Development and application of computational methods for NGS-based microbiome research

Yaxin Xue

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2020

UNIVERSITY OF BERGEN

Development and application of computational methods for NGS-based microbiome research

Yaxin Xue

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 05.11.2020

Year: 2020

Title: Development and application of computational methods for NGS-based microbiome research

Name: Yaxin Xue

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

This work presented in this thesis was funded by a PhD grant from the **University of Bergen (UIB)**, and carried out at the *Computational Biology Unit* (CBU*)* in *Department of Informatics* (II). I was also affiliated with the National Research School in Bioinformatics, Biostatistics and Systems Biology (NORBIS). My work was supervised by Professor Inge Jonassen employed at CBU, together with two co-supervisors: Professor Lise Øvreås at Department of Biological Science, UIB, and Dr Anders Lanzén at Marine Research Division, AZTI.

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor – Inge Jonassen – for giving me this great opportunity to work with him and continuous support with his patience, motivation and expertise during my entire PhD period. Thanks for introducing me into multiple projects and your collaborators all over the world. You always have penetrating insight into science, and could take the time to discuss with me despite a busy schedule of you. What impressed me most is every time when I was stuck in problems, you were able to catch key points and inspire me with new ideas. Moreover, you are a cheerful and friendly person with a great love for life and work, you have made an excellent working and scientific environment in our group, I feel so lucky to be part of it.

Thanks to Lise Øvreås and Anders Lanzén, who are my active co-supervisors, for supervision and encouragement. To Lise, thanks for letting me involving your permafrost metagenomic study and giving me a lot giddiness with your expertise, which has greatly expanded my knowledge into this new field. Also, I'm grateful for getting me in contact with Neslihan and facilitating my visit of her lab in Berkeley. To Anders, thanks for your hard work and great contribution to the MetaRib project: you proposed it as one of my projects focusing on bioinformatics, spent time to discuss with us and helped to address many issues with your rich experience in bioinformatics.

In addition, I would like to thanks for all my collaborators. To Neslihan Taş, thanks for the opportunity to visit your inspiring group in Berkeley Lab. We had great collaborations started from there, and you gave me lots of advice and guidance. I'm glad to be linked with your fantastic team members, Yaoming and Megan, and really looking forward to more collaborations in the future. A big thank to Rune Nielsen and Tomas Eagan who bring me into the exciting Bergen COPD microbiome study, and put forward many good suggestions to improve our analysis from the clinicians' perspective.

Thank all colleagues at CBU and Informatics Department for creating a great work environment, for many events and social activities. Special thanks to my office mate and friend, XiaoKang, for the relaxed food parties and interesting discussions. Adam and Gunnar thanks for helping with getting started and organizing many CBU activities, I missed our relaxed gathering together with Kornel, Kasia, Takaya and other friends. A huge thank you to Christine for organizing many NORBIS activities and encouraged me to participate them, which has greatly expanded my network, you did a very good coordinator work. I am grateful to professor Pinar and Jan, who teached me how to be a good teacher and communicate with students. Thanks for the rest of CBUers.

Finally, I want to thank my family and friends. Heartfelt thanks to my family for being supportive and encouraging as always. To my mother and father, for concerning about me all the time with their loves. To my dear big sisters, for taking care of our parents so I could focus on my study abroad. A special thank to my dear cousin Ying Xue and her husband Xing Zhe, who introduced me to UIB, helped me for application and settlement, and took care of me during last 4 years in Bergen. Also, many thanks to my friends Yue, Mei, Kenneth, Merry, Jayme and many others who have supported me the PhD studies and thesis writing. (最后，由衷的感谢家人对我长久以来无私的支持和爱，没有你们我无法完成这一切。)

Yaxin Xue

August 2020

Summary

The advance of DNA sequencing technologies has dramatically expanded our knowledge of microbial community composition and their functions from diverse environments. The most common Next Generation Sequencing (NGS)-based methods used for this purpose are marker genes (16S ribosomal RNA (rRNA), 18S rRNA and Internal transcribed spacer (ITS)), metagenome and metatranscriptome, which all have wide applications with different prominence. Meanwhile, numerous bioinformatic tools and workflows have been developed for a complete and comprehensive analysis of above approaches, which makes it relatively easy to achieve basic results with standard procedure. However, current workflows can only provide generic analyses for well-studied environments, and the choice of methods affect results significantly. In this thesis, I explore best analytical practices and address bioinformatic challenges in NGS-based microbiome research, with emphasis on low-biomass and poorly characterized environments.

Paper I and **Paper II** investigated microbial community composition in human obstructive lung diseases through marker gene sequencing. First, we established robust methods for marker gene sequencing analysis in Chronic Obstructive Pulmonary Disease (COPD) microbiome research both experimentally and *in silico*. Second, we investigated the stability of airway microbiota in COPD patients and healthy control subjects over time using our procedures. In **Paper I**, we evaluated susceptibility of oropharyngeal contamination with three bronchoscopic sampling techniques: small-volume lavage (SVL), protected bronchoalveolar lavage (PBAL), bilateral protected specimen brush (PSB). We emphasized the impact of laboratorial and bronchoscopic contamination in COPD microbiome study; and demonstrated that protected approaches (PBAL and PSB) could discover more unique operational taxonomic units (OTUs) than unprotected lavage through the bronchoscope working channel. Due to the rapid advancement of microbiome analysis methods, **Paper II** further improved our bioinformatic processing, including replacing OTUs with amplicon sequence variants (ASVs) and removing potential contamination in silico. In **Paper II** we also

evaluated how microbial composition changed among groups by comparing both alpha and beta diversity quantitatively with advanced statistical methods. We observed that diversity between the two procedures was higher in the airway samples than in the oral samples and more so in the PSB samples than in the PBAL samples, which indicated the variance of microbiota between examinations. However, we found a significantly lower diversity within-individuals than between-individuals, supporting the existence of a core airways-residing microbiota.

In **Paper III** and **Paper IV,** we investigated microbial community composition and their functional potential from permafrost soil at Svalbard Norway, through a deep Whole Genome Metagenomics (WGMS) analysis. **Paper III** reported 56 metagenome-assembled genomes (MAGs) from 13 phyla recovered from Svalbard permafrost cores. **Paper IV** focused on revealing the key microbial community composition and combined this with metabolic potential in Svalbard permafrost by using novel bioinformatic methods. First, we explored the best practice of MAG refinement for complex environments like permafrost, proposing an improved workflow which could recover more MAGs that would otherwise be discarded due to the high contamination level. Second, we developed a novel computational approach for comparing functional potential across multiple samples from a MAG centric view, which integrated coverage distribution and KEGG module (MO) information. This approach enabled a deeper understanding of functions linked with soil depth and MAGs, in addition to discover new trends between active layer (AL) and permafrost layer (PL). Through these approaches, we found that microbial community composition shifted markedly with depth; we highlighted key metabolic characteristics in Svalbard MAGs, such as aerobic respiration and soil organic matter decomposition, that may play a crucial role in Svalbard permafrost. Our findings provided a novel view of how microbiome survive and acquire resources in an extremely limited resource condition like permafrost.

In **Paper V** we introduced a novel bioinformatic tool – MetaRib – for rRNA gene assembly. Accurate reconstruction of rRNA genes is essential to taxonomic identification within a microbial community. However, current rRNA assembly tools

are restricted to metagenomics or marker gene analysis, similar tools are lacking in total RNA metatranscriptomics due to the increasing size and complexity of the sequence data generated. In this work we developed MetaRib, aiming to fast and accurate reconstructing full-length rRNA sequences optimized for total RNA metatranscriptomic data. MetaRib implements an iterative process to reconstruct rRNA genes, and a post-assembly process to reduce false-positive sequences and estimate relative abundance. We applied it to both simulated and real-world total RNA metatranscriptomic datasets. Compared with other existing tools, we show that using MetaRib we are able to perform fast rRNA reconstruction across multiple samples with a low false positive rate, even in very large datasets, in addition it provides accurate taxonomy-independent relative abundance estimation.

List of publications

Paper I

Protected sampling is preferable in bronchoscopic studies of the airway microbiome (Grønseth, R.*, Drengenes, C., Wiker, H. G., Tangedal, S., **Xue, Y.**, Husebø, G. R., Svanes, Ø., Lehmann, S., Aardal, M., Hoang, T., Kalananthan, T., Hjellestad Martinsen, E. M., Orvedal Leiten, E., Aanerud, M., Nordeide, E., Haaland, I., Jonassen, I., Bakke, P., & Eagan, T.) (2017). *ERJ open research, 3(3), 00019-2017*. https://doi.org/10.1183/23120541.00019-2017.

Contribution

I performed the alpha and beta diversity analysis, investigated the potential of minimizing contamination issue with bioinformatics in downstream analysis, including remove common reagent and laboratory contamination and predict the contamination with negative controls.

Paper II

Repeated bronchoscopy in health and obstructive lung disease: Is the airway microbiome stable? (Grønseth, R.*, **Xue, Y.***, Jonassen, I., Haaland, I., Kommedal, Kommedal O., Wiker, H. G., Drengenes, C., Bakke, P., & Eagan, T.) (submitted)

Contribution

I contributed to the bioinformatic downstream analysis and method section of the manuscript. I assisted to improve bioinformatic workflows of alpha and beta diversity, perform the statistical analysis, generated figures, and write method section of the original manuscript.

Paper III

Bacterial and Archaeal Metagenome-Assembled Genome Sequences from Svalbard Permafrost. (**Xue, Y.***, Jonassen, I., Øvreås, L., & Taş, N.) (2019). *Microbiology resource announcements, 8(27), e00516-19*. https://doi.org/10.1128/MRA.00516-19.

Contribution

I organized the data, submit it to public available repository, performed the analysis and wrote the original manuscript.

Paper IV

Metagenome-assembled Genome Distribution and Key Functionality Highlight Importance of Aerobic Metabolism in Svalbard Permafrost. (**Xue, Y.***, Jonassen, I., Øvreås, L., & Taş, N.) (2020). *FEMS microbiology ecology, 96(5), fiaa057.* https://doi.org/10.1093/femsec/fiaa057.

Contribution

I contributed to the bioinformatic analysis and methods of MAG refinement and comparative functional analysis, implemented the code, and wrote the original manuscript.

Paper V

Reconstructing Ribosomal Genes From Large Scale Total RNA Meta-Transcriptomic Data. (**Xue, Y.***, Lanzén, A., & Jonassen, I.) (2020). *Bioinformatics (Oxford, England), 36(11), 3365–3371.* https://doi.org/10.1093/bioinformatics/btaa177.

