Metatranscriptomics:

A complete guide to characterizing metabolic activity in complex microbial communities with high-throughput RNA sequencing

David Levy-Booth

Metatranscriptomics

MICROBIOME INSIGHTS

Metatranscriptomics:

A complete guide to characterizing metabolic activity in complex microbial communities with high-throughput RNA sequencing

Version 0.1 (2018-07-05)

Prepared for:
Microbiome Insights Inc.
2950 Tolmie St.
Vancouver, BC
V6R 4K6

Prepared by:
David Levy-Booth
312-4028 Knight St.
Vancouver, BC
V5N 5Y8

CONSULTANT

David Levy-Booth is a research scientist in the fields of microbial ecology and biotechnology, with over 15 years of experience investigating microbial communities and their activity. He holds a PhD from the University of British Columbia and an MSc in Microbiology and Biotechnology from the University of Guelph. Dr. Levy-Booth consults on the adoption of microbial multi-omics in the agro-ecology and human health sectors.

"Nothing is more exciting for a scientist than to be caught among contradictory facts; nothing is more rewarding for him than to find the right way out of the contradiction."

- Jean Brachet, pioneer of RNA research[1]

[1] Jean Brachet. Rewriting the Book on Nucleic Acids. The Scientist Magazine. Jun 15, 1987. https://www.the-scientist.com/perspective/rewriting-the-book-on-nucleic-acids-63706

Contents

Glossary

- **RNA-Seq** – High-throughput RNA sequencing.
- **Genome** – The full, intact DNA sequence in a cell, individual, strain or species. Can be organized as a single chromosome, multiple chromosomes or extrachromosomal elements such as plasmids.
- **Transcriptome** – The full suite of expressed RNA in an organism including messenger RNA (**mRNA**), small RNAs (miRNA, siRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). Here, transcriptome will refer primarily to the mRNA fraction transcribed from protein coding genes.
- **Metatranscriptomics** – The characterization, sequencing, assembly and quantification of the full suite of RNAs, but primarily **mRNA**, in a whole-community. Often performed using **shotgun sequencing** approaches.
- **mRNA** – Messenger RNA the short-lived intermediate between DNA and protein. It is synthesized by RNA polymerase as the complementary sequence to a protein coding gene during transcription. In eukaryotes, pre-mRNA is matured by removing introns, before shuttling from the nucleus to ribosomes for translation to proteins. In prokayrotes, ribosomes attach directly to mRNA strands during synthesis from DNA by the RNA polymerase enzyme.
- **rRNA –** Ribosomal rRNA comprises the structural and catalytic subunits of the ribosome, where tRNA shuttles amino acids for elongation into protein chains, which is encoded by mRNA. In eukaryotes, large and small rRNA subunits are termed 28S LSU and 18S SSU, respectively. In prokaryotes, the ribosomal LSU and SSU are 16S and 23S, respectively. rRNA can also refer to the genes that encode these subunits, the sequences of which are frequent used as phylogenetic markers.
- **Prokaryote – Bacteria** and **Archaea**. These distantly-related domains of life have some similarities**:** they are unicellular and are lacking membrane-bound nucleus or organelles, and importantly to metatranscriptomics have fundamentally different transcriptional machinery than the eukaryotes. Prokaryotes are the most diverse and abundant organisms in most host-associated and environmental **microbiome** samples.
- **Eukaryote -** From fungal yeasts and plants, to mammals, eukaryotes are defined by their membrane-enclosed nucleus, and have complex transcriptional machinery including poly-A tails, pre-mRNA maturation by **intron** removal, multiple splicing patterns, mRNA export from the nucleus and a lack of multi-gene operonic transcripts.
- **Intron –** In eukaryotes, sections of pre-mRNA that are removed during mRNA maturation prior to translation.
- **Exon** – In eukaryotes, sections of pre-mRNA that are retained during mRNA maturation.
- **Shotgun sequencing –** Cataloguing the full range of bases of all DNA or RNA in a sequencing reaction. Often used with randomly-organized short strands of nucleic acids. Must be computationally assembled to obtain full-length DNA or RNA molecules.

Executive Summary

Metatranscriptomics is the characterization and quantification of the total messenger RNA (mRNA) content in a microbial community. It made possible by advancements in massively-parallel shotgun sequencing. This report serves as a comprehensive guide to the cutting-edge methods involved in end-to-end metatranscriptomics and is specifically designed to facilitate the adoption of metatranscriptomics-as-a-service (MTaaS) by Microbiome Insights Inc.

Metatranscriptomics offers unprecedented knowledge of how microorganisms respond to their environment. Sequencing novel transcripts can reveal new organisms, genes and pathways involved in biological transformations, and can show how environmental factors structure these complex expression networks. In all, metatranscriptomic analysis reveals the complex machinery the drives how microbial communities function in diverse ecosystems from host guts to deep-sea-vents and all environments in between.

MTaaS offers academic and corporate clients looking for the next breakthrough in microbial characterization access to a cutting-edge technology without having to invest in expensive infrastructure and advanced technical knowledge. However, many technical barriers exist to the adoption of high-throughput and cost-effective metatranscriptomics required to offer such analysis as a service. Many of these limitations emerge from the properties of RNA itself: RNA is a fragile and transient molecule, the isolation of which requires great care and technical capacity. As such, robust quality control/quality assurance (QA/QC) is involved in every step of RNA extraction and library preparation, expanding an already cumbersome and time-consuming process.

MTaaS is expensive. Material cost alone of an RNA extraction for a single sample can run between $60 and $100 (including QA/QC steps), with library preparation adding an additional $120 to $200. Sequencing costs for environmental samples can range between $200 and $400 per sample to obtain the high coverage required for assembly of all transcripts in a community. Multiple days of labour are required for this process. It is also computationally demanding, with assembly of complex meta-transcripts potentially overwhelming even the 264 GB sever used by Microbiome Insights. Less demanding options for “shallow,” unassembled short-read metatranscriptomes are possible but would require testing and validation.

Despite the sizable barriers to adoption, offering MTaaS would provide a strong competitive advantage. To the best of my knowledge, no other sequencing centre or microbiome company offers full, end-to-end metatranscriptomics: sample RNA extraction, sequencing and analysis. In the following report, I make 16 recommendations to facilitate the adoption of MTaaS and provide a detailed theoretical and practical underpinning of the entire metatranscriptomics process from sampling to analysis.

Recommendations

1 Any metatranscriptomics experiments should undergo an extensive design process, including power analysis. Experimental designs should be kept simple, with pairwise comparisons offering the most

straightforward analysis without losing statistical power due to multiple comparisons (**Section 3.1. Experimental design).**

2 Stool samples should be preserved using RNAlater or equivalent stabilization solution and frozen at -70^{o}C. Preserved samples should be extracted for RNA no longer than one week after collection, extended to one month if frozen. Soil samples should be flash-frozen in liquid N_2 or dry ice. Stabilization in LifeGuard Soil Preservation Solution recommended only if flash freezing is unavailable (**Section 3.2 Sampling**).

3 Conducting MTaaS starts before samples have been shipped, with clear communication of sample requirements and shipping protocols to clients. A client-focused LIMS should support semi-automated validation of client-submitted sample information prior to authorization to ship samples. Laboratory management should ensure that appropriate materials and trained technicians are available prior to accepting samples. A strict protocol should be in place to reject samples that fail QC, and this protocol should be communicated to clients prior to sample shipping. Clients should be encouraged to resubmit fresh material from failed samples (**Section 4.1 Prior to sample shipping**).

4 MTaaS laboratories should expect to receive two types of samples: raw material for RNA extraction, and pre-extracted RNA, each requiring standard operating procedures (SOP) be written for handling and sample QC. Specific QC at receiving includes: ensuring cold chain through the use of temperature loggers or presence of dry ice in the shipment, Ensuring sample mass, volume, and if RNA, concentration and purity are sufficient for library preparation and analysis (**Section 4.2 Sample receiving**).

5 MTaas laboratories require specialized facilities. This includes a fume hood for working with toxic organic solvents (e.g., phenol, chloroform), a biological safety cabinet for working with stool and other potentially infectious material, isolated bench space for aliquoting soil, dedicated lab space free from DNA or potential RNase contamination, ample -70^{o}C freezer space, a heat block capable of accurate heating to 62^{o}C, a fluorometer to measure RNA concentration, and other equipment such as a thermal cyclers and an autoclave (**Section 5.0 Facility and equipment requirements for RNA laboratory techniques**).

6 Soil RNA extraction is one of the most technically demanding aspects of metatranscriptomics. I have designed a modular protocol for the high-throughput extraction of a variety of soil types. Regular soil samples including agricultural soils and forest mineral layers can be extracted through the rapid lysis protocol, while samples containing “indestructible” bacteria can be extracted using a specialized high-temperature, high-force extraction method in “T-1000” buffer. Nucleic acid purification by phenol-chloroform liquid-liquid extraction requires the use of toxic solvents, but produces high yield and purity. Organic contaminants can be removed from samples with dissolved organic carbon concentrations > 10 mg/L by a series of flocculation and filtration steps. All methods feed into the same final purification and DNA removal protocols to provide intact, high-quality RNA suitable for sequencing (**Section 6.8 Extraction methods**).

7 RNA quality should be assessed using multiple methods. Gel electrophoresis to ensure that ribosomal RNA (rRNA) bands are intact and that genomic DNA has been removed. Fluorometric quantification is required even if clients provide sample concentration, as RNA may have degraded during shipping.

Optical purity ratios will reveal potential contamination issues including organic matter, phenol or salts. Finally, digital Bioanalyzer traces are recommended after library preparation to ensure that rRNA has been successfully removed from total RNA, leaving an appropriate concentration of mRNA for complementary DNA (cDNA) synthesis and sequencing (**Section 6.7 Post-extraction quality control/quality assurance**).

8 A variety of rRNA removal methods and RNA library preparation kits exist. I recommended using the Illumina ScriptSeq Complete Kit (Bacteria) kit, which combines these steps into a single cost-effective library preparation method suitable for partially-degraded RNA and low-concentration libraries (≥100 ng total RNA). (**Section 6.10 rRNA removal and library preparation**).

9 At this time, Illumina HiSeq2500 (1 lane), NextSeq 550 or HiSeq2500 (8 lane) are the recommended platforms for cost-effective small, medium and large RNA sequencing runs, respectively (**Section 7.1 Illumina RNA Sequencing**).

10 Samples per sequencing flow-cell should be optimized to provide at least 20 million reads per sample aiming for a total coverage of 1 Gb for host-associated experiments and 30 Gb for environmental experiments. Assembly requirements may be less strict than metagenomic or genome sequencing, as resolving the sequences of repetitive intergenic regions is not required when sequencing transcripts. On the other hand, a small number of transcripts, primarily housekeeping genes, will take up the majority of coverage. Coverage requirements that approach those of metagenomes can ensure that relevant genes in biochemical transformation pathways (aka "functional genes") are detected. Shallow, unassembled metatranscriptomes may be possible with 10-100x less coverage, but this approach would require extensive validation (**Section 7.2 Estimating coverage requirements for metatranscriptome assembly**).

11 I recommend using the de Bruijn graph assembler *Trinity* for transcript assembly due to its extensive documentation, ease of installation and use, low memory footprint, overall accuracy and useful add-ons for transcript quantification and annotation. However, dedicated metatranscriptome assemblers such as *IDBA-MT* can reduce chimeric assembly compared to *Trinity* and should be investigated as a potential replacement (**Section 7.2.1 Software for *de novo* metatranscriptome assembly**).

12 As assembly is a time-consuming and computationally-expensive process, host-associated metatranscriptomes recovered from microbial communities that are well represented in genome databases should instead be annotated and quantified by alignment to reference libraries. Further, no assembly is required for short-read annotation pipelines. Short-read approaches should maximize read length (e.g., using 150 bp-paired-end sequencing).

13 Long reads, specifically using Oxford NanoPore platforms, may offer a cost-effective (at ~$150 per sample) alternative to expensive but high-quality Illumina sequencing. In addition to cost, benefits of NanoPore sequencing includes speed (1 to 2 days per set of 12 libraries for a total of ~18 Gb) and the ability to obtain full-length transcripts. To my knowledge, no data exist on the use of NanoPore sequencing for quantification of expressed genes in complex metatranscriptomes, an approach that would require extensive validation prior to commercial application (**Section 7.3 Long reads: PacBio vs. Oxford NanoPore**).

14 *MetaTrans* offers an attractive option for an existing general-purpose short-read metatranscriptome annotation and quantification pipeline. It is recommended that the initial steps be swapped out with lightweight *Trimmomatic* filtering and trimming, and that annotation use *RefSeq* and other reference databases over M5nr. For human host-specific short-read annotation and quantification, the HUMAnN pipeline is recommended (**Section 8.0 Bioinformatics pipelines for metatranscriptomics**).

15 Rather than relying on pre-built bioinformatics pipelines, custom metatranscriptomics pipelines can be constructed using modular principals to swap-out assembly, annotation and quantification strategies. This will allow control over the bioinformatics process for different sample types or research objectives. Many downstream analysis methods exist. The most fundamental to metatranscriptomics is differential expression (DE) analysis. *DeSeq2* or *edgeR* are recommended for this purpose.

16 Overall, it is recommended to offer MTaaS only if the following conditions are met:

a. Validation testing to ensure that each element of sample receiving, RNA extraction and analysis can be performed with speed and accuracy by trained laboratory personnel. It is difficult to achieve 100% of RNA extractions to pass QC thresholds; a threshold of at least 90% of samples passing QC should be achieved prior to offering MTaaS.

b. Implementation of a comprehensive and client-focused LIMS to ensure smooth client communication including sample validation, communication of QC results, and downloading of metatranscriptome sequencing results and analysis.

c. A fast, accurate and modular metatranscriptome bioinformatics pipeline developed to handle both host-associated and environmental samples. Each sample type should be capable of processing using assembled and unassembled read annotation and quantification.

Client focus: Microbiome Insights Inc.

Microbiome Insights Inc. is a small biotechnology enterprise started at the University of British Columbia in Vancouver, Canada. They offer sample DNA extraction, taxonomic marker (e.g., 16S rRNA gene) amplification, and sequencing for microbiome applications. Although additional services include metabolite analysis (e.g., short chain fatty acids) and shotgun metagenome sequencing, the company primarily provides bio-medical, cosmetics and environmental research clients with end-to-end taxonomic marker sequencing, bioinformatics and statistical analysis.

Microbiome Insights has several key advantages as a company including strong leadership and technical teams. Scientific personnel include world-leading principal investigators and bioinformatics technicians, and a highly-trained laboratory operations group. However, several vulnerabilities also emerge from being a relatively small enterprise in the emerging and competitive field of microbiome analysis. It is difficult to match the sample throughput of large, established competitors, or the costs of not-for-profit sequencing centres. **Leveraging their capacity to innovate in emerging technological applications can serve a key avenue for growth for Microbiome Insights.**

Metatranscriptomics, the massively-parallel sequencing of RNA from microbial communities, is one such emerging technology. Background information about metatranscriptomics and its application is provided in detail in the following sections of this report. Most sequencing centres offer one component of metatranscriptomics: RNA library preparation and sequencing. However, few provide the most technically-challenging components of this technology: RNA extraction optimized for host-associated or environmental samples, or bioinformatic analysis of metatranscriptomics data.

The objective of this report is to provide the theoretical and technical expertise required for complete end-to-end metatranscriptomics analysis. The ultimate goal in adopting **metatranscriptomics-as-a-service (MTaaS)** is to support clients with an unserved need in the highly-competitive space of microbiome analysis with an emerging technology capable of characterizing the activity and environmental response of microbial communities.

1.0 Introduction to RNA

High-throughput ribonucleic acid (RNA) sequencing (**RNA-Seq**) provides an unprecedented understanding of gene expression—from single organisms to complex communities. Unlike the DNA content of the **genome**, the **transcriptome** of an organism is in constant flux. Transcriptomics seeks to capture all mRNA present in a cell in a given moment in time. Since mRNA is short-lived, with a half-live measured in minutes, capturing transcriptional changes through sequencing is a powerful tool to investigate the cellular response to conditions experienced by microorganisms.

Key discoveries in RNA biology can elucidate the differences between RNA and DNA molecules, and therefore clarify the requirements of successful RNA-Seq experiments. Over the past century, many advances in understanding RNA chemistry and associated enzyme biology have created the conditions where we now can use simple kits to isolate RNA and begin massively parallel sequencing in under a week. These developments are fueling an explosion of unprecedented discoveries. But it can be valuable to start with the basics to appreciate why RNA sequencing works, and why things can go wrong during an experiment.

1.1 Discovery of RNA and differentiation from DNA

In the early 1930s, the prevailing wisdom was that RNA or "yeast nucleic acid" occurred in plants, while "thymus nucleic acid" (DNA) was found in animals. Pioneering work by Belgian chemist Jean Brachet and Danish-German biologist Joachim Hämmerling, connected RNA with the cytoplasm of eukaryotic cells, while DNA was associated with the nucleus [1–3]. By the 1950s it was understood that several key differences between RNA and DNA existed:

- Messenger RNA (**mRNA**) is single-stranded, although double-stranded forms exist (e.g., in some RNA viruses)
- RNA can form secondary structures including ribosomal subunits (**rRNA**) and transfer RNAs (tRNA). Ribosomal RNA accounts for the majority (90-95%) of cellular RNA in prokaryotes. Separating mRNA from rRNA is a non-trivial component of accurate transcriptomic sequencing.
- RNA is the intermediate structure in protein synthesis, transporting genetic information from DNA to the ribosomes
- mRNA is readily denatured, e.g., by high pH and is transient in the cell due to abundant intercellular RNases
- RNA contains the nucleobase uracil (U) instead of thymine (T)
- RNA nucleobase order can be determined using sequencing by synthesis

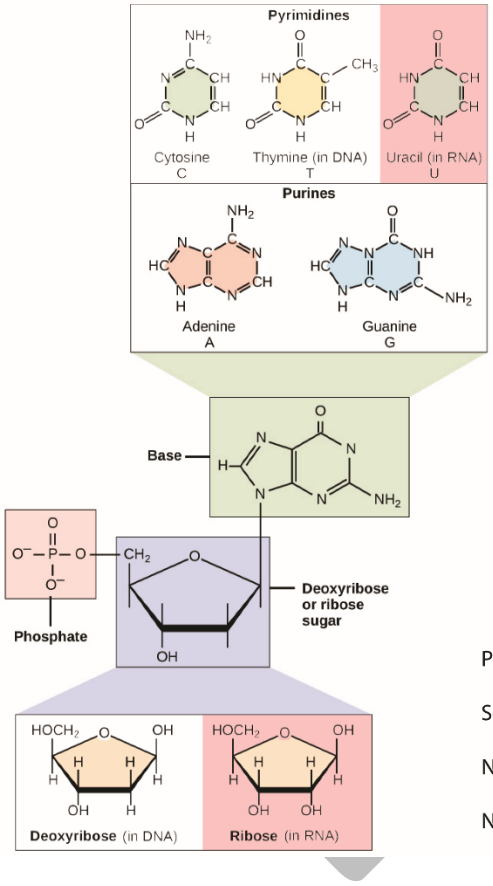
1.2 Building on pioneering RNA biology

While RNA handling and sequencing are routine today, we owe a substantial debt to these pioneers. It can be appreciated, for example, that when something goes wrong in an RNA-Seq experiment it is usually because of the short half-life of the mRNA molecule first discussed in 1941 by Frank Allen at the University of California—Berkley [4]. Or that viral **reverse transcriptases**, discovered by Howard Temin

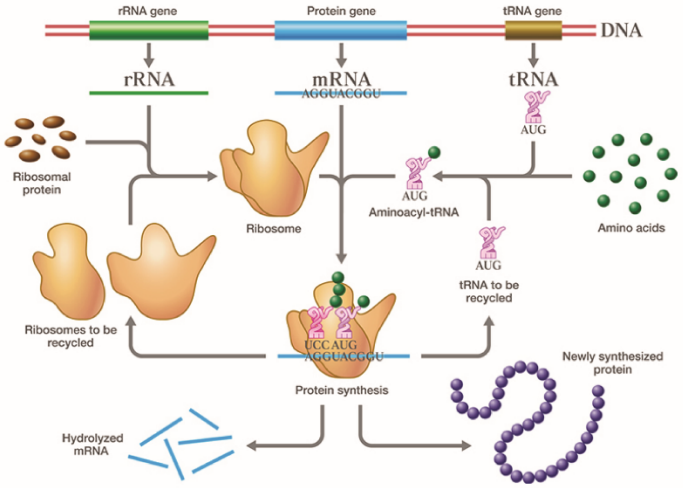
at the University of Wisconsin—Madison in 1970 [5], can produce complementary DNA (**cDNA**) from RNA in a stunning reversal of **central dogma of molecular biology**. We apply this advance to produce sequenceable cDNA from isolated RNA.

