CellPress
OPEN ACCESS

Trends in
Genetics

Review

Current concepts, advances, and challenges in deciphering the human microbiota with metatranscriptomics

Teija Ojala,[1] Aino-Elina Häkkinen,[2] Esko Kankuri,[1] and Matti Kankainen[2,3,*]

Metatranscriptomics refers to the analysis of the collective microbial transcriptome of a sample. Its increased utilization for the characterization of human-associated microbial communities has enabled the discovery of many disease-state related microbial activities. Here, we review the principles of metatranscriptomics-based analysis of human-associated microbial samples. We describe strengths and weaknesses of popular sample preparation, sequencing, and bioinformatics approaches and summarize strategies for their use. We then discuss how human-associated microbial communities have recently been examined and how their characterization may change. We conclude that metatranscriptomics insights into human microbiotas under health and disease have not only expanded our knowledge on human health, but also opened avenues for rational antimicrobial drug use and disease management.

Highlights

Metatranscriptomics has emerged as an effective methodology to characterize the human microbiome and evaluate its transcriptional activity in a culture-independent manner.

Analysis of human-associated microbial communities varies depending on the tissue type and its physiological features and no single best method exists that suits all situations.

An integrated metagenomics and metatranscriptomics approach can be used to distinguish transcriptome alternations emerging from DNA abundance changes from those driven by transcription.

The CRISPR/Cas9 system can be used to effectively deplete unwanted rRNA sequences and is an enticing alternative for the commonly used subtractive hybridization approaches.

Third-generation sequencing technologies can capture full-length transcripts and represent an intriguing alternative for current sequencers, although have not yet been applied to human-associated microbiotas.

Metatranscriptomics is a powerful tool for studying human microbiota

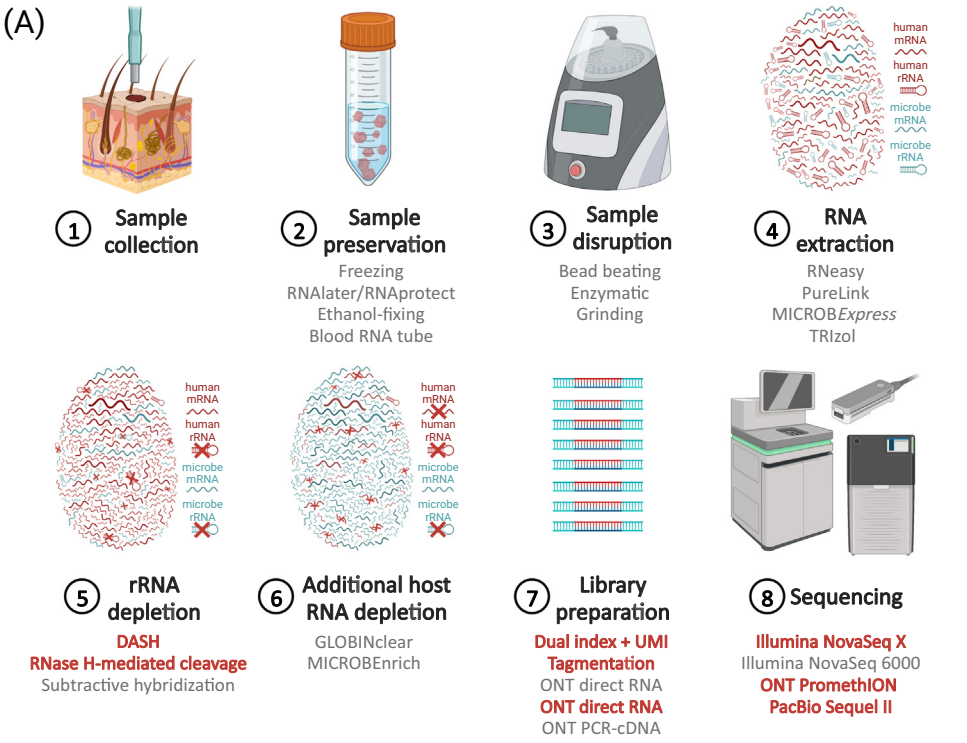
Metatranscriptomics (see Glossary) has emerged as a key technique to study gene expression in microbial communities. It circumvents the need for cultivation by relying on direct analysis of the community RNA and can be used to decipher active community members as well as their expressed genes and associated functions [1–3]. Thus, metatranscriptomics reveals biological information not readily available from conventional genomic profiling methods and complements **metataxonomic** (marker gene amplicon sequencing) and **metagenomic** (shotgun DNA sequencing) assessments that usually do not distinguish between active, inactive, and dead community members.

In the context of human microbiota research, metatranscriptomics has frequently been applied to intestinal and vaginal communities, among others [2,4]. For instance, analysis of intestinal communities revealed that only a subset of its functional potential is active at the same time [5]. The gut metatranscriptome is also temporally more dynamic and subject-specific compared with the gut metagenome [6,7]. In turn, analyses focusing on vaginal metatranscriptomes have described transcriptional activity changes associated with the imminent death and growth of bacteria and identified potential biomarkers for vaginal dysbiosis [8,9]. Emerging evidence also suggests that metatranscriptomics detects infectious diseases more efficiently compared with metagenomics [10]. Additional details on these and other key biological discoveries that have emerged from metatranscriptomics studies of human-associated microbial communities are comprehensively covered in another recent review [4].

In this review, we present the current state of metatranscriptomics analysis of human-associated microbial samples. We describe commonly used sample collection, sample preparation, **RNA-**

[1]Department of Pharmacology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[2]Hematology Research Unit, University of Helsinki, Helsinki, Finland
[3]Laboratory of Genetics, HUS Diagnostic Center, Hospital District of Helsinki and Uusimaa (HUS), Helsinki, Finland

*Correspondence:
matti.kankainen@hus.fi (M. Kankainen).



(See figure legend at the bottom of the next page.)

Glossary

Batch effects: variations in data that are unrelated to the biological question of interest. They can have negative impacts on statistical results and can cause false positive and negative findings. Batch effects can be corrected using batch adjustment methods or adjusted during statistical modeling, if uncorrelated with the phenotypes of interest.

Illumina short-read sequencing: high-throughput sequencing technology that uses bridge amplification to amplify templates attached to a solid surface and sequencing by synthesis with reversibly 3′-blocked fluorescently labeled nucleotides to determine their base order. It has massive throughput and partly configurable read length.

Long-read sequencing: sequencing techniques that can determine long nucleotide sequences up to several megabases in size. Reads suffer sequencing errors and cost-related throughput limitations but are in some solutions derived from single molecules without amplification. Representative examples include Pacific Biosciences and Oxford Nanopore Technologies.

Metagenomics: analysis of the collective genomes of all microbes present in a sample in a culture-free manner; provides information on the taxonomic composition and functional potential of microbial communities.

Metataxonomics: characterization of the microbial community through the amplification and sequencing of conserved marker genes. It is cost-effective and relevant even for samples rich in host cells but low in microbial biomass.

Metatranscriptomics: analysis of the collective transcriptomes of all the microbes present in a sample in a culture-free manner; provides information on the active microbes of the community and their expressed functions at a given moment and condition.

RNA-sequencing (RNA-seq): technique that allows examination of the quantity and sequences of RNA in a sample.

RNA-stabilizing solution: aqueous RNA stabilization and storage reagent that provides immediate RNase, DNase, and protease inactivation and stabilization of RNA. High-quality and intact RNA can be obtained from samples immerged in it even after

sequencing (RNA-seq), and data-analysis solutions (Figure 1A) and explain how these can be used to generate snapshots of the transcriptomic statuses of the microbes in a sample. We also outline challenges and advantages of different methods and provide insights into the contexts in which they are useful (Figure 2, Key figure). We conclude by discussing emerging opportunities that may transform metatranscriptomics research.

Isolation and preparation of RNA from human-associated microbial samples

The prerequisite of any metatranscriptomics experiment is appropriate sample handling and processing. This can be achieved by extracting RNA instantly after sampling. However, this is seldom practical or doable within the timeframe of swiftly responding [11] and degrading [12] microbial transcriptomes. Accordingly, human-associated metatranscriptomics studies have rarely relied on this strategy. Instead, snap-freezing and **RNA-stabilizing solutions** are popular (Figure 2). Compared with snap-freezing, RNA-stabilizing solutions allow storage at ambient temperature for several days [13,14]. This facilitates sample collection, transport, and storage. Use of these solutions can yield similar fecal metatranscriptomes as freezing and ethanol-fixation protocols [14]. However, different RNA-stabilizing solutions can differ in their preservation abilities (Box 1). For example, a comparison of commercial products reported RNAlater to preserve the RNA integrity and profile of fecal microbial samples better during storage compared with RNAProtect [13]. Additionally, some cancer studies have used formalin-fixed and paraffin-embedded (FFPE) tissues; although this preservative may obliterate transcriptional activity, it might be at the cost of RNA integrity [15].