Contribution

I participated in the methodological development of MetaRib, implemented the code, evaluated the workflow in both datasets, and wrote the original manuscript.

Abbreviations

AL	Active Layer
ASV	Amplicon Sequence Variant
COPD	Chronic Obstructive Pulmonary Disease
GHG	Green House Gas
ITS	Internal transcribed spacer
mRNA	Message RNA
MAG	Metagenome Assembled Genome
MGS	Metagenomics
MO	KEGG Module
MTS	Metatranscriptomics
NCS	Negative Control
NGS	Next Generation Sequencing
OTU	Operational Taxonomic Unit
OW	Oral Wash
PBAL	Protected Bronchoalveolar Lavage
PCR	Polymerase Chain Reaction
PL	Permafrost Layer
PSB	Protected Specimen Brush
rRNA	Ribosomal RNA
SOM	Soil Organic Matter
SVL	Small Volume Lavage
WGMS	Whole Genome Metagenomics

Contents

1. Introduction

1.1 Microbiome research methods

1.1.1 Early history

The history of microbiology can be tracked back to the 1670s, when Antonie van Leeuwenhoek, known as 'the father of microbiology', studied microbes with his self-made microscope [1]. Since then, diverse microbes have been found to play crucial roles in the environment and in human health. Microbiome refers to all genetic material of microbes (bacteria, archaea, protists, fungi and virus) that live in a given ecosystem. Methods for investigating microbiomes could be either culture-dependent or culture-independent. Culture-dependent methods, such as physiological characterisation, isolation and cultivation, were dominant over a long period in the past. However, the microbial universe is enormous; still it is estimated that less than 1% natural indigenous microbes could be cultivated using standard techniques [2]. Other restrictions include biased growth during culturing and fail to capture symbiotic and diverse relationships in complex environments [3].

Culture-independent techniques are mostly based on the sequences of ribosomal RNA (rRNA), a type of non-coding RNA with prevalent and conserved nature across all organisms because of its fundamental role in translation of transcribed genes. In the 1970s Woese et al. discovered that the sequences of rRNA genes could be used as an efficient evolutionary chronometer to analyse the phylogenetic relationships among all living organisms [4]. Since then, culture-independent methods have been further developed to overcome the drawbacks of culture independent methods. They have been widely used in investigating microbial communities, especially with the application of polymerase chain reaction (PCR) to amplify targeted rRNA genes. Several such PCR-based methods have been developed, including terminal restriction fragment length polymorphisms (T-RFLP) [5], denaturing gradient gel electrophoresis (DGGE) [6] and quantitative PCR (qPCR) [7]. Others are PCR-independent, such as fluorescence in situ hybridization (FISH) [8] and microarrays [9]. Although those

approaches have been widely used and propelled the field greatly, some limitations still remain. For example, those techniques lack the detailed genomic information on the whole microbial community and their individual members, making it difficult to obtain a deep understanding of diverse and or complex communities. Furthermore, these methods are primarily low-throughput techniques. However, the advent and application of next-generation sequencing (NGS) methods have revolutionized microbial research and given birth to many exciting new fields, such as metagenomics, metatranscriptomics and single-cell metagenomics [10] .

1.1.2 The rise of sequencing technology

DNA sequencing is the process of determining the order of nucleotides (A, T, C, G) in a given DNA. From the discovery of DNA structure by Watson et.al in 1953 [11], there have been incredible improvements in sequencing technologies.

Sanger sequencing

In 1977, Frederick Sanger and colleagues published the first-generation sequencing technology [12]. It is based on sequencing by replication of DNA and the incorporation of dideoxynucleotides (ddNTPs: ddATP, ddCTP, ddGTP, ddTTP) that will stop the replication once a ddNTP has been incorporated, so each fragment will end with a labeled ddNTP. This was for many years the dominant sequencing method until the next generation methods were developed. Yet even now Sanger method remains a popular technique in many laboratories, especially for targeting and validating short sequences.

Next-generation sequencing

Several new methods were developed in the mid to late 1990s as alternatives to Sanger Sequencing. These so called 'next-generation' (NGS) methods are massively parallel, allowing the entire genome to be fragmented and sequenced in one sequencing run by generating large number of short reads (typically 100~300 base-pairs) for each genome fragment. There are many differences between NGS technology and Sanger sequencing, but a key distinguishing characteristic is multiplexing. Multiplexing

allows large numbers of DNA fragments to be pooled and sequenced simultaneously during a single run, by using attached barcode (sample marker) sequences. The main advantage of this technology is high-throughput of samples without drastically increasing cost or time.

The 454 DNA sequencer was the first commercial NGS instrument released in 2005, with the re-sequencing of the *Mycoplasma genitalium* genome [13]. It was based on a pyrosequencing approach [14], which amplifies fragmented DNA in water-in-oil beads with PCR. 454 instruments could generate up to a million reads with average read length of 400 bases, but each run is expensive and generates significant homopolymer errors [15].

The Illumina platform is based on 'sequencing by synthesis' (SBS) method [16]. The principle is to use a reversible chain-terminating reaction. Nucleotides are fluorescently labelled and can be used to sequence DNA base by base. A library is constructed by adding universal adapter to both ends of each DNA fragment, then loaded onto the sequencing flow-cell. Each library fragment is amplified by bridge PCR to form a cluster. SBS is used during the sequencing step: each cyclic reaction can only extend one correct complementary base that is identified by imaging to determine four different fluorescent signals. The complete nucleic acid sequence (200~300 bp) is detected after corresponding cycles matched with sequence length.

The rapid development of NGS platforms, including 454, Complete Genome, SOLiD, Ion torrent and Illumina, led to a wide application of NGS and continuous reduction of sequencing cost. Therefore, the pace of advances in genome sequencing technology has accelerated. The speed of genome sequencing has more than doubled every two years since 2003 while the cost of DNA sequencing is dropped significantly [17]. Accompanied by the pace of improvement of NGS has slowed down, 454 and SOLiD are no longer supported, and Illumina platform is dominant nowadays. Their latest sequencer model, Illumina Novaseq, can generate over one billion reads in two days for a few thousand dollars with 99.9% accuracy.

However, NGS has also some disadvantages. One of the main limitations is the short-read lengths. Illumina sequencers can only produce short reads (up to 500bp): the accuracy of nucleotide identification drops due to the error accumulation and signal degradation [18]. The information and variation in repetitive regions are missed as well, as it cannot cover the whole region. This problem can be partially overcome by paired-end sequencing which is the most common sequencing strategy. Compared with single-read data, paired-end sequencing enables more accurate alignment and the ability to detect more variations type such as insertion/deletion [19], and it allows correction of sequencing artefacts such as apparent insertions or deletions. It produces two paired-end reads with a known distance that can span a larger region of genomes than single-end reads in order to include more unique sequences. Another limitation is that almost all of NGS platforms require an amplification step, which could introduce potential problems, like errors, amplification biases and information loss [17].

Third-generation sequencing

To overcome previous issues in NGS, several groups have explored alternative approaches. Single-molecule real-time (SMRT) sequencing developed by PacBio is one of the representatives that may revolutionize the field again. The PacBio platform is based on the properties of zero-mode waveguides (ZMW) [20]. ZMW is a very small hole less than half the light wavelength, which creates a tiny volume to observe only a single nucleotide of DNA being incorporated by DNA polymerase. Four different fluorescent dyes are used to represent four DNA nucleobases. A detector will detect fluorescently labelled nucleotides incorporated into the growing DNA chain, and the base call is made according to the corresponding fluorescence. The PacBio sequencer is able to produce extremely long reads (10kb - 100kb) that allows easier de novo genome assemblies, especially for many species which have long repetitive regions. Besides that, PacBio has other advantages: minimal bias (no amplification step, tolerance of high GC content), random errors distribution, and direct detection of base modification like methylation [21]. These characteristics enable broad applications of PacBio sequencing, although some drawbacks remain, including higher

error rate, lower throughout and higher cost compared to NGS platforms such as Illumina. In practice, hybrid sequencing strategies are more affordable and scalable making use of both accurate short Illumina reads and PacBio long reads instead using PacBio sequencing alone [21].

Another promising approach is nanopore sequencing. The idea is detecting the primary sequence when a single-strand DNA molecule passes through a nanopore channel using electrophoresis transportation [22]. It is most developed by Oxford Nanopore Technologies (ONT), founded in 2005. Nanopore can generate extremely long reads up to 900 kb. Other advantages include miniaturization, amplification free, fast detection and low sample materials preparation. Compared to other platforms, a major difference is the extreme portability of nanopore devices which can be placed in a USB stick as the detection is based on electronic single rather than reaction or optical. Although some challenges remain (lower accuracy and efficiency), it shows great potential in many fields, like DNA methylation, structural variation calling, pathogen surveillance and bacterial/viral outbreak investigation [23].

1.1.3 Sequencing Methods in microbiome research

Environmental genomics is the research of genetic material recovered from samples containing microbes of different species. Handelsmann et al. raised the term 'Metagenomics' for the first time by cloning the DNA fragments of collective soil genomes into BAC vectors and exploring the metabolic functions [24]. Metagenomics has had a rapid development since the emergence of NGS and the number of published metagenomics papers has an exponential growth (Figure 1.1).

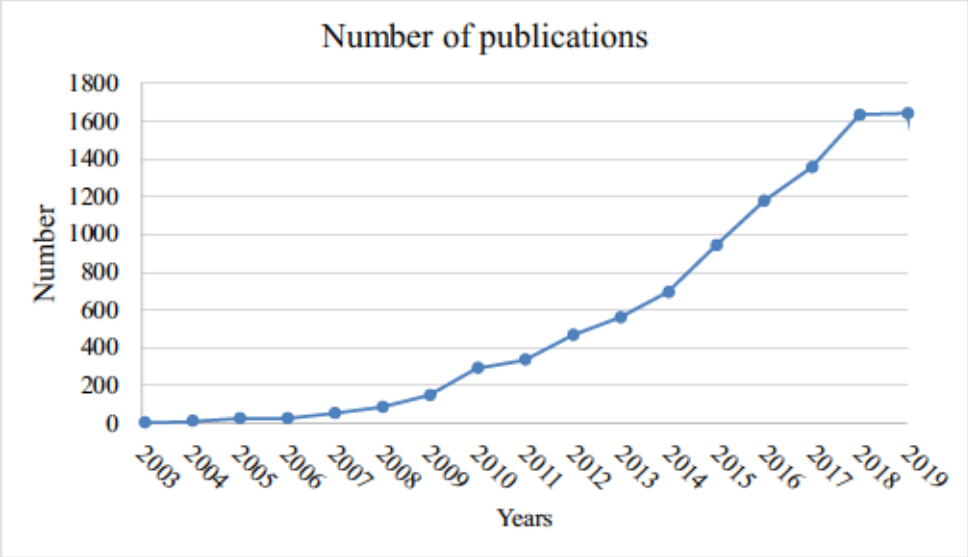


Figure 1.1: Number of Published papers which contain 'metagenome' or 'metagenomics' in PubMed (https://www.ncbi.nlm.nih.gov/pubmed).

