Briefings in Bioinformatics, 19(6), 2018, 1415–1429

doi: 10.1093/bib/bbx051
Advance Access Publication Date: 8 May 2017
Software Review

Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes

Sheng-Yong Niu,* Jinyu Yang,* Adam McDermaid, Jing Zhao, Yu Kang and Qin Ma

Corresponding author: Qin Ma, Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture, and Plant Science, South Dakota State University, Brookings, SD 57006, USA. Tel.: 1-706-254-4293; E-mail: qin.ma@sdstate.edu
*The first two authors should be regarded as joint first authors.

Abstract

Metagenomic and metatranscriptomic sequencing approaches are more frequently being used to link microbiota to important diseases and ecological changes. Many analyses have been used to compare the taxonomic and functional profiles of microbiota across habitats or individuals. While a large portion of metagenomic analyses focus on species-level profiling, some studies use strain-level metagenomic analyses to investigate the relationship between specific strains and certain circumstances. Metatranscriptomic analysis provides another important insight into activities of genes by examining gene expression levels of microbiota. Hence, combining metagenomic and metatranscriptomic analyses will help understand the activity or enrichment of a given gene set, such as drug-resistant genes among microbiome samples. Here, we summarize existing bioinformatics tools of metagenomic and metatranscriptomic data analysis, the purpose of which is to assist researchers in deciding the appropriate tools for their microbiome studies. Additionally, we propose an Integrated Meta-Function mapping pipeline to incorporate various reference databases and accelerate functional gene mapping procedures for both metagenomic and metatranscriptomic analyses.

Key words: metagenome; metatranscriptome; microbe; bioinformatics tools; functional and quantitative analysis; integrated mapping pipeline

Introduction

In addition to the human body, communities of microbes are found in other environments, such as the ocean, soil, plants and other animals [1]. These microbial communities each have their own complexities, diversities and competitions. Together with the environments, members of the communities interact and cooperate with their environments or hosts. Recently, human microbiota is rapidly linked to various diseases, with the development of sequencing technologies enabling the capability of researchers to study genomes of microbiota (a.k.a. microbiome). One of the most critical areas is human microbiome research associated with several well-known human diseases, such as obesity, inflammatory bowel disease (IBD) and lean or obese twins [2, 3]. Furthermore, growing evidence

Sheng-Yong Niu is an undergraduate student in the Department of Biochemical Science and Technology, National Taiwan University, Taiwan.
Jinyu Yang is an MS student in the Department of Mathematics and Statistics at South Dakota State University, Brookings, SD, USA.
Adam McDermaid is a PhD student in the Department of Mathematics and Statistics at South Dakota State University, Brookings, SD, USA.
Jing Zhao is a biostatistician in the Population Health Group at Sanford Research and an assistant professor in the Department of Internal Medicine at University of South Dakota Sanford School of Medicine.
Yu Kang is an associate professor in the Beijing Institute of Genomics of Chinese Academy of Sciences, with a research emphasis of metagenome.
Qin Ma is the director of the Bioinformatics and Mathematical Biosciences Lab and an assistant professor in the Department of Agronomy, Horticulture and Plant Science at South Dakota State University. He is also an adjunct faculty member of the Department of Mathematics and Statistics at South Dakota State University and BioSNTR, SD, USA.
Submitted: 13 February 2017; **Received (in revised form):** 12 April 2017

indicates the microbiota within the human body, especially the intestinal microbiome, plays critical roles in human physiology [4, 5].

Nowadays, various methods are applied to infer different levels of information about microbiome. These methods include 16S ribosomal RNA (rRNA) analysis, whole-genome shotgun (WGS; metagenome) analysis and whole-transcriptome shotgun (metatranscriptome) analysis. Analysis of 16S rRNA uses the conservation of the 16S rRNA gene to identify the microbes. The WGS analysis uses information of all genes to interpret microbial identities down to species or strain level. The whole-transcriptome shotgun analysis allows for the observation of gene expression patterns and functionality of microbial communities. The whole-metabolite analysis provides a comprehensive list of chemicals in the environment of interest and allows correlating abundance of microbes to the downstream chemicals.

Several population-based microbiome studies, for example the Human Microbiome Project [6], focus on the study of microbial communities that inhabit the human body of healthy individuals with emphasis on nasal, oral, skin, gastrointestinal and urogenital areas. Interactive Human Microbiome Project focuses on understanding human–microbiome interactions by longitudinal studies, which gather multiple omics data sets from both the microbiome and human [7]. Additionally, Metagenomics of the Human Intestinal Tract (MetaHIT) focuses on understanding the relationship between human intestinal microbiota and human health/disease [2]. MetaHIT also studies obesity and IBD. Earth Microbiome Project (EMP) focuses on characterizing the diversity, distribution and structure of microbial ecosystems across the planet and has already gathered >30 000 samples from diverse ecosystems including humans, animals, plants terrestrial, marine, constructed environments and many others [8]. EMP is one of the pioneer microbiome projects to set some standard protocols for other microbiome studies.

With the increased limitations in understanding an individual microbe's mechanisms on a global scale and the difficulties associated with culturing individual microbial species, metagenomics and metatranscriptomics have been investigated more frequently in recent studies [9–12]. Acknowledging the importance of microbiome studies, we endeavor to give a comprehensive review of existing metagenomic and metatranscriptomic technologies, specifically the data analysis methods. This review will provide complementary insights of previous reviews and prospective in this field [13, 14]. Additionally, a pipeline for general functional genetic mapping in various reference databases is developed. We hope the provided information will help more researchers in identifying the appropriate tools for their microbiome studies.

Metagenomic data analysis

16S rRNA analysis is one of the most popular and a relatively cheap method to profile the genus composition of microbiota. In this review article, we provide tools for 16S rRNA analysis as follows.

16S rRNA data analysis

The utilization of 16S rRNA gene sequences in understanding microbial taxonomy and phylogeny is one of the most common approaches for multiple reasons. The reasons include (i) the function of the 16S rRNA gene has not altered over time, suggesting random sequence changes are a more precise measure of time or evolution; (ii) 16S rRNA is present in almost all microbes; (iii) it is large enough for informatics purposes [15, 16]. Here, we introduce five bioinformatics tools for 16S rRNA analysis, as shown in Table 1, which are QIIME, UPARSE, MOTHUR, DADA2 and minimum entropy decomposition (MED).

QIIME aims to perform the downstream analyses because of the lack of library demultiplexing and taxonomy assignment tools as the development of high-throughput pyrosequencing [17–19]. UPARSE is developed for constructing Operational Taxonomic Units (OTUs) *de novo* from next-generation reads that achieves high accuracy in biological sequence recovery and improves richness estimates on mock communities [20]. MOTHUR intends to be a comprehensive software package that allows users to use a single piece of software to analyze community sequence data [18]. For example, it was mentioned that the generation of field-wide analysis standards has not been developed, making it difficult to perform meta-analyses. Hence, it is important for MOTHUR to be flexible and easily maintained. DADA2 is a model-based approach for correcting amplicon errors without constructing OTUs. It aims to identify fine-scale variations in 454-sequenced amplicon data while outputting few false positives [21]. The purpose of MED is to solve the constraint of fine-scale resolution descriptions of microbial communities resulting from pairwise sequence alignments for similarity assessment and *de novo* clustering with *de facto* similarity thresholds to partition high-throughput sequencing data sets [22].

Regarding the methods of these tools, QIIME uses the PyCogent toolkit to address the problem of interpretation and database deposition using raw sequencing data [19]. UPARSE works by quality-filtering reads, trimming them to a fixed length, optionally discarding singleton reads and then clustering the remaining reads [20]. MOTHUR implements the algorithms used in previous tools including DOTUR, SONS, TreeClimber, LIBSHUFF, ∫-LIBSHUFF and UniFrac [18]. DADA2 uses a new quality-aware model of Illumina amplicon errors to improve the DADA algorithm [21]. MED iteratively partitions a data set of amplicon sequences into homogenous OTUs that serve as input to alpha- and beta-diversity analyses [22].