At the next step, RNA should be extracted in an unbiased manner. In this process, cells are lysed to release their contents, including genomic DNA, RNA, and proteins. Generally, lysis can be achieved mechanically, chemically, or enzymatically, or with a combination thereof. Lysis should be efficient enough to release the intracellular contents of tough-to-lyse cells but gentle enough to keep nucleic acids of easy-to-lyse cells intact. It should also disrupt tissue structures. A recent review [16] summarizes the benefits of mechanical lysis for fecal samples. In particular, results from metagenomic studies suggest bead beating to increase nucleic acid yields and to result in a greater bacterial diversity and capture of Gram-positive bacterial species [16]. Thus, the popularity of mechanical lysis in metatranscriptomics studies is not surprising (Figure 2). As an alternative to total sample disruption and RNA isolation, the physiochemical properties of cells could be utilized to lyse cells in sequential order to enrich the microbial portion, similar to strategies used for cell line models [17,18]. However, this can introduce transcriptional shifts.

Regarding tailored solutions for blood samples, PAXgene and Tempus blood RNA tubes lyse blood cells and preserve RNA immediately upon sampling. Although shown to provide

several days of storage at ambient temperature.

Subtractive hybridization: technique for the enrichment of nucleic acid of interest with nucleic acid probes that are reverse complementary to the depletion target; used commonly for removing rRNA.

Total RNA: the mixture of coding and noncoding RNA molecules in a cell. The rRNA (primary component of ribosomes) and tRNA (adaptor molecule translating the genetic code of mRNA into an amino acid sequence) comprise most cellular RNA, while mRNA (the transcript carrying genetic information) and other noncoding RNAs account for a smaller portion. In blood samples, hgbRNA (a transcript encoding globins and highly expressed in erythrocytes) is a major player.

Unique molecular identifiers (UMIs): short random sequences added in some protocols to sequencing libraries before PCR amplification, enabling the identification of duplicates formed during PCR.

Figure 1. Metatranscriptomics workflow for human-derived samples. (A) Metatranscriptomics laboratory process. The experiment begins with sample collection and preservation. Following sample homogenization and extraction of total RNA, the desired RNA fraction can be enriched. This enrichment step can involve depletion of host rRNA, microbe rRNA, and/or other host RNAs by a multitude of methods. In long-read sequencing technologies, RNA needs to be poly(A)-tailed. A sequencing library is then prepared and sequenced to the desired depth. Short-read or long-read sequencing platforms can be used. The order of analysis steps can vary and some steps are optional. Recent methodological developments are highlighted in red. (B) Data-analysis procedure for metatranscriptomics data. The typical steps include preprocessing and quality control to remove adapter sequences, contaminants, low-quality bases, and other artifacts. This step may also include removal of unwanted sequence reads. In the case of long-read data, the preprocessing can also include error correction. Reads can then be *de novo* assembled into longer transcripts and their coding regions be identified using gene prediction programs. Taxonomic and functional assignments are then obtained. Finally, multivariate, network, dimension reduction, and functional enrichment tools are used to highlight differences and similarities between conditions and samples. The order of analysis steps can vary and some steps are optional. Recently improved or released bioinformatics methods are highlighted in red. Abbreviations: DASH, Depletion of Abundant Sequences by Hybridization; ONT, Oxford Nanopore Technologies; PacBio, Pacific Biosciences; UMI, unique molecular identifier. Created with BioRender (BioRender.com).

Key Figure

Overview of experimental approaches used in the characterization of human-associated metatranscriptomes

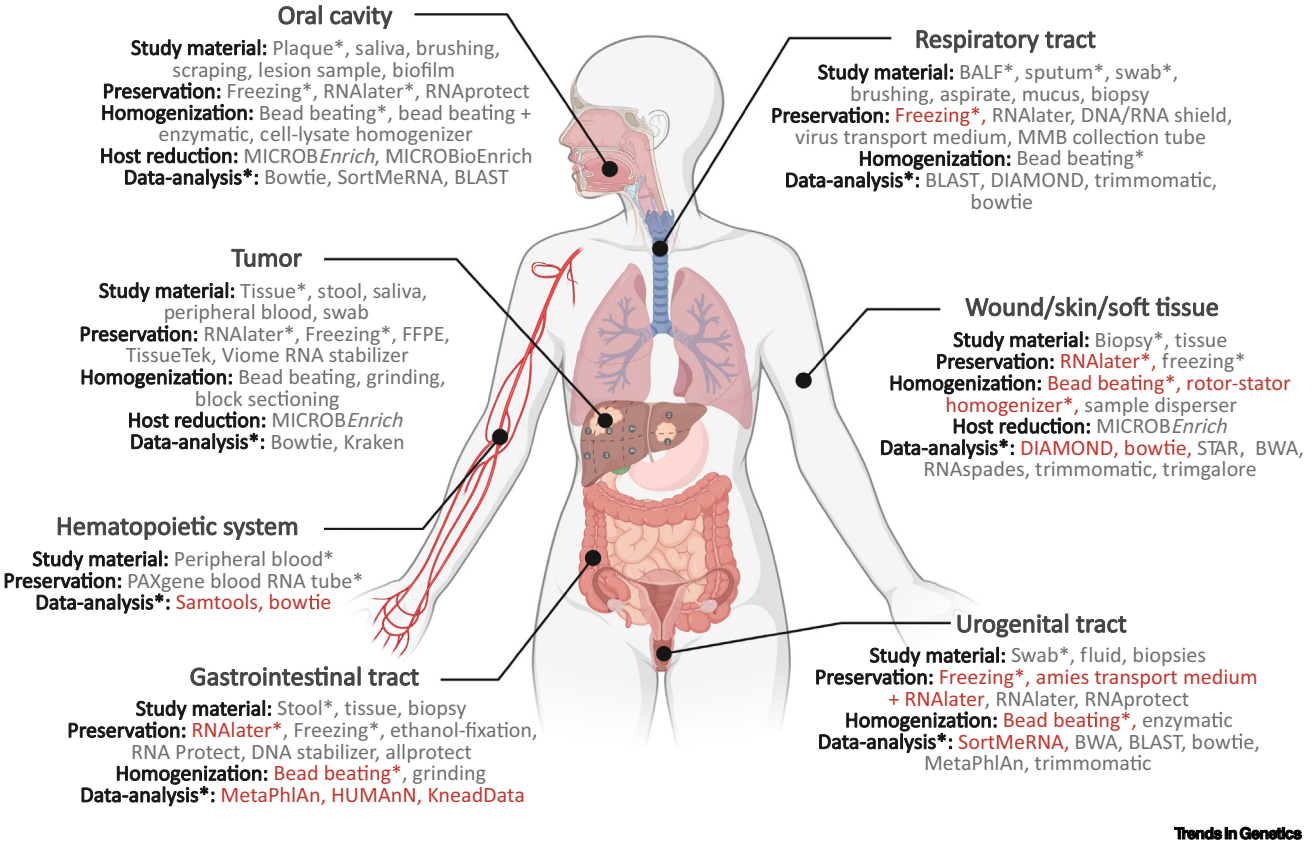


Figure 2. Many metatranscriptomics-profiling studies have used RNA-stabilizing solutions and/or snap-freezing for preservation and bead beating for homogenization. By contrast, metatranscriptome data processing has relied on a vast variety of bioinformatics software. *Experimental choices adopted by ≥25% of the studies of the category. For bioinformatics methods, only those adopted by at least two and ≥25% of the studies of the category are shown. Methods adopted by ≥50% of the studies of the category in the past 3 years are highlighted in red. The studies are listed in [4]. Abbreviations: BALF, bronchoalveolar lavage fluid; FFPE, formalin-fixed, paraffin-embedded. Created with BioRender (BioRender.com).

comparable RNA in terms of quality and yield from human whole blood [19], knowledge of their effects on metatranscriptomes is limited. However, use of multiple solutions in a single analytic setting should be avoided [19].