Sequencing technologies have a wide application in profiling of microbial communities, which provide the information about composition and dynamics of the total community from multiple perspectives (Figure 1.2), spanning from DNA to protein level. In this chapter, I will give a brief introduction of the most used techniques such as marker gene sequencing, whole genome metagenomics and metatranscriptomics, which also lay the foundation and are highly relevant with my projects.

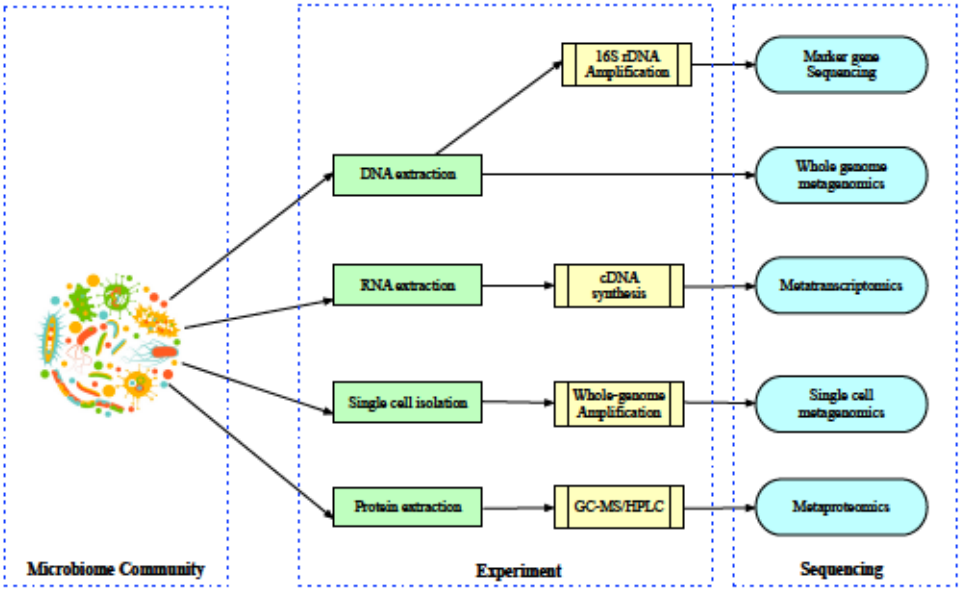


Figure 1.2: Overview of the application of sequencing technologies in microbiome research. Each approach reveals different layers of information (DNA, RNA, single cell, protein) of characterisation the microbiome community.

Marker gene sequencing

Marker genes represent special gene groups that could be used to distinguish between taxonomic lineages [25]. Most of them are from conserved genes, such as 16S ribosomal RNA (rRNA), 18S rRNA and internal transcribed spacer (ITS). Marker gene sequencing utilizes PCR to amplify specific marker gene regions, followed with NGS technologies to generate sequences of mixed samples. This approach provides a fast and cost-effective way to investigate microbial phylogeny and diversity, and has been well-tested and widely used in many studies [26]. 16S rRNA sequences is one the most commonly used marker genes. A typical 16S rRNA gene is approximately 1500 bases long and include 9 conserved regions (C1-C9) and hypervariable regions (V1–V9) (Figure 1.3). Generally speaking, a selected target hyper-variable region of 16S rRNA gene (normally V3-V4) will be amplified and sequenced, as shown in Figure 1.3.

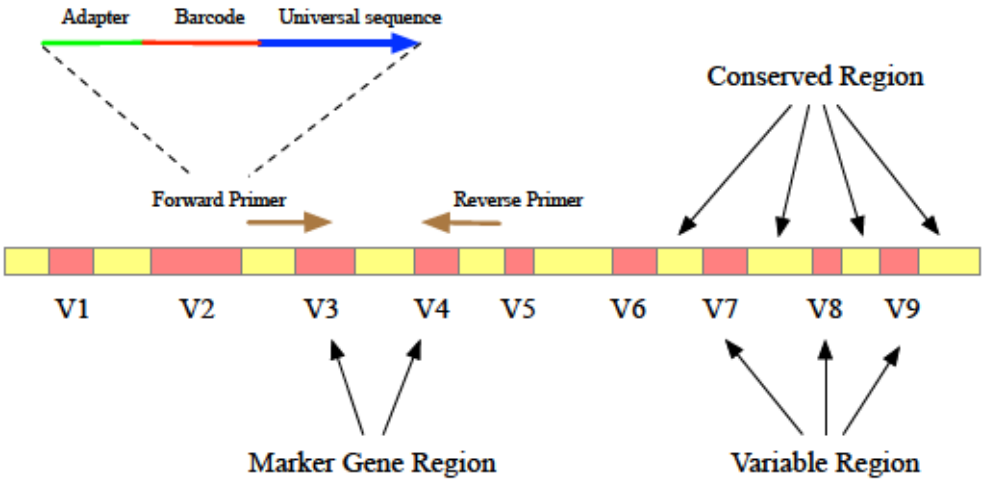


Figure 1.3: Conceptual representation of the 16S rRNA gene sequences. Yellow boxes indicate conserved regions and pink boxes variable regions.

A major advantage of marker gene sequencing is the ability to detect and target non-culturable microbiota. It also allows for the estimation of relative abundance of species in multiple samples simultaneously. Thus, it is widely used in taxonomy analysis of microbiome diversity as a cost-efficient method to assess different types of habitats [27–30]. However, this method also has some limitations. For example, primers used to amplify targeted sequencing regions will introduce biases as PCR efficiency varies and these regions are not totally conserved across all bacteria. Thus, marker gene sequencing has a relatively low resolution due to the high similarity of 16S rRNA genes in close species [31]. Particularly, low-biomass samples are susceptible to be affected with over-amplification: contaminating microorganisms become over-represented as the number of PCR cycles increase [32].

Whole genome metagenomics

Whole genome metagenomics (WGMS) refers to the application of NGS to sequence the whole DNA content in a community directly without marker gene amplification. WGMS shears DNA extracted from habitat randomly, then sequences and assembly into long contigs and scaffolds. Comparing with marker gene sequencing, it enables not only a deeper taxonomic identification but also additional functional knowledge [33]. The first WGMS study conducted using NGS was published in 2006 using 454

pyrosequencing [34]. With the decreasing sequencing cost and improved throughput, WGMS has been applicable in many large scale investigation of complex microbiomes [35–38].

The advantage of WGMS is to investigate the general diversity of all microbiomes, however, it has some limitations [39]. The main challenge of WGMS approach is the large amount of sequence data generated and complexity of computational analysis. Besides, the lack of reference databases makes it is challenge to interpret results biologically.

Metatranscriptomics

There are some limitations of WGMS and marker gene analysis. For example, they cannot discriminate if sequences that are observed in a community are from active members or just merely present. Metatranscriptomics (MTS) uses RNA sequencing to record expressed transcript within a microbial community at a given time point, which provides a more direct measurement of functional activity and actively expressed genes in a community. Studies with MTS have dramatic increase with a wide range of applications, such as active member characterization [40], Antisense RNA detection [41] and host-parasite integration. Some adapters for third-generation sequencers like Nanopore also allow the direct sequencing of RNA. An rRNA depletion step is typically included in MTS studies in order to focus on expressed message RNA (mRNA) encoding proteins, but a more direct alternative is “total RNA sequencing”, where this is not carried out [42]. More detailed information of total RNA sequencing is described in section 1.3. While WGMS focuses on cataloging the present microbiomes within a community, MTS is able to quantify the expression level and monitor the variance of functionality of microbial communities, which provides detailed information in understanding the interaction between a microbial community and its host [43]. A special advantage of MTS is studying different active functionalities with similar microbiome composition [44].

Like general transcripts, MTS has the disadvantages since there is a gap between actively expressed genes and final metabolic products, and it will lose the information of those microbiome which were not active in that time point.

Summary

All approaches introduced above are widely used from surveying microbial communities with their strengths and weaknesses. Table 1.1 summarizes the advantages and disadvantages of different sequencing approaches. In practice, the choice of methods depends on your research question, hypothesis, sample type and resources.

Methods	**Advantages**	**Limitations**
Marker gene sequencing (Who is there?)	• Quick and cheap for sample preparation and sequencing • Many available public datasets and bioinformatic tools • Verifiable ability for detecting both abundant and rare taxa	• Potential biases: amplification, selected variable regions) • Low resolution best to genus level • Limited functional information • Unable to identify microbiota states(live/dead/active) • Contaminations from host/laboratory may affect microbial signatures
Metagenomics (What are they doing?)	• High resolution to species and strain level • Detect novel species/genes • Infer relative abundance and functional potential simultaneously • Avoid PCR-related biases	• More complex and expensive for sample preparation and sequencing • Require heavy computational resource and bioinformatic analysis • Unable to identify microbiota states(live/dead/active) • Contaminations from host/laboratory may affect microbial signatures
Metatranscriptomics (How do they respond?)	• Provide information of active functions directly • Identify microbiota states • Capture dynamic variations among samples	• Most complex and expensive for sample preparation and sequencing • Potential biases from host contamination and high transcription rate of microbiota • Requires high quality sample collection and storage

Table 1.1: Comparison of different sequencing methods in microbiome study. Here are the main advantages and disadvantages of NGS approaches applied in my thesis, based on previous publications [45,46].

1.2 Bioinformatics

In this section we will give an overview of the most common bioinformatic steps and tools involved in marker gene, WGMS and MTS analysis.

1.2.1 Marker gene analysis

Quality control

The first step before starting analysis is to assess the quality of the reads. Removing or trimming of low quality reads is the fundamental process to output reliable results, as most biased diversity analysis are caused by sequencing errors [47]. Several tools are available: some are general quality control (QC)-filter tools for NGS data, like FastQC [48], FASTX-Toolkit [49]; some are specifically developed for marker gene sequencing, such as AmpliconNoise [50] or PRINSEQ [51].

