths and weaknesses, and go for successful collaborations; welcome to and good luck in the vibrant meta-omics road.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by the UNAM-DGAPA-PAPIIT grant IA200514, and the CONACyT Ciencia Básica grant 0237387.

Author details
[1]1. Departamento de Ciencias Naturales, Unidad Cuajimalpa, Universidad Autónoma Metropolitana, Av. Vasco de Quiroga 4871, Col. Santa Fe Cuajimalpa, 05348 Mexico city, MX. [2]1. Departamento de Ciencias Naturales, Unidad Cuajimalpa, Universidad Autónoma Metropolitana. 2. Laboratorio Nacional de Ciencias de la Sostenibilidad (LANCIS). Instituto de Ecología, Universidad Nacional Autónoma de México, AP 70-275, Ciudad Universitaria, Coyoacán, 04510 Mexico city, MX.

References
Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25(17):3389–3402
Alberti A et al (2014) Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. BMC Genomics 15(1):912
Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol11(10):R106
Benítez-Páez A et al (2014) Microbiota diversity and gene expression dynamics in human oral biofilms. BMC Genomics 15(1):311
Berman HM (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
Cole JR et al (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37(November 2008):141–145
Denman RB (1993) Using RNAFOLD to predict the activity of small catalytic RNAs. Biotechniques 15(6):1090–1095
DeSantis TZ et al (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72(7):5069–5072
Finn RD et al (2008) The Pfam protein families database. Nucleic Acids Res 36(Database issue):D281–D288
Franzosa EA et al (2014) Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A 111(22):E2329–E2338
Frias-Lopez J et al (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A 105(10):3805–3810
Giardine B et al (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15(10):1451–1455
Gilbert JA et al (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One 3(8):e3042
Gilbert JA, Hughes M (2011) Gene Expression Profiling: Metatranscriptomics. Methods in Molecular Biology 733:195–205
Glass EM, Meyer F (2012) 13. Analysis of metagenomics data. In: Rodríguez-Ezpeleta N, Hackenberg M, Aransay AM (eds) Bioinformatics for high throughput sequencing. Springer, New York, NY, pp 219–229
Gosalbes MJ et al (2011) Metatranscriptomic approach to analyze the functional human gut microbiota. PLoS One 6(3):e17447
Grabherr MG et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652
Güell M et al (2011) Bacterial transcriptomics: what is beyond the RNA horiz-ome? Nat Rev Microbiol 9(9):658–669

Hardcastle TJ, Kelly KA (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11(1):422
Hewson I et al (2009) Microbial community gene expression within colonies of the diazotroph, Trichodesmium, from the Southwest Pacific Ocean. ISME J 3(11):1286–1300
Huang Y et al (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics (Oxford, England) 26(5):680–682
Huber W et al (2015) Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12(2):115–121
Hunter S et al (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40(Database issue):D306–D312
Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
Karp PD et al (2002) The EcoCyc database. Nucleic Acids Res 30(1):56–58
Kelley DR et al (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res 40(1):e9
Knight R et al (2012) Unlocking the potential of metagenomics through replicated experimental design. Nat Biotechnol 30(6):513–520
Kosakovsky Pond S et al (2009) Windshield splatter analysis with the Galaxy metagenomic pipeline. Genome Res 19(11):2144–2153
Langmead B et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25
Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25(14):1754–1760
Li R et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics (Oxford, England) 25(15):1966–1967
Li S et al (2014) Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol 32(9):915–925
Li S-K et al (2013) Organism-specific rRNA capture system for application in next-generation sequencing. PLoS One 8(9):e74286
Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550
Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5):955–964
Lozupone C et al (2011) UniFrac: an effective distance metric for microbial community comparison. ISME J 5(2):169–172
Luo H et al (2014) The importance of study design for detecting differentially abundant features in high-throughput experiments. Genome Biol 15(12):527
Luo R et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience 1(1):18
Markowitz VM et al (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36(October 2007):534–538
Meyer F et al (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386
Moriya Y et al (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35(Web Server issue):W182–W185
Nawrocki EP (2009) Structural RNA homology search and alignment using Covariance Models. Washington University, St. Louis
Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics (Oxford, England) 25(10):1335–1337
Overbeek R et al (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 42(5):1–9
Parkhomchuk D et al (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res 37(18):e123
Paulson J, Pop M, Bravo H (2011) Metastats: an improved statistical method for analysis of metagenomic data. Genome Biol 12(Suppl 1):P17
Powell S et al (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res 42(Database issue):D231–D239
Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33(Database issue):D501–D504
R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.orgR
Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) 26(1):139–140
Sambrook J, Russell D (2012) Molecular cloning: a laboratory manual, 4th edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
Schloss PD (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput Biol 6(7):e1000844
Sorek R, Cossart P (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat Rev Genet 11(1):9–16
Szklarczyk D et al (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39(Database issue):D561–D568