A range of custom and commercial methods exist for isolation of RNA and removal of analysis-contaminating substances. These include several organic extraction methods, such as Thermo Fisher Scientific's TRIzol® reagent, as well as different solid-phase extraction solutions, such as Thermo Fisher Scientific's PureLink and Qiagen's RNeasy products. However, methods differ in performance, cost-effectiveness, and suitability for obtaining the required RNA types from the desired sample materials [20,21]. For example, the selective binding properties of solid-phase silica surfaces [21] and use of TRIzol on low biomass samples [22] can reduce the short RNA

fraction. If the interest is in regulatory and small structural RNAs, use of mirVana or some other approach optimized for small RNAs should be considered. Furthermore, notable variance in RNA yield, purity, and integrity has been observed among methods. For instance, a study of stool samples reported over fivefold differences in RNA yields between the tested RNA isolation strategies [13], emphasizing the need to determine the appropriate approach in each study.

Enriching relevant microbial RNAs optimizes cost-effectivity

The **total RNA** extracted from a human-derived sample is a mixture of different human and microbial RNA molecules, such as rRNA, tRNA, mRNA, small regulatory RNA (sRNA), and long noncoding RNA (lncRNA). Its most abundant transcript type is rRNA, accounting for over 80% of the total content [23]. However, because sequencing of high amounts of rRNA is often undesirable, rRNA elimination has become standard practice. This improves the cost-effectivity of the experiment and detection sensitivity of scientifically more alluring RNA types.

Typically, the elimination of rRNAs in metatranscriptomics studies has relied on their reduction, because universal transcript features are mainly lacking for the positive selection of relevant microbial RNA. The most used method for this by far has been **subtractive hybridization** with sequence-specific capture oligonucleotides. This approach has also been at the heart of many commercial kits that, at best, have excellent depletion efficiency [24,25]. These products also appear to retain RNA profiles intact [26,27]. However, not all products perform equally well [24–26]. Alternatively, sequence-specific DNA oligonucleotides can be hybridized onto rRNA, and the resultant rRNA–DNA hybrids digested by RNase H [28]. More recently, the CRISPR/Cas9 system

Box 1. Experimental factors to consider when planning metatranscriptomics studies

Metatranscriptome samples are often treated with common preservatives to avoid unwanted community changes and RNA fragmentation. Nevertheless, use of RNA stabilization reagents may introduce varying degrees of biases in specific taxa [13]. Delay in preservation and length of storage can also result in altered or fragmented profiles.

The efficiency of RNA extraction workflows varies among organisms, sample materials, and RNA species. For example, inefficient lysis of a specific taxon is possible, resulting in falsely reduced abundances. In turn, commonly used silica-based approaches can result in the loss of short RNAs [21]. If interested in regulatory and small structural RNA molecules, approaches optimized for small RNAs should be considered. Additionally, inefficient removal of DNA can introduce artifacts [90].

Depletion of undesired RNA improves the sequencing efficiency of relevant transcripts. For instance, efficient rRNA depletion facilitates the discovery of low abundance transcripts [91]. However, the depletion efficiency varies greatly, for instance, with products [24,25,91], rRNA species [24,25,91], and RNA integrities [13]. Nonspecific depletion can also cause bias.

Typical short-read sequencing library preparations involve fragmentation, cDNA synthesis, adaptor ligation, and PCR amplification [36]. However, each step can introduce biases [37]. For example, fragmentation or cDNA synthesis may not be random. Amplification can also introduce bias against long sequences and those with extreme GC content, which propagate to later cycles [37]. Amplification-free long-read sequencing is insensitive to amplification errors, but may suffer from other artifacts [46,47,92–94].

The choice of sequencing strategy can also affect results. For example, the depth of sequencing determines the resolution of the analysis. Higher sequencing depths improve the discovery of rare taxa and transcripts, while shallower miss them more easily. Additionally, different sequencing platforms introduce different types of coverage bias [46,95] and systematic error [46] to sequencing outcomes. Long- and short-read data also have distinct purposes and strengths. For instance, long-reads improve assembly contiguity and taxonomic and functional classifications, but suffer cost-related throughput limitations and sequencing errors.

Bioinformatic analyses can generate bias. For example, reference databases can comprise distinct organisms or genes. Databases can also contain misannotated references, causing false findings. Results are also incomparable between sequence abundance profilers that use genome references and taxonomy abundance profilers relying on clade-specific marker genes [70]. Host-derived and rRNA reads can also produce false assignments. To eliminate bioinformatic biases, the same workflows with matching software versions and databases should be used.

coupled with a library of guide RNAs has been used to reduce unwanted cDNA before library amplification [29,30]. This Depletion of Abundant Sequences by Hybridization (DASH) method appears to remove 56–86% of rRNA at only one-tenth of the cost of subtractive hybridization [30], providing an enticing alternative for the abruptly discontinued top-of-the-line Ribo-Zero subtractive hybridization product. Lastly, a pricy solution is to tackle the problem with brute-force sequencing and computational filtering, which is reversible and skew-free, if done properly.

The disproportion between host and microbial RNA types other than rRNA in many human-derived samples may make the elimination of the poly(A)-tail-containing human RNA fraction desirable. This can be achieved by using oligo(dT) capture, which is used for enrichment in human transcriptomics. For instance, Ambion's MICROB*Enrich*™ kit uses subtractive hybridization for the reduction of poly(A)-tailed RNA and human, mouse, and rat rRNA [31] and has been used in some human-associated metatranscriptomics studies. Other highly abundant transcripts can be eliminated in a similar way. For example, probe hybridization methods exist for the reduction of globin gene RNA (hgbRNA), which is abundant in blood and bone marrow [32].

Contamination controls help exposing false findings

Microbial contamination (Box 2) is a common issue in microbial sequencing studies and can also produce erroneous interpretations in metatranscriptomics studies. In extreme cases, microbial contaminants can even become the predominant feature [33]. Low microbial biomass samples are particularly vulnerable to contamination and require the most stringent precautions (Box 2).

Short-read sequencing is (still) the basic workhorse, but for how long?

Modern metatranscriptomics relies on sequencing. Initially, pyrosequencing [34] was mainly used, whereas **Illumina short-read sequencing** [35] is the current mainstay technology. In future, studies may rely on amplification-free **long-read sequencing** technologies.

Box 2. Contamination control

Microbial contaminants can emerge from external or internal sources. External contamination originates from the outside of samples and can occur at any step of the workflow. Sources of external contamination include surfaces, air, patients, researchers, and equipment [96,97]. Reagents are another well-known source [98]. By contrast, internal contamination relates to the exchange of genetic material between samples. It can occur during sample processing [99] and sequencing [38,100]. Furthermore, contamination-like errors may result from erroneous read classification.

The most effective strategy for contamination control involves the use of negative and positive controls [100]. These include sequencing blanks taken during sampling (negative sampling control), extraction samples from nucleic acid-free sample material (negative extraction controls), and template-free amplification samples (negative PCR amplification controls), which are useful to monitor external and cross-sample contamination. Positive controls comprise negative control matrix with quantitated spike-in mock organisms. They allow the detection of performance failures in the process. Ideally, both controls are based on a matrix that mimics the characteristics of the biological sample, included in each batch, and processed alongside biological samples using the same reagent lots. After being sequenced, the frequency of false positive findings among negative control samples and false negative findings among positive control samples should be quantified.

Efficient contamination control includes pretreatment of the reagents, labware, and surfaces to remove exogenous nucleic acids and aseptic handling of samples. Similar to other processes involving exponential amplification, it is recommended to maintain a unidirectional laboratory workflow and separate pre- and postamplification processes. It is also advisable to use modern sequencing protocols and instruments not prone to barcode index switching [38].

Contamination control can be accomplished bioinformatically. Decontam [101] and SourceTracker [102,103] are two popular solutions for this. In a recent comparison involving polymicrobial 16S rRNA profiling data [104], SourceTracker removed >98% of contaminants, but also erroneously some genuine findings. By contrast, Decontam eliminated 70–90% of the contaminants and no real findings. However, caution is needed when applying these tools to metatranscriptomics data, because no data on their performance in this context have been published to our knowledge. Alternatively, contamination removal can be based on blacklisting of known contaminants [98], filtering of microbes present in a negative-control samples, or filtering findings based on species or gene prevalence across tissue types [105]. Given that some contaminants may be genuinely present in research samples [99], caution is needed when using these strategies.