QIIME provides visualization functionalities that have been essential for several high-profile studies, including network analyses, histograms of within- or between-sample diversity, graphical displays that allow users to interact with the data, etc. [19]. UPARSE generates OTUs that were reconstructions of sequences representing known species in some samples and that were close to a known biological sequence in others, although a few chimeras and unclassified sequences may be retained [20]. The final product of DADA2 includes a sequence table, analogous to the ubiquitous 'OTU table', which records the number of times each ribosomal sequence variant was observed in each sample (http://benjjneb.github.io/dada2/tutorial.html). MED generates standard output files such as observation matrices, FASTA files for representative sequences and network descriptions. Here, we list and compare five bioinformatics tools for 16S rRNA data analysis [21].

WGS sequencing, which is an important method for microbiome studies, provides an integrated understanding of community structure, genetic population heterogeneity and potential metabolism pathways with relatively low costs, improved time requirements and higher quantities of data. Owing to the rapid advancement of sequencing technologies, WGS-based metagenomic studies are growing sharply [23]. However, exponential growth in sequencing data size generally requires more computational resources and efficient tools for

Table 1. Comparison among 16S rRNA amplicon analysis tools

Tool	QIIME	UPARSE	MOTHUR	DADA2	MED
PubMed	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3249058/	http://www.ncbi.nlm.nih.gov/pubmed/23955772	http://www.ncbi.nlm.nih.gov/pubmed/19801464	http://www.ncbi.nlm.nih.gov/pubmed/27214047	http://www.nature.com/ismej/journal/v9/n4/full/ismej2014195a.html
Core algorithm/tools		1. UPARSE-OUT 2. UPARSE-REF	1. DOTUR 2. SONS 3. TreeClimber 4. LIBSHUFF 5. ∫-LIBSHUFF 6. UniFrac	1. DADA 2. Divisive Amplicon Denoising Algorithm 3. Needleman–Wunsch algorithm	1. MED
Visualization			Venn diagrams, heat maps and dendrograms	plotErrors	betaisper and ADONIS R function in vegan package, Gephi with ForceAtlas2
Short description	QIIME is a pipeline for performing microbiome analysis from raw sequencing data through publication quality graphics and statistics. This includes demultiplexing and quality filtering, OTU picking, taxonomic assignment, phylogenetic reconstruction, visualizations, etc.	UPARSE is an approach for producing clusters (OTUs) from next-generation sequencing reads of marker genes such as 16S rRNA	MOTHUR incorporates the algorithms from previous tools and integrates additional features such as ecological parameters, visualization and screening sequence collections based on quality	DADA2 is an open-source R package that improves the DADA algorithm. DADA2 package implements the full amplicon workflow such as filtering, dereplication, sample inference, chimera identification, merging of paired-end reads, etc.	MED is an algorithm extending the principles of oligotyping to entire marker gene data sets. MED partitions high-throughput sequencing data sets into ecologically meaningful and phylogenetically homogeneous units
Web address	http://qiime.org/	http://drive5.com/uparse/	http://www.mothur.org/wiki/Main_Page	http://benjjneb.github.io/dada2/index.html	http://merenlab.org/software/med/
Availability	Open source. The basic command lines, tutorial and compressed packages are provided	Tutorials	Open source, source code and tutorial		The basic command lines and tutorial
Published date and citation counts	1 December 2011; 167	18 August 2013; 831	2 October 2009; 5605	23 May 2016; 4	17 October 2014; 32
Hardware requirement (OS, RAM, disk space, and CPU)	Mac OS X, Windows or Linux		Mac OS X, Windows or Linux		Mac or Linux

downstream analyses and in-depth interpretation. Thus, species-level and strain-level comparisons of various tools using metagenomic data are conducted in the following sections. The information provided will facilitate the selection of suitable tools meeting specific aims of each study.

Species-level metagenomic data analysis

Six metagenomic analysis tools are listed in Table 2, including MetaPhlAn2 [24], Kraken [25], CLARK [26], FOCUS [27], SUPER-FOCUS [28] and MG-RAST [29]. The core algorithm, required curated database (*de novo*), visualizations, brief introduction, availability, publish date, citation counts, hardware requirement [operating system (OS), random access memory (RAM), disk space and central processing unit (CPU)] are compared.

These tools are used to report the organisms present in metagenomic samples and profile their abundances. Specifically, MetaPhlAn2 uses Bowtie2 [23] and UCLUST [30] as its core algorithms, Kraken applies exact alignment of *k*-mers, CLARK uses reduced sets of *k*-mers (i.e. DNA words of length *k*), FOCUS makes use of nonnegative least squares (NNLS) to report the organisms' profile, SUPER-FOCUS uses a reduced reference database to report the subsystems present in metagenomic data sets and profile their abundances and MG-RAST extracts multiple features for users to assess sequence quality and address some of the common issues (e.g. high error rates,

Table 2. Comparison among species-level metagenome analyses tools

Tool	MetaPhlAn2	Kraken	CLARK	FOCUS	SUPER-FOCUS	MG-RAST
PubMed	http://www.ncbi.nlm.nih.gov/pubmed/26418763	http://www.ncbi.nlm.nih.gov/pubmed/24580807	http://www.ncbi.nlm.nih.gov/pubmed/25879410	http://www.ncbi.nlm.nih.gov/pubmed/24949242	http://www.ncbi.nlm.nih.gov/pubmed/26454280	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2563014/
Core algorithm/tools	1. Bowtie2 [23] 2. UCLUST [30]	1. Exact alignment of *k*-mers	1. Reduced sets of *k*-mers	1. NNLS	1. BLASTX 2. RAPSearch2 3. DIAMOND	1. BLAT [32] 2. FragGeneScan [33] 3. QIIME [19] 4. Bowtie [34] 5. SolexaQA [35]
Require curated database (*de novo*)	1. MetaPhlAn2 Database	1. Kraken database 2. RefSeq 3. NCBI Taxonomy Database	1. CLARK Database 2. RefSeq	1. FOCUS Database	1. Reduced SEED Database (by CD-HIT)	1. M5nr [36] 2. Greengenes [37] 3. SILVA [38] 4. RDP [39] 5. The SEED [40] 6. GenBank [41] 7. RefSeq [42] 8. IMG/M [43] 9. UniProt [44] 10. eggNOG [45] 11. KEGG [46] 12. PATRIC [47]
Visualization	Heatmap, GraPhlAn plots, Krona2 plots and single microbe barplot	Krona Visualization Package		STAMP		Barcharts, tree diagram, spreadsheet-like tables, heatmaps, Principal Coordinate Analysis (PCoA), rarefaction plots and KEGG maps
Short description	MetaPhlAn2 is a pipeline for microbial composition profiling compared with MetaPhlAn. Notably, strain fingerprinting and tracking are also provided by MetaPhlAn2	Kraken is a software usually used in metagenomics studies for assigning taxonomic labels to short DNA sequences	CLARK is a software tool for classifying any type of DNA/RNA sequences in any format (reads, contigs, scaffolds, etc.) based on a supervised sequence classification using discriminative *k*-mers	FOCUS is an approach using NNLS to identify the organisms present in metagenomics samples and profile their abundances	SUPER-FOCUS is a homology-based approach adopting a reduced SEED database to report the subsystems present in metagenomic samples and profile their abundances	MG-RAST is an open-source Web server to profile the function and composition of a microbial community from metagenomic or metatranscriptomic data
Web address	http://segatalab.cibio.unitn.it/tools/metaphlan2/	https://ccb.jhu.edu/software/kraken/	http://clark.cs.ucr.edu/	http://edwards.sdsu.edu/FOCUS/	http://edwards.sdsu.edu/superfocus/	http://metagenomics.anl.gov/
Availability	Open source. The basic command lines, tutorial and compressed packages are provided	Open source. Tutorial, databases, source code are provided	Open source. Tutorial, databases, source code are provided	Open source. Tutorial, databases, source code are provided	Open source. Tutorial, databases, source code are provided	Open source. Just create an account and login

Table 2. (Continued)

Tool	MetaPhlAn2	Kraken	CLARK	FOCUS	SUPER-FOCUS	MG-RAST
Published date and citation counts	29 September 2015; 32	3 March 2014; 240	25 March 2015; 46	5 Jun 2014; 23	9 Oct 2015; 4	19 September 2008; 1643
Hardware requirement (OS, RAM, disk space and CPU)	Unix-based OSs, MacOS or Windows	1. Linux-based OS 2. Disk space >160 GB 3. The disk used to store the database should be locally attached storage. Storing the database on a Network File System partition can cause Kraken's operation to be slow, or to be stopped completely 4. RAM >75 GB	1. 64 bit OS (Linux or Mac) 2. GNU GCC to compile version 4.4 or higher	1. Command line version that works on OS X, Unix and a Graphical User Interface (GUI) for Microsoft Windows users		Not applicable

contamination with adapter sequences, contamination with artificial duplicate reads, etc.) [31].