Researchers have also learned a great deal since this time about the presence of **introns** in immature eukaryotic mRNA, which can act as mobile genetic elements that silence the expression of other genes. Other small RNAs that can regulate gene expression were discovered as recently as the 1990s, including microRNA (miRNA) and small interfering RNA (siRNA). The role of antisense RNAs in carrying out similar regulatory functions in **prokaryotes** is an understudied topic that is currently gaining interest for researchers in bacteriology and microbial ecology [6].

A. Structural components of RNA



B. Role of mRNA, tRNA and rRNA in protein synthesis



C. Characteristics of of mRNA, tRNA and rRNA

	mRNA	tRNA	rRNA
Percent of cellular RNA:	3-5%	10-20%	~80%
Sedimentation coefficient:	8S	3.8S	28S-5S
Number of nucelobases:	900-1,500 bp	73-93 bp	120 bp / 5S
Number of unusual bases:	low	1 / 30-40 bp	1 / 150 bp

Figure 1. Characteristics of RNA. A) The structure of the RNA molecule is differentiated from DNA by its single-stranded ribose sugar backbone and by encoding a uracil nucleobase rather than thymine. B) Several types of RNA are involved in protein synthesis including messenger RNA (mRNA), encoding the code to produce amino acids, shuttled by transfer RNA (tRNA) to the ribosomal RNA (rRNA) subunits. C) Types of RNA can be differentiated by size, base methylation and percent of total cellular RNA. Adapted from https://cnx.org/ under Creative Commons Attribution 4.0 License.

Methods for the study of RNA have advanced rapidly. RNA extraction protocols have developed substantially since the early days of long CsCl ultracentrifugation runs. By leveraging the ability of phenol to separate RNA from DNA and proteins [7], researchers can perform rapid extraction and purification of RNA. Since phenol and other chemicals used in RNA extraction are highly toxic, new methods focused on reduced use of harsh organic solvents and the application of robotics-assisted, semi-automated workflows will provide researchers with even greater throughput. Northern blotting [8], mRNA microarrays [9] and RT-qPCR have been used to study gene expression by sequential generations of RNA researchers. Microarrays, which semi-quantitatively measure mRNA binding to arrays of fluorescent sequence probes, were the first tools for whole-genome expression but had to be painstakingly built from known sequences. Today, RNA-Seq offers a powerful tool for the detection and characterization of *all* mRNA in single organisms or complex communities with no *a priori* knowledge.

The **metatranscriptomics** revolution is poised to address longstanding questions about the activity of microorganisms in diverse environments from the surface of our skin, to soil, to deep ocean sediments. However, many challenges must be addressed to fully leverage the transformational capacity of metatranscriptomics.

2.0 RNA-Seq for metatranscriptomic research

2.1 Introduction to Metatranscriptomes

The **transcriptome** of a cell is the complete set of genes expressed under specific conditions. **Metatranscriptomes** are the complete set of expressed genes in a whole community (e.g., from a soil or fecal sample). **Shotgun** RNA-Seq can recover the full set of sequences in a transcriptome or metatranscriptome, and allow for the **semi-quantitative** characterization of expression rates.

Why do we characterize RNA-Seq data as semi-quantitative? RNA-Seq results in millions or billions of fragmented reads that can be assembled to produce full transcripts. A variety of tools are used to map individual reads to each transcript then estimate the number of mapped reads per transcript. These values reflect the sequencing depth of each transcriptomic library, the quality of the reads, the length of the transcripts and the software used to map and estimate read counts. They do not represent absolute expression (transcript copies per cell), and can over- or under-estimate expression rates relative to other transcripts. Therefore, it is of critical importance to verify key transcriptomic results using other methods, such as quantitative real-time PCR (**qPCR**). Nonetheless, transcriptomic sequencing is a powerful tool that can be used to address a number of research goals, including but not limited to:

- Recover sequences of expressed genes
- Relative differential expression patterns between treatments
- Detecting differential splicing and transcript isoforms in **eukaryotic** systems
- Elucidate potential activities of microbial communities and regulation mechanisms
- Linking ***in situ*** activity with process rates, phenotypes or treatments

Prior to initiating a metatranscriptomics experiment, researchers must be clear about the system they are studying. Transcripts are relatively short-lived, and therefore only reflect the immediate response of a cell or community to their direct environment. There will inevitably be an incongruence between

transcript abundance and protein levels, due to the longer life span of proteins, and potential post-transcriptional regulatory interference. DNA, proteins and mRNA reflect different time scales of information: DNA is essentially a stable, therefore genomics and metagenomics can reveal only the functional potential of a cell or community with little information regarding its response to environmental change. Proteins are moderately stable and provide a snapshot of the actual catabolic capacity of a cell. **Proteomics**, however, provides lower-resolution data at high cost relative to genomics or transcriptomics, and metaproteomics currently has an extremely low capacity to characterize unknown protein sequences. Metatranscriptomics, therefore, can offer a much more immediate and high-resolution snapshot of metabolic response, that is indicative, but not confirmatory, of catalytic potential.

2.2 Comparison with metagenomics

Metatranscriptomics is a direct extension of **metagenomics**, the sequencing of the full DNA content of an environmental or host-associated community. Metagenomics provides a profile of the partial **functional potential** of a sample, and when integrated with the metatranscriptomic profile can provide a semi-quantitative summary of expression rates of individual genes or organisms. A *de novo* metatranscriptomic approach can only be used to detect expression dynamics without normalization to the number of gene copies in a sample. Therefore, it is highly recommended that research that requires normalized expression values use paired metatranscriptomic and metagenomics sequencing.

Many of strengths, drawbacks, and workflows of metagenome sequencing apply to metatranscriptomics as well. Throughout this document, metagenomics/metatranscriptomics will be understood to mean whole-metagenome/transcriptome shotgun sequencing (WMS) unless otherwise noted. One of the key limitations of WMS is the vast sequencing coverage needed to accurately assemble and quantify genomes or transcriptomes, particularly in highly-diverse environments such as soil and sediments. A non-trivial task will be to correctly estimate the resources allocated to WMS. Handling this amount of sequencing data requires massive computational resources, and unlike metagenomes, there is no current understanding of appropriate coverage required for successful metatranscriptomics assembly.

DNA extraction for metagenomics is highly routine due to the small quantities of material required (1 ng DNA), the lack of post-extraction processing and the stability of DNA, both in communities and during laboratory handling. None of these attributes are true for RNA applications. Unlike the metagenome, the metatranscriptome is highly variable according to environmental conditions. Therefore, great attention must be paid to proper experimental design to maximize the likelihood of a successful experiment.

2.3 Challenges of metatranscriptomic experiments

- Sampling - Sample timing and storage poses non-trivial constraints. RNA*later* very effective for stability but liberates organic inhibitors in soil
- Difficult lysis conditions can degrade RNA, requires harsh chemicals
- Purification with flocculants/filters greatly reduce yields
- Host RNA contamination (e.g., biopsy samples)
- High diversity requires high sequencing depth and computational resources

- Assembly – ***de novo*** or assembly by genome/metagenome reference recommended for accurate quantification. Computationally intensive
- Count normalization – non-normalized, reference-normalized using metagenomes

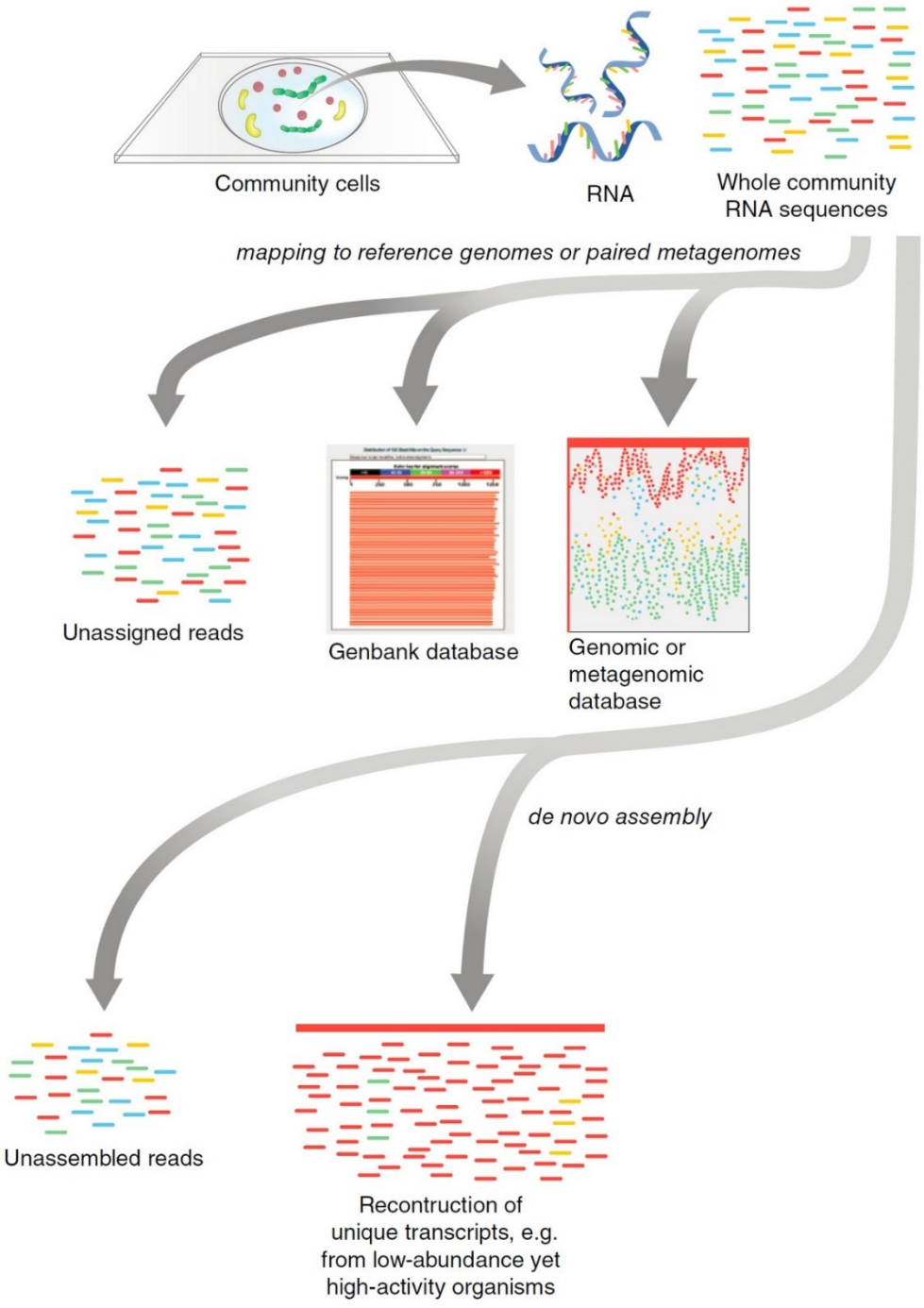


Figure 2. Overview of metatranscriptome analysis. Messenger RNA libraries from complex communities undergo shotgun sequencing to produce read fragments, which can be annotated by alignment to

databases of known genes, genomes or metagenomes. A variety of tools and approaches exist for the assembly, annotation and quantification of transcripts in a metatranscriptome. Adapted from [10].

3.0 Metatranscriptomics of soil

This document is primarily orientated to analyzing gene expression in **soil** communities. However, the information and protocols herein are also applicable to other environments such as seawater, sediments, or **host-associated** communities, including plant- and animal-**microbiomes**. A brief discussion of methods specifically for host-associated RNA processing and analysis will be included.

Soil is a unique and highly complex environment. Because many soil organisms are difficult to culture under typical laboratory conditions, *in situ* molecular characterization is required to understand how they respond to their environment through alteration of gene expression. Enzyme diversity and concentration have a more direct relationship with activity rates in soil than mRNA, and proteins have more stable concentrations in response to external influences, mRNA analysis provides a more immediate picture of the cells' responses to changing environmental conditions [11]. Three major hurdles of mRNA recovery and characterization in soil research are:

i) The high rRNA content relative to mRNA, particularly in cool or low-redox environments such over winter samples or deep soil
ii) The co-extraction of enzyme-inhibiting humic and fulvic acids
iii) The high diversity of soil communities, leading to large computational requirements and a substantial knowledge-base for interpretation of results

3.1 Experimental design

Properly designing an experiment to optimize for the recovery of relevant RNAs is a critical component of metatranscriptomics. First, developing a conceptual model of the conditions needed for the expression of key genes of interest can assist with sample collection strategies. The more that researchers control for variation in environmental conditions, the more likely they are to recover a relevant signal. For example, sequencing results from soil samples removed at different times of day could be affected by diurnal temperature and water-table fluctuations. Or fecal material stabilized immediately collection or several hours later could have greatly different mRNA profiles due to exposure to oxygen and removal from the host environment.

As gene expression patterns often reveal microbial response to changes in substrate metabolism, growth phase, or exposure to environmental conditions, sampling for mRNA extraction should be targeted to capture moments of maximum functional response. This can be done several ways. When adding a substrate to test functional response of a community, e.g., in stool, soil, or sediment, measuring respiration, biomass, cell density, substrate concentration, or transformation products can reveal optimum sampling time – typically periods of maximum growth rate or functional response.

When these types of data are unavailable, or researchers are interested in different phases of metabolism, a sampling time series can ensure capturing the full range of metabolic response.

To ensure adequate recovery of mRNA, preliminary extraction is highly recommended to optimize sample mass, extraction method and post-extraction purification. Several methods for RNA extraction are provide in the following section.

Experimental designs to detect the effects of multiple treatments can be complex. For example, a relatively simple 2^2 factorial experiment will contain four separate treatments (F_1F_1, F_1F_2, F_2F_1, F_2F_2). As most RNA-Seq statistical comparisons rely on differential expression analysis, this experiment will require five pairwise comparisons to resolve. Increasing the number of pairwise comparisons also increases the probability of committing type I error, necessitating false discovery rate (FDR) correction. The number of samples run for each treatment (n-value) can determine the number of differentially-expressed genes detected. Ensuring that sample size per treatment is sufficient to determine biological effects at this scale can benefit from some level of preliminary study to guide power analysis. Balancing sample number, coverage, statistical power, and cost is critical to the success of any metatranscriptomics experiment. Several tools exist for power analysis of RNA-Seq experiments, including:

- **PROPER (https://rdrr.io/bioc/PROPER/)** – R package to guide experimental design of RNA-Seq experiments, including issues of multiple testing, sequencing coverage and differential expression analysis. Uses simulated RNA-Seq data for performing power analysis.
- **Scotty (http://scotty.genetics.utah.edu/)** – Web-based application that uses pilot data or curated datasets to perform power analysis for RNA-Seq experiments.
- **ssizeRNA (https://cran.r-project.org/web/packages/ssizeRNA/index.html)** – An R package that uses the Voom method for RNA-seq data analysis, which was originally developed for microarray power analysis.

3.2 Sampling

Working with soil presents numerous challenges for sampling, transport, handling, purification, and analysis. For soil metatranscriptomics experiments that use *in situ* samples, as opposed to incubations, sampling strategies for RNA work must account for numerous characteristics of mRNA in complex communities. Note, this is not a complete sampling SOP. Instead, the aspects of sampling specific to metatranscriptomics will be addressed.

High-organic soil samples including from peat, permafrost and forest floor are poor candidates for routine, high-throughput extraction due to specific and custom purification requirements. Samples that result in routine extraction of high-quality RNA include forest mineral soil and agricultural soil.

3.2.1 Sample collection

It is imperative to balance collecting as much sample as possible with transportation, freezing and storage limitations. Push-coring is the most commonly used soil sampling method, which generally results in intact soil cores >1 m in length. Multiple cores are required for deep soil collection with this method. Some examples of sampling devices include:

- **Open-face augers** – General purpose soil sampling.
- **Screw augers** – For exceedingly rocky or rooty soil.
- **Russian peat borers** – For deep peat soil.
- **Hydraulic coring** – Requires vehicular access. A powerful tool for deep sampling.
- **Hand-pump** – Efficient method to sample shallow sediments

Sterilizing the corer before each sample is recommended. True sterilization is not realistic or efficient, but cleaning between samples is preferred to reduce cross-contamination. Some push coring devices can accommodate liners, disposable plastic tubes that can be capped to provide self-contained samples. These are recommended when intact soil cores are desired. If soil sub-samples are required, e.g., at different depths, the ubiquitous Axygen 5 ml blue screw top tube is recommended.

3.2.2 Stabilization solution

A variety of products claim to preserve RNA in samples for transport and storage prior to extraction. These solutions often contain pH-stabilization buffers and proteases to denature extracellular RNases. There are several commercial preparations on the market, although most are chemically indistinguishable. Examples include:

- ***RNAlater*** **(Invitrogen, Cat No. AM7021M)** – The market leader in RNA stabilization for good reason, *RNAlater* is proven to maintain RNA concentration and integrity for several hours at 37°C, several weeks at room temperature and indefinitely when frozen, and preserved unfrozen *Synechococcus* and *Pseudomonas* cells with no change in mRNA abundance [12]. However, care must be taken to ensure that preservation does not bias community composition [13]. Further, the ammonium salts in *RNAlater* liberate organic acids in environmental samples, making it a poor choice for preservation of soil and organic sediments. **Recommended sample types:** seawater, plant or animal biomass, stool, cell cultures.
- **RiboReserve (VWR, Cat No. N633)** – Similar to *RNAlater* but at a lower price point, RiboReserve uses pH of 6.4 and sodium citrate as a chelating agent to preserve cellular RNA in solution. **Recommended sample types:** seawater, plant or animal biomass, stool, cell cultures.
- ***LifeGuard*** **Soil Preservation Solution (QIAGEN, Cat No. 12868)** – An alternative stabilization solution for environmental samples, *LifeGuard* guarantees RNA integrity for up to 30 days at -20°C, 1 week at 4°C, and 3 days at room temperature. A major drawback of *LifeGuard* is the high cost. **Recommended sample types:** soil sediment, degraded plant or animal biomass, stool.