The Illumina technology has high accuracy and can generate up to 16 Tb per run. Given that a single sample seldom is sequenced this deeply, it is common to sequence multiple libraries with unique sample-specific barcodes in a single run. Typically, RNA-seq on the Illumina platforms requires fragmentation, cDNA synthesis, adaptor ligation, PCR amplification, and sequencing [36]. However, all these steps can introduce biases (Box 1) [37]. For example, fragmentation of source RNA or cDNA synthesis may not work in a random manner and amplification may not amplify all transcripts. The amplification can also introduce bias against long sequences and those with extreme GC content. High numbers of amplification cycles appear to propagate artifacts [37]. Different kits and input RNA amounts also exhibit varying degrees of bias and imprecision to metatranscriptomes [27]. Alarmingly, a major drawback of some old instruments and library preparations is false extension of library fragments with an oligo containing the wrong sample-specific barcode, causing barcode index switching [38]. However, modern library preparation methods that support dual indexes and **unique molecular identifiers (UMIs)** have resolved this defect. These identifiers also allow control of PCR artifacts.

Pacific Biosciences [39] and Oxford Nanopore Technologies (ONT) [40] are two popular long-read sequencing technologies that provide an alternative to short-read sequencing. These techniques routinely produce reads several kilobases in size and can capture entire transcripts in single reads [36,41]. This can improve taxonomic classification and functional annotation. Other benefits include rapid data generation and data analysis during runtime [41,42]. For instance, recording of unwanted templates can be aborted in ONT during sequencing to enrich other templates [43,44]. The major disadvantages involve low throughput, low read accuracy, high sequencing costs, and incompatibility with degraded RNA [36,42,45,46]. Some protocols are also prone to transcript truncation, which can be mitigated by improving base calling [47] and/or library preparation [48]. Currently, transcriptomes of a handful of bacterial species [49–53] and mixed communities [54] have been solved using long-read sequencing. In these investigations, bacterial RNA was subjected to poly(A)-tailing and then processed using: ONT direct RNA-seq enabling inference of native RNA without amplification, or cDNA conversion [49–52]; ONT amplification-free cDNA sequencing avoiding artifacts resulting from amplification [49]; ONT amplification-based cDNA sequencing requiring relatively small amounts of RNA [49]; or the Pacific Biosciences amplification-based cDNA sequencing protocol [53]. All these approaches were suitable for the task [50–54], with ONT amplification-based cDNA sequencing offering a particularly good yield and accuracy according to Grünberger *et al.* [49].

The optimal sequencing depth needed for a given condition is experiment-specific. In general, metatranscriptomes require several fold higher coverage compared with metagenomes [2]. Conceptually, assuming a 100-fold expression difference between the rarest and most common transcript and requiring at least five read-pairs for each, characterization of the transcriptome of a bacterium with 4000 genes needs 1 million read-pairs. If there are 100 such bacteria and assuming a 100-fold abundance difference between the rarest and most common organism, 10 000 million read-pairs are needed. Obviously, this depth is unrealistic, considering that current studies have used between 1 million [5] and 250 million [55] reads. Thus, if the aim is also to characterize low-abundance members, the greatest depth possible should be preferred. High depths also add robustness to contaminants if they can be distinguished computationally from relevant organisms.

Quality control is the foundation of metatranscriptomics data analysis

Computational analysis of metatranscriptomics data mimics that of metagenomics. It comprises a series of analyses (Figure 1B) and involves the use of various dedicated software packages (Table 1). A common first step in the analysis is quality control. This includes removing low-

Table 1. A selection of software tools for metatranscriptomics data

Software	Synopsis	Usage[a]	Website	Refs
Quality control				
Trimmomatic	Versatile read trimming and filtering tool for short-read sequencing data	****	www.usadellab.org/cms/?page=trimmomatic	[111]
fastp	Quality control and data filtering of sequencing data	**	https://github.com/OpenGene/fastp	[112]
FastQC	Field-standard toolkit for sequence quality control analysis and generation of summary statistics about the reads	**	www.bioinformatics.babraham.ac.uk/projects/fastqc/	c
PRINSEQ	Filters, reformats, and trims genomic and metagenomic sequence data	**	https://prinseq.sourceforge.net/	[113]
KneadData[b]	Performs quality control and removes host and rRNA-originating reads from metagenomic and metatranscriptomic sequencing data	**	https://huttenhower.sph.harvard.edu/kneaddata/	–
Cutadapt	Finds and removes adapters, primers, poly(A) tails, and other undesirable sequences from sequencing data	**	https://cutadapt.readthedocs.io/	[114]
Trim Galore	Wrapper around Cutadapt and FastQC for automated quality and adapter trimming analyses	**	https://github.com/FelixKrueger/TrimGalore	–
SortMeRNA[b]	Finds, removes, or clusters rRNA reads from metatranscriptomics data	***	https://bioinfo.lifl.fr/RNA/sortmerna/	[57]
Infernal	Finds structural RNAs; uses covariance models that combine primary sequence and secondary structure conservation information		http://eddylab.org/infernal/	[115]
Deconseq	Removes human reads from genomic data sets based on BWA-SW alignments	*	https://github.com/kiwiroy/deconseq	[59]
BMTagger	Removes human reads from genomic data sets	*	https://help.rc.ufl.edu/doc/Bmtagger	–
RiboDetector	Recurrent neural network approach for removing rRNA sequences		https://github.com/hzi-bifo/RiboDetector	[58]
RSeQC	Quality control tool designed for RNA-seq data		https://rseqc.sourceforge.net/	[116]
Gene prediction				
Prodigal	Protein-coding gene prediction for prokaryotic genomes and metagenomes	*	https://github.com/hyattpd/Prodigal	[117]
Assembly				
Trinity	Modular single-organism transcriptome assembler based on de Bruijn graphs of *k*-mers	**	https://github.com/trinityrnaseq/trinityrnaseq/wiki	[118]
Trans-Abyss	Single-organism transcriptome assembler based on distributed and scalable de Bruijn graphs of *k*-mers		https://github.com/bcgsc/transabyss	[119]
rnaSPAdes	Transcriptome assembly extension of SPAdes microbe genome de Bruijn graph assembler	*	https://cab.spbu.ru/software/rnaspades/	[120]
IDBA-MT[b]/IDBA-MTP[b]	Metatranscriptome assembler based on de Bruijn graph with multiple *k*-mers; IDBA-MTP guides assembly based on information of known microbial protein sequences	*	https://i.cs.hku.hk/~alse/hkubrg/projects/idba_mt/index.html	[121]
TAG[b]	Metatranscriptome assembly by mapping metatranscriptomics reads onto graphs of matched metagenome assembly		https://omics.informatics.indiana.edu/TAG/	[122]
STAble[b]	Transcriptome assembler based on de Bruijn graphs of full-length reads			[123]
MEGAHIT	Fast metagenome assembler based on succinct de Bruijn graphs	**	https://github.com/voutcn/megahit	[124]
Contamination analysis				
Decontam	Identifies and removes contamination in metagenomics data; requires DNA quantitation data or sequenced negative control samples		https://benjjneb.github.io/decontam/vignettes/decontam_intro.html	[101]
SourceTracker	Identify sources and proportions of contamination in		https://github.com/biota/sourcetracker2	[102,103]

(continued on next page)

Table 1. (continued)