Furthermore, it should be pointed out that identification and removal of possible contamination sequences is a necessary but easily neglected QC step in marker gene analysis. However, contamination sequences may obscure microbial signatures. It may come from various sources, including PCR reactions, reagent, cross-contamination and environment. Previous research demonstrated that contaminants could impact the result critically thus lead to inaccurate conclusion [32], especially for low-biomass environments. In addition to careful library preparation, several bioinformatic tools were developed to address this issue. For example, Decontam is an open-source R package to classify contaminations based on a statistical model of OTU frequency distribution in low-biomass and negative control samples [52]. It requires the use of negative controls, which uses the same procedures as a primary experiment with a placebo or no treatment and is always recommend in marker gene analysis. Other tools like SourceTracker [53] implement with a Bayesian approach that estimates the proportion of contaminants in a community.

Chimeras removing

Chimeras are sequences formed from two or more biological origins incorrectly joined together. These sequences can artificially change the microbiome composition thus need to be removed. There are two major approaches to detect chimeras. One is reference-based detection, all reads will be screened for chimeras using a well-established, non-chimeric reference database, like UCHIME [54] and ChimeraSlayer [55]. Another is *de novo* detection. A chimera-free reference database will be generated for each NGS data according to their abundance, assuming that the most abundance sequences are unlikely to be chimeras thus could be used as reference. UCHIME provides this approach too. UCHIME is the most widely applied tool as it supports two modes and is also implemented in comprehensive pipelines like QIIME (Quantitative Insights Into Microbial Ecology) [56] and MOTHUR [57]. DECIPHER [58] is another popular tool in chimeras detection, which is applicable for long sequences (>= 500 bp).

Sequence clustering

One common approach in marker gene sequencing is to cluster short sequences into Operational Taxonomic Units (OTUs) based on sequence similarities. Each OTU is intended to represent a taxonomic unit depending on the similarity threshold. The sample-by-OTU table can then be used to investigate microbial "species", diversity and composition, etc. Many available tools are proposed for OTU clustering, which can be categorized into reference-based OTU and *de novo* OTU approaches: a more detailed comparison is available in [59]. For the past years clustering reads into OTUs has been the standard process in marker gene sequencing analysis [60]. However, OTU clustering is typically used arbitrarily with limited resolution: the common 97% similarity can often only distinguish taxa at genus level.

Recently, new methods have been developed to address OTU issues. Amplicon sequence variants (ASVs) methods attempt to model the sequencing error and apply the model within clustering, which could distinguish single sequence variant [61,62]. Tools like Deblur [63] and DADA2 [64] already implement ASVs as standard

workflows. Considering both sequence similarity and abundance in a model, ASV methods have shown improved sensitivity and specificity in marker gene sequencing analysis compared to OTU-based methods in recent benchmark studies [65]. Another hybrid clustering methods are SWARM [66] and SWARM2 [67], which define an unit in between ASVs and OTUs, with consideration of abundance patterns.

Taxonomy classification

Taxonomy classification is to assign taxonomic names to biological sequences. This step is typically achieved either by aligning sequences against a reference database or using k-mer based techniques. There are several commonly used rRNA databases including Silva [68], Greengenes [69] and the Ribosomal Database Project (RDP) [70]. The choice of databases has been found to affect the final taxonomy result [71]. Silva, the largest database, includes the most taxonomic units and has the best overall performance, but it requires more computational resource [71].

Pipelines

Several marker gene pipelines allow the user to perform the whole analysis workflow, from raw DNA sequence data to publication-ready results. QIIME is one of the major packages for marker gene analysis published in 2010 that has been applied to many studies [56]. QIIME 2 is a updated version available since 2018 [72]. It addresses several limitations of QIIME 1 with many new features like improved methods, graphic interface, plugin architecture, etc.

MOTHUR is another open-source project aiming to analyze and compare microbial communities as a single piece software [57].The main difference is the philosophy: MOTHUR is a standalone executable program which has integrated many excellent algorithms into one, while QIIME is a python interface connecting a large number of disparate programs with great expansibility and freedom. A recent published benchmark study evaluated the performance of QIIME 2, MAPseq [73] and MOTHUR, demonstrating that QIIME 2 was optimal in marker gene profiling while also most computationally expensive tool [74].

1.2.2 Whole genome metagenomics

WGMS targets the complete sequences of all microbial genomes within a community, thus it yields board range of taxonomic, functional and evolutionary information. All shotgun reads are used to determine composition and function in a community, either by read-based or assembly-based analysis (Figure 1.4).

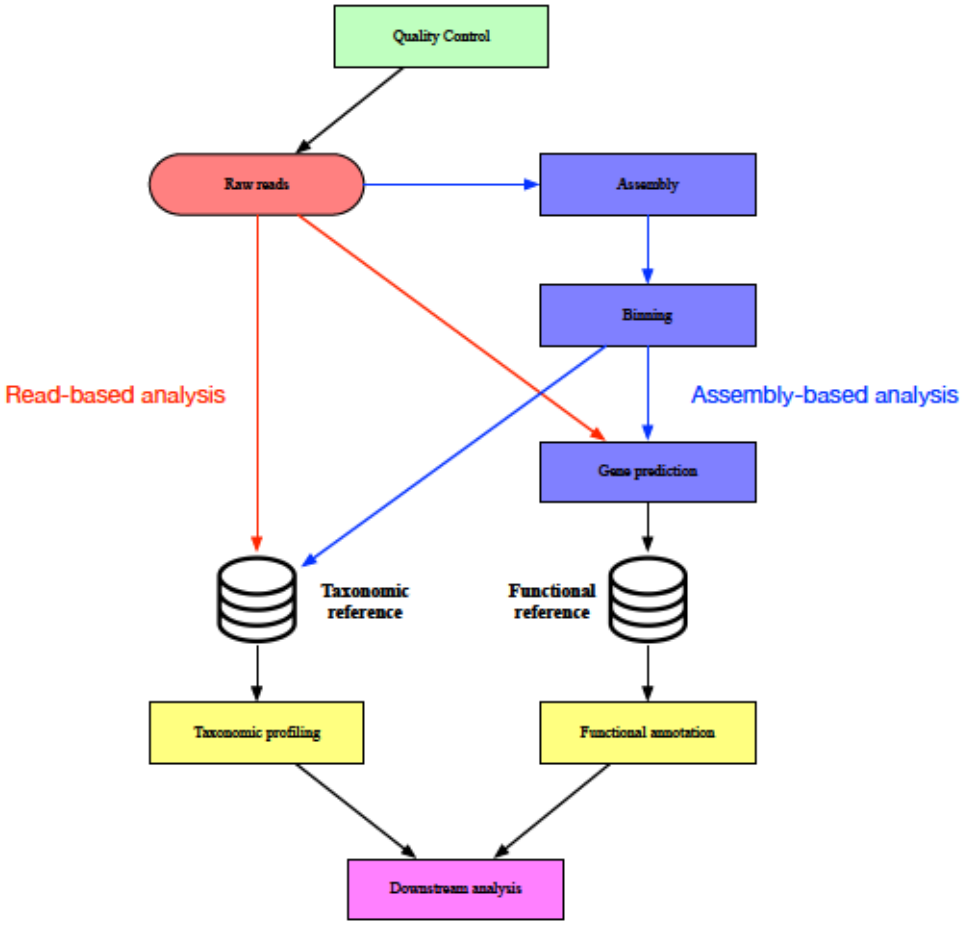


Figure 1.4: Summary of bioinformatic workflow in WGMS analysis. The WGMS data could be analysed using read-based approach or assembly-based approach, depending on the research objectives. Read-based analysis takes the unassembled reads and compares them with the reference directly; assembly-based analysis attempts to assemble and bin genomes firstly, then analyse the genes and contigs with reference databases.

Read-based analysis

Read-based analysis utilizes raw sequence reads after QC. The core idea is to map reads against reference databases and extract information based on alignment hits. Because each read is considered independently in this approach, it allows to perform large-scale metagenomic profiling efficiently and provide a rapid profiling of community composition and function. Furthermore, it can capture the information of reads that cannot be assembled.

Though plenty of reads alignment tools are available, mapping numerous reads directly to reference is typically not the best solution for such analysis: not only concerns about extensive CPU usage but also inevitable high false positive hits. To reduce computational resource and false positive rate, many tools utilize sequence character (like k-mer) or string compression (like Burrows-Wheeler transform: BWT) to preprocess reads and references [75].

Assembly-based analysis

Comparing with read-based analysis, the assembly-based approach is more complicated with several steps, including assembling the reads into contigs, 'bin' contigs into metagenome-assembled genomes (MAGs), gene prediction and functional annotation. This approach enables to reveal previously unknown and uncharacterized genomes and pathways and thus provide novel biological insights into complex communities, but it typically requires heavy computational resources (especially memory) and additional analytic processes.

Quality control

WGMS analysis needs careful QC as an initial step, which aim to identify and remove low-quality sequences and contaminants. There are several tools that are available to perform QC in WGMS, including FastQC [48] , MultiQC [76], FastQ Screen [77], BBDuk [78], Khmer [79], etc. Table 1.2 summarizes their key characteristics. In addition, as WGMS is the study of the entire environmental microbial community directly, identification and filtering of possible host/contamination sequences is a

necessary QC step. For example, it is useful to screen sequences against human reference in a human-related microbiome study, like human gut or skin microbiome.

Tools	Features	Website
FastQC	Provide several graphic QC statstics information	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC	Aggregate results from multipe samples into one single report	https://multiqc.info/
FastQ Screen	Screen sequences against a set of reference database	https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/
BBDuk	Decontaminate sequences using Kmer-based operations	https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/
Khmer	Trim and normalize sequences for Kmer-based analysis	https://khmer.readthedocs.io

Table 1.2: A list of tools for quality control in WGMS. The table contains some of the most commonly used tools for WGMS analysis.

Assembly

In bioinformatics, assembly refers to aligning and merging short reads into long DNA fragment called contigs, a set of overlapping segments that represent a consensus region of DNA. Two different algorithms are commonly used in assembly [80]: overlap based algorithm – including traditional overlap-layout consensus (OLC) method [81] or recent string graph [82], and de Bruijn graph [83]. Numerous approaches for assembly have been published and the choice of methods depends heavily on your research purpose and sample type. Overlap based algorithms are suitable for long sequencing reads, like Sanger or PacBio, but the computational requirements become impractical with enormous reads. De Brujin graph addresses this issue partially by splitting each read into overlapping subsequences of fixed length k (k-mer), that enables efficient assembly for Illumina short reads sequencing.