3.2.3 Snap-freezing

Recall that the half-life of mRNA is in the range of seconds to minutes [14], which differs between soil species and genes [11–13], with housekeeping genes generally having the most stable mRNA. To account for this, soil must be snap-frozen.

- **Liquid nitrogen (~-196°C)** – The best method of freezing is liquid N_2. The extremely low temperature ensures rapid freezing. Liquid N_2 requires a specialized dewar for handling. Transport of liquid N_2 is difficult, as most commercial transportation methods (ferries, float planes, etc.) forbid it. Therefore this method is only available when sampling soils with access by private vehicles. **Warning: improper liquid N_2 handling can result in severe injury. Always use appropriate protective equipment including safety glasses and insulated gloves.**

- **Dry-shipper (~-196°C)** – To transport frozen samples, liquid N_2-infused dry shippers are recommended. As no liquid remains in the dewar, they can be shipped on airplanes, although airline and customs personnel may not be familiar with dry-shippers and create complications during shipping. Dry-shippers can also be used for sample collection, with a freezing time close to that of liquid N_2. While various sizes of dry-shippers are available, they are often bulky and may not be suitable for use in remote locations.
- **Dry ice (~ –78°C)** – Dry ice is inexpensive, light, and can be found even in rural communities (this is particularly true in coastal British Columbia due to its use in the sport fishing industry). Most Canadian airlines have an allowance for 2.5 kg of dry ice, provided it is declared and properly labeled. Note that flying with water ice is not allowed due to weight shifting as the ice melts. In long-term remote sites, it is possible to make dry ice with a Snowpack dry ice maker and a canister of CO_2. Dry ice sublimates very quickly, so it needs to be refreshed constantly and is not appropriate for long periods of travel. Nonetheless, dry ice is an attractive option for remote sampling, despite the longer freeze time. For safety, always handle dry ice with gloves.

3.2.4 Transport and storage

Sample transport will ideally use a liquid N_2 infused dry shipper or dry ice to maintain **Cold Chain**. Samples should always be stored at ≤-70°C. Short-term storage at -20°C is possible but not recommended.

- **Temperature loggers** are highly recommended during sample transport or shipping. If a sample is being shipped by a client or collaborator, it is recommended that they have received a temperature logger ahead of time. A variety of temperature loggers and probes exist with different functionality and price points. Loggers must be capable of withstanding temperatures ≤-80°C, such as the USB-enabled FlashLink CT -80°C Dry Ice Data Logger. Not that loggers should be placed with the samples and not in direct contact with the ice to ensure accurate readings.

3.2.5 Sample handling

When aliquoting soil for RNA extraction, proceed with the following maxim in mind:

"Work fast, work cold, work clean."

Ensure that all material is either autoclaved (bead tubes, scoopulas, etc.) or wiped with RNase removal solution (balance, gloves, tube racks, benchtop, etc.) prior to handling soil. Gloves and scoopulas should be wiped or changed between samples. Aliquot into tubes on ice and work in a cold-room if possible. Note that this does not apply when aliquoting samples directly into solution containing SDS or CTAB detergents, as these reagents precipitate below 14°C.

All steps in sample collection, storage and handling should be validated by comparing to fresh-extracted RNA to understand and reduce sources of bias.

4.0 Receiving external samples for metatranscriptomics analysis

Metatranscriptomics laboratories should expect to receive two types of samples: raw material for RNA extraction, and pre-extracted RNA.

Sample shipment, receiving and validation are critically important elements in the metatranscriptomics analysis pipeline. If these are not performed with absolute rigor, the remainder of the analysis will be cast in doubt. Therefore, strict operating procedures must be in place to either validate or reject shipped samples.

The World Health Organization reports[2] that most laboratory errors occur in the pre-analytical stage, singling out sample reception as a phase requiring meticulous protocol adherence. It is vital to reduce shipping and receiving errors that communication is shared between clients and supervisory personnel at each stage of the process, including per-shipping.

Whether samples are being received from clients or collaborators, it must be made clear to the parties shipping raw material or extracted nucleic acids what sample amounts, concentrations and purity are expected, what shipping methods and packaging are expected, and what tests will be performed to verify these conditions. It must also be made clear that failure to adhere to these strict standards will result in sample return or disposal. Unlike samples for DNA extraction or DNA-amplicon analysis, there exists no "wiggle-room" for RNA samples.

4.1 Prior to sample shipping

The first point of contact for most clients or collaborators will be sales personnel. Therefore, these personnel must be able to ascertain relevant information including number of samples, sample type for RNA extraction, current sequencing instrument, desired coverage (number of samples per unit of sequencing) and desired of processing and analysis. Ideally, the sales team will have a firm grasp on what is, or is not, feasible so as to guide clients towards the best solutions for their inquiry and to initiate a service request. Each service request will be assigned an identifier that will continue with the associated samples for the duration of contracted work. Therefore, all samples received should correspond to a single quote, and the quote number and service request number should be used to track the sample through validation, handling, and processing.

All sample requests should be entered into the laboratory information management system (LIMS), where technical personnel can verify it prior to a quote being provided. Prior to signing off on a request the technical supervisor will be required to verify:

- The laboratory has the required capacity including to carry out the requested work, e.g., if RNA extraction from a biopsy sample is requested, ensure that valid **standard operating procedures (SOPs)** for this procedure are on file.
- Laboratory personnel trained on the procedures are available.
- All materials, reagents and kits necessary for the requested methods are in stock or can be ordered in a timely manner.

[2] WHO Laboratory Quality Stepwise Implementation Tool (Accessed 2018-06-25): https://extranet.who.int/lqsi/content/develop-sop-sample-reception-and-processing-and-start-registering-all-samples-received

- Necessary permits (if required) are available up to date, e.g., for receiving international soil.

In addition to a quote, all clients should receive a booklet of critical information regarding acceptable shipping guidelines, including sample quantity, RNA concentration and integrity guidelines (if shipping purified samples), package type, labeling, dry ice specifications and shipping methods. For example:

- When handling samples for RNA extraction or purified RNA, always work on ice and change gloves frequently.
- Always suspend RNA in commercially-available RNase free water. Avoid DEPC-treated water or buffers containing detergents (e.g, SDS, CTAB).
- Sample names of tubes, strips or plates much match sample request forms. A paper copy of the sample request form must be included with shipped samples.
- Any shipped material will be disposed of according to laboratory protocols after one month. Prepared sequencing libraries will be stored for one year.
- A minimum of 2000 ng total RNA in a 20-25 µl aliquot (at a concentration of 100ng μl^{-1}) with a RIN score > 6.5 is required for metatranscriptomics analysis.
- A minimum of 1 g stool or 5 g soil or sediment material is required for RNA extraction. Additional samples may be required for specialized sample types.
- Only provide sample aliquots, not sample stocks.

Prior to sample shipping, particularly for large shipments, it is highly recommended that **temperature loggers** be sent to clients or collaborators. Clients agreements should contain information about their responsibilities, including ensuring adequate dry ice for shipping, choosing shipping options that will ensure samples arrive in a timely matter (before dry ice melts), and verifying the correct address for shipping.

The sample shipping address and other required information, e.g., speedchart, purchase order, etc., should be clearly communicated in the pre-shipping document. Further, care should be taken to avoid communicating alternative addresses to avoid confusing clients.

Once the client has received all information relevant for sample shipping, and has had their service request entered into the LIMS, and accepted by technical personnel, the system should provide the clients with automated notice that they are welcome to ship samples.

4.2 Sample receiving

SOPs for sample receiving must contain detailed instructions to quality assurance specialists or laboratory personnel to check the integrity of each arriving sample. Key criteria will include:

- **Correct sample packaging** – Insulated containers with no apparent leaks or damage.
- **Cold chain integrity** – Download results from temperature logger, or simply assure that dry ice has not completely melted from container and that samples remain frozen.
- **Sample details** – The labelling of received samples must match request forms. Sample vessels must match requirements (e.g., tubes or plates must be proper volume and correctly sealed).

Once all of these criteria are met, this information will be entered into the LIMS, which should send automated e-mails to both the client and the manager responsible for the quote number noting that samples have been received and have passed initial screening. Clients will be notified of failing initial screening, or subsequent QA/QC assessment, with a **Sample Rejection Form**. It is easy to understand why commercial laboratories would be hesitant to reject a sample. However, samples that fail QA/QC will likely fail sequencing runs or provide biased results. It is paramount to the reputation of the lab that rejection criteria is followed as strictly as possible. The rejection form will list the reason for rejection and contain a request for a new sample (if possible).

4.2.1 Safely receiving external stool, soil and nucleic acids

All stool, soil, or other sample types should be handled as if it contains infectious material that could be harmful to technical staff and the surrounding ecosystems should it be handled improperly. Cold chain should be maintained at all times. Spillage should be minimized and cleaned according to proper protocols at all times. The site of spilled stool, soil or other environmental material should be decontaminated promptly with 70% ethanol, and spilled material should be autoclaved. In some cases, samples including stool or fine particulate matter will require the technique wear an appropriate ventilator in addition to standard **personal protective equipment (PPE)**, and operate in a biological safety cabinet**.**

Sample integrity is paramount. Freshly sterilized consumables and tools including microtubes and scoopulas must be used for each sample to avoid cross contamination. Gloves should be changed or sterilized between handling samples. While nucleic acids are less likely to contain infectious material, they should still be handled as if they do. Similarly, processes to avoid cross-contamination need to be strictly followed.

4.3 Quality control (QC) and quality assessment (QA)

Prior to initiating requested laboratory work, quality assessment is necessary. For material for RNA extraction, this is straightforward. Is the material still frozen, and is there sufficient material for at least two sets of extractions (RNA extraction will have a higher failure rate than nearly all other methods. This should be accounted for in the requested sample mass). Once total RNA is extracted, or when RNA is received the quality assessment methods discussed in section **5.7 Quality control** will be run:

- **Volume** – Ensure that the sample volume matches the requested amount (20-25 µl).
- **Agarose gel electrophoresis** – Ensure that rRNA bands are clearly visible and intact, ensure genomic DNA removed.
- **Fluorometric concentration** – Ensure that at least 2000 ng is available for downstream use.
- **Bioanalyzer** – Ensure peak integrity, RIN > 6.5
- **NanoDrop** – Ensure that 260:280 and 260:230 ratios are within acceptable parameters

QA should take about 2-4 µl of sample, leaving the remainder for downstream processing. Gel images, Bioanalyzer traces, and RNA concentrations will be uploaded to LIMS. If these pass QA/QC criteria, the client, manager and technical personnel should be notified that the samples have passed QA/QC and will proceed to the next stage of processing. For mRNA sequencing, this will be library preparation (see section **5.9 rRNA removal and library preparation**). It should be made clear to the client if specific QA

criteria are waved (e.g., RIN values or concentrations from environmental samples), then the laboratory cannot guarantee that libraries will produce quality sequencing results.

4.4 Providing sequencing results

Once a prepared sequencing library has passed internal QA/QC, it will be run on the requested sequencing instrument. The results of the sequencing should pass additional QA/QC criteria, chiefly, minimum average quality and number of reads. Sequencing results should then be summarized and communicated. For example, the base size of the sequencing library, the number of reads and the average quality (Q-score). Each demultiplexed samples should be contained within a password-protected download folder, with links in the client-facing LIMS to download, but also assessable for automated downloading (e.g, utilizing *wget* commands) by experienced users. Ideally, this system will be automatically enabled during the sample intake and processing phase by the LIMS.

4.5 Providing analysis results

Many commercial and non-profit sequencing centres exist. What will set a market-leading metatranscriptomics laboratory apart is the ability to carry out and communicate cutting-edge analysis. Initial communication with clients should seek to understand the key questions that they seek to address, and to help guide experimental design (section **3.1 Experimental Design**), coverage requirements and requested bioinformatics.

To communicate results, a standard suite of bioinformatics analysis can be provided in an autonomous or near-autonomous manner using the LIMS. For example, providing beta-diversity ordination, taxonomic composition, abundance of functional orthologs (e.g, KEGG or COG counts) and quality metrics can be provided in tabular and graphical reports. Additional analysis including environmental correlations, co-expression networks and differential expression testing can be included in this reporting at the request of the client. PDFs can provide standardized reporting formats, while web-based reports can allow of interactive graphics and the ability to download publication-ready figures.

5.0 Facility and equipment requirements for RNA laboratory techniques

There is a spectrum of appropriate space and facility requirements for RNA extraction. One extreme would include an RNA-only room separated from other laboratory facilities with a self-contained fume-hood, biological safety cabinet, and freezers. This is impractical and unnecessary. The following are a list of facilities and equipment required for RNA extraction, which highlights the essential elements of conducting RNA research.

5.1 Workspace

- **Soil-specific workspace** – Highly recommended to physically separate aliquoting soil into extraction tubes from remainder of RNA extraction work. Soils contain abundant RNases and the potential for cross-contamination is high.
- **RNA-specific room** – Will help reduce potential crossover contamination, including RNases. Should be regularly and thoroughly cleaned with RNase removal solutions. Very useful to store RNA-specific reagents to increase ease of extraction. Room temperature can be set cooler than ambient laboratory air (14-18°C). Recommended only for facilities processing a high volume of RNA extractions.
- **RNA-specific bench space** – As above, can reduce contamination and spread of RNases, provided thorough and regular cleaning with RNase removal solutions. The main benefit is to increase the efficiency of laboratory personnel by consolidating equipment and reagents into a single space. Ensure separation from areas where genomic DNA preparation and other procedures using RNase are conducted. Recommended for most use cases.
- **Shared bench space** – A general purpose bench space can be used for RNA extraction provided it is thoroughly decontaminated of potential RNases. Not recommended.

5.2 Major equipment

- **Freezer and fridge space** – A tiered freezer system is recommended where reagents are stored at -20°C, separate from incoming samples at -70°C and processed RNA at -70°C. Most -70°C freezers have multiple compartments suitable for separating samples from purified nucleic acids to allow for a single freezer to accommodate both. Care should be taken to physically separate other reagents that require refrigeration at 4°C from bacterial cultures, DNA kits or other potential contaminants.
- **Fume-Hood –** Essential for working with toxic and carcinogenic reagents, e.g., phenol, chloroform. Also useful for chemicals with strong odours, e.g., 2-mercaptoethanol. **The majority of sample lysis and RNA isolation will take place in a fume-hood, so it should be located close to centrifuges, the Fast-Prep instrument, water baths or heat blocks, and other required equipment.**
- **Centrifuges** – Virtually all RNA extraction methods will require a temperature-controlled, high-speed micro-centrifuge, such as the Eppendorf 5417R capable of holding 24 or 30 micro-centrifuge tubes. In addition, a large temperature-controlled, bench-top centrifuge (e.g., Eppendorf 5810R) may be required for RNA extractions in 15 or 30 ml Falcon tubes, or pelleting cultures, wet soil or sediment.
- **Homogenizer –** High-speed cell disruption is highly recommended over vortex-adaptor methods for fast and efficient cell lysis. The MP Biomedicals Fast-Prep 24 is the best choice for this application.
- **Laminar-flow hood** – Non-essential but useful to prevent contamination.
- **Biological safety cabinet** – Essential for working with samples containing potential infectious material such as stool.
- **Ice machine** – Working on ice prevents enzymatic degradation of intracellular and liberated RNA.

- **Autoclave** – For reagent preparation.
- **Fluorometer** – Essential for accurate quantification of RNA concentration. The Invitrogen Qubit system is a fast and accurate option. In general, non-fluorometric quantification methods (e.g., NanoDrop) should be avoided. However, NanoDrop measurement of purity (260:280 and 260:230 wavelength ratios) are useful and required by some facilities.
- **Thermal cycler** – A conventional thermal cycler is required for in-house cDNA library preparation. A quantitative real-time thermal-cycler is required for RT-qPCR. While either conventional or quantitative PCR can be used to verify DNA removal from purified RNA, qPCR is highly recommended for this assay.
- **Agilent BioAnalyzer** – For accurate sizing, quantitation, and purity assessments of nucleic acids. Essential to assess integrity following mRNA enrichment.
- **Gel doc** – To image agarose electrophoresis gels.

Other equipment

- Heat block or water bath
- Vortex
- Micropipettes (1-1000 µl)
- Scoopulas
- Hotplate with stirring capability
- Glassware (25-1000 ml) **Can be autoclaved or treated with DEPC and baked prior to reagent preparation.*
- Tube racks
- Gel molds, power-pack, and electrophoresis cells

Consumables

- Filtered pipette tips (1-1000 µl)
- Syringes and 0.2 µm filters
- 15 and 50 ml Falcon tubes
- 1.7 ml micro-tubes

6.0 Messenger RNA (mRNA) isolation, purification, and enrichment

RNA extraction and purification methods will differ depending on sample type, treatment, and downstream analysis. The following section covers the chemistry and conditions for RNA liquid-liquid extraction, assuming technical ability of laboratory personnel is sufficient to prepare and handle the required solutions. Due to the moderate difficulty in preparing solutions, and the adverse health effects

of several solvents used (e.g., phenol, chloroform), options to avoid liquid-liquid extraction in favour of kit-based RNA isolation, including the use of magnetic beads for purification should be considered.

The first step in choosing the right extraction method for your application is understanding the underlying chemical components used in an RNA extraction, and their relative strengths and weaknesses:

6.1 Solvents

- **Guanidine thiocyanate** – A **Chaotropic** agent used to disrupt hydrogen bonds, therefore, denaturing proteins, including RNases, in solution. Useful for RNA extraction of cell cultures. Not recommended for environmental extraction.
- **2-Mercaptoethanol** – Does double duty by denaturing proteins by reducing disulfide bonds, and protecting RNA by scavenging hydroxyl radicals. Useful but not essential during RNA purification. This thiol has a very strong odour, a fume hood is highly recommended during handling.
- **Phenol** – Used in most liquid-liquid extraction methods as a non-polar solvent that can dissolve and denatures proteins but not highly-polar nucleic acids. Phenol used for extraction is saturated with buffer (e.g., TRIS) to achieve target pH. Phenol (1.07 g cm^{-3}) is slightly denser than water (1.00 g cm^{-3}), creating a biphasic solution with DNA and RNA safely dissolved in the upper aqueous phase, and lipids and proteins in the lower organic phase. Partially denatured proteins may also locate within the interphase. At equilibrium, the aqueous phase will contain about 7% phenol, which must be removed by the addition of an additional solvent (e.g., chloroform). **WARNING: Phenol is highly toxic and corrosive to the eyes, the skin, and the respiratory tract. Only handle with proper training and protective equipment.**
- **Chloroform** – Increases density of organic phase (to 1.47 g cm^{-3}), which improves phase separation during liquid-liquid extraction. The use of chloroform is essential when using a high-salt buffer to account for the increased density of the aqueous phase, and to remove residual phenol. **WARNING: Chloroform is a strong irritant, can be toxic at high doses, and is a potential carcinogen. Handle with proper training and protective equipment.**
- **Isoamyl alcohol** – Primary role in liquid-liquid extraction is as an anti-foaming agent in chloroform. Can denature proteins and further prevent polyadenylated RNA from attaching to the interphase.
- **2-Propanol (Isopropanol)** - Used to precipitate nucleic acids following bulk extraction. Less volume is required than ethanol, but will co-precipitate salts. Best for initial precipitation.
- **Ethanol** – Used to precipitate nucleic acids following bulk extraction due to poor solubility of DNA and RNA in ethanol. Requires addition of a greater volume than aqueous nucleic acid solution. Will retain salts, and is often used as a wash following **2-Propanol** precipitation.