Software	Synopsis	Usage[a]	Website	Refs
	metagenomics data based on Bayesian approach			
Generic mappers and aligners				
Bowtie/Bowtie2	Ultrafast aligner for mapping reads to large reference database based on Ferragina–Manzini index	****	https://bowtie-bio.sourceforge.net/bowtie2/	[125]
BWA	Ultrafast aligner for mapping reads to large reference database based on backward search with Burrows–Wheeler transform	***	https://bio-bwa.sourceforge.net/	[126]
Minimap2	Versatile aligner for mapping long and error-prone Oxford Nanopore Technologies and PacBio reads to large reference database		https://github.com/lh3/minimap2	[127]
STAR	Splice-aware mapping of RNA-seq reads to genomic references	**	https://github.com/alexdobin/STAR	[128]
hisat2	Ultrafast aligner for mapping reads to a population of human genomes and single reference genome	**	http://daehwankimlab.github.io/hisat2	[129]
DIAMOND	Ultrafast aligner for protein and translated DNA searches	****	https://github.com/bbuchfink/diamond	[130]
BLAST	Heuristic sequence alignment algorithm for highly sensitive local nucleotide, protein, and translated-nucleotide protein alignment	****	www.ncbi.nlm.nih.gov/BLAST/	[131]
CD-HIT	Clusters and compares protein or nucleotide sequences	**	http://cd-hit.org/	[132]
Taxonomic characterization				
MetaPhlAn	Taxonomic profiler against over 1 million prokaryotic genome sequences using clade-specific microbial marker genes	****	https://huttenhower.sph.harvard.edu/metaphlan/	[74]
Kraken2	Fast taxonomic classifier against genome sequences based on exact *k*-mer matches	***	https://ccb.jhu.edu/software/kraken2/	[72]
Centrifuge	Fast taxonomic classifier against genome sequences based on Burrows–Wheeler transform and Ferragina–Manzini index	*	https://ccb.jhu.edu/software/centrifuge/	[133]
Kaiju	Fast taxonomic classifier against protein sequences		https://kaiju.binf.ku.dk/	[134]
mOTUs2[b]	Fast taxonomic profiler using universal phylogenetic marker genes	*	https://motu-tool.org/	[135]
CCMetagen[b]	Taxonomic classifier based on *k*-mer matches and scoring references based on all possible mappings of all reads	*	https://github.com/vrmarcelino/CCMetagen	[136]
Taxonomic and functional characterization				
HUMAnN[b]	Functional profiler based on a tiered sequence similarity search strategy	****	https://huttenhower.sph.harvard.edu/humann/	[76]
Megan	Uses DIAMOND output for generating taxonomic and functional bins	**	http://megan.husonlab.org/	[137]
MetaCLADE[b]	Profile-based domain annotation profiler for metagenomic and metatranscriptomics data		www.lcqb.upmc.fr/metaclade/	[138]
eggNOG-mapper	Functional profiler based on fast orthology assignment of reads to clusters and phylogenies from the eggNOG database		http://eggnog-mapper.embl.de/	[139]
Statistical analysis				
edgeR	R package for differential expression analysis based on negative binomial distribution and empirical Bayes estimation		https://bioconductor.org/packages/edgeR/	[140]
DESeq2	R package for differential expression analysis based on negative binomial distribution and estimation of variance–mean dependence		https://bioconductor.org/packages/DESeq2/	[80]
limma	R package for linear model analysis of single-organism expression data; functions for transforming RNA-seq count data for linear modeling		https://bioconductor.org/packages/limma/	[110]
Sequencingeffort	R package for estimating required depth of sequencing		https://github.com/amonleong/Sequencingeffort	[141]

Table 1. (continued)

Software	Synopsis	Usage[a]	Website	Refs
NOISeq	R package for differential expression analysis without parametric assumption		https://bioconductor.org/packages/NOISeq/	[109]
Phyloseq	R package for analyzing and graphically displaying taxonomically classified sequencing data	*	https://bioconductor.org/packages/phyloseq	[142]
LefSe	Linear discriminant analysis of effect size method to determine features explaining class differences	*	https://huttenhower.sph.harvard.edu/lefse/	[143]
RUVseq	R package to remove unwanted variation based on control genes, replicate samples, or residuals from RNA-seq data		https://bioconductor.org/packages/RUVSeq/	[144]
ALDEx2	R package for differential abundance analysis that uses a Dirichlet-multinomial model to infer abundance from counts		https://bioconductor.org/packages/ALDEx2/	[85]
eBay	Empirical Bayes normalization method for microbiome data		https://github.com/liudoubletian/eBay	[86]
ANCOM	R package for detecting abundance differences based on additive log ratio transformation and heuristic strategy		https://bioconductor.org/packages/ANCOMBC/	[87]
ANCOM-BC	R package for detecting abundance differences by estimating unknown sampling fractions, correcting bias through log linear regression model, and identifying taxa according to the variable of interest		https://bioconductor.org/packages/ANCOMBC/	[81]
MaAsLin2[b]	Assesses multivariable associations of microbial community features with complex metadata in population-scale observational studies		https://huttenhower.sph.harvard.edu/maaslin/	[84]
Pipelines				
MetaTrans[b]	Pipeline for rRNA read filtering, gene prediction, gene clustering, functional assignment, and differential expression analysis		https://github.com/KavrakiLab/MetaTrans	[145]
MOCAT2	Pipeline for read quality control, assembly, gene prediction, gene clustering into reference gene catalogs, functional annotation against eggNOG database, and generation of taxonomic profiles based on reference marker gene mapping	*	https://mocat.embl.de/	[146]
SAMSA2[b]	Pipeline for preprocessing and read trimming, filtering of rRNA sequences, functional annotation using protein-similarity search, and differential expression analysis	*	https://github.com/transcript/samsa2	[147]
Databases				
SILVA	Resource for high-quality rRNA sequence data	*	www.arb-silva.de/	[148]
Rfam	Database of noncoding RNA and structured RNA elements		https://rfam.xfam.org/	[149]
UniProt	Database of protein sequence and functional information		www.uniprot.org/	[65]
VFDB	Database of virulence factors in bacterial pathogens		www.mgc.ac.cn/VFs/	[66]
CARD	Database of antimicrobial resistance genes, proteins, and phenotypes		https://card.mcmaster.ca/	[67]
KEGG	Database of metabolic pathways, diseases, drugs, and chemical substances	*	www.genome.jp/kegg/	[69]

[a]Usage indicates how frequently the software has been used in human-associated metatranscriptomics studies, ranging from zero (very seldom mentioned) to four (frequently mentioned) asterisks. The studies used for assessing usage are listed in [4].
[b]Method developed originally for metatranscriptomics data.
[c]www.bioinformatics.babraham.ac.uk/projects/fastqc/.

quality sequences and contaminants and trimming sequencing adapters and low-quality bases. In the case of long-read data, it can also include error correction [56]. Several tools exist for these tasks (Table 1). Additionally, human and rRNA-derived reads can be removed by the use of dedicated software tools [57–59] or by performing similarity searches against relevant reference collections. This is advisable because human and rRNA-derived reads can cause false assignments

in downstream analyses. Removal of rRNA reads is particularly desirable if rRNA has been depleted, because depletion protocols can deplete rRNAs from different organisms to varying degrees [26]. If total RNA was sequenced, rRNA may be exploited in taxonomy assignment, but not in functional characterizations. Finally, a rich repertoire of tools is available to generate quality metrics (Table 1), which are useful in identifying quality-compromised samples.

Assembly can benefit analysis of short-read data

Quality controlled short reads have been subjected, in increasing amounts, to *de novo* assembly in human-associated metatranscriptomics studies. The advantage of this is that it enhances detection of novel expressed genes and removes redundancy among read data. The long contiguous sequences are also more informative in subsequent analyses. Indeed, assembly has increased the yield of functional annotations by one-fifth in human metagenomics [60] and by one-third in murine metatranscriptomics data [61]. However, metatranscriptome assemblies often miss and contain partial transcripts because sufficient read coverage is rarely obtained for all transcripts. Complexity of metatranscriptomes also limits assembly quality. Moreover, not many assemblers are available for the task. The few native metatranscriptome assemblers that exist have not been updated often since release and have demonstrated a worse performance compared with their single-organism competitors [61]. The value of binning analysis, a common procedure in metagenomics data analysis [62], is another open topic. Binning of sequences into groups based on sequence content and coverage may not be fruitful, because chromosomally separated transcripts can have the same read coverage and sequence content.

Assembly is also beneficial for long-read sequencing data

Current long-read sequencing technologies offer opportunities to generate individual reads that can cover almost the entire transcript and have sufficient accuracy for read alignment analyses. Yet, long reads can benefit from error correction and assembly into even more complete transcripts. This can, for example, resolve errors related to truncated reads. The accuracy of an assembly can then be further improved by removing errors from the contigs using polishing algorithms [46].