Both advantages and disadvantages of these tools have been demonstrated in their original papers. Kraken can achieve genus-level sensitivity and precision that are similar to that obtained by the fastest BLAST program, Megablast [48]. However, its memory usage of the default database requires 70 GB, a value that will grow in linear proportion to the number of distinct *k*-mers in the genomic library. Although CLARK can classify short metagenomic reads with high accuracy at multiple taxonomic ranks (i.e. species and genus level), it cannot take advantage of taxonomic tree structures. Besides, as other methods are created to profile metagenome sequences, FOCUS depends on a curated microbial reference genome database to predict a specific genus. If a reference genome is absent, FOCUS will predict the closest available reference. MG-RAST provides post-annotation analysis and visualization directly through the Web interface, or by tools like matR (metagenomic analysis tools for R) [31] that use the MG-RAST API to easily download data from processing pipeline. In addition, MG-RAST is a useful tool for metatranscriptomic analyses, which will not be discussed again in the following section focusing on metatranscriptomic data.

The output results of these tools are similar in many aspects but still have a few key differences. The output of MetaPhlAn2 includes a matrix containing relative abundances of the identified species, genus, family, order, class, phylum and kingdom. In Kraken (http://ccb.jhu.edu/software/kraken/MANUAL.html), each sequence classified by Kraken results in a single line in the output, which contains letter code of classification, sequence ID, taxonomy ID, length of the sequence in bp and Lowest Common Ancestor (LCA) mapping results of each *k*-mer. In CLARK, its results contain hit count in target, length of object, gamma ratio, target that obtained the highest hit count, confidence score, etc. FOCUS can generate the output in STAMP format (http://kiwi.cs.dal.ca/Software/STAMP), which has a graphical interface permitting easy exploration of statistical results and generation of publishable quality plots for inferring the biological relevance of features in a metagenomic profile. Then, relative abundances and microbial profiles can be used to investigate associations of metadata in subsequent applications. For example, principal component analysis (PCA) [49] or machine learning methods have been used to understand or predict the relative abundance in different profiles. MG-RAST can display results in commonly used formats, including bar charts, trees incorporating abundance information, heatmaps and raw abundance tables. The raw or intermediate data can be recovered via downloading pages or within the matR package [31].

Strain-level metagenomic data analysis

A good number of human diseases, such as cancer, are known to be related with a single or a group of microorganisms [3]. In addition, different strains in the same species may have different influences on human health, such as *Escherichia coli* O157:H7 [50], which is a seriously virulent *E. coli* strain. However, most strains in the same species are nonpathogenic. Thus, the identification and profile of microbial strains in the environment and human hosts are crucial to revealing human–microbial interactions [51]. In this circumstance, we underscore the latest bioinformatics tools for strain-level metagenomics analysis. As there is no ubiquitously accepted tool for strain-level analysis, such as the standard pipeline as QIIME in 16S analysis [52], we are eager to introduce the current proposed methods to help users compare and select the most suitable tools based on their

Table 3. Comparison among strain-level metagenome analyses tools

Tool	StrainPhlAn	PanPhlan	Constrains	Sigma	LSA
PubMed	https://www.ncbi.nlm.nih.gov/pubmed/28167665	http://www.ncbi.nlm.nih.gov/pubmed/26999001	http://www.ncbi.nlm.nih.gov/pubmed/26344404	http://www.ncbi.nlm.nih.gov/pubmed/25266224	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4720164
Core algorithm/tools	1. MetaPhlAn2 [24]	1. Bowtie2 2. SAMtools	1. ConStrains	1. Sigma 2. Bowtie2 3. SAMtools	1. LSA 2. Deconvolution algorithm
Require curated database (*de novo*)	1. MetaPhlAn2 database	1. PanPhlan Database 2. Ref. DB 3. KEGG		1. RefSeq	
Visualization	Phylogenetic trees and ordination plots	Heatmap, hierarchical clustering, PCoA plots and network analysis		HTML output	
Short description	StrainPhlAn is a strain-level metagenome profiling tool by profiling microbes from known species with strain-level resolution and producing comparative and phylogenetic analyses of strains retrieved from metagenomic samples	PanPhlAn is a strain-level metagenomic profiling tool for reporting the gene composition and *in vivo* transcriptional activity of individual strains in metagenomics samples	Constrains is an open-source algorithm that using SNP patterns in a set of universal genes to infer within-species structures that represent strains. Constrains identifies conspecific strains from metagenomic sequence data and produces the phylogeny of these strains in microbial communities	Sigma is a sequence similarity-based approach for strain-level identification of genomes from metagenomic analysis. Sigma uses maximum likelihood estimation to estimate the relative abundances and likelihood ratios of each genome, given a list of reference genomes	LSA is a pre-assembly tool for partitioning metagenomic reads. LSA uses streaming SVD and a hyperplane hashing function to find covariance relations between *k*-mers
Web address	http://segatalab.cibio.unitn.it/tools/strainphlan/	http://segatalab.cibio.unitn.it/tools/panphlan/	https://bitbucket.org/luo-chengwei/constrains	http://sigma.omicsbio.org/	http://latentstrainanalysis.readthedocs.io/en/latest/
Availability	Open source. The basic command lines, tutorial and compressed packages are provided	Open source. The basic command lines, tutorial and compressed packages are provided	Open source. The basic command lines, tutorial and compressed packages are provided	Open source. The basic command lines and tutorial	Open source. The basic command lines, tutorial and compressed packages are provided
Published date and citation counts	6 February 2017; 0	21 March 2016; 13	07 September 2015; 25	29 September 2014; 14	14 September 2015; 22
Hardware requirement (OS, RAM, disk space and CPU)	MacOS or Linux platforms	Linux or Ubuntu		Linux or Unix	

own specific requirements. We introduce five tools for strain-level metagenome analysis as shown in Table 3, which are StrainPhlAn [53], PanPhlan [54], Constrains [55], Sigma [56] and LSA [57]. They aim to identify microbes and characterize their functional potential, both of which are essential for pathogen discovery, epidemiology, population genomics and biosurveillance [53].

Various read mapping techniques have been applied in the tools shown in Table 3. StrainPhlAn uses MetaPhlAn2 to map the given reads to the MetaPhlAn2 marker database. PanPhlan uses Bowtie2 and SAMtools [58] in its metagenomic read mapping, quality filtering and per-base coverage calculating. Constraint, which is an open-source algorithm that identifies conspecific strains from metagenomic sequence data, applies single-nucleotide polymorphism (SNP) patterns on a set of universal genes to infer within-species structures that represent strains. The Sigma algorithm uses read mapping approach with a probabilistic model for sampling reads with mismatches from genomes at unknown abundances. LSA uses a deconvolution algorithm to identify clusters of hashed *k*-mers that represent potential variables.