6.2 Detergents

- **Sodium dodecyl sulfate (SDS)** – A synthetic surfactant. Denatures proteins by disrupting non-covalent bonds and solubilizes cell membranes. SDS is anionic and can therefore denature DNA, so is not recommended when extracting DNA and RNA from the same reaction. Note that SDS should not be used with buffers containing potassium (e.g, KH_2PO_4/K_2HPO_4), as SDS will precipitate from solution.
- **Hexadecyltrimethylammonium bromide (CTAB)** – Used for cell lysis and precipitating polysaccharides. Does not denature DNA and can bind and remove polysaccharides, making it useful for RNA and DNA extraction from difficult to lyse gram-positive cells.

6.3 Buffers

- **Tris(hydroxymethyl)aminomethane (TRIS)** – A poly-amine buffer that is toxic to cells. Buffers in pH range useful for biological reactions (pH 7.0 to 9.0; *pKa* = 8.07 at 25°C). However, buffering capacity of TRIS is temperature sensitive, and less effective at low temperatures.
- **Ethylenediaminetetraacetic acid (EDTA)** – Ligand and chelating agent to sequester metals, reduces damage to nucleic acids by metal-dependent enzymes. Frequently used in combination with TRIS.
- **Phosphate Buffer** – Useful for RNA extraction because of the lower pH range (pH 5.8 to 8.0) and the reduced effect of temperature on buffering capacity. Phosphate buffers (either K or Na) maintain cellular osmolality and are non-toxic. Recall that potassium buffers should not be used with SDS. Further, potassium buffers may result in greater liberation of humic substances from soil than sodium buffers.
- **Diethyl pyrocarbonate (DEPC) -** Inactivates RNase enzymes in water and glassware by disrupting covalent bonds of several amino acid residues. Following application of 1% DEPC, solutions must be autoclaved and glassware baked to prevent denaturing of downstream enzymes. DEPC is generally not recommended for most solutions as commercially-available, certified RNase-free water is inexpensive and convenient. As autoclaving is sufficient for irreversible denaturation of most RNases, and DEPC can interfere with downstream treatment and sequencing, there are few reasons to apply DEPC.
- **RNase free water –** Commercially-prepared water that is certified RNase-free is highly recommended for reagent preparation and dissolving purified RNA.

6.4 Conditions

- **pH -** RNA is frequently extracted at pH 4.5, including with acid-saturated phenol tri-solution (e.g., TRIzol). This method takes advantage of the 2' hydroxyl group in RNA, which increase the polarity of the molecule relative to DNA, thus selecting for RNA in the aqueous phase while relegating partially-denatured DNA to the interphase during liquid-liquid extraction. However, RNA recovery is improved with increasing alkalinity up to pH 8.0. Wang et al. [18] suggest using a buffer at pH 7.0 to balance RNA recovery with the liberation of humic acids from soil, which also increases with pH. Since pH does not influence humic acid concentrations as much as other

factors (e.g., temperature), either pH 7.0 or 8.0 result in similar RNA recovery and purity at similar temperatures.

- **Temperature** – While we strive to conduct most of the RNA extraction process at low temperature (either on ice or at 4°C for incubations and centrifugation), initial use of low temperatures will result in the precipitation of SDS or CTAB surfactants. The lowest recommended temperature for the initial steps of RNA extraction is 14°C [18]. Although temperatures up to 65°C can improve lysis of resilient gram-positive cells, temperature greatly increases humic acid liberation in soil. Elevated temperature should be avoided when handling high-organic soils.
- **Lysis method** – When extracting genomic DNA, chemical lysis using detergents, proteases, and lysozyme are used to prevent shearing of nucleic acids. In RNA extraction, these methods are not recommended as they are less efficient than bead beating [19] and require several hours for lysis. Optimal bead beating will use a mix of bead types including 1.4 ceramic beads, 0.1 mm silica beads, and/or one 4 mm glass bead to liberate cells and nucleic acids from soil and lyse cells. Either commercial bead tubes such as MP Biomedicals Lysing Matrix E or in-house bead preparations can be used.
- **Time** – In addition to working clean and cold, working quickly is paramount in RNA extraction. Many RNA researchers use protocols based on a simplified extraction method proposed by Griffiths [20], that combines cell lysis and phenol-chloroform purification in a single step, using a simple, phosphate-based extraction buffer. The simplicity and speed of this method make it highly attractive for high-throughput extraction. It can also co-purify DNA and RNA, making it ideal for applications requiring metagenomic and metatranscriptomic characterization.

6.5 Purification

RNA purification is the separation of total RNA from other molecules in the aqueous extract, including DNA and humic substances (when working with soil). Several options for RNA purification exist.

- **Humic acids** – During the extraction process, use of flocculants such as aluminum amino sulphate ($AlNH_4(SO_4)_2$) can be used to selectively bind and precipitate humic acids. Post-extraction flocculation using purification kits (e.g., QIAGEN RNeasy PowerClean Pro Cleanup Kit; Cat. No. 13997) are highly effective at removing humic substances but diminish total RNA yields. Therefore, optimizing the purification process for each soil type is highly recommended; it is possible to add lower volumes of the PowerClean Pro floculants prior to filtration to streamline the extraction protocol and reduce RNA losses with a single filtration step. Filtration is typically necessary to remove humic and fulvic acids when handling high organic soil. The organic carbon content will determine the type of filtering approach required. Typical soil such as forest mineral layer or agricultural soil with organic carbon < 5% can be purified using spin-columns, while more robust gravimetric purification columns may be required for soil with organic carbon > 5%.

- **DNA** – DNA removal is fairly straight-forward. Several excellent DNase preparations are commercially available, including DNase treatments for on-column (e.g., QIAGEN RNase-Free DNase set; Cat No. 79254) or in solution (e.g., Invitrogen Turbo DNase; Cat No. AM2238) genomic DNA removal. Column DNase treatment is rapid and convenient, but can produce inconsistent reaction conditions. Solution-based reactions require additional inactivation and

removal of the DNase enzyme. Heat-based inactivation can degrade RNA. DNase inactivation solutions can remove the enzyme and Mg+ cations necessary for its activity, and do not appear to affect downstream RNA modification or sequencing. However, for fully purified RNA, additional phenol-chloroform or filtration treatments are required after DNase treatment. As kit-based DNase treatment provides only sufficient Tris-MgCl buffer for low volume reactions, these treatments are typically carried out in post-filtration reactions of 20-60 µl with addition of 10x buffer. As DNase I buffer is relatively easy to make, experimenting with adding "homemade" buffer for DNase treatment of pre-filter volumes (400-1000 µl) may streamline RNA purification and allow for filter-based enzyme inactivation and purification. This approach has yet to be thoroughly tested.

Always read the preparation and handling information for each DNase preparation, as the enzyme can denature easily if handled incorrectly, for example, by vortexing or improper storage temperature. It is important to verify DNA degradation. Quantitative PCR on DNase treated samples using 16S rRNA primers will accurately determine whether all DNA is removed.

6.6 Filtration

As filtering can remove humic acids and DNA, the type of filter used is balance to optimize purification and yield. Two examples are provided here, although several vendors provide additional filtering systems that may provide equivalent results:

- **Spin columns**: For purification of most RNA extractions, centrifugation columns (e.g., QIAGEN RNeasy MinElute Spin Columns) are generally recommended. These small columns fit into 2 ml collection tubes and can therefore withstand centrifugation, making them a fast and efficient method of RNA purification. However, they cannot remove high concentrations of humic acids. Several QIAGEN kits contain these columns and will differ based on application. Choosing the right kit to handle the RNA volume and concentration is essential, and will be discussed in the following section.
- **Gravimetric columns:** Anion exchange columns including Mo Bio RNA Capture Columns (Cat No. 12866-25-SF) are thick, gravity-fed columns to remove organic contamination from RNA preparations. They can be used as part of the RNeasy PowerSoil Total RNA Kit (Cat No. 12866-25) or ordered separately. One benefit of using the entire kit is the application of Mo Bio's Inhibitor Removal Technology (IRT), which is generally excellent at removing organic contamination. The downside is the 15 ml format, which is time-consuming and may result in inefficient RNA recovery.

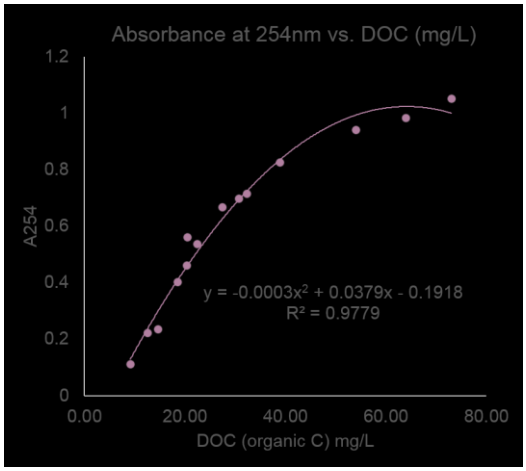


Figure 3. *Estimated DOC content of a peat bog soil by absorbance at 254nm. Soil with A254 > 0.2 can be considered "high-organic."*

QA/QC Step: During the extraction process, or prior to sample processing, the absorbance of buffer-extracted soil organic matter (NaH_2PO_4/Na_2HPO_4; 0.2 M, pH 7.4) can

provide a measure of the expected organic matter co-extraction. Absorbance of 2 µl lysate at 254nm measured on a NanoDrop instrument can be used to estimate dissolved organic matter content (**Figure 3**). A value of A_{254} of 0.2 or above indicates that additional purification of raw lysate using either chemical flocculation (e.g., RNA PowerSoil Pro Clean-up kit) or gravimetric filtration (e.g., RNA PowerSoil capture column) will be required.

6.7 Post-extraction quality control/quality assurance

- **Agarose gel electrophoresis** – Following total RNA extraction and purification it is essential to perform agarose electrophoresis. As single-stranded RNA can conform to a variety of secondary structures, a variety of agarose gel electrophoresis methods have been developed, including temperature or chemical denaturation [21]. These methods are beyond the scope of this document, but denaturation is recommended for publication-quality gel images. For routine diagnostic gel imaging, no changes to electrophoresis protocols for DNA are required, other than ensuring fresh, RNase-free agarose and TAE preparations. Gel imaging is critical to ensure that:
    - i) rRNA peaks are observed,
    - ii) mRNA smears are present,
    - iii) genomic DNA is removed

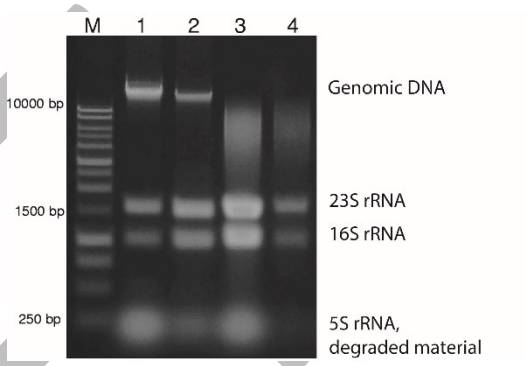


Figure 3. 1% Agarose gel showing genomic DNA and rRNA positions. Smear from around 10,000 bp to 250 bp includes mRNA. Modifed from [22].

- **Fluorometric quantification** – Using the Qubit RNA HS Assay Kit or a similar fluorometric process, accurate RNA quantification is essential. Downstream processing will remove ~95% of total RNA, so ensure that at least 2000 ng per sample has been recovered (e.g., 100 µl sample with Qubit concentration of 20 ng μl^{-1}). However, for samples are only capable of producing ≥100 ng RNA, low-input library preparation is available.
- **NanoDrop spectrophotometry** – NanoDrop does not accurately quantify nucleic acids, particularly at low concentrations. However, since RNA absorbs at 260 nm, purity ratios can be useful in diagnosing potential contaminants:
    - i) 260:280: Pure RNA has a ratio of ~ 2.0. Lower values indicate protein or phenol contamination. Note that acid samples, such as those prepared in acid buffers or even with RNA dissolved in pure water can reduce 260:280 ratios by 0.2-0.3, while alkaline buffers can overestimate purity. Sample dilution can reduce the effect of pH. GC-rich samples can have lower 260:280 as well.
    - ii) 260:230: Some salts and solvents including phenol can absorb at 230 nm. This value is expected to be 2.0-2.2.
- **RT-qPCR** – Removing a small (1-11 µl) aliquot of total RNA for cDNA conversion and qPCR of a common marker gene (e.g., 16S rRNA V4 region) can determine if residual DNA is present in the RNA preparation, and if additional DNase treatment is required.

- **Bioanalyzer** – Digital nucleic acid integrity traces is recommended following total RNA purification. Bioanalyzer produces a useful metric known as RNA Integrity Number (RIN) to asses total RNA integrity by expert categorization of degraded RNA, and by the ratio of Eukaryotic 18S rRNA and 25S rRNA peaks. As the majority of rRNA in soil will be Prokaryotic, which has a lower ratio between large and small rRNA subunits, this number can underestimate RNA integrity of soil RNA. It is recommended that samples proceed with library preparation if RIN > 6.5. However, samples with failed RIN values can result in usable metatranscriptomes but would require authorization by internal supervisors, clients and/or collaborators to proceed following communication that the sample is likely to fail.

6.8 Extraction methods

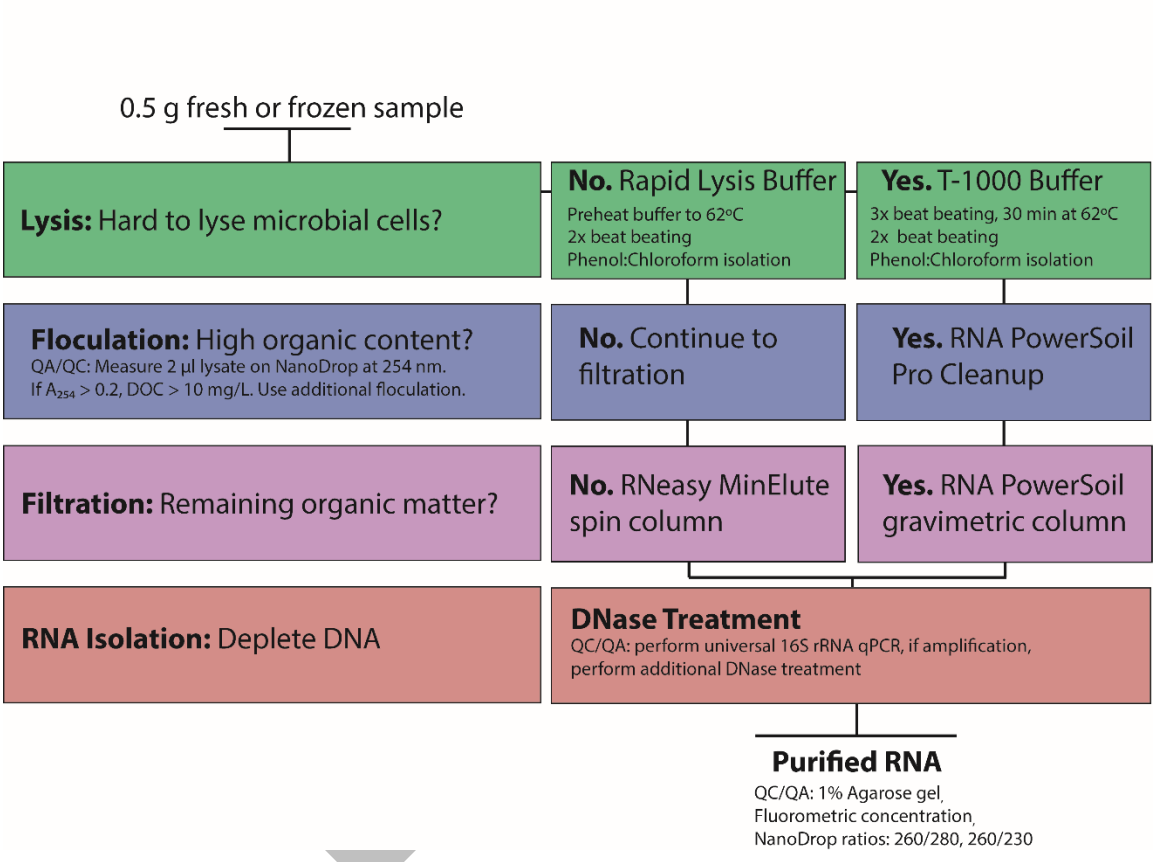


Figure 4. RNA extraction flow chart. Samples for RNA extraction, specifically soil and sediment, require a variety of decisions to be made during their processing. Difficult to lyse and high-organic samples must undergo specific treatments. This suite of methods allows the modular extraction where each step is optimized to produce the correct characteristics (e.g., volume and salt concentration) for the proceeding step.

Taking advantage of the above knowledge, three main extraction methods are recommended for environmental samples. These methods can also be applied to extraction RNA from host-associated tissue samples, stool or cultures:

1) **TRIzol-based RNA extraction from environmental strains and cell culture** (3 hours, see manufacturer's recommended protocol)
2) **Phosphate-buffer-CTAB extraction for RNA extraction and purification from high-organic samples** (8 hours, Appendix A)
3) **Guanidine thiocyanate-SDS (T-1000 lysis buffer) RNA extraction for difficult environmental samples** (6 hours, Appendix B)

The lysis, purification and filtration steps of the above extraction methods are designed to be modular and can be selected based on the properties of the sample. For example, a difficult-to-lyse, high-organic sample can use the lysis method from Protocol 3, with the filtration method from Protocol 2. A relatively routine soil sample will use the rapid lysis buffer described in Protocol 2 with the spin-column based centrifugation method in Protocol 3. **Figure 4** provides a flow-chart of key decisions in the RNA extraction process. Material costs associated with RNA extraction provided in **Appendix C**.

6.9 Alternative RNA extraction methods

The presented RNA extraction methods all use liquid-liquid purification with toxic organic solvents. To improve safety in the lab, and potentially improve consistence, alternative extraction methods should be investigated.