Understanding data through taxonomic and functional characterization

One of the main goals of metatranscriptomics is to understand what organisms and functions are active in the community. This can be achieved by querying reads and/or transcript contigs against multiorganism reference databases providing taxonomic [63], functional [64–68], or metabolic [69] information. Alternatively, transcript contigs can be annotated for coding genes and these coding sequences be used in similarity searches. This may be beneficial if contigs feature multiple genes. Typically, accelerated nucleotide and protein-level alignment tools are used (Table 1). Of these, protein-space aligners detect more distantly related sequences and, thus, are recommended for communities with species lacking a reference genome. However, they are more prone to false findings than are nucleotide aligners [70]. For both, ambiguous homologs sparsely distributed over taxonomy and repetitions are problematic. Ideally, the used reference database should include both microbial and human host sequences to mitigate false assignments of human-originating sequences. If splice-agnostic aligners are used, host transcripts should also be included to ensure subtraction of exon-spanning host sequences. Microbe references may suffer human contamination, which is an issue especially among draft genomes [71]. Finally, mapping information is converted into an abundance matrix describing absolute or scaled frequencies of taxon or gene models. During the process, multi-mapping reads are often resolved choosing the lowest unambiguous taxonomic rank/gene family [72] or dividing the assignment proportionally with all the best hits. More recently, assignment-making based on maximum likelihood estimates of transcript abundances [73] has become common. It is also commonplace to aggregate assignments over taxonomic or gene similarity levels.

The use of dedicated taxonomical and functional profilers in metatranscriptomics studies has increased steadily over the years. Widely used software includes MetaPhlan [74], which uses phylogenetically conserved clade-specific marker genes and reports relative taxon abundances, and Kraken2 [72], which compares sequences with reference genome collections and reports the fraction of reads assigned to each taxon. Notably, both were among the highest ranked software in their categories in a recent critical assessment of metagenome interpretation [75]. An example of a widely used functional profiling tool is HUMAnN [76]. As an example of the importance of software updates, the newest version of HUMAnN improves accuracy of functional annotations by over 20% compared with previous versions [74]. However, because many of the dedicated tools have been developed for use with metagenomic data and rely on assumptions not necessarily valid for community transcriptome data, they should be applied with caution to metatranscriptomics data [2].

Converting data into biological hypotheses

The evaluation of completeness of the sequencing is a common downstream analysis. The traditional technique for this involves generation of a rarefaction curve, which involves taking a predefined number of sequences at random and plotting the number of unique genes and/or organisms found in each rarefaction subset. Visual interpretation of the plot or its mathematical modeling allows inference of the completeness of the sequencing. Several R packages are available for the rarefaction analysis (Table 1).

Differential abundance analysis allows detection of differences in community transcription across conditions (Box 3). In many respects, it follows the form of single-organism transcriptomics

Box 3. Statistical aspects of metatranscriptomics analyses

Power analysis helps identify misleading and ungeneralizable experiments. Its four parts are: sample size, probability of finding an effect that is not there (significance level), probability of finding a true effect (power), and magnitude of the effect. Given three of these, the fourth is computed. However, the result only holds if the chosen test and the power analysis have the same assumptions. Given a lack of power calculators for metatranscriptomics, those designed for metataxonomic [106] or single-organism transcriptomic [107] experiments could be considered.

Metatranscriptomes can be studied using various study designs. The simplest involves comparison between groups of unrelated samples. While popular, its results are sensitive to interindividual variation and sample sizes. Ideally, groups would include individuals with similar characteristics. Repeated measures design involves measuring the same subjects across multiple conditions. Its error term is smaller than that of unrelated samples. A longitudinal study design helps distinguish temporal from stable activities. The number of serial samples needed is unclear, but recent metagenomic research has suggested to include between five and nine [108]. Metatranscriptomics may require more owing to its higher within-subject longitudinal variation [7].

Transcript abundance data tend to involve an excessive number of zeros, large dynamic range, high skewness, and mean-variance dependency. They also are sensitive to sequencing depth and transcript length, are not inherently normally distributed, and are initially discrete in nature. Therefore, it is typical to normalize data [107] and use nonparametric tests [109] or count distributions [80] to find expression alterations. Methods using normal distribution can also be used after appropriate transformation of data [110]. In differential analysis, linear models and generalized linear models are frequently adapted, because they accommodate arbitrarily complex designs.

The strong correlation between metagenomic and metatranscriptomics composition represents an analytical challenge that can prevent true expression alterations being distinguished from changes in genomic copy number. In the presence of metagenomic data, this can be addressed by incorporating DNA-level gene abundance as a covariate [88]. In the absence of paired metagenomic data, taxon-level total RNA and DNA estimates can be used as a proxy for gene abundance [88].

Batch effects are technical or biological sources of variation that are unrelated to the biological question of interest. For example, use of multiple RNA isolation methods could generate batches that confound real biological signal. Creation of batch effects should be avoided, but if existing and uncorrelated with the phenotypes of interest, they can be tackled in statistical testing.

analysis. The typical input of the analysis is an organism or gene family abundance matrix. As a common first step, systematic effects are removed. Noteworthy normalization approaches with good performance in past metagenome evaluations [77,78] include the trimmed mean of m-value (TMM) normalization method [79], relative log expression method [80], and analysis of compositions of microbiomes with bias correction (ANCOM-BC) [81]. Taxon-specific scaling is another enticing method [82]. Following normalization, abundance alterations are detected. This is often achieved using methods developed for use with metagenome or single-organism transcriptome data. A recent assessment of nine approaches for detecting abundance differences in metagenomic data [83] pinpointed MaAsLin2 [84], ALDEx2 [85], eBay [86], and ANCOM [87], while another comparison [78] highlighted ANCOM-BC [81] and ANCOM [87].

The strong correlation between RNA and DNA abundances represents a specific analytical challenge for metatranscriptomics differential abundance analyses. If not taken into account properly, the statistical analysis becomes prone to mistaking changes in underlying genomic copy number for expression-level changes. For example, loss of a taxon will abolish expression of its genes without downregulation. As a solution, metatranscriptomes can be analyzed jointly with metagenomes (e.g., [2,5–7,9,14]). While RNA/DNA abundance ratios [14] have frequently been used to tackle the dependency between DNA and RNA abundances in such studies, linear models with feature-specific gene-copy numbers as a covariate may provide a better approach [88]. In the absence of paired metagenomes, the species-total RNA or DNA abundances serve as proxies for per-gene gene-copy numbers. In addition to the use of paired metagenomes, study power, study design, and **batch effects** involve other statistical aspects to be considered before setting up the experiment (Box 3).

Concluding remarks

Metatranscriptomics is an invaluable tool in modern microbiology. It enables the investigation of microbial community transcription in a culture-free manner and reveals biological information not obtainable by more conventional metagenomic profiling techniques. Over the past few years, numerous human microbiota studies have relied on metatranscriptomics [4]. Most have focused on easily collectable specimens, such as feces, rich in bacteria, whereas use of metatranscriptomics on human tissues rich in host cells but low in microbial biomass is still in its infancy. We believe that the metatranscriptomics exploration of such samples will greatly benefit from improved microbial mRNA, sRNA, and lncRNA enrichment solutions and real-time long-read sequencing. However, several other challenges on sample preparation remain to be addressed (see Outstanding questions).

The current studies on human-associated metatranscriptomes have used rather similar laboratory workflows, often involving snap-freezing or RNAlater, bead beating, subtractive hybridization by rRNA probes, and Illumina short-read sequencing (Figure 2). In future, amplification-free long-read RNA-seq may replace short-read sequencing because its long reads benefit subsequent analyses. With reduced bias, this technology may provide missing clues on microbiotas. By contrast, the data analysis protocols have renewed over time and have included varying mixtures of software and databases (Figure 2). Related to data analysis, we expect to see performance advances through the emergence of deep-learning solutions. For example, deep learning-based protein structure modeling can characterize genes unattainable by sequence similarity approaches [89]. Deep learning may also help identify complex and nonlinear interactions between genes, transcripts, metabolites, and phenotypes. In general, we envision growing use of multiomics data in the future. Such approaches may overcome limitations of studies relying on a single approach, help to formulate regulatory networks, and improve understanding of microbe–microbe and host–microbe interactions.

Outstanding questions

Various sample collection, sample preservation, RNA extraction, RNA enrichment, and sequencing choices are available for characterization of metatranscriptomics. What is the ideal procedure for each sample type and how can it be identified?

To what extent do quality control, assembly, functional characterization, taxonomic profiling, and differential expression analysis solutions affect metatranscriptomics results? Is there a universal computational pipeline that suits all samples?

Which methods are optimal for studying microbial samples rich in host but poor in bacterial cells? Could the DASH method be used to reduce the entire human transcriptome?