It is known that genome assembly is challenging due to many factors, like heterogeneity, sequencing errors and long repeat regions, which may lead to mis-assembly and fragmentation. However, WGMS assembly is more challenging due to its particularity. Firstly, a metagenomic sample represents a group of species with different abundance rather than uniform distribution in typical single genome assembly, meaning that low abundant genomes may barely assemble due to insufficient data. Picking a lower k-mer may help, but it will have a higher chance of getting repetitive k-mers that leads to mis-assembly of the genomes. Thus, there is a tradeoff between covering low abundant genomes with accurate assembly for high abundant genomes. Another problem is the phylogenetic distance. A metagenome sample may include some highly similar sequences – such as different strains from same species – that only a few nucleotide variances. It can cause the assembly to generate many fragmented contigs instead of complete drafts.

Several metagenome-specific assemblers have been developed to tackle these challenges. For example, Meta-IDBA attempts to cover for both high and low abundant genomes by iterating with multiple k-mer size [84]. Its extension, IDBA-UD, uses similar strategy with special optimization to handle uneven distributed sequences [85]. MetaVelvet-SL is an extension of Velvet that integrating a Support Vector Machine (SVM) – is trained by a similar population of samples – to increase the performance [86]. MEGAHIT uses increasing k-mer strategy with succinct de Bruijn data structure to reduce computational cost [87]. metaSPAdes is a mode of the assembly software SPAdes for metagenomic assembly, using a heuristic method to distinguish interspecies repeats [88]. It was reported that MEGAHIT had the best overall performance based on their benchmark data sets in Critical Assessment of Metagenome Interpretation (CAMI) challenge [89]. Another benchmark study suggested that MEGAHIT together with metaSPAdes may be the best choice [90]. In general, different programs have their own strengths and weaknesses with specific datasets. Picking the proper tool is depending on several factors (e.g, research purpose, sample type, platform, coverage).

Binning

The aim of binning is to group contigs into MAGs: each bin will represent one single species or strain ideally. In short, two types of binning methods are available: supervised approaches align contigs against reference databases and assign them into related taxonomy labels, unsupervised methods cluster contigs into groups based on sequence characteristics (e.g., tetranucleotide frequency, homology, GC content). Supervised methods rely on search homology against known genomes, however, only a small fraction of microbiome have been sequenced, thus most contigs derived from a metagenomic sample, especially novel genomes, may barely map to the reference. Therefore, currently most binning methods are developed based on the sequence composition, especially k-mer frequency [91].

MetaBAT uses a k-medoid clustering method to bin contigs by calculating pair-wise distance based on tetranucleotide frequency [92]. Maxbin and its updating version, MaxBin2, employs an Expectation-Maximization (EM) algorithm to cluster contigs after co-assembly of multiple metagenomic datasets [93]. Recently, coverage were found as a very strong characteristic in binning contigs when multiple sequence samples were produced in a WGMS [94]. CONCOCT [95] and GroopM [96] are two automated binning tools that utilize both sequence composition and coverage information.

Existing binners are developed based on different clustering methods and sequence features, and evaluated with their respective benchmark data sets. DASTool is an post binning method that integrates output from a flexible number of current binning tools to calculate an optimized, non-redundant set of bins [97].

CheckM is common used to evaluate the quality of recovered MAGs, like completeness and contamination, based on the frequency of single-copy marker genes [98]. Genomic Standards Consortium (GSC) present a standard for assessing and estimating the quality of MAGs [99]. Comparing with contigs, recovered MAGs could provide more detailed information for downstream analysis, such as phylogenetic analysis, functional profiles and abundance estimation across samples.

Gene prediction

Predicting genes or other features, like CRISPR repeats, tRNA, and non-coding RNA, is a common prerequisite step to functional annotation. It is noted that assembly is not a necessarily precondition for this procedure: raw reads can also be used to predict genes directly. For example, FragGeneScan [100] is a tool that predicts genes from short reads incorporating a sequence error model and codon usage statistics. This approach is able to provide an overview of functionality quickly. However, it is highly depending on the length and quality of raw reads, which is impractical to obtain detailed functional information from a biological perspective. Therefore, most gene prediction tools focus on long contigs deriving from assembly or binning procedure.

Metagenomic gene prediction is more challenging compared with single genomes, due to the diversity of microbial composition, sequencing errors and fragmented contigs [101]. Several tools have been specific developed to address this issue, such as Glimmer-MG [102], MetaProdigal [103], MetaGeneMark [104], Orphelia [105] and Prokka [106]. In summary, most of them are context-based methods that uses different models to detect inherent variations between coding and non-coding regions by selected sequence properties, like GC-content, codon usage, k-mer frequency [107]. Advantages of context-based approaches include reference-free, fast process and detection of novel genes. Others are similarity-based approaches searching for similar existing gene sequences in reference, like BLAST [108]. However, this approach is computational expensive and cannot discover novel genes, thus is not recommended in most situations.

Taxonomic profiling

The aim of taxonomic profiling is to identify the composition and abundance of microbiome in a community. Compared with marker gene analysis, WGMS can perform comparative analysis across samples in a better resolution with less bias either by read-based or assembly-based approaches. All of these approaches heavily rely on reference catalogs, against which either reads or assembled contigs are matched. BLAST [108] was widely used to assign taxonomy with NCBI GenBank in the early

stage of WGMS, while it is no longer applicable due to the dramatic increasing size of reference databases and sequencing data. Taxonomic profilers are developed to overcome the above-mentioned difficulty. To do so, several sequential approaches are introduced to reduce query time and potential hits in the reference, like k-mer analysis, Burrows-Wheeler Transform (BWT) and full-text index (FM index). As a result, they are able to provide taxonomic assignment in a much faster way but usually less sensitive than BLAST. Depending on the type of databases, these tools can be further divided into two categories: nucleotide (blast**n**) and protein (blast**p**) classifiers, to search against reference databases of DNA sequences or protein sequences respectively.

Most nucleotide-based classifiers utilize k-mers to assign taxonomy. In brief, these tools search k-mer hits against a predefined database which stores k-mer with corresponding taxonomic identifier of every genome. The selection of k-mer number reflects the trade-off between sensitivity and specificity: short k-mers may generate many non-specific matches (multiple hits) while long k-mer may fail to match. Several tools are available for fast taxonomic profiling in nucleotide database. For example, Kraken [109] and its derivative tools (Kraken2 [110], KrakenUniq [111], Bracken [112]) identifies a sequence's taxa by searching exact k-mer match with lowest-common ancestor (LCA) records in the database. CLARK [113] (and CLARK-S [114]) is a similar approach but using discriminative k-mers at genus/species level only. k-SLAM [115] is a novel approach that uses k-mer to find assignment firstly and then peforms local alignment and pseudo-assemble to increase specificity. Except k-mer based technique, other tools like Centrifuge [116] utilizes BWT and FM index to compress the database to reduce redundancy and increase specificity.

Protein sequences are composed of 20 amino acid characters rather than 4 nucleotides. Proteins are also more conserved compared to the DNA sequence that encodes them. Thus, protein-based classifiers can be more sensitive. However, this approach is normally more computationally intensive due to the six frame translations from DNA to protein, and information of non-coding sequences is absent [117]. DIAMOND [118] uses double index for both a protein reference database and translated sequences firstly

and finds potential matches in parallel after sorting the index. Kaiju [119] utilizes different indexing strategy: it indexes a protein reference using BWT while translated sequence with FM index. This approach enables Kaiju to search against large protein database efficiently.

The choice of classifier depends on a number of factors, including research subject, data size, computational resource, etc. Broadly speaking, nucleotide classifiers have better performance than protein classifiers for well-characteristic environments, as the absence of non-coding sequences in protein databases [117]. On the other hand, protein classifiers provide more sensitive hits thus could be considered for poorly characterized environments (like soil) with the usage of large databases (like NCBI nr database). Computational burden is another limitation. For most classifiers, there is a trade-off between computing cost and accuracy. For example, Kraken series will have a very good performance and fast identification if the server has very large memory (> 100Gb); similar with CLARK-S compared to CLARK [117]. If it is not the case, Centrifuge is a good alternative with limited computational resource [116]. Table 1.3 summarizes a few common tools and their key features for taxonomic profiling, a more comprehensive evaluation could be found here [117].

Database	Tools	Key features	Reference
Nucleotide	Kraken	Exact k-mer search in memory	[109]
	Kraken2	New version of Kraken with improvement of speed and memory	[110]
	KrakenUniq	Special version of Kraken using the stream sketching algorithm HyperLogLog (HLL)	[111]
	Bracken	Compute relative abundance of species using Bayesian estimation	[112]
	CLARK	Supervised sequence classification using discriminative k-mers	[113]
	CLARK-S	CLARK version with spaced k-mers. It requires more RAM but offers a higher sensitivity	[114]
	k-SLAM	K-mer search with additional validation using pseudo-assembly	[115]

	Centrifuge	Fast and memory-efficient tools for taxonomic profiling using BWT	[116]
Protein	DIAMOND	Protein homology search using spaced seeds with a reduced amino acid alphabet	[118]
	Kaiju	Fast for large-scale profiling in protein database	[119]

Table 1.3: A list of commonly used classifiers and their key features in taxonomic profiling. These tools can be mainly divided into nucleotide-based and protein-based classifiers, using different sources of reference databases and indexing strategies.

Functional annotation

Microbial communities are not only a group of taxonomic species, but also represent a collection of biochemical functions that interact with environment. WGMS provides a novel approach to answer the question 'What are they doing?' by preforming functional profiling depend either reads or predicted genes within assembled contigs.

Read-based functional profiling utilizes raw reads to map to functional reference databases. In this approach, each read is considered independently to find best hits in annotated genes, proteins or pathways, and then it could be used to provide an aggregated picture of community function. For example, HUMAnN (HMP Unified Metabolic Analysis Network) [120] and its updated version HUMAnN2 [121] utilizes a tiered search strategy that aggregates single protein hits into higher-level functional units (metabolic modules or pathways), providing comprehensive reports of metabolic presence, absence and abundance. The read-based approach is efficient to perform a quick functional profiling for a complex dataset, but it depends heavily on hits of homologous genes thus choose a proper database is the most critical step. In general, curated databases – like RefSeq [122] and Uniref [123] – are more applicable to well-studied samples such as human microbiome considering the accuracy and efficiency; whereas large databases such as NCBI NR could be considered for poorly annotated samples like permafrost soil. Besides, several specialized tools have been created for specific annotation: like FOAM [124] with environmental focus, PHASTER [125] for identifying putative prophages, and Resfams [126] for antibiotic resistance function.