Various types of output are provided by these tools. StrainPhlAn generates subspecies OTUs and sequence alignment with visualization of the results using several common

Table 4. Comparison among different metatranscriptome analyses tools

Tool	Leimena-2013	HUMAnN2	MetaTrans	SAMSA
PubMed	https://www.ncbi.nlm.nih.gov/pubmed/23915218		https://www.ncbi.nlm.nih.gov/pubmed/27211518	https://www.ncbi.nlm.nih.gov/pubmed/27687690
Core algorithm/tools	1. SortMeRNA 2. BLASTN [66] 3. MegaBLAST 4. KAAS [67]	1. Bowtie2 [23] 2. MetaPhlAn2 [24] 3. MinPath [68] 4. DIAMOND [69]	1. Kraken [25] 2. SortMeRNA [70] 3. UCLUST [30] 4. SOAP2 [71] 5. FragGeneScan [33] 6. DESeq2 [72]	1. Trimmomatic [73] 2. FLASH [74] 3. MG-RAST [29] 4. DESeq2 [72]
Require curated database (*de novo*)	1. SILVA [38] 2. COG [75] 3. MetaHIT [2] 4. Human small intestinal metagenome database [76] 5. KEGG [77]	1. UniRef 2. MetaCyc [78] 3. ChocoPhlAn	1. SILVA-115 [79] 2. Greengenes-13.5 [37] 3. Rfam-11.0 [80] 4. tRNA-all [81] 5. PhiX genome 6. MetaHIT-2014 [82] 7. M5nr-20130801[36]	1. NCBI Reference Database [83] 2. SEED Subsystems reference database [84]
Visualization		PCoA plots, abundance plot of pathway	Metabolic pathways network (iPath2 tool), PCA plot	Barplots and dendograms
Short description	This comprehensive metatranscriptome analysis pipeline is designed for function assignment and mapping based on given RNA-seq data. The performance of this pipeline has been evaluated using human small intestine microbiota data	HUMAnN2 is a pipeline equipped with community function profiles, expanded databases, simple user interface and accelerated reads mapping for microbial pathways profiling	MetaTrans is an open-source pipeline that integrates quality control, rRNA removal and read mapping for taxonomic and gene expression analysis	SAMSA is a comprehensive pipeline for metatranscriptome analyses by working with MG-RAST, providing an ability to fully analyze the expression activity within microbial communities
Web address		http://huttenhower.sph.harvard.edu/humann2	http://www.metatrans.org/	https://github.com/transcript/SAMSA
Availability		Open source. Tutorial, databases, source code and demos are provided	Open source. Executable scripts, databases and third-party tools are provided	Open source. Tutorial and executable programs are provided
Published date and citation counts	2 August 2013; 53	TBD	23 May 2016; 10	29 March 2016; 0
Hardware requirement (OS, RAM, disk space and CPU)		Linux or Mac. RAM ≥16 GB, disk space ≥10 GB	Linux x86_64 bits. At least 16 GB RAM and 10 CPUs are recommended. In addition, 3 GB disk space is required for software and 66.1 GB for databases	Unix/Mac OS/Windows environment. RAM ≥8 GB

tools, including ETE, GraPhlAn and ggtree. PanPhlan yields a matrix that includes the presence/absence profile of each gene family for all strains detected in samples. Constraint outputs a table of position on the universally conserved genes, reference base defined by the phylophlan seed DB [59] and genotypes of each strain within the species (https://bitbucket.org/luo-chengwei/constrains). Sigma provides genome alignment results, estimated relative abundances and percentage chances of genomes in the HTML format for result visualization and text formatting for further data analysis. LSA uses a hyperplane hashing function and streaming singular value decomposition (SVD) to generate matrices of covariance relations between *k*-mers. Although some tools, for example PanPhlan, do not include abundance estimation, users can still apply them with complementary software to understand microbial profile and abundances for future analysis.

Metatranscriptomic data analysis

Although metagenomic analysis has achieved an unprecedented performance in microbial community profiling, it only allows researchers to access the composition of microbial communities [60]. Conversely, metatranscriptomic analysis provides complementary insight into the gene expression profiles and even regulatory mechanisms, which will significantly contribute to drug discovery and human health [61]. Meanwhile, several efficient pipelines and Web servers for metatranscriptomic analysis have been developed in the past few years [29, 62–65]. In this section, we introduce and summarize four metatranscriptomics analysis tools including Leimena-2013 (the pipeline does not have a specific name, and we use the first author's name followed by the publishing year to represent it), HUMAnN2, MetaTrans and SAMSA as shown in Table 4.

Another tool, MG-RAST, can be used to analyze metatranscriptome data sets and has been summarized in Table 2.

HUMAnN2 is designed for the analysis in both metagenomics and metatranscriptomics and accelerating the functional profiling and translated searches by using MetaPhlAn2, ChocoPhlAn pangenome database and DIAMOND [69]. Leimena-2013 uses SortMeRNA and BLASTN in transfer RNA (tRNA) for reads removal and alignment, uses MegaBLAST in mRNA reads assignment and classifies the assignment bit scores to predict the phylogenetic origin [65]. MetaTrans uses multi-threading computers to accelerate the taxonomic and gene expression analyses of active microbial communities [62]; SAMSA is a comprehensive pipeline including four phases: preprocessing phase, annotation phase, aggregation phase and analysis phase for gut microbiome data analysis [63].

Citation of metagenome and metatranscriptome analysis tools

To demonstrate the usage of existing metagenome and metatranscriptome analysis tools, we have collected the citing information of 14 tools from Google Scholar, and their relative annual citations are showcased in Figure 1. It is obvious that MOTHUR [18] and MG-RAST are the two most popular tools in 16s rRNA-level and species-level metagenome analyses tools, respectively. In strain-level metagenome analyses, PanPhlan, Constrains, Sigma and LSA share most of the citations. No metatranscriptome analyses tools are included in this figure, as almost all of them are published after 2015, and the citing information might have bias in the first 2 years.

Functional annotation analysis

The state-of-the-art technologies

Once the complicated tasks of read mapping and assembly have been accomplished, the remaining immense challenge is related to functionally annotating the genes derived from the previous analysis. Owing to the inherent large size of metatranscriptomic data, there are vast computational requirements for functional annotation [85]. Within the radar of functional annotation of genes, there are two main approaches: stand-alone tools and Web-based servers. Each of these two approaches has significant drawbacks. For the stand-alone annotation tools, the downfall comes in the form of enormous computational requirements for any level of accuracy. The nature of these tools requires homology searches of the metadata sets against other databases before functional analysis can be conducted [85]. The current alternative, Web-based servers, also has a significant bottleneck in the form of size limitations on data being uploaded to the server. Even some of the best available programs, such as MG-RAST, have the issue of size limitations that do not allow for analyzing large data sets. However, the recent development of new stand-alone tools can overcome the accuracy/computational requirement trade-off. One such program, COGNIZER, has been shown to perform at a highly accurate level while not requiring the computationally intensive steps of many other stand-alone programs [85]. Although there may still be plenty of room for improvement with these types of programs, COGNIZER provides a pathway to success that can be examined and modified for further improvement.

Prospective

Once the annotation of the metatranscriptomic data has been conducted, there are a variety of different analyses that can be performed. Differential expression analysis can illustrate how different microbial communities adapt to different environments, such as between two different human hosts or between two drastically different extreme habitats. This allows for an understanding of which genes or gene families are being expressed at what levels to help determine evolutionary steps from members of the same species. Functional enrichment analysis can also be performed once annotation is completed [86]. This analysis, which is sometimes referred to as gene set enrichment, aims to assign functional determination for specific genes of interest. These analyses allow for an understanding of how various environmental factors and even host genome can affect gene expression and consequently functional expression of various genes [87]. More information regarding metatranscriptomic data assembly and specific applications can be found in the Supplementary Materials file.

Integration of metagenomic and metatranscriptomic approaches

One of the final destinations of microbiome research is to figure out the community responses to environmental changes. As RNAs surely respond more rapidly than DNAs, metatranscriptome will give more strong signals than metagenome data for the responsible pathways. Understanding the role of a microbial community within a microbiome goes further than any single analysis. Incorporating metagenomic and metatranscriptomic approaches to a single microbial community can give a comprehensive understanding of the microbiome and its reaction to the environment, and provide vast amounts of information that are not available through any current metagenomic or metatranscriptomic pipeline. Incorporating species-level metagenomic tools in the pipeline allows for the understanding of what species are being observed within the microbiome, while strain-level metagenomic analyses provide insight into what specific strains are present. This information is important to the analysis, but it does not provide the whole story of the community. However, the inclusion of metatranscriptomic analysis tools within the pipeline allows for a further understanding of the gene expression levels present in the identified species. As many of the organisms within a microbial community work in conjunction with the other organisms in the community, it is important to understand not only what species are present but also what the expression levels of genes or gene families are within the community. The integration of these two approaches allows for a much fuller understanding of not only the component species present but the microbial community as a whole.