Tarazona S et al (2011) Differential expression in RNA-seq: a matter of depth. Genome Res 21(12):2213–2223
Tatusov RL et al (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28(1):33–36
The Gene Ontology Consortium (2014) Gene Ontology Consortium: going forward. Nucleic Acids Res 43(D1):D1049–D1056
Trapnell C et al (2013) Differential analysis of gene regulation at transcript resolution with RNAseq. Nat Biotechnol 31(1):46–53
UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36(Database issue):D190–D195
Wattam AR et al (2014) PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 42(Database issue):D581–D591
Westermann AJ, Gorski SA, Vogel J (2012) Dual RNA-seq of pathogen and host. Nat Rev Microbiol 10(9):618–630
Wilke A et al (2012) The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. BMC Bioinformatics 13:141
Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13(5):329–342
Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de bruijn graphs. Genome Res 18(5):821–829
Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38(12):e132
Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9(1):133–148

Table 1. Metatranscriptomic Experimental design (hints)

Technique	Protocol	Control library	Recommended starting material (ng)	Number of Replicates	Sequencing depth	Recomended sequencing platform & run
Metatranscriptomics	RNAseq	cDNA Input Control	Minimum of 1 ng (TruSeq®) Tipically 100 ng cDNA RNA integrity, check nanodrop UV spectrums and if possible RNA degradation with Bioanalyzer®	2 minimum for each library, for being able to estimate variances.	If possible, estimate species diversity (16S rRNA amplicons). Then, calculate expected average genome sizes, then calculate an 8X minimum coverage.	Illumina's MiSEQ 2 x 300 bp (0.3 – 15 Gb, HiSeq 2500 (10 – 1000 Gb). First check with your provider to be able to use the latest technology. Considerations: Quality Coverage Read length Sample number Budget

Table 2. Web Resources. This is a list of basic tools and resources that could help to get through the bioinformatic and statistical analysis required for metatranscriptomics.

Software / Resource	Notes	Method / Language / Platform	Input	Results output	Results format	URL References
Basic Linux Tutorial	Start here	Web resource Linux/UNIX, Windows	-	-	-	http://www.ee.surrey.ac.uk/Teaching/Unix/ -
Bioconductor training workflow for RNA-seq	Want to try the experience of statistical analysis ahead of getting your own sequence? This is the place to start. It has a guide with publicly available data to perform the whole analysis. Highly recommended	Web resource R packages	Example datasets are provided in the workflow	Tables Plots	Raw text FASTA/FASTQ plots (png, pdf).	http://www.bioconductor.org/help/workflows/rnaseqGene/ - (Huber et al. 2015)
BIOM	Helpful format if dealing with multiple samples and probably the next standard for reporting abbundances results for metagenomics	Python scripts Linux/UNIX	Clustering files	BIOM	BIOM matrix could be converted to tabular output (biom convert).	http://biom-format.org/
CD-HIT-est	The preferred selection for gene clustering at established cut-off thresholds	Executables Linux/UNIX	MultiFASTA file	Clusters file, and representative sequences	Raw text Multi FASTA file	http://weizhongli-lab.org/cd-hit/ (Huang et al. 2010)
Choosing Sequence Technology / provider	This web sites offer free help to choose sequencing technology and provider	Web Any OS	Declare number of samples, sequencing library preparations,	Provider comparsions	Html	http://allseq.com and, https://genohub.com

			sequencing coverage, and budget			
Experimental Design in Differential Abundance analysis (EDDA)	Here you can design your RNAseq experiment	Web R packages Any OS Linux/UNIX	Number of samples, replicates,	Raw text	Raw text / html	http://edda.gis.a-star.edu.sg/dad/ (Luo et al. 2014)
FastX Toolkit	A useful software suite that allows to process initial quality control, screening, and filtering of sequencing files. This is the place to start if you just received sequences. It is possible to integrate FastX to a local installed Galaxy server	Executables Linux/UNIX	FASTQ	FASTA plot images	MultiFASTA png, pdf images (for quality boxplots).	http://hannonlab.cshl.edu/fastx_toolkit/
Galaxy Server / Tutorial	The galaxy servers allows to process and manipulate from raw sequences / data / mapping files to plots, sequences.	Source Web server Linux/UNIX Any OS	FASTA FASTQ and virtually all the known formats for alignments, mappings, and biological files	Plots, alignments, mappings.	Plots (png, pdf), raw text, alignments (SAM, BAM)	https://usegalaxy.org/ https://usegalaxy.org/u/aun1/p/galaxy101 (Giardine et al. 2005)
Glimmer-Gm	Metagenomic gene prediction	scripts Linux/UNIX	FASTA	MultiFASTA file with gene predictions Gene calling scores file	MultiFASTA Raw text	http://www.cbcb.umd.edu/software/glimmer-mg (Kelley et al. 2012)
Bioconductor	Defined by the Bioconductor community: “provides tools for the analysis and comprehension of high-throughput genomic data”	R packages Linux/UNIX never tried it but it is stated that it could run on	FASTA FASTQ and virtually all the known formats for alignments,	Raw text Html multiFASTA alginments	Txt html multiFASTA alignment files	http://bioconductor.org/install/ (Huber et al. 2015)