- **Sodium trichloroacetate (NaTCA) method** [22] – NaTCA is the strongest-known chaotrophic agent and can be used to preserve RNA integrity during long periods of bead beating. This method does not use any phenol or chloroform, instead using *N*-lauroylsarcosine (SLS) and bead beating for cell lysis and isopropanol precipitation for purification. Co-extracted proteins and organic matter will likely require additional flocculation and filtering to remove. This method should be tested specifically for samples containing "indestructible" bacterial cells.
- **Magnetic bead method** – While magnetic beads are routinely used to extract DNA, this method has not been widely adopted for RNA due to the high yields and preservation of intact RNA offered by liquid-liquid extraction. Few magnetic kits exist specifically for soil RNA extraction, although Bioclone Inc. and Macherey-Nagel offer kits claiming to purify total RNA from soil using magnetic beads.

6.10 rRNA removal and library preparation

Following total RNA extraction and purification, including removal of residual DNA by DNase treatment, mRNA must be enriched to obtain adequate sequencing coverage. Note that DNase I must be removed with column purification or phenol-chloroform extraction, as it interferes with rRNA removal. In rare cases, sequencing total RNA to produce rRNA and mRNA sequencing pools that can be bioinformatically

separated may be desired. Since mRNA represents 1-5% of the total RNA pool [23], several kits have been produced to remove rRNA for efficient sequencing:

- **Ambion MICROBExpress Bacterial mRNA Enrichment kit (Cat No. AM1905)** – Subtractive hybridization to capture oligonucleotides specific to 16S and 23S rRNAs. Uses magnet beads. Suitable for gram positive and negative bacteria.
- **Epicentre mRNA-ONLY Prokaryotic mRNA Isolation kit (Cat No. MOP51010/MOP51024)** – Exonuclease digestion of processed RNAs with 5' monophosphate including rRNAs.
- **Illumina Ribo-Zero rRNA Removal Kit (Bacteria) (Cat No. MRZB12424**) – Subtractive hybridization to capture oligonucleotides specific to 16S and 23S rRNAs. Uses magnet beads. Suitable for gram positive and negative bacteria.
- **MagJET mRNA Enrichment Kit (Cat No. K2811)** – Hybridization to capture Eukaryote poly(A) mRNA. Uses magnet beads. Suitable for Eukaryotic organisms.
- **NEBNext Poly(A) mRNA Magnetic Isolation Module (Cat No. E7490)** – Hybridization to capture Eukaryote poly(A) mRNA. Uses magnet beads. Suitable for Eukaryotic organisms.

Of these approaches, bacterial rRNA removal using subtractive hybridization is the most effective, removing >80% of rRNA from total RNA from mock bacterial communities [23].

It is recommended to use at least **2 µg total RNA** for rRNA removal, resulting in ~1 ng mRNA for library preparation and sequencing.

Additionally, NEB and Illumina offer library preparation kits that include rRNA removal. This is undoubtedly the most cost-effective method of mRNA enrichment for metatranscriptomics including the recommended:

- **Illumina ScriptSeq Complete Kit (Bacteria) (Cat No. BB1224)**
- **Illumina ScriptSeq Index PCR Primers (required for multiplexed library preparation) (Cat No. RSBC10948)**

This kit is capable of library preparation with as little as **100 ng total RNA** and is suitable for partially-degraded samples, and therefore presents an attractive option when RNA recovery from soil is poor. Index primers are only required when multiplexing metatranscriptomes. Full rRNA removal and library preparation are projected to take **1-2 full days**.

Bioanalyzer analysis is essential following mRNA enrichment, and before library preparation, to ensure mRNA quantity and integrity. Synthesis of cDNA is required for library preparation and is included in modern library preparation kits.

7.0 Sequencing platforms for metatranscriptomics

7.1 Illumina RNA Sequencing

Prior to sequencing, RNA is generally fragmented, size-selected and converted to double-stranded cDNA using single-strand cDNA conversion and second-strand preparation. This process eliminates strandedness; analysis depending on strand selection, e.g., anti-sense RNA characterization, will require

strand labeling during the conversion processes. Depending on the sequencing platform adaptors and barcodes are then ligated to the cDNA fragments during library preparation. Direct RNA sequencing technology is emerging in several forms but is currently beyond the scope of this document.

Cluster generation - To achieve adequate fluoresce signal per molecule, each bound strand must be identically-copied up to 1000 times per cluster using “bridge amplification” by sequential binding to reverse-strand oligos, clonal amplification, then repeating this process on forward-strand oligos. When optimum cluster density is reached, all fragments bound to reverse-strand oligos are removed. This frees these oligos for reverse strand sequencing later in the process The majority of RNA-Seq uses **Illumina** technology, e.g., HiSeq, NextSeq. Detailed information about these platforms is provided in the section on choosing a sequencing platform. In general, Illumina sequencing uses a single flow-cell (NextSeq) or an eight-lane flow-cell (HiSeq) containing lawns of forward- and reverse- strand oligos that match the forward and reverse strand of the prepared cDNA fragment libraries. Adaptor-ligated cDNA libraries are bound to positive-strand oligos on the flow-cell floor for cluster generation and sequencing:

- **Cluster generation** - To achieve adequate fluoresce signal per molecule, each bound strand must be identically-copied up to 1000 times per cluster using “bridge amplification” by sequential binding to reverse-strand oligos, clonal amplification, then repeating this process on forward-strand oligos. When optimum cluster density is reached, all fragments bound to reverse-strand oligos are removed. This frees these oligos for reverse strand sequencing later in the process.
- **Sequencing by synthesis –** Second, Illumina platforms use **sequencing by synthesis (SBS)** technology, which measures fluoresce of labeled deoxynucleoside triphosphate (dNTP) as they are incorporated into the cDNA strand. This method is highly accurate, without problems of differentiating multiple placements of the same dNTP (homopolymers), or GC deletions that cause insertion-deletion (indel) errors other platforms such as 454-pyrosequencing and PacBio RSII, respectively. As a result error rates for Illumina technology are generally the lowest of any platform, primarily substation error around 0.1%, compared to 1% or 16% for 454-pyrosequencing and PacBio RSII, respectively [24].
- **Reverse strand sequencing** – Following the first sequencing pass, the reads are re-amplified once, after which the forward strand is removed. The reverse strand can now be sequenced. Their location on the flow-cell is embedded in the sequencing data, allowing forward and reverse stands to be computationally-paired during analysis. One of the benefits of Illumina sequencing methods is the reverse strand sequencing. This can extend the length of the sequenced region, which is important since one of the main drawbacks of Illumina sequencing is the short insert size. For RNA-Seq this is less of an issue since assembly of short Kmers is efficient and accurate at the transcript level. However, reverse-strand sequencing also improves error rates of the overlapping region by confirming nucleotide calling. Reverse-strand quality quickly deteriorates through the length of the read, so full overlap is recommended during library size selection.

7.1.2 Illumina HiSeq 2500 vs. NextSeq 550

The principle of Illumina sequencing by synthesis is described in detail in **2.2 RNA Sequencing**. In practice, the low price per base and error rates of Illumina sequencing make it the primary platform for

metatranscriptomics. Several Illumina instruments exist, including multiple versions of the NextSeq platform, which utilizes a single flow-cell per run, and the HiSeq, which uses 8 lanes per flow-cell. Here we only consider the dual flow-cell approach. There are several HiSeq instruments including HiSeq2500, HiSeq3000, HiSeq4000 and HiSeqX. The **HiSeq2500** remains the workhorse of shotgun sequencing, but the **NextSeq550** presents a **fast** and **cost-effective** option for small to medium runs. For RNA-Seq short reads (75-100 bp) are recommended over longer reads, as they provide the best price per base, and short reads can be effectively assembled into transcripts. However, short reads can result in chimeric assemblies, particularly in read sets from highly-complex soil samples. The main choice when sequencing, then, is how to choose the best Illumina platform, output mode, and library chemistry to optimize coverage and cost.

Comparison of NextSeq500 and HiSeq2500 Specifications. Options that present the best balance of coverage, quality, read-length, and cost are highlighted in bold. (From Illumina.com)

Instrument	Run Mode	Chemistry	Read-Length	Time	Output	Reads >Q30	Unit
NextSeq 550	High-Output	TruSeq V3	1 × 75 bp	11 hr	25-30 Gb	> 80%	1 Run
			2 × 75 bp	**18 hr**	**50-60 Gb**	**> 80%**	**1 Run**
			2 × 150 bp	29 hr	100-120 Gb	> 75%	1 Run
	Mid-Output	TruSeq V3	**2 × 75 bp**	**15 hr**	**16-19.5 Gb**	**> 80%**	**1 Run**
			2 × 150 bp	26 hr	32.5-39 Gb	> 75%	1 Run
HiSeq2500	High-Output	HiSeq V4	2 × 50 bp	2.5 days	360-400 Gb	> 85%	8 Lanes
			2 × 100 bp	**5 days**	**720-800 Gb**	**> 80%**	**8 Lanes**
			2 × 125 bp	6 days	900 Gb-1 Tb	> 80%	8 Lanes
	High-Output	TruSeq V3	2 × 50 bp	5.5 days	270-300 Gb	> 85%	8 Lanes
			2 × 100 bp	**11 days**	**540-600 Gb**	**> 80%**	**8 Lanes**
	Rapid	HiSeq Rapid V2	2 × 50 bp	16 hr	50-60 Gb	> 85%	8 Lanes
			2 × 100 bp	**27 hr**	**100-120 Gb**	**> 80%**	**8 Lanes**
			2 × 150 bp	40 hr	150-180 Gb	> 80%	8 Lanes
			2 × 250 bp	60 hr	250-300 Gb	> 75%	8 Lanes

7.2 Estimating coverage requirements for metatranscriptome assembly

Estimating Illumina sequencing coverage for host-associated or environmental metatranscriptomes is non-trivial. To obtain ≥95% coverage of a single mammalian transcriptome, about 17.5 Gb of sequencing is required [25]. As bacterial genomes are considerably smaller, simpler and produce fewer transcript variants we can safely assume that less coverage is required for *de novo* assembly of predominately prokaryotic communities. Haas et al. [26] suggest around 100 to 200 Mb per bacterial genome to obtain ≥95% transcriptome coverage, or around 20x to 40x coverage per genome. **Soil contains an average of around 10^4 species** [27]**, which would require between 50 and 100 Gb coverage to accurately *de novo* assemble a near-complete metatranscriptome, assuming an average genome size of 5 Mb** [28]**.**

An experiment with multiple samples multiplexed across two lanes of HiSeq2500 100bp-PE (high-output mode) would achieve this coverage level. This is similar to the amount of coverage required for metagenome assembly [29]. Only a portion of the full metagenome is expressed, but highly-expressed transcripts can have tens of thousands times more coverage than low-abundance transcripts. A human intestinal microbiome can contain up to 1000 genomes [30], which would require about 1 Gb to adequately cover its transcriptome. The actual value may be lower, as host-associated genomes are typically smaller than environmental strains, or higher to account for strain-level variation. Even environmental samples may require less coverage than estimated if the assembly is based on existing metagenome or genomic references, or we are interested in only the most-abundant members of the community or most-expressed genes. However, these coverage estimates may provide a reasonable "rule of thumb."

7.2.1 Software for *de novo* metatranscriptome assembly

There are too many assembly programs to provide a comprehensive list. Many rely on similar principles of Kmer de Bruijn graph construction and were developed for genome or metagenome assembly. Priority will be given to assemblers developed for transcriptomic or metagenomic application. Best assembly results will require throughout quality control and read filtering, described in the next section.

- **Trinity (v2.6.6) [24-25] –** Specifically designed for *de novo* reconstruction of transcriptomes from RNA-Seq data, Trinity relies on three modules for efficient de Bruijn graph creation from 25-mer sequences: *Inchworm* for assembly of unique isoforms, *Chrysalis* for isoform clustering and de Bruijn graph construction, and *Butterfly* for producing polished full-length transcripts from graph data. The low memory footprint, comprehensive documentation, and ease of use make trinity an attractive option for in-house computation. Trinity outperformed other assemblers (Oasis, IDBA-MT) in terms of the number of metatranscriptome reads that could be aligned to a contig with known function, but suffered from higher assembly error [33] https://github.com/trinityrnaseq/trinityrnaseq/wiki
- **Oasis (v0.2.08)** [34] – A dedicated transcriptome assembler built on top of **Velvet (v1.2.1)** [35], An early short read *de novo* assembler using de Bruijn graphs. This program is extremely memory intensive and should be run on distributed computing clusters. https://github.com/dzerbino/oases
- **MetaVelvet (v1.2.01)** [36] – An extension of Velvet assembler to carry out *de novo* metagenome assembly from short sequence reads. Requires a distributed computing cluster for optimum performance. https://metavelvet.dna.bio.keio.ac.jp/
- **IDBA-MT (v1.0)** [37] – An iterative de Bruijn graph de novo short read assembler specifically for metatranscriptomes. Reduces chimeric contigs compared to Trinity or Oasis. https://i.cs.hku.hk/~alse/hkubrg/projects/idba_mt/index.html
- **Trans-ABySS (v2.0.1)** [38] – Developed at the University of British Columbia, this dedicated transcriptome assembler is built on the **ABySS (v2.0)** [39] platform. Also based on de Bruijn graph construction with a default Kmer length of 32, Trans-ABySS can also merge multiple Kmer lengths for a consensus *de novo* transcriptome. http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss

7.2.2 Estimating coverage requirements for short-read annotation and counting

Note that the above coverage estimates only apply when *de novo* metatranscriptome assembly is required. Several metagenomic and metatranscriptomic bioinformatics pipelines annotate and count short reads without assembly [24–26]. For these methods, it is recommended to achieve coverage of only 1-2 Gb per sample [42]. Further, the concept of shallow metagenomic sequencing is emerging as a method for higher sample throughput, with binning and assembly being guided by abundance across samples [43]. Shallow metatranscriptomes may be leveraged to assess taxonomic activity levels across many samples by annotating conserved and constitutively-expressed housekeeping orthologs, but no studies have yet been published validating such an approach.

As part of the initial design phase of an experiment, the required replication, sample number, and sequencing depth should be evaluated using the above guidelines. Previously-described design tools such as *PROPER* (https://rdrr.io/bioc/PROPER/) should be used to estimate coverage requirements for assembly and power required for differential expression analysis.

7.2.3 Assessing assembly quality

A good-quality assembly will have several consistent characteristics regardless of sample type. Assembled transcripts and transcript fragments should be several-hundred to several-thousand base pairs in length, have the majority of input reads map back to the assembled transcripts, and represent the nearly-complete suite of genes expressed by a community in a given environment or treatment. Several software packages exist to assess these elements of assembly quality. Outputs of these tools should be consulted before moving on to read quantification and analysis, or incorporated into automated pipelines for flagging assembly quality level.

- **Read mapping** – *bowtie* or *bowtie2* can be used to map short sequencing reads to the assembled transcripts. Even for a complex metatranscriptome, about 70-80% of input sequences should map back to the assembly as reads pairs.
- **N_{50} and ExN_{50} statistics** – n_{50} is typically used to assess the length of contiguously-assembled sequences (contigs). This value represents the half contig length, or the length of at least half of all transcripts. ExN_{50} is the top most highly expressed transcripts that represent 50% of the total normalized expression data. As such this value excludes low-expression contigs, and may be more useful for transcript data. Both N_{50} and ExN_{50} can be calculated with a variety of tools, including helpful scripts that are included in the *Trinity* assembly software.
- **Conserved orthologs –** Active organisms will constitutively express housekeeping genes. These are often highly conserved within genomes, and as such can be used to determine the completeness of a transcriptomics dataset. Tools such as *BUSCO* (http://gitlab.com/ezlab/busco) can compare expressed genes to a database of near-universal single-copy orthologs.
- **Contig completeness –** Ideally, assembled metatranscriptomes will provide full-length transcripts that can be isolated and characterized. To assess contig length, a simple *blast* search against protein databases such as *SwissProt* can be used to examine the percent of target sequences aligned to assembled contigs.

7.3 Long reads: PacBio vs. Oxford NanoPore

Long-read technology is emerging as a powerful tool to address the main limitation of Illumina sequencing: read length. While high-throughput RNA-Seq rarely leverages these platforms, integration into metatranscriptomics pipelines may greatly improve de novo assembly, particularly when Illumina short reads are combined to error-correct long read scaffolds.

- **PacBio** – Single-molecule, real-time sequencing platforms by Pacific BioSciences (PacBio) produce reads >10,000 bp, which makes them well-suited for genome or transcriptome de novo assembly applications. However, the high error rate (10-15%) is a considerable shortcoming. To address this, PacBio differentiates subreads (induvial reads) from CSS circular reads, which result from a circular template being sequenced multiple times. CSS reads have substantially better quality (~1% error rate), but with far fewer overall reads. PacBio sequencing and assembly mainly relies on proprietary and convoluted file formats and software, making it difficult to integrate into existing pipelines built on open-source tools. The PacBio RSII, roughly two-thirds the size of a Smart car, produces 0.5-1 Gb of sequencing data per SMRT cell using the latest P6-C4 chemistry. The newest PacBio instrument, the Sequel (about half the size of the RSII), produces 10-20 Gb per SMRT cell using Sequel 2.1 chemistry. Both instruments can be run with a variable number of SMRT cells, a flexibility missing from Illumina platforms. They are also faster than Illumina SBS, as real-time sequencing can complete a run in ~30 mins, as it does not require a pause between each nucleotide incorporation. It remains to be seen if PacBio technologies can be efficiently integrated into metatranscriptomic pipelines, but they hold great promise for improving *de novo* transcript assembly.
- **Oxford NanoPore** – In contrast to the monolithic instruments produced by PacBio, the Oxford NanoPore MinIon is roughly the size of a beer bottle, costs $1000 dollars for academic laboratories to begin using, and produces reads averaging 5 kb (up to 900 kb) with 5-10% error rates [44]. High-throughput NanoPore sequencing using the PromethION instrument has been introduced but little data exists about its sequencing output. Due to the footprint of the MinIon, field use is an attractive possibility. Currently, Oxford NanoPore is outperformed by PacBio in terms of quality (compared to CSS reads) and average read length [45]. Both represent opportunities to introduce long read scaffolding into metatranscriptomic workflows. The price and speed of NanoPore sequencing, however, are major factors driving its adoption for producing full-length transcriptomic sequences. **It remains to be seen if NanoPore can be leveraged to produce near-complete, full-length transcripts from complex communities in a high-throughput manner.**

7.3.1 Software for long-read assembly

- **HGAP** – Part of the PacBio SMRT analysis pipeline, hierarchical genome-assembly process (HGAP) can yield single-contig bacterial genomes from a single SMRT cell exceeding 99.999% accuracy [46].HGAP is a pipeline including a PreAssembler, the Celera Assembler and Quiver polishing, run through a Python-based scripting service, SMRTpipe. Installation is complex and requires SMRT Analysis ≥2.3 (https://www.pacb.com/documentation/smrt-analysis-software-installation-v2-3-0/). Recent versions require the SMRT Link server (https://www.pacb.com/support/software-downloads/). Additionally, launching pre-packaged

Amazon instances of SMRT analysis at $0.4 per hour can be a simple and cost-effective method of accessing HGAP assembly.