A case study: IMF mapping pipeline

Motivation

To understand the role of microbiome, we need to investigate not only what and how many organisms are composed of microbiota but also the biological functions using the microbiome data, including metagenomic and metatranscriptomic sequencing data. Depending on the functions of interest, we can use different compiled gene sets to mine their functional profiles and activities. Currently, there are diverse and abundant gene sets in public domain, including antibiotic-resistant genes [88, 89], drug target genes [90, 91], human homologs [92] and virulence genes [47, 93].

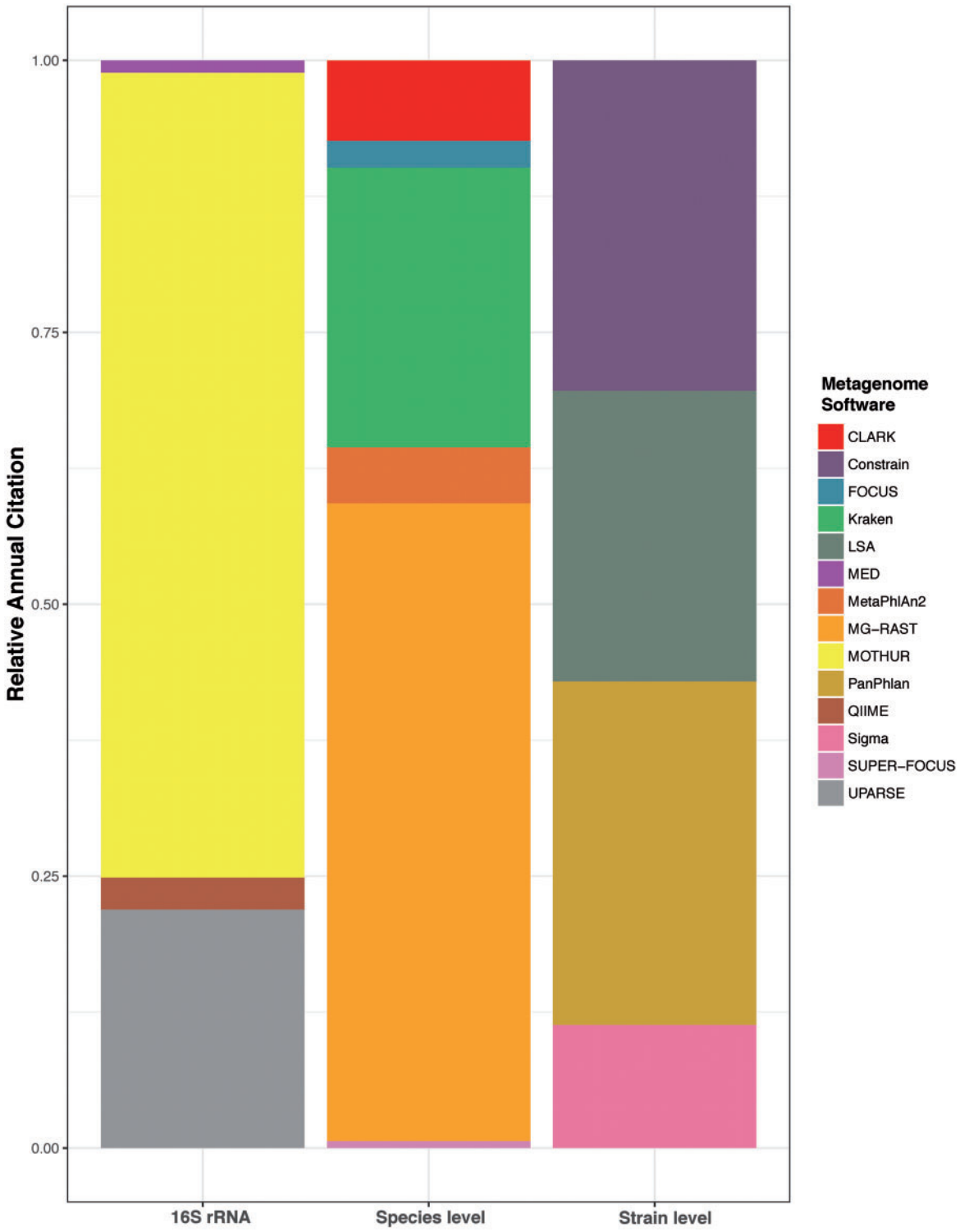

Figure 1. Relative annual citation information of 14 tools according to the numbers from Google Scholar.
Note: All the tools summarized in this figure must be published for at least 2 years.

With a gene set along with matching metagenome and metatranscriptome data, we can understand microbiota's role in diseases or important biological mechanisms. However, there is a missing gap for detection and mapping of these data sets by general mapping pipeline. We have developed an IMF mapping pipeline to incorporate various databases, e.g. DrugBank [91], Kyoto Encyclopedia of Genes and Genomes (KEGG) [94], Comprehensive Antibiotic Resistance Database (CARD) [88], Pathosystems Resource Integration Center (PATRIC) [95], Virulence Factor DataBase [93], Antibiotic Resistance Genes Database [89] and Therapeutic Target Database [96], for efficient mapping of metagenomic and metatranscriptomic data sets together (see the flowchart in Figure 2A).