		Windows (risks are on your own).	mappings, and biological files			
R	Is a programming language, intended for statistical computations and analyses.	Source, and executables Any OS	Tables, raw text, csv, JSON, BIOM, FASTQ	Raw text plots	Raw text PDF, png	http://cran.r-project.org/ (Team 2008)
KEGG’s Automatic Annotation Server	Automatted annotation server using Kyoto Encyclopedia of Genes and Genomes	Web server Any OS	MultiFASTA files	MultiFASTA files with annotated datasets Metabolic pathways maps	MultiFASTA raw text / html png images	http://www.genome.jp/tools/kaas/ (Moriya et al. 2007)
Metagenemark	Metagenomic gene prediction	Web server Any OS	FASTA	MultiFASTA file with gene predictions Gene calling scores file	MultiFASTA Raw text	http://exon.gatech.edu/meta_gmhmmp.cgi (Zhu et al. 2010)
Metagenomics analysis server MG-RAST	An amazing web server that allows to annotate high-throughput metagenomic experiments in both taxonomic and functional features. This server integrates information features from multiple databases into its M5NR DB. there is an API (application program interface) which allows to access server capabilities from terminal.	Web server Any OS, works best with Firefox,	FASTQ, FASTA files,	MultiFASTA files Abbundance tables Heatmaps Ordination plots	MultiFASTA BIOM, txt, html png, pdf	http://metagenomics.anl.gov/ (Meyer et al. 2008)
Metagenomeseq Metastats	Determine differential abundant species, genes and features	Bioconductor R package Web server Linux/UNIX Any OS	Abundance table (raw text, csv file) R object	Tables with p-values and False Discovery Rate (FDR) corrected values.	Raw text png, pdf	http://metastats.cbcb.umd.edu/software.html (Paulson et al. 2011)

				Plots		
NCBI all bacteria	Predicted proteomes of finished bacteria genomes	Database Any OS	Format DB to use it with BLAST	BLAST DB	Raw text / html	ftp.ncbi.nih.gov/genomes/Bacteria/all.faa.tar.gz
NCBI NR DB	The reference DB to perform annotation, it includes all the known proteins deposited on GenBank. The size of this DB is huge and you will need a computer with enough RAM to handle it.	Database	Already formated DB ready to use with BLAST standalone versions	BLAST DB	Raw text / html	ftp.ncbi.nih.gov/blast/db/
NCBI's Transcriptome Shotgun Assembly (TSA)	TSA hosts assembled sequences of transcriptomes, by any method from traditional cDNA clone / sequencing to NGS datasets	Database	BAM unannotated assembly. Sequences of at least 200 b. No more than 10% ambigous bases.	TSA record (Contigs)	GenBank file FASTA file ASN.1 file	http://www.ncbi.nlm.nih.gov/genbank/tsa/
NCBI's Short Read Archive SRA	Stores raw sequence data from NGS, is the primary archive of NGS data.	Database	FASTQ BAM qseq srf	SRA files raw sequencing reads	Dump to FASTQ is available (FASTQ-dump)	http://www.ncbi.nlm.nih.gov/sra
RNA-seqlopedia	This is an amazing resource for understanding step by step the RNA-seq processing. We highly recommend it.	Web resource	-	-	-	http://rnaseq.uoregon.edu/
RNAfold Web Server	RNA secondary structure prediction	Web server Standalone version Linux / UNIX	RNA sequence in FASTA format	Minimum free energy structure calculation	raw text, html PDF, png	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi (Zuker & Stiegler 1981)

SOAPdenovo	Short Oligonucleotide Analysis Package (assembler). Designed as an Illumina's short read genome assembler.	Executables 64-bit Linux, minimum 5 G RAM	FASTA FASTQ BAM	Contigs Scaffolds mappings pregraph	Raw text	http://soap.genomics.org.cn/soapdenovo.html (Luo et al. 2012)
Trinity / Trinity Galaxy server	Transcriptome assembler.	Scripts Web server Linux/UNIX Any OS	FASTQ	Contig file	FASTA format	http://trinityrnaseq.github.io/ https://galaxy.ncgas-trinity.indiana.edu/root (Grabherr et al. 2011)
tRNAscan-SE	tRNA prediction software	Web server Scripts Linux/UNIX	FASTA GenBank EMBL GCG IG Raw sequence	tRNA predictions Run statistics Predicted tRNA structures.	Raw text	http://lowelab.ucsc.edu/tRNAscan-SE/ (Lowe & Eddy 1997)
Velvet assembler	Short read assembler	Scripts Linux/UNIX	FASTA FASTQ	Contigs file Stats file Velvet assembly file	FASTA raw text asm file (open with AMOS)	https://www.ebi.ac.uk/~zerbino/velvet/ (Zerbino & Birney 2008)