Assembly-based annotation requires a gene prediction step (see above), followed with assigning predicted genes to functional categories by either homology-based or pattern-based search. This approach is able to reveal previously uncharacterized functions or pathways with constructed genomes, which provides novel biological insights into complex communities. Specific tools like AntiSMASH [127] could be used to predict biosynthetic or metabolic pathways for bacteria and fungi. However, this approach does not apply to all studies. Firstly, many factors – such as low-coverage data, mixture of close species – will affect the performance of assembly, generating many fragmented and mis-assembled contigs, which could obscure functional annotation for downstream analysis. Next, only part of the metagenomic genomes can be captured by assembly and still many genes are uncharacterized in reference databases.

Regardless of whether read-based or assembly-based approach is adopted, the crucial step in functional analysis is the choice of databases. They could be divided into two categories: homology-based and pattern-based. Homology-based approach is the most common method with relative high accuracy [128] by searching homologous genes between query sequences and existing reference databases. Several public resources are available with different priorities. Non-redundant protein databases are most common, including NCBI NR [122], SMART [129] and UniProt [130]. Besides, orthologous genes represent special gene clusters that exist in different species while are originated by a common ancestor gene, thus often have similar functions [131]. Several specialized databases – like COG [132] and eggNOG [133] – have been organized to infer sets of orthologous groups. KEGG [134] and SEED [135] databases are usually used to annotate pathway and subsystem information. Gene ortholog (GO) [136] provides a set of hierarchical graphs describing the functions of genes related with biological process, cellular component and molecular function. Conserved domains reveal aspects of functional and/or structural units of a protein, therefore may pinpoint precise function transfer across species. Databases like Pfam [137] and CDD [138] provide comprehensive annotation and evaluation for conserved domains of proteins. Pattern-based approach could be considered when protein sequences show poor results using homology search. InterPro [139] is an integrated database of protein

domains, motifs and functional units that can be applied to characterize functions. It consists of many diagnostic signatures with different scopes: from high-level superfamily (like CATH-Gene3D [140]) to specific subfamily (like PRINTS [140]).

Both homology-based and pattern-based approach needs tools to search against databases. These searches generally come in two algorithms: BLAST-like [108] (BLASTX , BLASTP, etc) and HMMER-like [141] (Hidden Markov Model). The former is the most preferred algorithms while the latter is used in protein families and pattern-based search. Web-based servers such as MG-RAST [142] and IMG/M [143] enable users to access multiple database and perform comparative analysis with other published studies easily. IMG/M integrates most comprehensive analysis, but it requires strict data organization and has long waiting time; while MG-RAST is able to provide a fast feedback with selected databases. Other tools like MEGAN [144] is a standalone software that supports functional analysis using InterPro2GO, SEED, eggNOG or KEGG. Figure 1.5 summarizes commonly used databases and tools in functional annotation. However, it is noted that the main limitation now in functional profiling of a community is still the lack of annotation in most microbiomes except a few selected model species, which means only highly conserved pathways and genes could be detected in functional profiling [128]. It is partially explained confusing results between high taxonomic diversity and low functional variance in metagenomic studies.

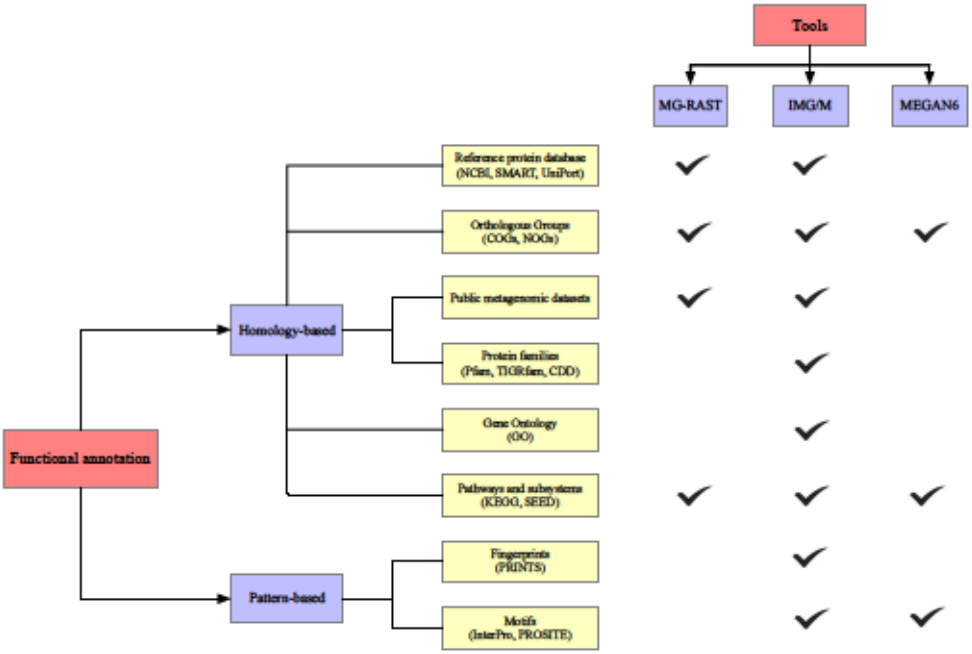


Figure 1.5: Summary of common databases and tools in metagenomic functional annotation. Most tools are achieved using homology-based approaches against specific domains of databases; pattern-based approaches provide an alternative method when homology-based approaches have poor similarities.

Pipelines

WGMS analysis is a complicated process involving many different steps. Several pipelines have been developed to facilitate the whole analytic workflow. Tools like EBI metagenomics [145] and MG-RAST [142] are web-based services could provide a basic overview analysis but may be short of detailed information. MetAMOS [146] and MOCAT2 [147] are command line based pipelines that integrate major steps in WGMS analysis with different tools. Anvi'o platform [148] has an user-friendly interface that allows users to optimize the assembly and binning approach with an major advance of flexibility and visualization. KBase [149] is developing a suite of microbiome analysis apps with a graphic interface that has a handy way to import, extract, and edit data with various apps.

1.2.3 Metatranscriptomics

Bioinformatics in MTS generally could be divided in two types: mRNA analysis and rRNA analysis.

mRNA analysis

To date, most MTS studies have focused on mRNA analysis, as transcriptomes could offer more detailed functional variance within individuals than WGMS [150].

Pre-processing

Similar to WGMS datasets, QC step is necessary to minimize errors. A specific step that should be taken into consideration is the removal of rRNA, as they often dominate in samples (~ 90%). In addition to physical removal in library preparation, some bioinformatic tools, like SortMeRNA [151], can be used to identify and remove rRNA reads after QC.

Transcripts assembly

Preprocessed reads can be assembled into full length transcripts. It is noted that MTS assembly has some unique challenges like uneven sequencing depth and conserved regions of mRNA across species. Thus, traditional single-genome transcript assemblers may have poor performance. Several tools are designed specifically for MTS, such as IDBA-MTP [152], Trinity [153] and TAG [154], but still efficient assembly tools in MTS are lacking and have many remaining issues, especially for complex and large volume datasets [155].

rRNA analysis

Although rRNA reads are frequently removed in MTS, it has a few advantages in accessing taxonomy diversity of a community. Firstly, rRNA in MTS is able to detect species in all three domains of life, meanwhile avoiding amplification bias, compared to PCR-based marker gene surveys. Further, it allows for the reconstruction of full length rRNA sequences, enabling a higher resolution for taxonomy profiling. Moreover, rRNA is also essential for protein synthesis in all organism, therefore its relative abundance across taxa generally reflects overall community structure.

rRNA sequence reconstruction

To have a high resolution of taxonomic classification, recovering the full-length rRNA is the crucial step in rRNA analysis. There are many tools to extract SSU rRNA sequences from total reads, such as SortMeRNA [151]and phyloFlash [156], more challenges would come in the reconstruction process. Existing *de novo* assembly tools were designed primarily for genomic or metagenomic data and do not perform well on rRNA due to their complex structure including both conserved and hypervariable regions [157].

Specific tools for reconstructing rRNA sequences can be divided into two groups: reference-based and assembly-based approaches. For example, EMIRGE [158] utilizes an expectation maximization approach with known rRNA sequences to reconstruct rRNA genes from a community. REGAO [157] is an optimized *de novo* assembly tool that reconstructs rRNA sequences with overlaps between reads using a suffix/prefix array. However, these tools were designed for analysis of smaller datasets with limitations in terms of high error rates as well as computational resources, thus cannot be used directly to analyze rRNA in total metatranscriptomics due to the extreme high volume of rRNA reads in total RNA data (97-98%).

1.2.4 Downstream analysis

Downstream analysis uses statistical tools to investigate the relationship between sample metadata and microbial features, mainly including taxonomic and functional matrices generated by primary analysis. Main challenges come from highly dimensional and sparse property of microbiome dataset, which requires careful statistics to avoid wrong conclusions. Common downstream analyses include alpha and beta diversity, differential analysis, machine learning approaches and omics data integration.

Alpha diversity

Alpha diversity quantifies mean diversity of microbiomes within specific sites or habitats. It answers the question ‘how many species in a microbial community’ by calculating features of data frame within individual samples. A variety of

measurements have been created to infer the diversity. For example, species richness is the count of species in a community and evenness represents how evenly the individuals in a community are distributed among the different species. Species richness and Faith's phylogenetic diversity are closer to real diversity whereas they are susceptible to sequencing depth. In practice, Shannon index is one of the most common used measurements with consideration of both richness and evenness [159]. Notably, alpha diversity measurement is usually applied in marker gene analysis.

Beta diversity

Beta diversity compares the differentiation between two sites or communities by generating a distance matrix between pairs of samples. It answers the question 'how different is the microbial composition in one community compared to another' by quantifying the dissimilarity metrics between sample pairs. Beta diversity metrics can be divided into two groups: quantitative and qualitative. Quantitive metrics, like Bray-Curtis or weighted UniFrac, uses feature abundance to calculate; while qualitative metrics like binary Jaccard or unweighted UniFrac only consider the feature's presence–absence. Bray–Curtis is calculated based on abundance or read count data while UniFrac is based on the fraction of branch length of sequence distance. Currently there is no consensus best metric. For example, some studies showed that metrics like UniFrac outperformed others as they also consider biological phylogeny [160] while others indicated the opposite [161]. Differences in metrics will lead to a performance trade-off between sample size bias and rare species turnover. More detailed comparison and evaluation is described elsewhere [162].