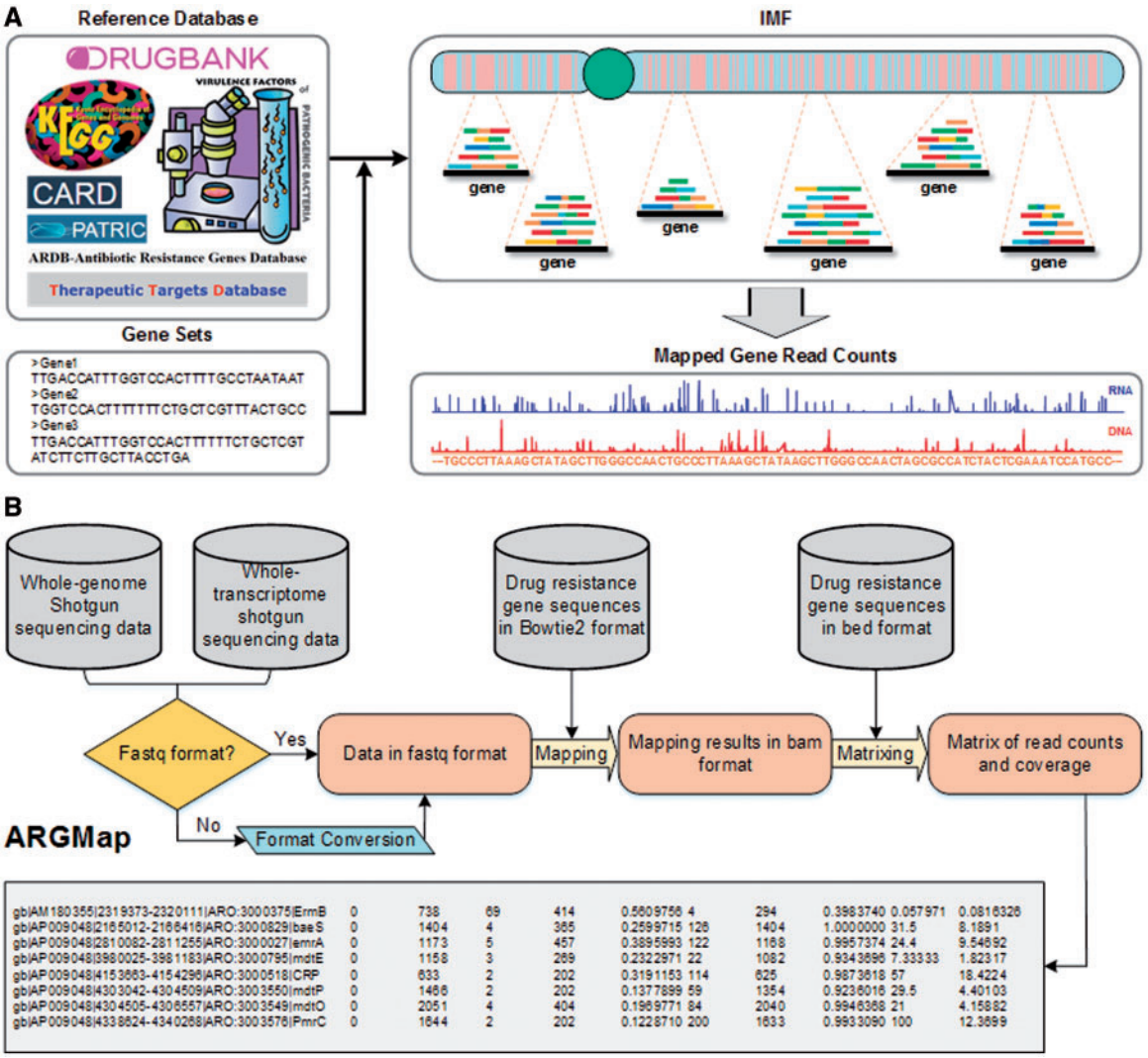

Figure 2. (**A**) Workflow of the IMF pipeline. IMF uses reference databases, e.g. DrugBank, KEGG, CARD, PATRIC, Virulence Factor DataBase, Antibiotic Resistance Genes Database and Therapeutic Target Database, to map with input gene sets. It can produce mapped DNA and RNA read counts for each of the given genes, in support of other downstream analyses. (**B**) Flow chart of pipeline construction of ARGMap. It takes metagenomic or metatranscriptomic sequencing data pair-ended file in FASTQ format as input files. If the input files are not in FASTQ format, user should convert them into the FASTQ format. For example, if the original formats are in BAM format, then user should use the function 'bamToFastq' in Bedtools to convert them into FASTQ formats. Our pipeline will download CARD database by default. User will obtain CARD reference database in FASTA format in the CARD directory. Then, it will use Bowtie2 tool to map between the CARD reference database and input files to generate mapping results in BAM format. Finally, it will use Bedtools to generate read counts tables.

Here, we show an application of this tool, antibiotic resistance gene mapping pipeline (ARGMap, https://github.com/s18692001/ARGMap), for investigating gene content and gene expression of antibiotic-resistant genes in metagenomic and metatranscriptomic data (Figure 2B).

Methods and results

In the ARGMap pipeline, our example input metagenomic and metatranscriptomic sequencing data (SRR769533 and SRR769401, respectively) were downloaded from NCBI SRA database as of 23 January 2017 [97]. The antibiotics-resistant genes will be automatically downloaded from the CARD database (CARD's v.1.1.0 collection). Our primary scripts for execution are drug_resistance_pipeline.sh and pipeline_config.sh. The former file contains all necessary pipeline shell commands, and the latter is for pathway configuration purpose. In the drug_resistance_pipeline.sh file, we first installed the CARD reference database by default for ARG mapping. Second, it uses Bowtie2 to map our metagenomic or metatranscriptomic sequencing data (user could choose input file by editing the pathway in the configuration file) with CARD reference database and generate mapping results in BAM format. Third, it uses Bedtools to generate a table/matrix of (normalized) read counts and coverage over each gene in both DNA and RNA levels. Finally, users can obtain the mapping results in BAM and SAM formats and the final read count table in output directory by default. More details of the data collection and methods could be found in the Supplementary file. A WGS sequencing data set (SRR769533, 3.1 GB) and a metatranscriptome data set (SRR769401, 2.5 GB) were used to test our pipeline on a machine with two CPUs (12 cores, 2.93 GHz). It took around 50 min to generate all the final results in Figure 2B. A more comprehensive comparison with other tools, in terms of both efficiency and accuracy performance, will be carried out in the future based on large-scale benchmark data sets.

The output files of the ARGMap pipeline include mapped results in BAM and SAM formats, and a read count table containing (normalized) read counts and coverage of each gene in both RNA and DNA levels. In each row of the read count table, as shown in Figure 2B, we have 11 columns: (i) ID of an antibiotic resistant gene, (ii–iii) start and end positions of the gene, (iv) number of DNA reads mapped to the gene, (v–vi) number of nucleotides of the gene covered by DNA reads and fraction of the gene covered by DNA reads, (vii) number of RNA reads mapped to the gene, (viii–ix) number of nucleotides of the gene covered by RNA reads and fraction of the gene covered by RNA reads, (x) a normalized score of (column-vii/column-iv, and (xi) a normalized score of (column-vii/column-ix)/(column-iv/column-vi). The full table can be found in Supplementary Table S1 and the output file for the metagenomic and metatranscriptomic read mapping results can be found at https://github.com/s18692001/ARGMap. To gain an insight of species-level composition, we carried out an overall profiling of our WGS sequencing example using MetaPhlAn2. The abundance heatmap and cladogram of the identified species have been showcased in Supplementary Figures S1 and S2. In addition, we provide the species, containing the antibiotic-resistant genes in Table S1, based on the built-in species annotation in the CARD database (see details in Table S2). *Bacteroides fragilis* is identified as the most enriched species among all the identified 19 species (Figure S3), which is reported as a part of the normal flora of the human colon but can cause infection if displaced into the bloodstream or surrounding tissue following surgery, disease or trauma.

Intuitively, there will be a large portion of ambiguous reads when short reads in a metagenome sample are mapped back to a set of homologous genes, leading to various bias in downstream functional analysis. We plan to build a *k*-mer database for all the homologs of an ARG, which can capture some unique features in each of the homologous gene sets, and then map the sequenced reads back to the *k*-mer database to accurately retrieve the species and strain information in this pipeline.

Applications and broader impact

To speed up the detection and mapping procedures of metagenomics and metatranscriptomics data sets, we are eager to accelerate the procedures using our proposed pipeline instead of traditional time-consuming analysis pipelines, aiming to support some specific gene-level interpretation. As shown in Figure 2A, the IMF pipeline could not only be the mapping pipeline for antibiotic-resistant genes but also be an efficient mapping tool in general for other important gene sets, e.g. virulence factors, drug targets and human homologs. Users could easily apply different reference databases based on their usage purposes, for instance, virulence factor mapping with drug target gene mapping with DrugBank to obtain an overview of target gene reads.

It is noteworthy that this ARGMap analysis can fully support pharmacogenetics studies and will be completely developed in our computational pipeline in the future. Users could use the read counts information for further analysis, especially when they have multiple metagenome and metatranscriptome data sets corresponding to various samples/conditions. In such a situation, an integrated read count matrix could be generated, with each row representing a specific gene and each column representing a specific sample. Each element of this matrix could be a normalized score from metagenomic and metatranscriptomic read counts (column x–xi in generated read count table), representing a normalized RNA expression level of a target gene in the corresponding sample. Then, biclustering methods [98, 99] can be applied to such a matrix and used to identify drug-resistant genes sharing similar activities among multiple data sets, leading to systematic regulatory mechanism elucidation in the metagenome level.

Conclusion and discussion

With the growing importance of metagenomic and metatranscriptomic analysis for microbial profiles and abundances, we review the latest bioinformatics tools in this article. We first collect information from the publications and official Web sites of these tools and compare their differences in terms of methods, results, applications, etc. In the metagenome data analysis tool sections, because WGS sequencing provides a comprehensive understanding of community structure, genetic population heterogeneity and potential metabolism pathway with relatively lower-cost and higher-throughput data than in the past, the metagenomics tools emerge rapidly. Furthermore, instead of species-level analysis, strain-level metagenomics analysis is of crucial importance resulting from the fact that different strains in the same species may have drastically different influences on human health. As opposed to metagenomic approaches, metatranscriptomics data analysis could provide insights into the gene expression profiles, or even regulatory mechanisms, which will significantly contribute to new drug discovery and human health. In addition, in this article, to accelerate the mapping procedures from current metagenomics and metatranscriptomics tools, we provide an IMF mapping pipeline to incorporate various databases.