Oxford Nanopore sequencing is a third-generation sequencing technology that can capture entire transcripts in single reads in real time. Could its selective sequencing concept enhance our ability to study microbial samples with high host content and aid diagnosis of infectious diseases? Moreover, generation of long-reads requires high-quality RNA, but how can this be achieved?

A large portion of microbial transcripts present in human-associated metatranscriptomes remains uncharacterized. Could deep-learning solutions provide important advances for their annotation, given that it has improved prediction of antimicrobial resistance gene and taxonomic profiling?

Many software tools used in the analysis of metatranscriptomics were originally developed for use with metagenome and single-organism transcriptome data. Would thorough evaluation of methods for metatranscriptomic assembly, functional annotation, taxonomic profiling, and differential expression analysis reveal analysis bottlenecks?

The integrated analysis of metagenomics and metatranscriptomics data is advantageous in distinguishing transcriptome alternations emerging from DNA abundance alterations from those resulting from transcriptional changes. Which methods are best to achieve this in the absence of paired metagenome data?

In conclusion, the field of metatranscriptomics has developed steadily over the past few years. Much of this progress has stemmed from advances in metagenomic and single-organism transcriptomic technologies, leaving room for improvement. We foresee that future advances will strengthen the role of metatranscriptomics as an indispensable method for assessing microbial functions in health and disease and pave the way for the broader clinical use of this promising and culture-independent method.

Acknowledgments

This work was supported by the Academy of Finland (grant numbers 292635 and 292605); Business Finland (grant number 6113/31/2016); and the Finnish Cultural Foundation.

Declaration of interests

The authors declare no conflicts of interest.

References

1. Filiatrault, M.J. (2011) Progress in prokaryotic transcriptomics. *Curr. Opin. Microbiol.* 14, 579–586
2. Zhang, Y. *et al.* (2021) Metatranscriptomics for the human microbiome and microbial community functional profiling. *Annu. Rev. Biomed. Data Sci.* 4, 279–311
3. Shakya, M. *et al.* (2019) Advances and challenges in metatranscriptomic analysis. *Front. Genet.* 10, 904
4. Ojala, T. *et al.* (2023) Understanding human health through metatranscriptomics. *Trends Mol. Med.* 29, 376–389
5. Abu-Ali, G.S. *et al.* (2018) Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nat. Microbiol.* 3, 356–366
6. Heintz-Buschart, A. *et al.* (2016) Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* 2, 16180
7. Mehta, R.S. *et al.* (2018) Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* 3, 347–355
8. Ma, B. *et al.* (2020) A comprehensive non-redundant gene catalog reveals extensive within-community intraspecies diversity in the human vagina. *Nat. Commun.* 11, 940
9. France, M.T. *et al.* (2022) Insight into the ecology of vaginal bacteria through integrative analyses of metagenomic and metatranscriptomic data. *Genome Biol.* 23, 66
10. Tao, Y. *et al.* (2022) Diagnostic performance of metagenomic next-generation sequencing in pediatric patients: a retrospective study in a large children's medical center. *Clin. Chem.* 68, 1031–1041
11. Tollerson 2nd, R. and Ibba, M. (2020) Translational regulation of environmental adaptation in bacteria. *J. Biol. Chem.* 295, 10434–10445
12. Deutscher, M.P. (2006) Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Res.* 34, 659–666
13. Reck, M. *et al.* (2015) Stool metatranscriptomics: a technical guideline for mRNA stabilisation and isolation. *BMC Genomics* 16, 494
14. Franzosa, E.A. *et al.* (2014) Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* 111, E2329–E2338
15. Fujii, T. *et al.* (2020) Evaluation of DNA and RNA quality from archival formalin-fixed paraffin-embedded tissue for next-generation sequencing - retrospective study in Japanese single institution. *Pathol. Int.* 70, 602–611
16. Gangadoo, S. *et al.* (2021) The multiomics analyses of fecal matrix and its significance to coeliac disease gut profiling. *Int. J. Mol. Sci.* 22
17. Raynaud, S. *et al.* (2018) Selective recovery of RNAs from bacterial pathogens after their internalization by human host cells. *Methods* 143, 4–11
18. Eriksson, S. *et al.* (2003) Unravelling the biology of macrophage infection by gene expression profiling of intracellular *Salmonella enterica*. *Mol. Microbiol.* 47, 103–118
19. Skogholt, A.H. *et al.* (2017) Gene expression differences between PAXgene and Tempus blood RNA tubes are highly reproducible between independent samples and biobanks. *BMC Res. Notes* 10, 136
20. Thatcher, S.A. (2015) DNA/RNA preparation for molecular detection. *Clin. Chem.* 61, 89–99
21. Ali, N. *et al.* (2017) Current nucleic acid extraction methods and their implications to point-of-care diagnostics. *Biomed. Res. Int.* 2017, 9306564
22. Kim, Y.K. *et al.* (2012) Short structured RNAs with low GC content are selectively lost during extraction from a small number of cells. *Mol. Cell* 46, 893–895
23. Westermann, A.J. and Vogel, J. (2021) Cross-species RNA-seq for deciphering host-microbe interactions. *Nat. Rev. Genet.* 22, 361–378
24. Petrova, O.E. *et al.* (2017) Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. *Sci. Rep.* 7, 41114
25. Bhagwat, A.A. *et al.* (2014) Evaluation of ribosomal RNA removal protocols for salmonella RNA-Seq projects. *Adv. Microbiol.* 4, 25–32
26. Giannoukos, G. *et al.* (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* 13, R23
27. Alberti, A. *et al.* (2014) Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* 15, 912
28. Huang, Y. *et al.* (2020) Scalable and cost-effective ribonuclease-based rRNA depletion for transcriptomics. *Nucleic Acids Res.* 48, e20
29. Gu, W. *et al.* (2016) Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* 17, 41
30. Prezza, G. *et al.* (2020) Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. *RNA* 26, 1069–1078
31. Bikel, S. *et al.* (2015) Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Comput. Struct. Biotechnol. J.* 13, 390–401
32. Jang, J.S. *et al.* (2020) Comparative evaluation for the globin gene depletion methods for mRNA sequencing using the whole blood-derived total RNAs. *BMC Genomics* 21, 890
33. Glassing, A. *et al.* (2016) Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* 8, 24
34. Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380
35. Bentley, D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59