- **PBcR** – The PacBio Corrected Reads (PBcR) pipeline, can also assemble NanoPore data, and includes options for error correction using Illumina short reads [47]. PBcR ≥8.3 (https://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/wgs-8.3/) can be run from the command line. PBcR is a suite of open-source software, with most components under GNU General Public License or equivalent.
- **SPAdes** – Supports hybrid assembly of long and short reads for scaffolding and error correction, respectively.
- **SSPACE-LongRead** - Novel hybrid assembly method to iteratively scaffold pre-assembled contigs using PacBio RS long reads.

8.0 Bioinformatics pipelines for metatranscriptomics

The processing and analysis of metatranscriptomic data will ultimately depend on the questions to be addressed in a particular study. However, several consistent elements will occur in any pipeline:

- Read filtering and quality control
- Read or contig annotation
- Transcript quantification

Here, the objective is to provide robust, customizable and modular pipeline elements that can be interchanged to produce the desired data formats for downstream analysis. Finished bioinformatics workflows must be able to work with unassembled reads, or reads assembled by a variety of strategies:

- Unassembled short read annotation and quantification
- *De novo* assembly and quantification
- Mapping and quantification by reference genomes
- Mapping and quantification by assembled reference metagenomes

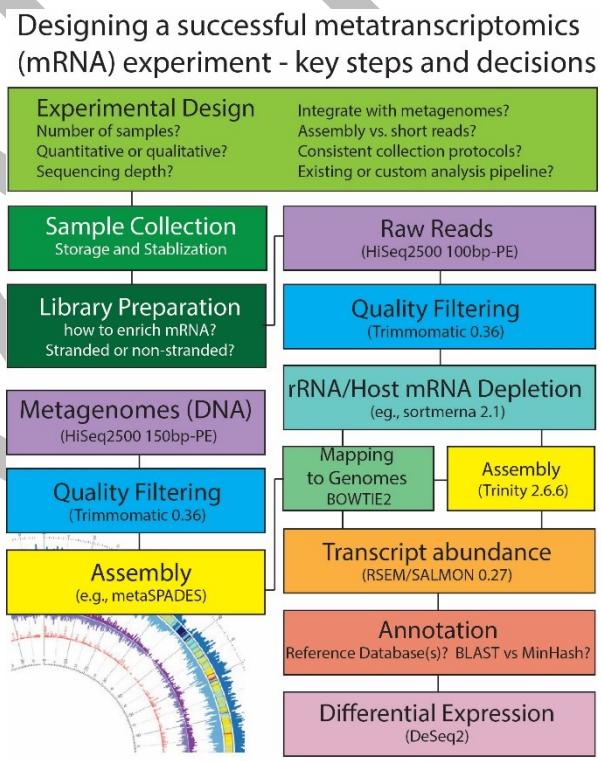


Figure 5. *Example of a metatranscriptomics analysis pipeline for assembled and unassembled reads.*

A review by Lott and co-authors [48] highlights the importance of developing modularized data analysis workflows for metatranscriptomics, and provide a comprehensive list of computational tools to develop metatranscriptomic analysis bioinformatics pipelines. A list of recommend software for metatranscriptomic is additionally provided in **Appendix D**.

Before developing custom analysis workflows, consider existing pipelines, including those originally developed for metagenome applications. The following pipelines exist for metatranscriptome or metagenome short-read annotation and quantification. Metatranscriptome reads can be normalized by metagenome counts when RNA and DNA shotgun sequencing libraries are paired. Many metagenomics or transcriptomics pipelines exist. Here, we examine three that are particularly useful for metatranscriptomics workflows, and one that is better than nothing.

- **MetaTrans (v1.0)** [42] – A short-read clustering, annotation and quantification pipeline for metatranscriptomes. MetaTrans does not require assembled transcripts or reference metagenomes. This pipeline is also capable of rRNA/tRNA separation and taxonomic analysis. However, as virtually all library preparation require rRNA reduction to obtain sufficient mRNA coverage, this functionality will be ignored. One drawback of this pipeline is that reads are only compared between samples based on functional annotation. The steps and software used are as follows:
  i) **Input** – Paired-end ***fastq*** files
  ii) **Quality check** – FastQC (v0.11.7)
  iii) **Quality control –** Kraken (v13.274): **Q** ≥ 10, read length ≥ 30 bp. Note that Kraken can be replaced by running the lightweight and fast alternative Trimmomatic (v0.38) [49] first, then calling *-nqc* when initiating the MetaTrans pipeline. Recommended settings include using a minimum length of 50 bp and a minimum 4-base sliding quality window of Q20.
  iv) **rRNA sorting –** SortMeRNA (v2.1) [50]: Including the PhiX genome in the SortMeRNA database can remove residual spiked reads from metatranscriptomes. **Paired-end joining –** Fastq-Join: Minimum overlap of 8 bp and a maximum difference of 10%.
  v) **Isolate coding mRNAs** - FragGeneScan (v1.17). Converts nucleotide sequences to amino acids.
  vi) **Read clustering –** CD-HIT (v4.6) [51]: identity ≥ 95%, gene overlap ≥90%. Note that clustering selects the longest amino acid chains per cluster, which serves as a rudimentary form of assembly.
  vii) **Functional annotation –** EggNOGv3, MetaHIT-2014 (human-associated microbial genes), and M5nr (MG-RAST) using massively-parallel DIAMOND **blastp** [52]. It is highly recommended to skip this step and perform DIAMOND annotation on clustered reads with up-to-date databases such as RefSeq nr, KEGG, COG, CAZy, etc. Reference databases are covered in **Section 9.3.1**.
  viii) **Read quantification -** DESeq2 [53], based on the number of reads per cluster sharing identical annotation. Allows between-sample comparison. See **Section 9.4** for more detail.

- **MG-RAST (v4.0.3)** [55] – Developed for metagenome annotation, this platform is fairly user-friendly owing to the web-based GUI and processing of samples on dedicated remote servers. A glaring downside is that processing speed is based on how soon the data will be made public. Groups that do not want their data public (e.g., commercial or medical entities) will have to wait several months per sample for their data to work through the MG-RAST servers. It is also made purposely difficult to work with the hash-based M5nr database, unlike standardized *accession* numbers in RefSeq protein and GenBank nucleotide reference databases. MG-RAST takes as input un-assembled short reads from a variety of instruments.

- **COMAN** [54] – The COmprehensive Metatranscriptome ANalysis (COMAN) web-server (http://sbb.hku.hk/COMAN/) offers a similar pipeline as MetaTrans, but with a web-based interface. Read filtering is performed using custom scripts including BLAST to remove reads that match rRNA databases. Unlike MetaTrans, reads are mapped to reference genomes rather than performing BLAST against a comprehensive database such as *RefSeq* or M5nr. Functional annotation is performed using COG and KEGG databases. Co-expression network and pathway analysis are built into the pipeline. The use of reference genomes for annotation suggestions that COMAN may be more suitable for host-associated metatranscriptomes including human microbiomes, rather than environmental libraries. While more robust pipelines such as HUMAnN2 have been developed for microbiome analysis, the metatranscriptome-specific pipeline components and the downstream analysis present are attractive.

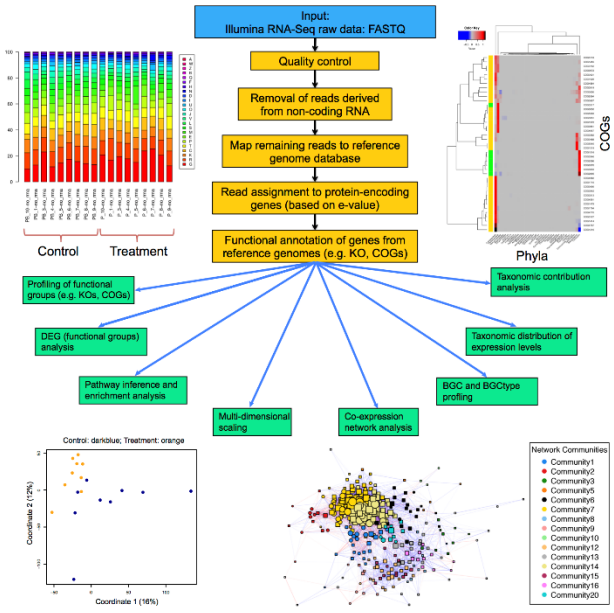


Figure 6. *The COMAN metatranscriptomics workflow. One benifit of this web-based tool is the downstream analysis including expression network*

- **HUMAnN2** – The second-generation Human Microbiome Project Unified Metabolic Analysis Network (HUMAnN2) and its predecessor [40] were developed for high-throughput screening of human microbiome metagenomes and metatranscriptomes [56] for functional profiling through pathway annotation and quantification. The use of highly-curated and relatively complete genomic reference databases for human-associated microorganisms allows high-confidence annotation to species level. Such confidence is lacking for soil and other environmental samples, and the lowest common ancestor (LCA) approach is recommended instead for these sample types. The HUMAnN2 pipeline uses the MetaPhlAn2 [41] and ChocoPhlAn pangenome database for taxonomic profiles, to which unassembled reads are mapped to using the gapped read aligner, *bowtie2*. Unmapped reads are subsequently annotated by DIAMOND blast against the universal protein reference database, UniRef50, and against the MetaCyc pathway database. Normalized gene and pathway abundances are produced. Model-based normalization and statistical testing in programs such as DeSeq2 require non-normalized integer count values, although can work using HUMAnN2 output with slight modification.

- **Anvi'o** [57] **–** A powerful and customizable genomics analysis and visualization platform developed by A. Murat Eren, Anvi'o is capable of a wide range of advanced analysis pipelines suitable for metagenomics and metatranscriptomics workflows, including read assembly, annotation, quantification, binning and visualization. While the full suite of functions available in

Anvi'o is beyond the scope of this document, the recommended metagenomic pipeline maps shorts reads in *bam* files to assembled contigs contained in a *fasta* file, which are annotated using **Hidden Markov Model (HMM)** and **Clusters of Orthologus Groups (COG)** functional profiling. Custom functions can be included in this workflow, and it can accommodate combined analysis of metagenomes and metatranscriptomes. The main drawback of this pipeline is that the high degree of customization available also creates a moderate conceptual barrier to entry, even for advanced bioinformaticians.

8.1 metatranscriptomic bioinformatics pipeline recommendations

In practice, the pipeline of choice will depend on the level and type of the desired assembly, the reference material, and the sample environment:

- For **human microbiome samples**, particularly those with only **moderate number of reads per sample**, **low total coverage**, or **lacking paired metagenomes**, **HUMAnN2** will likely be the best option.
- For similarly **low-coverage**, **short-read environmental samples**, the **MetaTrans** pipeline would be suitable, although it is recommended that this pipeline is customized to annotate reads using reference databases other than the M5nr.
- Metatranscriptomes that assemble successfully or that are paired with assembled metagenomes can take advantage of the existing workflows in **Anvi'o**, including the compelling visualization options.
- MG-RAST should be avoided, except for laboratories lacking any computational infrastructure or the ability to pay for relatively inexpensive server time i.e., **Amazon EC2** instances. **COMAN** may be a useful alternative web server for metatranscriptomics analysis.
- Custom pipelines can be implemented in a modular fashion to address domain-specific needs. Workflow systems such as Galaxy (https://usegalaxy.org/) can help scaffold custom bioinformatics pipelines.

9.0 Analysis of metatranscriptomic data

9.1 Read quantification

An unresolved problem in metatranscriptomics analysis is how can we most accurately reflect true expression rates using RNA-Seq read counts?

- **Eukaryotes** – Alignment is often the first step of read quantification. The standard tool for this is *bowtie* [58] or *bowtie2*, with the latter generally providing superior results. Alignment typically produces a *sequence alignment map (sam)* file, typically converted to a *binary alignment map* (*bam*) that can be sorted and indexed. Creating a *bam* file by aligning quality filtered reads to assembled transcripts allows quantification software such as *Cufflinks* [59] to estimate the expression levels of each transcript. However, this method requires single, delineated transcripts for each gene, typical of Eukaryotic genetics. An additional benefit of this system in

Eukaryotic transcriptomics is that gapped-aligners such as *bowtie2* are capable of splitting reads while mapping to genomic regions containing introns, in addition to mature transcripts.

  - **HTSeq** [60] **–** A Python library to develop custom workflows for analysis of high-throughput sequencing (HTS) data. The *htseq-count* tool preprocesses RNA-Seq data for differential expression analysis by counting the overlap of reads with genes. Can be used for both eukaryotic and prokaryotic transcript analysis and quantification.
  - **featureCounts** [61] – A general-purpose program for counting mapped reads for genomic features such as genes, exons, promoter, gene bodies, genomic bins and chromosomal locations.

- **Prokaryotes –** In prokaryotic systems, single transcripts may be used to express multiple genes organized in operons. Further, terminator sequences can be absent due to horizontal transfer of operons, or alteration of transcription reading frames by insertion sequences. This can also cause transcription through regions on the opposite strand to coding sequences, creating anti-sense RNAs. Several methods exist to account for these differences, although there is currently no leading process of prokaryotic transcriptome or metatranscriptome quantification. The most straightforward solution is to align and map reads to the extracted gene sequences from a genome or transcriptome. This method is leveraged by bacterial transcriptomic tools such as *SPARTA* [62] and *Rockhopper* [63].
  - **FADU** [64] **–** Bacterial genomes and genes are frequently smaller than their eukaryotic counterparts, and multiple bacterial genes can be organized in **operons**. Operons provide coordinate regulation of all genes contained within. However, small genes within operons, particularly those that are antisense to other coding regions, are miscounted by most read counting software. *HTSeq* and *featureCounts* discard fragments that map to multiple genes as ambiguous. By considering the regulation of genes in operons, *FADU* can accurately quantity expression of small, overlapping bacterial genes, and can detect expression of genes missed by programs built for eukaryotic transcriptional models.

9.2.1 Read alignment versus "quasi-mapping"

There are hundreds of RNA-Seq normalization and quantification tools; several are even useful. Here, we compare two handy packages for read quantification that differ based on their use of alignment. Aligning reads, whether to assemblies or reference genomes, is a lengthy and memory-intensive process, but one thought necessary to achieve accurate quantification. Recent advances in "quasi-mapping" suggest otherwise.

- **RNA-Seq by Expectation-Maximization (RSEM) –** Expectation–maximization iteratively finds the maximum likelihood estimate of gene and isoform expression levels from reference-aligned reads. The output is in both TPM and FPKM, which are discussed in the next section.
- **Salmon** [65] – A "wicked-fast" and bias-aware quasi-mapping algorithm. Uses a model-based approach accurately quantify transcript expression, taking about $1/10^{th}$ of the time of alignment-based approaches. However, *Salmon* can also quantify reads using an expectation–maximization algorithm when presented with an alignment file. The output is in TPM and raw read counts, which can be normalized to desired specifications using the output of effective transcript length, also included. Here, *Salmon* is the **obvious recommendation** for quantifying metatranscriptomic expression rates.

One workflow for metatranscriptomics that is conceptually feasible but lacks robust validation is coding sequence prediction from assembled transcripts, followed by *FADU* fragment counting or *Salmon* quantification.

9.2 Normalization of metatranscriptomic feature counts

Whether an existing or customized metatranscriptomics pipeline is used, the data output should be one or more count tables, where the number of each feature (gene, ortholog, pathway, etc.) in every sample is represented by a raw integer value, although various methods of normalization can be applied. The count table is the basis of most downstream analysis methods, which will be discussed in the next section.

First, note that for assembled metatranscriptomes, unlike metagenomes, can produce multiple isoforms of a gene, particularly for Eukaryotic transcripts. It is critical that the sequence form used for analysis is consistent. A general guideline is that analysis, e.g., differential expression, can be performed at the gene level to reduce noise and computational complexity unless specifically analyzing for alternative splicing or other artifacts of Eukaryotic transcription. Multiple count types can be used to represent quantified genes to reduce biases due to library size, gene length, taxon abundance and other sources of variation:

- **Raw counts** – The exact number of time a feature (gene, pathway, taxonomic unit) is identified in the sequencing reads. For assembled metatranscriptomes, the number of reads that map to a specific gene or feature. Suited for model-based statistical comparisons requiring integer values as inputs, but does not account for the effect of gene length read mapping.
- **Reads per kilobase million (RPKM)** – This method normalizes reads to the library size (reads per million) and the length of each gene in kb (reads per kilobase), creating a robust normalization metric that was originally developed for single-ended short-read data. The closely-related **fragments per kilobase million (FPKM)**, accounts for both paired-end reads mapping to genes as a single fragment.
- **Transcripts per million (TPM)** – A simplified read normalization method for modern metatranscriptomics, TPM first normalizes the number of reads mapped to a gene per kilobase of length (reads per kilobase), then uses the sum of these values to normalize library sizes, with an arbitrary denominator of a million reads per kilobase. The benefit of this method versus FPKM is that the sum of TPMs in two samples is always identical, allowing direct comparisons between libraries. The shortcoming is that TPM is as a measure of relative abundance, and therefore eliminates information regarding quasi-absolute quantification.
- **Trimmed mean of M values (TMM)** [66] – A quasi-absolute normalization scheme that reduces gene-length bias and is used to normalize count data for several differential expression analysis tools (e.g., edgeR [67]).

Other normalization schemes exist that have yet to be adopted widely owing to their complexity and lack of systematic validation. When using existing metatranscriptomic pathways decisions about normalization will have been made by the package authors. It is vital to understand what normalization procedures have been applied to your data and how this may affect downstream analysis.

9.3 Annotation

A number of reference databases can be used to determine the function or taxonomic source of short reads or assembled transcripts. It is highly recommended to perform all annotation on amino acid sequences rather than nucleotide sequences. To ensure accuracy of the predicted reading-frame, programs such as *Prodigal* and *FragGeneScan* can be used to predict protein sequences from a nucleotide sequence. Annotation can be performed with the following tools using either direct short reads or assembled transcripts:

- **Basic Local Alignment Search Tool (BLAST)** – BLAST-based heuristic methods can align a query sequence to all sequences in a nucleotide or protein reference database by seeking short, local areas of near-exact matches (typically only a few letters), then extending the alignment until it falls below a cutoff score. Each region is assigned a score based on the length of the aligned region, the number of exact letter matches and the presence of gaps in the alignment. Translated nucleotide sequences can be compared to protein references using the *blastx* function. While this method can search all possible reading frames, this method is time consuming. DIAMOND blast allows for rapid, multi-processor blast-annotation and is the recommended method for annotation of massively-parallel sequencing libraries. Examples of reference databases frequently used with BLAST annotation are NCBI *RefSeq* non-redundant (nr) and GenBank.
- **Hidden Markov models (HMMs)** – In addition to BLAST, a commonly used method for annotating predicted protein sequences is the use of hidden Markov models (HMMs). HMMs present a sensitive methods to examine the probability that a sequence matches the position-specific alignment scores in a profile of a reference protein model. Current HMM annotation tools are approaching the speed of BLAST-based annotation. Examples of HMM-based models for metatranscriptome annotation include *Pfam* (https://pfam.xfam.org/), a database of protein-function profiles. Custom HMMs can be built for the sensitive detection of protein families of interest.
- **MinHash** – Min-wise independent permutations locality sensitive hashing scheme (MinHash) compares Jaccard similarity coefficients between hashes built from query and reference sequences. A *hash* function simply produces a short, consistent-length digital fingerprint of a given string of characters, such as a nucleotide or amino acid sequence. By comparing similarity of hashes, MinHash provides an extremely fast method to compare unknown sequences to a reference library. *Sourmash* (http://sourmash.readthedocs.io) is an example of a MinHash tool specifically designed for annotation of metagenomes against *RefSeq* or *GenBank* databases, or against the full microbial genome library for lowest common ancestor (LCA) taxonomic identification. As this is a developing method of annotation, care should be taken to compare and validate its use against BLAST and HMM methods prior to wide-spread implementation.