In metagenomics data analysis, various algorithms have been developed to identify taxonomic profile of microbial community. The following algorithms are most often used: (i) *K*-mer approach: compares *k*-mer frequency profiles with those of organisms representing a wide range of clades. This approach is used by Kraken and CLARK. (ii) Marker gene approach: metagenomic reads are aligned with preselected clade-specific marker genes, and taxonomic classification is inferred from phylogenetic distances to marker genes. This approach is used by MetaPhlAn. Both approaches can produce high-quality results at species level or above. However, these two methods may not be proper for strain-level classification. As has been discussed, however, it is of great importance to resolve different strains within the same species. Alternatively, (iii) the read mapping approach: search metagenomics reads against a database of reference genomes. This approach can provide high-resolution taxonomic classification represented by reference genomes, which is used by Sigma. It is suitable for known pathogens that generally have reference genomes available for various strains. However, it is hard to detect novel pathogens with this method. If the new strain is not present in current database, its related strains need to be identified and their sequence variations need to be determined [56]. Hence, in current approach of metagenomic data analysis for microbial taxonomic profile, identification of novel strains absent from existing database is one of the recent bottlenecks. In addition, although some tools for metatranscriptomic data analysis, such as MetaTrans, can operate paired-end RNA-Seq analysis and be adaptable to other high-throughput experiments, current approaches for metatranscriptomics analysis depend on restricted databases and a dedicated computing cluster, or metagenome-based methods that have not been fully evaluated for metatranscriptomic data sets [65].

It is of note that many of the metatranscriptomic processes that have been developed and evaluated focus mainly on prokaryotic bacterial communities [100], and some studies focus on

analyses of interaction between microbiome and its host [101]. As eukaryotic species make up a significant portion of the biological world, there is a wide area of need within the field of metatranscriptomics with respect to the analysis of eukaryotic communities. Without processes that are designed to use metatranscriptomic analyses to eukaryotic communities, entire taxa of organisms, such as fungi, cannot be optimally studied. This fact makes it difficult to understand the happenings within soil-based microbial communities in which fungi are a main component [100]. The main reason for the apparent void in programs developed for eukaryotic metatranscriptomic data analysis is the complexity involved. With metatranscriptomics being a relatively new field, the initial analyses focus on much simpler prokaryotic organisms for refinement. As with many of approaches, once the field of metatranscriptomics continues to develop, better approaches designed for eukaryotic communities will be developed as well.

Recently, there has been much evidence to suggest the inclusion of additional analyses to more fully understand the mechanisms behind microbial communities [102–107]. These analyses, including whole-metabolite analysis and metaproteomic data analysis, provide another dimension to the biological processes involved in microbiomes. While it is true that integrating these additional types of omic data analysis would aid in elucidating a complex biological system, we are inclined to focus on these additional aspects in future publications because of the requisite amount of information to properly discuss these topics. It is our hope to integrate these analyses within our IMF pipeline in the near future.

Key Points

- This article is to give a comprehensive review of existing technologies, especially the computational data analysis methods, for metagenomic and metatranscriptomic studies.
- MOTHUR is the most popular tool, with comprehensive functionalities, to analyze community sequence data in the 16S rRNA level. MetaPhlAn2 appears to be a good complete software that can be used for species-level analysis; alternatively, the MG-RAST Web server is a good option for those seeking a Web-based program to provide similar analysis. StrainPhlAn appears to be an acceptable option for performing strain-level analysis because of its foundation in the MetaPhlAn2 software and a variety of available visual outputs for interpreting the results of the analysis.
- Analyses of metatranscriptomic data could provide insights into gene expression profiles and even regulatory mechanisms. MetaTrans would be a preferred program for use because of its specific design for metatranscriptomic data. However, if the resources are not available to support MetaTrans, then HUMAnN2 is an acceptable alternative and can perform either metatranscriptomic or metagenomic data analyses.
- Integration of metagenomic and metatranscriptomic tools into a single pipeline is important to understand what species and what strain of the specific species are present and what genes are expressed at given levels. This combination will allow for a more comprehensive understanding of the microbial community of interest; hence, we developed an IMF mapping pipeline for general functional genetic mapping in diversified reference databases.
- An ARGMap (https://github.com/s18692001/ARGMap) is provided, for understanding gene content and expression of antibiotics-resistant genes in metagenomic and metatranscriptomic data.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The authors would like to thank Qingren Meng from Beijing Institute of Genomics of Chinese Academy of Sciences for his technical assistance in IMF case study using MetaPhlAn2. Especially, the authors want to thank Dr Wen-Chi Chou from Broad Institute of MIT and Harvard for his valuable comments and consulting.

Funding

This work was supported by the State of South Dakota Research Innovation Center, the Agriculture Experiment Station of South Dakota State University. Support for this project was also provided by the Sanford Health - South Dakota State University Collaborative Research Seed Grant Program. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

References

1. Shade A, Peter H, Allison SD, *et al*. Fundamentals of microbial community resistance and resilience. *Front Microbiol* 2012;**3**:417.
2. Qin J, Li R, Raes J, *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;**464**:59–65.
3. Turnbaugh PJ, Hamady M, Yatsunenko T, *et al*. A core gut microbiome in obese and lean twins. *Nature* 2009;**457**:480–4.
4. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**:207–14.
5. Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, *et al*. A catalog of reference genomes from the human microbiome. *Science* 2010;**328**:994–9.
6. Aagaard K, Petrosino J, Keitel W, *et al*. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *FASEB J* 2013;**27**:1012–22.
7. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* 2014;**16**:276–89.
8. Larsen PE, Field D, Gilbert JA. Predicting bacterial community assemblages using an artificial neural network approach. *Nat Methods* 2012;**9**:621–5.
9. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 2004;**68**:669–85.
10. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 2004;**38**:525–52.

11. Streit WR, Schmitz RA. Metagenomics–the key to the uncultured microbes. *Curr Opin Microbiol* 2004;**7**:492–8.
12. Handelsman J, Rondon MR, Brady SF, *et al*. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 1998;**5**:R245–9.
13. Teeling H, Glockner FO. Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective. *Brief Bioinform* 2012;**13**:728–42.
14. Prakash T, Taylor TD. Functional assignment of metagenomic data: challenges and applications. *Brief Bioinform* 2012;**13**:711–27.
15. Patel JB. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Mol Diagn* 2001;**6**:313–21.
16. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol* 2007;**45**:2761–4.
17. Cole JR, Wang Q, Cardenas E, *et al*. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 2009;**37**:D141–5.
18. Schloss PD, Westcott SL, Ryabin T, *et al*. Introducing MOTHUR: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009;**75**:7537–41.
19. Caporaso JG, Kuczynski J, Stombaugh J, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;**7**:335–6.
20. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 2013;**10**:996–8.
21. Callahan BJ, McMurdie PJ, Rosen MJ, *et al*. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;**13**:581–3.
22. Eren AM, Morrison HG, Lescault PJ, *et al*. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* 2015;**9**:968–79.
23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.
24. Truong DT, Franzosa EA, Tickle TL, *et al*. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015;**12**:902–3.
25. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;**15**:R46.
26. Ounit R, Wanamaker S, Close TJ, *et al*. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 2015;**16**:236.
27. Silva GG, Cuevas DA, Dutilh BE, *et al*. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* 2014;**2**:e425.
28. Silva GG, Green KT, Dutilh BE, *et al*. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics* 2016;**32**:354–61.
29. Meyer F, Paarmann D, D'Souza M, *et al*. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008;**9**:386.
30. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;**26**:2460–1.
31. Keegan KP, Glass EM, Meyer F. MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol Biol* 2016;**1399**:207–33.
32. Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res* 2002;**12**:656–64.
33. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 2010;**38**:e191.
34. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
35. Cox MP, Peterson DA, Biggs PJ. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 2010;**11**:485.
36. Wilke A, Harrison T, Wilkening J, *et al*. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics* 2012;**13**:141.
37. DeSantis TZ, Hugenholtz P, Larsen N, *et al*. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006;**72**:5069–72.
38. Pruesse E, Quast C, Knittel K, *et al*. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007;**35**:7188–96.
39. Cole JR, Chai B, Marsh TL, *et al*. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res* 2003;**31**:442–3.
40. Overbeek R, Begley T, Butler RM, *et al*. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 2005;**33**:5691–702.
41. Benson DA, Cavanaugh M, Clark K, *et al*. GenBank. *Nucleic Acids Res* 2013;**41**:D36–42.
42. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007;**35**:D61–5.
43. Markowitz VM, Ivanova NN, Szeto E, *et al*. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* 2008;**36**:D534–8.
44. Magrane M, UniProt C. UniProt Knowledgebase: a hub of integrated protein data. *Database* 2011;**2011**:bar009.
45. Jensen LJ, Julien P, Kuhn M, *et al*. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 2008;**36**:D250–4.
46. Kanehisa M. The KEGG database. *Novartis Found Symp* 2002;**247**:91–101. discussion 101-103, 119-128, 244-152.
47. Snyder EE, Kampanya N, Lu J, *et al*. PATRIC: the VBI PathoSystems Resource Integration Center. *Nucleic Acids Res* 2007;**35**:D401–6.
48. Morgulis A, Coulouris G, Raytselis Y, *et al*. Database indexing for production MegaBLAST searches. *Bioinformatics* 2008;**24**:1757–64.
49. Dinsdale EA, Edwards RA, Bailey BA, *et al*. Multivariate analysis of functional metagenomes. *Front Genet* 2013;**4**:41.
50. Karch H, Tarr PI, Bielaszewska M. Enterohaemorrhagic *Escherichia coli* in human medicine. *Int J Med Microbiol* 2005;**295**:405–18.
51. Tu Q, He Z, Zhou J. Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Res* 2014;**42**:e67.
52. Brito IL, Alm EJ. Tracking Strains in the microbiome: insights from metagenomics and models. *Front Microbiol* 2016;**7**:712.
53. Truong DT, Tett A, Pasolli E, *et al*. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res* 2017;**27**:626–38.
54. Scholz M, Ward DV, Pasolli E, *et al*. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods* 2016;**13**:435–8.