Regardless of metric selection, the result of a beta diversity analysis will be a multi-dimensional complex matrix, which cannot be interpreted immediately. Ordination methods, such as principal coordinates analysis (PCoA) or principal component analysis (PCA), are commonly used to reduce the metrics into a low-dimensional (2D or 3D) representation. Later it can be integrated or visualized with various categories of meta data to investigate the correlation between samples phenotype and microbial diversity using unsupervised clustering. More information about ordination and visualization of beta diversity is described in this review [163].

Differential analysis

Differential analysis is to identify if there are specific microorganisms or functional elements (genes or annotations) over- or under-abundant in some interesting groups (like disease) relative to a reference group (like control). For example, one of the main tasks of mRNA analysis in MTS is to study the differential gene expression patterns after transcript assembly. Several tools originally developed for single genome RNA-seq can be leveraged for differential analysis in MTS, such as edgeR [164] and DESeq2 [165]. Similarly, these approaches have also been applied to identify significantly differentially abundant OTUs [166].

Statistical methods such as multivariate analysis of variance (ANOVA) could be further used to test if differences between groups are statistically meaningful, but ANOVA requires a normal distribution of the data which restricts its widespread application. Non-parametric tests, such as ANOSIM (analysis of similarity) [167] and PERMANOVA (permutational ANOVA) [168], are distribution-free methods therefore more widely used and robust in microbial ecology [169].

Data mining in microbiome dataset is quite challenging due to the sparse and high dimension of the input. Therefore, classical statistical methods may lead to erroneous conclusions due to compositionality and sparsity of microbiome datasets [170]. Previous research pointed out that NGS-based microbiome studies should always be considered as compositions at all stages of analysis [171]. The main idea is that the total number of counts (reads, OTUs, or genes) is constrained: an increase of one variable implies the decrease of another one so that the total number does not exceed to 1. Several compositional methods have been developed and applied in analyzing microbiome datasets. One approach is to use isomeric log-ratio transformation (ILR) to transform the data into relative abundance, then test with standard statistical tools [172]. Tools like SPARCC [173] and SPIEC-EASI [174], assume a sparse data matrix that few species are correlated. BAnOCC is a novel Bayesian framework to estimate correlations assuming a log-normal on datasets [175]. ANCOM makes no distributional assumption and can be used to compare microbiome differential abundance in consideration of compositional constraints [176]. In summary, the

composition nature of microbiome data should be emphasized to researchers, a detailed guidance of compositional microbiome analysis is available here [171].

Machine learning approaches

The emerging machine learning approaches have shown great potential in microbiome downstream analysis, especially for classifying current status or predicting future status by combining microbiome data with their metadata category [177,178]. For example, 16S Classifier [179] uses Random Forests (RF) to perform taxonomy classification with Greengenes database while TAC-ELM [180] is a kmer-based method that uses a Neural Network (NN) to assign taxonomy. Pasolli [181] systematically evaluated the performance of the machine learning methods SVM, RF, Lasso, and ENet in predicting the status of six different diseases using metagenomic datasets from eight studies, which demonstrated that RF obtained the best overall performance followed by SVM. DeepARG [182] is another tool developed for identifying potential novel antibiotic resistance genes in metagenomics data with a Deep Learning (DL) approach. Regardless of methods, one particular consideration in machine learning is to avoid overfitting. The dataset needs to be substantial and representative, and the analysis needs to be combined with cross-validation and independent test.

Omics data integration

The integration of multi-omic approaches – including marker gene analysis, WGMS, MTS, metaproteomics, metabolomics and other techniques – enable a more comprehensive understanding of a microbiome community. However, integrating omics datasets has immense challenges in many aspects. For example, there are different time scales between mRNA expression and metabolite, as well as protein [183]. In addition, metaproteomics and metabolomics is still quite low-throughput comparing with NGS-based approaches: the latter is much sparser and high-dimensional.

Correlation analysis, such as Pearson or Spearman, is the most straightforward and commonly used approaches for omics data integration. However, they are error-prone

due to the sparsity and high dimensionality of the omics dataset. Advanced statistical approaches like Procrustes analysis[184], calculating correlation at the low-dimensional space rather than using raw data matrix. For example, McHardy et.al [185] utilized Procrustes analysis to investigate the relationship between metabolome and microbiome, and observed a stronger inter-omic connection in cecum than sigmoid colon. Other methods such as co-inertia analysis [186], which consider not only datasets correlation but also their relevant metadata categories. It should be noted that critical corrections, such as Bonferroni or Benjamini-Hochberg correction, in conjunction with statistical models can further increase the overall performance in multi-omic comparisons [187].

Several integrative analysis tools are available, including easy-to-use online tools as well as versatile tools with computational experience. For example, web-based tools like XCMSOnline [188], which allows integration with metabolomic, transcriptomic and proteomic data. OmicsIntegrator [189] is another online tool that applies network analysis to identify interpretable pathways by combining transcriptomic data with protein interaction data.

Several studies have shown great success of omics data integration in characterizing the composition, functional, and metabolic activity of microbiomes. For example, Heintz-Buschart et al. [190] demonstrated a correlated variation between gastrointestinal microbiomes and families with type 1 diabetes mellitus (T1DM) through a multi-omics approach. Jason et al. [191] detected some novel clades which were highly associated with inflammatory bowel disease (IBD) by integrative analysis of multi-omics measurements. In summary, these achievements highlight the importance of omics data integration in microbiome research.

1.3 Applications of NGS-based approaches in microbiome research

The optimal sequencing strategy depends on the complexity of samples and your scientific goal, both single and mult-omic approaches have wide applications in different microbial ecosystems, such as human, soil, water, food, plant and animal among others. In this section I give a brief scientific and biological introduction of studies related with our projects during my PhD period.

1.3.1 Bergen COPD microbiome study

Chronic Obstructive Pulmonary Disease (COPD) is a progressive lung disease affecting about 2.4% of the global population [192]. Although the risk factors of developing COPD – such as smoking, occupational exposures, air pollution and asthma – have been characterized many years, we have limited knowledge about why only some individuals will develop the disease under similar exposed risk conditions. Genetics can only explain a small fraction (1-5%) [193] of the risk.

Our understanding of the human microbiome has rapidly expanded in the past decade. Especially the gut microbiota has turned out to be associated with many human diseases – such as obesity, colorectal cancer, and inflammatory bowel disease [194–196]. These studies highlight the importance of the microbiome in human disease and encourage similar applications in COPD research. Previous studies observed the correlation of respiratory microbiome both between healthy/COPD subjects and COPD exacerbation stages [197–199], indicating a close association between the lung microbiome and COPD development. However, many of previous studies were restricted by either the low number of samples or contamination-prone sampling methods. The low biomass in the lung and airways increases the importance of the latter issue: contamination is likely to occur during both sampling of the airway microbiome and in the laboratory, including PCR reagents, DNA extraction kits and ground water [32]. Compared with gut microbiota, the understanding of lung microbiota is still in an earlier stage of development, and many uncertainties remain.

For example, previous studies have demonstrated the complexity of lung microbiome about different respiratory microbiota between COPD and healthy subjects [192,193][200–202], however, some of their findings were partly contradictory due to the limited number of samples and different sampling materials [203]. So far, we are still lacking established methods and large-scale datasets to obtain a deep understanding of the relationship between lung microbiome and COPD.

To address current issues in COPD microbiome research, Bergen COPD microbiome study (MicroCOPD) has been designed to investigate the compositional and functional roles of airways microbiome in COPD development [203]. MicroCOPD is an on-going cohort study aiming to provide a large-scale dataset with a particular focus on minimizing contamination in a follow-up periods, the detailed design of the entire study has been published previously [203].

1.3.2 Svalbard permafrost metagenomic study

Permafrost, a type of soil with a temperature that remains at or below 0 °C at least two consecutive years, constitutes almost a quarter of the northern hemisphere. Meanwhile, it represents a unique ecosystem of extreme cold and low nutrient condition for cold-adapted microbiomes [204]. Permafrost contains almost half of soil organic matter (SOM), that consists of plant or animal detritus in various stages of decomposition and tissue of soil microbes as a major carbon sink on Earth [205]. As a result of rises in global temperature, permafrost thaw becomes a serious concern due to the increasing soil microbial activity may lead to release more greenhouse gases (GHGs), such as carbon dioxide (CO_2) and methane (CH_4), thus amplifying the effects of global warming in a positive-feedback loop [206]. Consequently, a better understanding of microbial composition and activity in permafrost is essential to predictive global climate change models. However, we had limited knowledge on permafrost microbiomes until recent years, as most of the microbes fail to be cultivated under laboratory conditions. Advances in NGS-based approaches has thus significantly strengthened our skills to investigate the compositional and functional trait of microbiomes from permafrost.

The Svalbard archipelago is the largest permafrost area in Europe outside of Russia. Svalbard permafrost is proposed to be more sensitive to climate change due to the young age (i.e. Holocene) of the area and the effect of North Atlantic Current [207]. In a previous publication from the same Svalbard permafrost core, we observed a diverse and gradual shift of microbial community spanning from the active layer (AL) at the surface into the deeper permafrost layers (PL) via 16S rRNA analysis [208]. However, the previous study focused on the measuring soil characteristics and identifying taxa via 16S rRNA analysis, which had some limitations. Currently most of our current knowledge of permafrost microbiome is still based on studies using marker gene analysis [209–212], which is informative for describing a microbial community in a low-resolution view but not really applicable for exploring functional potential and novel species [213]. In this project, we performed a deep comparative study through one permafrost core from Svalbard via whole genome metagenomic analysis.

1.3.3 Reconstructing ribosomal genes from total RNA metatranscriptomic data

Metatranscriptomics, the direct sequencing and analysis of all RNA in a microbial community, has been widely used in determining microbiota gene expression and regulation [214–216]. Compared to genomic approaches (marker gene and metagenomics), it offers a more informative perspectives of direct functional output in a given context and is extremely useful in understanding the environment-microbe interactions [155]. The whole microbial RNA pool is dominated by rRNA and tRNA (95-99%) while only small fractions is mRNA (1-5%). So far, most metatranscriptomic studies have focused on mRNA only, depleting rRNA both experimentally and *in silico*. However, rRNA and its abundance could provide novel insights into dynamic structure of microbial community and specific function of protein synthesis [217]. “Total RNA metatranscriptomics” involves the isolation and sequencing of total RNA pools – including mRNA, rRNA, tRNA and other non-coding RNA – from samples directly without any PCR or cloning step. It enables us to obtain both structural (rRNA) and functional information (mRNA) simultaneously of a microbial community in one experiment [218].