9.3.1 Reference databases

Annotation is a very time consumer process. The database of choice will affect the type of annotation (taxonomy, protein family, ortholog group, etc.) and its speed. As indicated in the previous section, a number of databases against which query sequences can be annotated have been released:

- **RefSeq** – The back-bone of any annotation should be the NCBI *RefSeq* database (https://www.ncbi.nlm.nih.gov/refseq/). *RefSeq* is a curated, comprehensive and non-redundant set of sequences from all domains of life. A specific prokaryote *RefSeq* database has also been curated. It contains fewer sequences than *GenBank*, but the sequences that are present are more likely to have good-quality annotations to resolve putative protein function. *RefSeq* can be used with BLAST tools, and a custom *RefSeq MinHash* library has been released for *SourMash*.
- **Uniprot** – The Universal Protein Resource (*UniProt*) (https://www.uniprot.org/) is a comprehensive resource for protein sequence and annotation data, and includes a number of inter-related databases for assessing protein function such as *Swiss-Prot.*
- **Pfam** – A method to annotate sequences using HMM models created from *UniProt* protein families. Can search motifs in protein sequences to reveal functionality missed by *RefSeq* annotation.
- **KEGG** – The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a system for understanding how genes and proteins contribute to metabolic pathways. At its core, the ortholog-based protein identification system can be used with HMM or BLAST annotation to provide consistent annotation of protein function than can be missed by *RefSeq*. While the up-to-date KEGG database requires paid subscription, the online metagenome annotation tool, GhostKOALA (https://www.kegg.jp/ghostkoala/), which uses the BLAST-like GhostX tool to search sequence homology using suffix arrays.
- **MetaCyc** – MetaCyc (https://metacyc.org/) annotates proteins based on conserved metabolic pathways. It can be difficult to use and offers similar functionality as KEGG. MetaCyc is currently offering a paid RNA-Seq interpretation service.
- **COG** – The Clusters of Orthologous Groups (COGs) was designed by comparing predicted and known proteins in all completely sequenced microbial genomes to infer sets of orthologs. Searching conserved orthologs is much faster than searching the full set of protein sequences in *RefSeq*, and counting COGs can offer a simplified view of the functional capacity of a metatranscriptome. COGs are organized in hierarchal categories similar to KEGG orthologs. GO terms are a similar method for classifying ortholog groups.
- **CAZy** – The Carbohydrate-Active enZYmes (CAZy) (http://www.cazy.org/) database is a curated group of protein families involved in carbon transformation and cycling. A number of tools exist to search the CAZy database including the automated *DBcan2* meta server (http://cys.bios.niu.edu/dbCAN2/) that uses BLAST, HMMs, or a combined approach to annotate metagenome or metatranscriptome protein sequences.
- **Custom functional gene databases** – Several databases that have been custom curated to annotate genes of interest have been compiled including the *nifH* database for nitrogen fixation (https://blogs.cornell.edu/buckley/nifh-sequence-database/) and *FunGene* (http://fungene.cme.msu.edu/) that contains databases of many genes of interest including nitrogen cycling and antibiotic resistance genes.

Recommendations: It is advised that *RefSeq* annotation be used as a baseline of protein identification. Currently DIAMOND BLAST is the fastest and most accurate tool for this annotation. However, approaches such as *MinHash* should be investigated. Further, Pfam and KEGG are powerful tools for understanding protein function and metabolic pathways, although the KEGG paid service may prove prohibitive for adoption. The free KEGG web-service prevents full automation but serves as a fast and

accurate functional annotation method. Beyond basic annotation, custom databases such as CAZy and FunGene are highly-useful for domain-specific annotation and are trivial to convert to BLAST databases.

9.4 Differential expression (DE) analysis

Model-based differential expression (DE) analysis is a powerful tool to detect changes in gene expression in metatranscriptomes. Several programs exist to carry out this analysis including *DeSeq2* and *edgeR* in the R programing language, and the stand-alone *Rockhopper* software. This section will focus on *DeSeq2* exclusively, as the *edgeR* and *Rockhopper* algorithms are both based on this method. DeSeq2 seeks to test differential expression by applying negative binomial generalized linear models to raw integer count data. It estimates count dispersion and logarithmic fold changes in this approach. This approach avoids using rarefaction, which can remove rare sequence counts and bias the predicted expression count. However, DeSeq2 can internally correct for library size, which can lead to scaling artifacts when comparing widely different sized sequence libraries. Care must be taken to validate the approach (e.g., using external or internal standards including blanks).

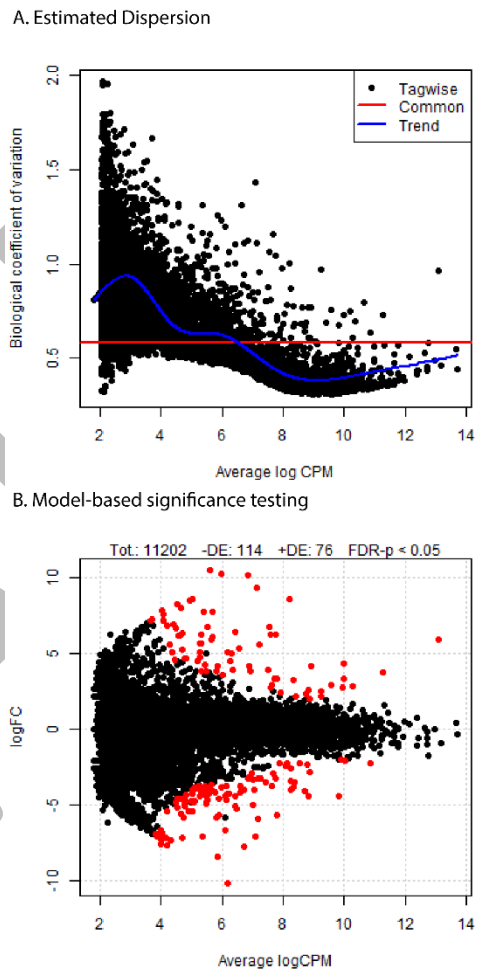


Figure 8.*Negative binomial modeling of dispersion and differential expression. Genes with lower counts have higher thresholds of log2-fold-change to be considered significantly different (red) due to the uncertainty in low-abundance counts.*

DeSeq2 is also designed to input data directly from transcript abundance quantification software profiled in **Sections 9.1** and **9.2.1** including *Salmon* and *RSEM* using the *tximport* package. This presents a method for automating DE analysis directly after read quantification.

DE analysis models the transcript counts in each sample using a negative binomial distribution and calculates gene specific dispersion parameters. This allows the algorithm to calculate the expected "true concentration" of each gene in a sample, and using the dispersion parameter, estimate the variance between the observed gene count and its expected mean. These values are used in a general linear model (GLM) to test the pairwise difference between two treatments using the Wald test for significance of GLM coefficients. In short, DeSeq2 presents a final result of the $\log_2$ fold change of a gene between two treatments, and an estimated p-value of the significance of this change. This p-value is automatically adjusted using the false-discovery-rate (FDR) algorithm to account for the multiple comparisons of all genes in the dataset. Several other analysis methods and algorithms exist within the

DeSeq2 package that are beyond the scope of this document. An example of the pairwise comparison of expressed genes is shown in **Figure 8**.

DE analysis can provide clear evidence for genes involved in the metabolism of specific substrates, metabolic responses during environmental stress or selection of specific active communities between environments or treatments. The normalized read counts can be extracted from the *DeSeq2* object for downstream analysis.

10. Concluding remarks

Metatranscriptomics is a technically-challenging but immensely powerful method to understand microbial activity in complex communities. The primary source of failure in metatranscriptomics is the extraction of sufficient quantities of intact high-quality RNA for library preparation and sequencing. Therefore, considerable care must be taken to develop and validate a reliable and high-throughput method of extraction. This requires a deep understanding of the chemistry and process of RNA extraction, which is why so much space has been devoted to the elements of RNA extraction.

Validating that extracted RNA meets rigorous quality standards is essential to producing reliable RNA sequencing libraries. Rejecting submitted RNA that does not meet these standards can be a difficult call, and every effort should be made to recover sufficient RNA for downstream use. However, attempting to produce sequencing libraries from poorly preserved samples will undermine the metatranscriptomics service, and therefore rejecting samples and requesting fresh material will be essential in these instances.

Clients may also balk at the high costs of metatranscriptomics. And with good reason. However, these costs are primarily material and overhead, and are difficult to avoid. Some suggestions on lowering the cost are included, including switching to novel sequencing technologies such as those offered by Oxford Nanopore. There are few, if any, sequencing centres or microbiome companies offering end-to-end metatranscriptomics, which despite high costs creates the potential to capture market share, and to innovate to develop reliable, high-throughput metatranscriptomics preparation, sequencing, and analysis pipelines.

This report is merely a blueprint. Tough decisions will still have to be made regarding whether to build the capacity to offer metatranscriptomics services, or how best to achieve this goal. I hope this document presents clear information to guide these difficult decisions. Metatranscriptomics has been a powerful tool in my own research, allowing me to understand how active microbial communities are structured by environmental forces, and how these changes affect important metabolic pathways. I see great value in opening up this technology to academic and industry researchers by offering metatranscriptomics-as-a-service.

Cited Literature

[1] J. Brachet, “La detection histochimique et la microdosage des acides pentose-nucleique.,” *Enzymologia*, vol. 10, pp. 87–96, 1941.

[2] J. Hämmerling, “Nucleo-cytoplasmic Relationships in the Development of Acetabularia,” in *International Review of Cytology*, vol. 2, G. H. Bourne and J. F. Danielli, Eds. Academic Press, 1953, pp. 475–498.

[3] J. Brachet, “Action of ribonuclease and ribonucleic acid on living amoebae,” *Nature*, vol. 175, no. 4463, pp. 851–853, May 1955.

[4] F. W. Allen, “The Biochemistry of the Nucleic Acids, Purines, and Pyrimidines,” *Annu. Rev. Biochem.*, vol. 10, no. 1, pp. 221–244, Jun. 1941.

[5] H. M. Temin and S. Mizutani, “RNA-dependent DNA polymerase in virions of Rous sarcoma virus,” *Nature*, vol. 226, no. 5252, pp. 1211–1213, Jun. 1970.

[6] M. K. Thomason and G. Storz, “Bacterial Antisense RNAs: How Many Are There, and What Are They Doing?,” *Annual Review of Genetics*, vol. 44, no. 1, pp. 167–188, 2010.

[7] P. Chomczynski and N. Sacchi, “The single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction: twenty-something years on,” *Nature Protocols*, vol. 1, no. 2, pp. 581–585, Aug. 2006.

[8] J. C. Alwine, D. J. Kemp, and G. R. Stark, “Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.,” *Proc Natl Acad Sci U S A*, vol. 74, no. 12, pp. 5350–5354, Dec. 1977.

[9] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray,” *Science*, vol. 270, no. 5235, pp. 467–470, Oct. 1995.

[10] D. Pérez-Pantoja and J. Tamames, “Prokaryotic Metatranscriptomics,” in *Hydrocarbon and Lipid Microbiology Protocols*, Springer, Berlin, Heidelberg, 2015, pp. 69–98.

[11] A. T. Tveit, T. Urich, and M. M. Svenning, “Metatranscriptomic Analysis of Arctic Peat Soil Microbiota,” *Appl. Environ. Microbiol.*, vol. 80, no. 18, pp. 5761–5772, Sep. 2014.

[12] D. S. Bachoon, F. Chen, and R. E. Hodson, “RNA recovery and detection of mRNA by RT-PCR from preserved prokaryotic samples,” *FEMS Microbiol Lett*, vol. 201, no. 2, pp. 127–132, Jul. 2001.

[13] A. McCarthy, E. Chiang, M. L. Schmidt, and V. J. Denef, “RNA Preservation Agents and Nucleic Acid Extraction Method Bias Perceived Bacterial Community Composition,” *PLOS ONE*, vol. 10, no. 3, p. e0121659, Mar. 2015.

[14] M. P. Deutscher, “Degradation of RNA in bacteria: comparison of mRNA and stable RNA,” *Nucleic Acids Res*, vol. 34, no. 2, pp. 659–666, Jan. 2006.

[15] J. A. Bernstein, A. B. Khodursky, P.-H. Lin, S. Lin-Chao, and S. N. Cohen, “Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays,” *PNAS*, vol. 99, no. 15, pp. 9697–9702, Jul. 2002.

[16] G. Hambraeus, C. von Wachenfeldt, and L. Hederstedt, “Genome-wide survey of mRNA half-lives in<Emphasis Type="Italic"> Bacillus subtilis</Emphasis> identifies extremely stable mRNAs,” *Mol Gen Genomics*, vol. 269, no. 5, pp. 706–714, Aug. 2003.

[17] D. W. Selinger, R. M. Saxena, K. J. Cheung, G. M. Church, and C. Rosenow, “Global RNA Half-Life Analysis in Escherichia coli Reveals Positional Patterns of Transcript Degradation,” *Genome Res.*, vol. 13, no. 2, pp. 216–223, Feb. 2003.

[18] Y. Wang, S. Morimoto, N. Ogawa, T. Oomori, and T. Fujii, “An improved method to extract RNA from soil with efficient removal of humic acids,” *Journal of Applied Microbiology*, vol. 107, no. 4, pp. 1168–1177, 2009.

[19] F. M. Lakay, A. Botha, and B. A. Prior, “Comparative analysis of environmental DNA extraction and purification methods from different humic acid-rich soils,” *Journal of Applied Microbiology*, vol. 102, no. 1, pp. 265–273, 2006.
[20] R. I. Griffiths, A. S. Whiteley, A. G. O’Donnell, and M. J. Bailey, “Rapid Method for Coextraction of DNA and RNA from Natural Environments for Analysis of Ribosomal DNA- and rRNA-Based Microbial Community Composition,” *Appl Environ Microbiol*, vol. 66, no. 12, pp. 5488–5491, Dec. 2000.
[21] T. Masek, V. Vopalensky, P. Suchomelova, and M. Pospisek, “Denaturing RNA electrophoresis in TAE agarose gels,” *Anal. Biochem.*, vol. 336, no. 1, pp. 46–50, Jan. 2005.
[22] S. McIlroy, K. Porter, R. J. Seviour, and D. Tillett, “Simple and Safe Method for Simultaneous Isolation of Microbial RNA and DNA from Problematic Populations,” *Appl. Environ. Microbiol.*, vol. 74, no. 21, pp. 6806–6807, Nov. 2008.
[23] S. He *et al.*, “Validation of two ribosomal RNA removal methods for microbial metatranscriptomics,” *Nature Methods*, vol. 7, no. 10, pp. 807–812, Oct. 2010.
[24] T. C. Glenn, “Field guide to next-generation DNA sequencers,” *Molecular Ecology Resources*, vol. 11, no. 5, pp. 759–769, 2011.
[25] B. J. Blencowe, S. Ahmad, and L. J. Lee, “Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes,” *Genes Dev.*, vol. 23, no. 12, pp. 1379–1386, Jun. 2009.
[26] B. J. Haas, M. Chin, C. Nusbaum, B. W. Birren, and J. Livny, “How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes?,” *BMC Genomics*, vol. 13, p. 734, Dec. 2012.
[27] L. F. Roesch *et al.*, “Pyrosequencing enumerates and contrasts soil microbial diversity,” *ISME J*, vol. 1, no. 4, pp. 283–290, Aug. 2007.
[28] M. Land *et al.*, “Insights from 20 years of bacterial genome sequencing,” *Funct Integr Genomics*, vol. 15, no. 2, pp. 141–161, 2015.
[29] L. M. Rodriguez-R and K. T. Konstantinidis, “Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets,” *Bioinformatics*, vol. 30, no. 5, pp. 629–635, Mar. 2014.
[30] J. Lloyd-Price, G. Abu-Ali, and C. Huttenhower, “The healthy human microbiome,” *Genome Med*, vol. 8, Apr. 2016.
[31] M. G. Grabherr *et al.*, “Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data,” *Nat Biotechnol*, vol. 29, no. 7, pp. 644–652, May 2011.
[32] B. J. Haas *et al.*, “De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity,” *Nat Protoc*, vol. 8, no. 8, Aug. 2013.
[33] A. Celaj, J. Markle, J. Danska, and J. Parkinson, “Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation,” *Microbiome*, vol. 2, p. 39, Oct. 2014.
[34] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney, “Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels,” *Bioinformatics*, vol. 28, no. 8, pp. 1086–1092, Apr. 2012.
[35] D. R. Zerbino and E. Birney, “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs,” *Genome Res.*, vol. 18, no. 5, pp. 821–829, May 2008.
[36] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara, “MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads,” *Nucleic Acids Res.*, vol. 40, no. 20, p. e155, Nov. 2012.
[37] H. C. M. Leung, S.-M. Yiu, J. Parkinson, and F. Y. L. Chin, “IDBA-MT: de novo assembler for metatranscriptomic data generated from next-generation sequencing technology,” *J. Comput. Biol.*, vol. 20, no. 7, pp. 540–550, Jul. 2013.
[38] G. Robertson *et al.*, “*De novo* assembly and analysis of RNA-seq data,” *Nature Methods*, vol. 7, no. 11, pp. 909–912, Nov. 2010.