55. Luo C, Knight R, Siljander H, *et al*. ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol* 2015;**33**:1045–52.
56. Ahn TH, Chai J, Pan C. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* 2015;**31**:170–7.
57. Cleary B, Brito IL, Huang K, *et al*. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat Biotechnol* 2015;**33**:1053–60.
58. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009; **25**:2078–9.
59. Segata N, Bornigen D, Morgan XC, *et al*. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 2013;**4**:2304.
60. Simon C, Daniel R. Metagenomic analyses: past and future trends. *Appl Environ Microbiol* 2011;**77**:1153–61.
61. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of metatranscriptomics in microbiome research. *Bioinform Biol Insights* 2016;**10**:19–25.
62. Martinez X, Pozuelo M, Pascal V, *et al*. MetaTrans: an open-source pipeline for metatranscriptomics. *Sci Rep* 2016; **6**:26447.
63. Westreich ST, Korf I, Mills DA, *et al*. SAMSA: a comprehensive metatranscriptome analysis pipeline. *BMC Bioinformatics* 2016;**17**:399.
64. Abubucker S, Segata N, Goll J, *et al*. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 2012;**8**:e1002358.
65. Leimena MM, Ramiro-Garcia J, Davids M, *et al*. A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genomics* 2013;**14**:530.
66. Johnson M, Zaretskaya I, Raytselis Y, *et al*. NCBI BLAST: a better web interface. *Nucleic Acids Res* 2008;**36**:W5–9.
67. Moriya Y, Itoh M, Okuda S, *et al*. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007;**35**:W182–5.
68. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol* 2009;**5**:e1000465.
69. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.
70. Kopylova E, Noe L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012;**28**:3211–17.
71. Li R, Yu C, Li Y, *et al*. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 2009;**25**:1966–7.
72. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
73. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.
74. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 2011; **27**:2957–63.
75. Tatusov RL, Galperin MY, Natale DA, *et al*. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000;**28**:33–6.
76. Zoetendal EG, Raes J, Van Den Bogert B, *et al*. The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *ISME J* 2012;**6**:1415–26.
77. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30.
78. Caspi R, Altman T, Billington R, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 2014;**42**:D459–71.
79. Quast C, Pruesse E, Yilmaz P, *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–6.
80. Burge SW, Daub J, Eberhardt R, *et al*. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* 2013;**41**:D226–32.
81. Chan PP, Lowe TM. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* 2009;**37**:D93–7.
82. Zan Y, Wu J, Li P, *et al*. SICR rumor spreading model in complex networks: counterattack and self-resistance. *Physica A* 2014;**405**:159–70.
83. Tatusova T, Ciufo S, Fedorov B, *et al*. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* 2014;**42**:D553–9.
84. Overbeek R, Olson R, Pusch GD, *et al*. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 2014;**42**:D206–14.
85. Bose T, Haque MM, Reddy C, *et al*. COGNIZER: a framework for functional annotation of metagenomic datasets. *PLoS One* 2015;**10**:e0142102.
86. Bao G, Wang M, Doak TG, *et al*. Strand-specific community RNA-seq reveals prevalent and dynamic antisense transcription in human gut microbiota. *Front Microbiol* 2015; **6**:896.
87. Wu M, McNulty NP, Rodionov DA, *et al*. Genetic determinants of *in vivo* fitness and diet responsiveness in multiple human gut Bacteroides. *Science* 2015;**350**:aac5992.
88. McArthur AG, Waglechner N, Nizam F, *et al*. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 2013;**57**:3348–57.
89. Liu B, Pop M. ARDB–Antibiotic Resistance Genes Database. *Nucleic Acids Res* 2009;**37**:D443–7.
90. Yang H, Qin C, Li YH, *et al*. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res* 2016;**44**:D1069–74.
91. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**:D668–72.
92. 1000 Genomes Project Consortium, Abecasis GR, Auton A, *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.
93. Chen L, Yang J, Yu J, *et al*. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 2005; **33**:D325–8.
94. Kanehisa M, Furumichi M, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.
95. Wattam AR, Abraham D, Dalay O, *et al*. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 2014;**42**:D581–91.
96. Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. *Nucleic Acids Res* 2002;**30**:412–15.
97. Franzosa EA, Morgan XC, Segata N, *et al*. Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci USA* 2014;**111**:E2329–38.

98. Zhang Y, Xie J, Yang J, *et al*. QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics* 2017;**33**:450–2.
99. Li G, Ma Q, Tang H, *et al*. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res* 2009;**37**:e101.
100. Yadav RK, Bragalini C, Fraissinet-Tachet L, *et al*. Metatranscriptomics of soil eukaryotic communities. *Methods Mol Biol* 2016;**1399**:273–87.
101. Waldor MK, Tyson G, Borenstein E, *et al*. Where next for microbiome research? *PLoS Biol* 2015;**13**:e1002050.
102. Biteen JS, Blainey PC, Cardon ZG, *et al*. Tools for the microbiome: nano and beyond. *ACS Nano* 2016;**10**:6–37.
103. Ma W, Huang C, Zhou Y, *et al*. MicroPattern: a web-based tool for microbe set enrichment analysis and disease similarity calculation based on a list of microbes. *Sci Rep* 2017;**7**:40200.
104. Ma W, Zhang L, Zeng P, *et al*. An analysis of human microbe-disease associations. *Brief Bioinform* 2017;**18**:85–97.
105. Zhou H, Jin J, Wong L. Progress in computational studies of host-pathogen interactions. *J Bioinform Comput Biol* 2013;**11**:1230001.
106. Coelho ED, Santiago AM, Arrais JP, *et al*. Computational methodology for predicting the landscape of the human-microbial interactome region level influence. *J Bioinform Comput Biol* 2015;**13**:1550023.
107. Coelho ED, Arrais JP, Matos S, *et al*. Computational prediction of the human-microbial oral interactome. *BMC Syst Biol* 2014;**8**:24.