36. Stark, R. *et al.* (2019) RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656
37. Shi, H. *et al.* (2021) Bias in RNA-seq library preparation: current challenges and solutions. *Biomed. Res. Int.* 2021, 6647597
38. Costello, M. *et al.* (2018) Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* 19, 332
39. Eid, J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138
40. Jain, M. *et al.* (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 239
41. Ciuffreda, L. *et al.* (2021) Nanopore sequencing and its application to the study of microbial communities. *Comput. Struct. Biotechnol. J.* 19, 1497–1511
42. Tedersoo, L. *et al.* (2021) Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Appl. Environ. Microbiol.* 87, e0062621
43. Marquet, M. *et al.* (2022) Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing. *Sci. Rep.* 12, 4000
44. Martin, S. *et al.* (2022) Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome Biol.* 23, 11
45. Jain, M. *et al.* (2022) Advances in nanopore direct RNA sequencing. *Nat. Methods* 19, 1160–1164
46. Amarasinghe, S.L. *et al.* (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21, 30
47. Workman, R.E. *et al.* (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* 16, 1297–1305
48. Bayega, A. *et al.* (2022) Improved Nanopore full-length cDNA sequencing by PCR-suppression. *Front. Genet.* 13, 1031355
49. Grunberger, F. *et al.* (2022) Nanopore sequencing of RNA and cDNA molecules in *Escherichia coli*. *RNA* 28, 400–417
50. Pust, M.M. *et al.* (2022) Direct RNA nanopore sequencing of *Pseudomonas aeruginosa* clone C transcriptomes. *J. Bacteriol.* 204, e0041821
51. Pitt, M.E. *et al.* (2020) Evaluating the genome and resistome of extensively drug-resistant *Klebsiella pneumoniae* using native DNA and RNA Nanopore sequencing. *GigaScience* 9, giaa002
52. Fang, Y. *et al.* (2022) Nanopore whole transcriptome analysis and pathogen surveillance by a novel solid-phase catalysis approach. *Adv. Sci. (Weinh)* 9, e2103373
53. Yan, B. *et al.* (2018) SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat. Commun.* 9, 3676
54. Yang, M. *et al.* (2020) Direct metatranscriptome RNA-seq and multiplex RT-PCR amplicon sequencing on Nanopore MinION - promising strategies for multiplex identification of viable pathogens in food. *Front. Microbiol.* 11, 514
55. Ojala, T. *et al.* (2021) Metatranscriptomic assessment of burn wound infection clearance. *Clin. Microbiol. Infect.* 27, 144–146
56. Zhang, H. *et al.* (2020) A comprehensive evaluation of long read error correction methods. *BMC Genomics* 21, 889
57. Kopylova, E. *et al.* (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217
58. Deng, Z.L. *et al.* (2022) Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic Acids Res.* 50, e60
59. Schmieder, R. and Edwards, R. (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6, e17288
60. Lloyd-Price, J. *et al.* (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550, 61–66
61. Celaj, A. *et al.* (2014) Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. *Microbiome* 2, 39
62. Breitwieser, F.P. *et al.* (2019) A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* 20, 1125–1136
63. Schoch, C.L. *et al.* (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020, baaa062
64. Finn, R.D. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285
65. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212
66. Chen, L. *et al.* (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* 33, D325–D328
67. Alcock, B.P. *et al.* (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 48, D517–D525
68. Feldgarden, M. *et al.* (2021) AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.* 11, 12728
69. Kanehisa, M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462
70. Sun, Z. *et al.* (2021) Challenges in benchmarking metagenomic profilers. *Nat. Methods* 18, 618–626
71. Breitwieser, F.P. *et al.* (2019) Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* 29, 954–960
72. Wood, D.E. *et al.* (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257
73. Bray, N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527
74. Beghini, F. *et al.* (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* 10, e65088
75. Meyer, F. *et al.* (2022) Critical assessment of metagenome interpretation: the second round of challenges. *Nat. Methods* 19, 429–440
76. Franzosa, E.A. *et al.* (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15, 962–968
77. Jonsson, V. *et al.* (2016) Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* 17, 78
78. Lin, H. and Peddada, S.D. (2020) Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms Microbiomes* 6, 60
79. Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25
80. Love, M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550
81. Lin, H. and Peddada, S.D. (2020) Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* 11, 3514
82. Klingenberg, H. and Meinicke, P. (2017) How to normalize metatranscriptomic count data for differential expression analysis. *PeerJ* 5, e3859
83. Cappellato, M. *et al.* (2022) Investigating differential abundance methods in microbiome data: a benchmark study. *PLoS Comput. Biol.* 18, e1010467
84. Mallick, H. *et al.* (2021) Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput. Biol.* 17, e1009442
85. Fernandes, A.D. *et al.* (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2, 15
86. Liu, T. *et al.* (2020) An empirical Bayes approach to normalization and differential abundance testing for microbiome data. *BMC Bioinforma.* 21, 225
87. Mandal, S. *et al.* (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* 26, 27663
88. Zhang, Y. *et al.* (2021) Statistical approaches for differential expression analysis in metatranscriptomics. *Bioinformatics* 37, i34–i41
89. Lin, Z. *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130

90. Bihl, S. *et al.* (2022) When to suspect contamination rather than colonization - lessons from a putative fetal sheep microbiome. *Gut Microbes* 14, 2005751
91. Wahl, A. *et al.* (2022) Comparison of rRNA depletion methods for efficient bacterial mRNA sequencing. *Sci. Rep.* 12, 5765
92. Soneson, C. *et al.* (2019) A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* 10, 3359
93. Mikheenko, A. *et al.* (2022) Sequencing of individual barcoded cDNAs using Pacific Biosciences and Oxford Nanopore Technologies reveals platform-specific error patterns. *Genome Res.* 32, 726–737
94. Viscardi, M.J. and Arribere, J.A. (2022) Poly(a) selection introduces bias and undue noise in direct RNA-sequencing. *BMC Genomics* 23, 530
95. Browne, P.D. *et al.* (2020) GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *GigaScience* 9, giaa008
96. Gu, W. *et al.* (2019) Clinical metagenomic next-generation sequencing for pathogen detection. *Annu. Rev. Pathol.* 14, 319–338
97. Nearing, J.T. *et al.* (2021) Identifying biases and their potential solutions in human microbiome studies. *Microbiome* 9, 113
98. Salter, S.J. *et al.* (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12, 87
99. Minich, J.J. *et al.* (2019) Quantifying and understanding well-to-well contamination in microbiome research. *mSystems* 4, e00186-19
100. Hornung, B.V.H. *et al.* (2019) Issues and current standards of controls in microbiome research. *FEMS Microbiol. Ecol.* 95, fiz045
101. Davis, N.M. *et al.* (2018) Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6, 226
102. Knights, D. *et al.* (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* 8, 761–763
103. McGhee, J.J. *et al.* (2020) Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics. *PeerJ* 8, e8783
104. Karstens, L. *et al.* (2019) Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. *mSystems* 4, e00290-19
105. Dohlman, A.B. *et al.* (2021) The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe* 29, 281–298
106. Ferdous, T. *et al.* (2022) The rise to power of the microbiome: power and sample size calculation for microbiome studies. *Mucosal Immunol.* 15, 1060–1070
107. Conesa, A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17, 13
108. Poyet, M. *et al.* (2019) A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* 25, 1442–1452
109. Tarazona, S. *et al.* (2015) Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* 43, e140
110. Law, C.W. *et al.* (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15, R29
111. Bolger, A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120
112. Chen, S. *et al.* (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890
113. Lim, Y.W. *et al.* (2013) Metagenomics and metatranscriptomics: windows on CF-associated viral and microbial communities. *J. Cyst. Fibros.* 12, 154–164
114. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12
115. Nawrocki, E.P. and Eddy, S.R. (2013) Computational identification of functional RNA homologs in metagenomic data. *RNA Biol.* 10, 1170–1179
116. Wang, L. *et al.* (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28, 2184–2185
117. Hyatt, D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* 11, 119
118. Grabherr, M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652
119. Robertson, G. *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods* 7, 909–912
120. Bushmanova, E. *et al.* (2019) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience* 8, giz100
121. Leung, H.C. *et al.* (2013) IDBA-MT: de novo assembler for metatranscriptomic data generated from next-generation sequencing technology. *J. Comput. Biol.* 20, 540–550
122. Ye, Y. and Tang, H. (2016) Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. *Bioinformatics* 32, 1001–1008
123. Saggese, I. *et al.* (2018) STAble: a novel approach to de novo assembly of RNA-seq data and its application in a metabolic model network based metatranscriptomic workflow. *BMC Bioinforma.* 19, 184
124. Li, D. *et al.* (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676
125. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359
126. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760
127. Li, H. (2021) New strategies to improve minimap2 alignment accuracy. *Bioinformatics* 37, 4572–4574
128. Dobin, A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21
129. Kim, D. *et al.* (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915
130. Buchfink, B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60
131. Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
132. Fu, L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152
133. Kim, D. *et al.* (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 26, 1721–1729
134. Menzel, P. *et al.* (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7, 11257
135. Milanese, A. *et al.* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* 10, 1014
136. Marcelino, V.R. *et al.* (2020) CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biol.* 21, 103
137. Bagci, C. *et al.* (2021) DIAMOND+MEGAN: fast and easy taxonomic and functional analysis of short and long microbiome sequences. *Curr. Protoc.* 1, e59
138. Ugarte, A. *et al.* (2018) A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. *Microbiome* 6, 149
139. Cantalapiedra, C.P. *et al.* (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38, 5825–5829
140. Robinson, M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140
141. Monleon-Getino, T. and Frias-Lopez, J. (2020) A priori estimation of sequencing effort in complex microbial metatranscriptomes. *Ecol. Evol.* 10, 13382–13394
142. McMurdie, P.J. and Holmes, S. (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8, e61217
143. Segata, N. *et al.* (2011) Metagenomic biomarker discovery and explanation. *Genome Biol.* 12, R60
144. Risso, D. *et al.* (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902

145. Martinez, X. *et al.* (2016) MetaTrans: an open-source pipeline for metatranscriptomics. *Sci. Rep.* 6, 26447
146. Kultima, J.R. *et al.* (2016) MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* 32, 2520–2523
147. Westreich, S.T. *et al.* (2018) SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC Bioinforma.* 19, 175
148. Pruesse, E. *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196
149. Burge, S.W. *et al.* (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* 41, D226–D232