[39] S. D. Jackman *et al.*, "ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter," *Genome Res.*, vol. 27, no. 5, pp. 768–777, May 2017.
[40] S. Abubucker *et al.*, "Metabolic reconstruction for metagenomic data and its application to the human microbiome," *PLoS Comput. Biol.*, vol. 8, no. 6, p. e1002358, 2012.
[41] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes," *Nature Methods*, vol. 9, no. 8, pp. 811–814, Aug. 2012.
[42] X. Martinez *et al.*, "MetaTrans: an open-source pipeline for metatranscriptomics," *Scientific Reports*, vol. 6, p. srep26447, May 2016.
[43] B. Hillmann *et al.*, "Evaluating the information content of shallow shotgun metagenomics," *bioRxiv*, p. 320986, May 2018.
[44] M. Jain *et al.*, "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nature Biotechnology*, vol. 36, no. 4, pp. 338–345, Apr. 2018.
[45] J. L. Weirather *et al.*, "Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis," *F1000Res*, vol. 6, p. 100, 2017.
[46] C.-S. Chin *et al.*, "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data," *Nature Methods*, vol. 10, no. 6, pp. 563–569, Jun. 2013.
[47] S. Koren *et al.*, "Hybrid error correction and *de novo* assembly of single-molecule sequencing reads," *Nature Biotechnology*, vol. 30, no. 7, pp. 693–700, Jul. 2012.
[48] S. C. Lott *et al.*, "Customized workflow development and data modularization concepts for RNA-Sequencing and metatranscriptome experiments," *Journal of Biotechnology*, vol. 261, pp. 85–96, Nov. 2017.
[49] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014.
[50] E. Kopylova, L. Noé, and H. Touzet, "SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data," *Bioinformatics*, vol. 28, no. 24, pp. 3211–3217, Dec. 2012.
[51] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006.
[52] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nat Meth*, vol. 12, no. 1, pp. 59–60, Jan. 2015.
[53] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, p. 550, Dec. 2014.
[54] Y. Ni, J. Li, and G. Panagiotou, "COMAN: a web server for comprehensive metatranscriptomics analysis," *BMC Genomics*, vol. 17, p. 622, Aug. 2016.
[55] F. Meyer *et al.*, "The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC Bioinformatics*, vol. 9, p. 386, Sep. 2008.
[56] E. A. Franzosa *et al.*, "Relating the metatranscriptome and metagenome of the human gut," *Proc Natl Acad Sci U S A*, vol. 111, no. 22, pp. E2329–E2338, Jun. 2014.
[57] A. M. Eren *et al.*, "Anvi'o: an advanced analysis and visualization platform for 'omics data," *PeerJ*, vol. 3, p. e1319, Oct. 2015.
[58] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.
[59] C. Trapnell *et al.*, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nature Protocols*, vol. 7, no. 3, pp. 562–578, Mar. 2012.
[60] S. Anders, P. T. Pyl, and W. Huber, "HTSeq—a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015.
[61] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–930, Apr. 2014.

[62] B. K. Johnson, M. B. Scholz, T. K. Teal, and R. B. Abramovitch, “SPARTA: Simple Program for Automated reference-based bacterial RNA-seq Transcriptome Analysis,” *BMC Bioinformatics*, vol. 17, p. 66, Feb. 2016.
[63] R. McClure *et al.*, “Computational analysis of bacterial RNA-Seq data,” *Nucleic Acids Res.*, vol. 41, no. 14, p. e140, Aug. 2013.
[64] M. Chung *et al.*, “FADU: A Feature Counting Tool for Prokaryotic RNA-Seq Analysis,” *bioRxiv*, p. 337600, Jun. 2018.
[65] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression,” *Nature Methods*, vol. 14, no. 4, pp. 417–419, Apr. 2017.
[66] M. D. Robinson and A. Oshlack, “A scaling normalization method for differential expression analysis of RNA-seq data,” *Genome Biol*, vol. 11, no. 3, p. R25, 2010.
[67] Chen Y, McCarthy D, Richie M, Robinson M, Smyth GK, “edgeR: differential expression analysis of digital gene expression data.,” 2017.
[68] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.
[69] B. Giardine *et al.*, “Galaxy: A platform for interactive large-scale genome analysis,” *Genome Res*, vol. 15, no. 10, pp. 1451–1455, Oct. 2005.
[70] D. E. Wood and S. L. Salzberg, “Kraken: ultrafast metagenomic sequence classification using exact alignments,” *Genome Biol*, vol. 15, no. 3, p. R46, 2014.
[71] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, “Prodigal: prokaryotic gene recognition and translation initiation site identification,” *BMC Bioinformatics*, vol. 11, p. 119, Mar. 2010.
[72] T. Seemann, “Prokka: rapid prokaryotic genome annotation,” *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014.
[73] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, “VSEARCH: a versatile open source tool for metagenomics,” *PeerJ*, vol. 4, p. e2584, 2016.
[74] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nature Methods*, vol. 9, no. 4, pp. 357–359, Apr. 2012.
[75] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.
[76] B. Li and C. N. Dewey, “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome,” *BMC Bioinformatics*, vol. 12, p. 323, Aug. 2011.
[77] B. D. Ondov, T. J. Treangen, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy, “Fast genome and metagenome distance estimation using MinHash,” *bioRxiv*, p. 029827, Oct. 2015.
[78] K. D. Pruitt, T. Tatusova, and D. R. Maglott, “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins,” *Nucleic Acids Res*, vol. 35, no. Database issue, pp. D61–D65, Jan. 2007.
[79] A. Bateman *et al.*, “The Pfam Protein Families Database,” *Nucleic Acids Res*, vol. 30, no. 1, pp. 276–280, Jan. 2002.
[80] S. R. Eddy, “Accelerated Profile HMM Searches,” *PLOS Computational Biology*, vol. 7, no. 10, p. e1002195, Oct. 2011.

Appendix A. Rapid phosphate-buffer-CTAB extraction for routine RNA extraction from environmental samples

D. Levy-Booth, 2015.

Adapted from: A. Hahn, 2012; K.M. DeAngelis' 2008; E. Brodie, 2003; R. Griffiths et al., 2000

Materials:

For DNA/RNA extraction

- 1.5 ml eppendorf tubes
- Filter tips
- Lysing matrix E glass/zirconia/silica beads Tubes (MP Biomedicals, Cat# 116914100)
- 10% CTAB in 1M NaCl
- 0.24 M phosphate buffer, pH 8
- 0.2 M Aluminum amino sulphate ($AlNH_4(SO_4)_2$)
- Phenol:chloroform:isoamylalcohol (25:24:1)
- Chloroform:isoamylalcohol (24:1)
- Phase-Lock gel (Heavy) tube (Phase Lock Gel Tube 2 ml, Heavy, Fisher Cat# 2302830)
- 15 ml falcon tubes
- RNase-Free H_2O
- RNase inhibitor
- RNase Away or equivalent for cleaning
- 2-Propanol (Isopropyl alcohol)
- 75% ethanol (ice cold)
- MoBio PowerSoil RNA capture columns (12866-25-SF), solution S5 (12866-25-5) and solution S6 (12866-25-6).

CLEAN CLEAN AND THEN RNA CLEAN

Preparation:

1. All vessels and reagents should be DNAse/RNAse free, or treated with DEPC to denature proteins. Bake glassware at 180°C for 3 hours before making reagents.
2. Cool centrifuge to 4°C.
3. Prepare 75% ethanol fresh and store at -20°C.
4. Centrifuge Phase-Lock gel (Heavy) tube for 30 s to concentrate gel at bottom of tube.
5. Heat water-bath to 42°C.

Method:

To extract DNA/RNA: <u>work on ice!</u>

1. Soil should be frozen on liquid N_2, or at -80°C. Add 0.5g still-frozen soil into bead tubes, placing tubes on ice.

2. Quickly add 0.350 ml of pH 8.0 phosphate buffer, 0.15 ml CTAB extraction buffer and 50 ul $AlNH_4(SO_4)_2$. Vortex immediately to mix. Add 0.5 ml phenol:chloroform:isoamylalcohol (25:24:1).

3. Shake tubes in the FastPrep at 6 m/s for 30 s. Place on ice for 5 s, then centrifuge at 16000 x g for 10 min at 4°C. Save bead tubes on ice.

4. Remove the aqueous top layer to a pre-spun Phase-Lock gel (Heavy) tube and add an equal volume (0.5-0.8 ml) of chloroform:isoamylalcohol (24:1). Mix tubes well and place on ice.

5. Repeat steps two more times:
    a. Add additional 0.350 ml of pH 8.0 phosphate buffer and 0.15 ml CTAB extraction buffer to used bead tubes containing soil.
    b. Shake tubes in the FastPrep at 6 m/s for 30 s. Place on ice for 5 s, then centrifuge at 16000 x g for 10 min at 4°C.
    c. Remove the aqueous top layer to a pre-spun Phase-Lock gel (Heavy) tube and add an equal volume (0.5-0.8 ml) of chloroform:isoamylalcohol (24:1). Mix tubes well and place on ice.
    d. The second and third extractions can be placed into the same tube. Use less chloroform:isoamylalcohol (24:1) in the second gel tube if necessary for equivalent volumes.

6. Centrifuge all Phase-Lock gel (Heavy) tubes at 16000 x g for 10 min at 4°C.

7. Combine extracts from the same soil sample in the two Phase-Lock gel (Heavy) tubes into 15 ml falcon tubes by removing the aqueous top layer from the Phase-Lock gel (Heavy) tube and moving it into a labeled 15 ml falcon tube. Add 5 ml 2-propanol. Incubate for 30 min at room temperature. Can be incubated longer if necessary.

8. Spin at 2500 g at for 30 min at 4°C. Remove as much 2-propanol as possible by pouring it out then pipetting. If there is any oily organic-phase carry over (there shouldn't be) carefully pipette to remove. Beware that the RNA pellet can sit on the surface of the oily residue.

9. Pellets from high-organic soil will be large and dark due to humic co-extraction. Wash with ice-cold 75% ethanol (add ethanol to pellet, mix gently and centrifuge 10 min at 4°C. Remove ethanol and dry pellet upside down on paper towel or upright in biological cabinet for about 5 minutes (do NOT totally dry).

10. The remaining protocol continues from MoBio RNA PowerSoil protocol step 12. Re-suspend pellets in 1 ml Solution SR5 from MoBio RNA PowerSoil Kit. Suspension may be aided by heating

for 10 min at 42°C in a water-bath and mixing by pipetting.

11. Prepare one RNA Capture Column for each RNA Isolation Sample: Remove the cap of a 15 ml Falcon tube and place the RNA Capture Column inside the tube. Add 2 ml of Solution SR5 to the RNA Capture Column and allow it to gravity flow through the column and collect in the15 ml Collection Tube. Allow Solution SR5 to completely flow through the column. Do not let the column dry out.

12. Add the 1 ml RNA solution to RNA Capture Column and allow it to gravity flow through the column. Add an additional 1 ml of Solution SR5 to the column to wash RNA.

13. Transfer the RNA Capture Column to a new 15 ml Falcon tube. Shake Solution SR6 to mix and then add 1 ml of SolutionSR6 to the RNA Capture Column to elute the bound RNA into the 15 ml Falcon tube. Allow Solution SR6 to gravity flow into the 15 ml Falcon tube.

14. Transfer the eluted RNA to a 2.2ml microtube and add 1 ml of 2-Propanol. Invert at least once to mix and incubate at -20°C overnight or at -80°C for at least 30 minutes.

15. Centrifuge the 2.2 ml microtube at 16,000 g for 30 minutes at 4°C to pellet the RNA.

16. Decant the supernatant and invert the 2.2 ml microtube on a paper towel or place upright and open in biological cabinet for 10 minutes to air dry the pellet. Do not over-dry pellet.

17. Resuspend the RNA pellet in 60 µl of RNase-Free H_2O. Quantify RNA and DNA concentrations using QUBIT, and *DNase treat if necessary*.

18. Add RNase inhibitor (1/40 volume) to final solution.

Storage and down-stream use:

1. To store the RNA long-term, add three volumes (150 ul) of 100% ethanol and 2 ul of NaOAc to each aliquot. All samples should be stored at -80°C.

2. To use stored sample: spin sample at 10, 000 rpm for 30 minutes at 4°C. Wash with 500 ul ethanol for 15 minutes. Spin sample at 10, 000 rpm for 30 minutes again to remove residual ethanol. Air dry pellet and resuspend sample in RNAse free water (pipette up and down).

Appendix B. Guanidine thiocyanate-SDS RNA extraction for difficult environmental samples

D. Levy-Booth, 2018.

Adapted from: Y. Wang et al., 2008, 2009; R. Hurt et al., 2001

Materials:

- 1.5 ml eppendorf tubes
- Filter tips
- Lysing matrix E glass/zirconia/silica beads Tubes (MP Biomedicals, Cat# 116914100)
- Phenol:chloroform:isoamylalcohol (25:24:1)
- Chloroform:isoamylalcohol (24:1)
- RNase-Free H_2O
- RNase inhibitor
- RNase Away or equivalent for cleaning
- **Phosphate buffer**: Na_2HPO_4/NaH_2PO_4 (240 mM, pH 8.0)
- **Guanidine solution:** 4M guanidine isothiocyanate in 10 mM TRIS-HCl and 1 mM EDTA. Add 0.5% fresh 2-mercaptoethanol prior to use.
- **QIAGEN RNeasy MinElute Cleanup Kit (Cat No. 74202)**
- **QIAGEN RNeasy PowerClean Pro Cleanup Kit (Cat No. 13997) (Optional)**

Preparation:

1. Prepare 80% ethanol fresh and store at room temperature.
2. Heat water-bath to 62°C.

Method:

1. Using autoclaved scoopulas, aseptically transfer 0.5 g sample to bead tubes containing Lysing Matrix E, 0.9 ml phosphate buffer, 0.1 ml 20% SDS, and 0.1 ml guanidine solution.
2. Bead beat using Fast-Prep at 6.0 m/s for 30 s (3x) cooling on ice between runs. Incubate at 62°C for 30 minutes, vortexing to mix every 10 mins. Repeat bead beating 2x.
3. Remove supernatant to new 2ml micro-centrifuge tube. Add 0.5 ml phenol:chloroform, vortex for 5 s, and centrifuge for 5 mins at 16,000 g and 4°C.
4. Remove supernatant to new 2 ml micro-centrifuge tube. Add 0.5 ml phenol:chloroform:isoamyl alcohol, vortex for 5s, and centrifuge for 5 mins at 16,000 g and 4°C.
5. Remove supernatant to new 2 ml micro-centrifuge tube.
6. ***Optional***: for high-organic samples: Add 20 to 50 µl RNeasy PowerClean Pro solution CU and vortex. Add 20 to 50 µl RNeasy PowerClean Pro solution IR and vortex. CU and IR volume can be adjusted based on optical purity of post-purified supernatant. Centrifuge for 2 min at 16,000 g and remove supernatant to new 2 ml micro-centrifuge tube.
7. ***Optional***: On-column DNase treatment can be applied at this step.
8. Add 250 µl of 96-100% ethanol to supernatant, mix by pipetting without vortexing. Transfer 700 µl

of sample to RNeasy MinElute spin column in a 2 ml collection tube. Centrifuge 30 s at 10,000 g at room temperature. Discard flow-through. Add any remaining sample up to 700 µl and repeat.

9. Add 500 µl 80% ethanol to column. Centrifuge 2 min at 10,000 g at room temperature to wash the spin column. Discard flow-through.
10. Open the lid of the column, centrifuge at full speed for 5 mins to dry the membrane. Discard flow through and place column in a new 1.5 ml collection tube.
11. Add 30 µl RNase-free water directly to the center of the spin column. Close lid and centrifuge for 1 min at full speed to elute RNA.
12. **Pause point:** RNA can be stored with 1 µl RNase inhibitor, or proceed with separate DNase treatment (e.g., Turbo DNase).

Appendix C. Material costs associated with RNA extraction

Item	Cat. No.	Cost (CAD)	Per 0.5 g sample
RNA Extraction			
Phenol-Chloroform	ThermoFisher 15593031	$204 / 400 ml	$0.59
Chloroform	Sigma C0549	$137 / 1000 ml	$0.07
Lysing matrix E	MP Biomedical 116914050	$369 / 100	$3.69
RNeasy MinElute	QIAGEN 74204	$412 / 50	$8.24
Turbo DNase	ThermoFisher AM1907	$149 / 50	$2.98
Qubit RNA HS	ThermoFisher Q32852	$390 / 500	$0.78
		Total:	**$16.35**
Library Preparation			
ScriptSeq Complete Kit (Bacteria)	Illumina BB1224	$4,319 / 24	$179.96
ScriptSeq Index PCR Primers (Set 1)	Illumina RSBC10948	$368 / 48	$7.67
		Total:	**$187.63**

Appendix D. List of recommended bioinformatics software for metatranscriptomics analysis

Software	Description	License	URL
General purpose platforms			
R [68]	Statistical computing	GPL2[1]	https://www.r-project.org/
Python	Programing language	GPL-ish[2]	https://www.python.org/
Bioconda	bioinformatics management system	GPL3/ MIT[3]	https://bioconda.github.io/
Galaxy [69]	Bioinformatics workflow system	AFL3[4]	https://usegalaxy.org/
Read filtering			
Trimmomatic [49]	Trimming and QC	GPL	https://github.com/timflutre/trimmomatic
SortMeRNA [50]	rRNA/host read filtering	GPL	http://bioinfo.lifl.fr/RNA/sortmerna/
Kraken [70]	Taxonomic classifier	GPL	https://ccb.jhu.edu/software/kraken/
Read processing			
Prodigal [71]	CDS prediction	GPL3	https://github.com/hyattpd/Prodigal
Prokka [72]	CDS prediction and annotation	GPL3	http://www.vicbioinformatics.com/software.prokka.shtml
CD-HIT [51]	Read clustering	GPL	http://www.bioinformatics.org/cd-hit/
VSEARCH [73]	Read processing and paired joining	GPL3	https://github.com/torognes/vsearch
Transcript assembly			
Trinity [31]	Transcriptome assembler	AFL3	https://github.com/trinityrnaseq/trinityrnaseq
Trans-ABySS [38]	Transcriptome assembler	GPL3	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss

Alignment and quantification			
Bowtie2 [74]	Gapped aligner	GPL3	https://github.com/BenLangmead/bowtie2
Samtools [75]	Read aligner and mapping	MIT	http://www.htslib.org/
Salmon [65]	Read quantification	GPL3	https://combine-lab.github.io/salmon/
RSEM [76]	Read alignment and counting	GPL3	http://deweylab.github.io/RSEM/
Cufflinks [59]	Gapped read quantification	BSL1[5]	http://cole-trapnell-lab.github.io/cufflinks/
DeSeq2 [53]	Differential expression	GPL3	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
Annotation			
DIAMOND [52]	Local alignment and annotation	GPL3	https://github.com/bbuchfink/diamond
SourMash [77]	MinHash annotation	BSD3[6]	http://sourmash.readthedocs.io/en/latest/
RefSeq [78]	Curated reference sequence library	Public domain	https://www.ncbi.nlm.nih.gov/refseq/
Pfam [79]	Protein family HMM library	CC0[7]	http://pfam.xfam.org/
HMMER [80]	HMM profile search	BSD3	http://hmmer.org/

Licenses:

1. GPL, Gnu Public License, https://www.gnu.org/licenses/gpl-3.0.en.html
2. Python GPL-compatible License, https://docs.python.org/3/license.html
3. MIT, Massachusetts Institute of Technology, https://opensource.org/licenses/MIT
4. AFL, Academic Free License, https://opensource.org/licenses/AFL-3.0
5. BSL, Boost Software License, https://www.boost.org/LICENSE_1_0.txt
6. BSD, Berkeley Software Distribution, https://opensource.org/licenses/BSD-3-Clause
7. CC0, Creative Commons zero, https://creativecommons.org/publicdomain/zero/1.0/