Many bioinformatic tools have been developed for metatranscriptomic data, such as IMP [219], SAMSA [220,221], MetaTrans [222], but they are mainly used for studying the functional profiling. However, structural profiling in total RNA metatranscriptomics brings its own advantages. Compared to PCR-based marker gene surveys, rRNA sequences obtained by this approach can access taxonomic diversity in all three domains of life while avoiding amplification bias [28]. Furthermore, it allows for the reconstruction of full length rRNA sequences, enabling a higher resolution for taxonomy profiling. This is typically not feasible in metabarcoding: using short read sequencing technologies results in amplicons with insufficient phylogenetic signal, while long read sequencing allows for longer amplicons but is currently restricted by higher error rates. Last, the relative abundance of rRNA sequences across taxa generally reflects the overall structural activity as well [217]. Existing de novo assembly tools for shotgun sequence reads are primarily applicable for genomic or metagenomic data while are not suitable for recovering rRNA genes [157]. Instead, there are several tools developed specifically for rRNA recovery – like EMIRGE [158], REAGO [157], RAMBL [223], and MATAM [224] – but they are designed for smaller datasets thus cannot be used directly to analyze total RNA metatranscriptomic data. To address those issues, we developed MetaRib, a novel tool for constructing full-length ribosomal gene sequences optimized for total RNA metatranscriptomic data. MetaRib is based on the popular rRNA assembly program EMIRGE [158], together with several improvements.

2. Aims of the thesis

The thesis aims to apply and develop state-of-art bioinformatic methods for NGS-based microbiome research. With the rapid development of bioinformatics and metagenomics, there are numerous tools available for each step of standard analysis, more challenges come from how to optimize current workflows and combine proper tools to meet specific requirements in microbiome study. To take full advantage of data and maximize the valuable information driven from samples, deep understanding of both biological questions and bioinformatic methods are needed. With a close collaboration between bioinformatians, biologists and clinicians, we addressed the following issues in my thesis:

1. Establish a customized bioinformatic workflow for marker gene analysis in Bergen COPD project (**Papers I and II**).

2. Evaluate susceptibility of contamination and methods of depleting contamination both experimentally and in silico (**Papers I and II**).

3. Investigate the stability of airway microbiota in health and obstructive lung disease with repeated bronchoscopy (**Paper II**).

4. Improve current bioinformatic workflow for recovering high quality Metagenome-Assembled Genomes (MAGs) with a new taxonomy-based refinement approach (**Papers III and IV**).

5. Develop a novel comparative strategy for assessing functional potential quantitatively based on genome coverage and KEGG modules in a MAG-centric view (**Papers III and IV**).

6. Develop a novel bioinformatic tool for reconstructing full-length ribosomal gene sequences from large-scale total RNA meta-transcriptomic data (**Paper V**).

3. Results and Discussion

3.1 Characterizing the role of airway microbiota in the development of pulmonary diseases

3.1.1 Conducting a robust experiment in COPD microbiome research

One of the main aims of Bergen MicroCOPD project is to establish a high-quality cohort dataset for human COPD microbiome research. Minimizing the contamination issues is critical in the process. To do so, we explored the best practice for human COPD microbiome research both experimentally and *in silico*. Collecting high quality biological samples is quite challenging in pulmonary disease research. Most of the previous COPD studies used sputum samples due to the cost and efficiency [225–227]. However, they are susceptible to contamination from the oral cavity [228]. Bronchoscopy (endoscopic examination of the airways) offers a method to obtain visually confirmed samples from the airways, but it is more invasive and is associated with technical issues as well. In particular the bronchoscope has to be inserted through the mouth or the nostrils, and unless protected sampling is applied, contamination may still stem from upper airways and/or the oral cavity [229]. However, how different sampling modes will affect the microbiome composition is largely unknown. In this project we evaluated both protected and unprotected methods and compared with OW samples of more than 120 participants. The MicroCOPD project is the largest single site study aiming to investigate the airway microbiota by bronchoscopic sampling.

In **paper I**, we investigated the impact of contamination with three sampling techniques– protected specimen brush (PSB), protected bronchoalveolar lavage (PBAL) and small volume lavage (SVL) – including healthy (67) and COPD (64) participants. Seven samples – negative controls (NCS), oral wash (OW), right lower lobe PSB (rPSB), left upper lobe PSB (lPSB), first fraction of PBAL of the right middle lobe (PBAL1), second fraction of PBAL of the right middle lobe (PBAL2), left

upper lobe SVL (SVL) –were evaluated by marker gene analysis per subject. In brief, we found that alpha diversity decreased in the order OW, SVL, PBAL and PSB while beta diversity showed a distinct distance between OW and other samples. Further PCA analysis indicated a closer similarity (beta-diversity) between OW/SVL and OW/PBAL than OW/PSB. Besides, sampling order (left/right) did not affect the diversity of PSB samples, again indicating that PSB samples enabled a clearly separation from OW samples than SVL and PBAL. Our results highlighted the issue of oropharyngeal contamination in COPD research, and showed that protected sampling approaches were preferred in airway microbiota investigation.

However, some remaining issues should be mentioned. A main limitation is the potential biases introduced by marker gene sequencing, like PCR, which could reduce the accuracy of the result to some extent. Second, to keep the results comparable among groups, we applied mostly standard settings of the QIIME pipeline with no special optimization. However, fine-tuning the QIIME default parameters with mock datasets have been proved a useful method to achieve better performance, especially for the Illumina datasets [230,231]. Third, we considered all OTUs in the NCS as contaminations and simply excluded all of them in the downstream analyses, which could discard some useful information (e.g., species /taxa present in the negative control also present in the lungs). In **paper II**, we made several improvements of our bioinformatic workflow, summarized and discussed in section 3.1.2.

3.1.2 Investigating the stability of airway microbiome by repeated bronchoscopy in healthy and COPD subjects

Previous studies have demonstrated a strong correlation between the shift of lung microbial community composition and the development of obstructive lung disease [229,232,233]. However, we have little knowledge about the stability of airway microbiome over time. In **paper II**, we tracked how microbial communities changed between healthy and COPD subjects who have completed two bronchoscopies during the MicroCOPD project.

21 subjects without 41 patients with COPD were subject to repeated examinations. Seven samples per subject were sequenced as described previously in section 3.1.1. In total 551 amplicon sequence variants (ASVs) were constructed from 19 million sequences in 727 samples. We observed a decreasing trend of alpha diversity in the second examination of OW and PBAL1, and larger differences in subjects not having received intercurrent antibiotics. Comparing within-individual and between-individuals beta-diversity using permutation-based tests indicated that within-individual diversity is significantly lower than between-individual diversity. Comparing sampling methods, non-parametric tests showed that beta diversity between the two examinations followed a pattern of PSB>PBAL2>PBAL1>OW. In summary, our results indicated that although airway microbiota varied over time and other key environmental factors, still a core microbial community may exist over time within each individual.

To date, only a few studies have investigated the stability over time of the lung microbiome with repeated samples. Sinha et al investigated the variability in sputum samples [234], showed differences in both diversity and more variation as sampling interval increased, but they were limited by low sample sizes (less than 10) and potential contaminations during sampling. Here we reported the first relatively large-scale investigation of the stability of airway microbiota with repeated bronchoscopies.

For this project we optimized our bioinformatic analysis in many aspects taking advantages of the rapid advancing of metagenomic analysis methods. First, recently there have been multiple studies recommending to use ASVs rather than OTUs in marker gene analysis [46,61]. Andrei et al evaluated performance between OTU-based workflows (QIIME, MOTHUR, etc) and ASV-based workflows (QIIME2, DADA2, etc), and concluded that ASV-based approaches offered better sensitivity, specificity and reproducibility [65]. In addition to default chimera removal with QIIME2, we performed additional chimera detection using VSEARCH [235] and excluded spurious ASVs according to their distribution across samples [236]. Second, we applied a specialized database – human oral microbiome database (HOMD) [237] – for taxonomic annotation instead of the default, which provided curated species

information in human respiratory tract. Third, we addressed the contamination issue *in silico* and manually as well. In addition to reduce potential contamination using Decontam [52], we further checked the top 50 most abundant taxa and removed species likely to be a be a contaminant judging by other publications [32]. More, advanced methods were applied in comparative diversity analysis. For example, we used Yue-Clayton index in beta diversity [238], which considers the number of bacterial species present and their relative abundances simultaneously. We also introduced random permutation test to assess whether samples from the same individual were significantly more similar than samples from different individuals.

In the current study, samples were dominated by *Firmicutes*, *Actinobacteria*, *Bacteroidetes* and *Proteobacteria* in phylum level, and by *Streptococci*, *Veillonella*, *Prevotella*, *Rothia* and *Haemophilus* in genus level, which was similar with previous observations [200,232,239]. However, we noticed some interesting shifts among groups. For example, OW samples were most stable while PSB samples fluctuated dramatically between two examinations in terms of diversity. This could be a result of more stable status in oral microbiota, or that airway microbiota is more susceptible to random fluctuations due to the low-biomass nature. Further, random permutation test indicated a significantly lower beta diversity within an individual than between individual. Dickson et al proposed an ecological modeling of the respiratory microbiome that the respiratory tract is comprising a continuous immigration of microbes rather than a stable residence [229]. Our results complement this model, in that there were indeed great changes by time, but there were also signs of a small stable core microbiome within individuals, especially in the lower airways.

Yet, some methodological weaknesses need to be mentioned. Although we did not observe covariations between the length of exam intervals and diversity, it does not necessarily mean that there is no correlation. The ideal interval of reexamination of COPD patients is unknown. In our dataset the interval varied substantially (88 - 349 days). No mock community analysis was performed in the current analysis but other authors in the MicroCOPD study have demonstrated fair performance when sequencing mock communities [240].

3.2 Disentangling the complexity of permafrost microbiota with metagenomics

3.2.1 Recovery and distribution of MAGs informed community composition patterns with depth

Recovering metagenomic assembled genomes (MAGs) from metagenomes, especially permafrost metagenomes, is quite challenging due to the extreme complexity and novelty of the microbiota [213,241,242]. Several bioinformatic tools, like MetaBAT2 [92] and MaxBin2 [93], have been widely used for binning contigs into MAGs (bins). DASTool [97], a recently published bin refinement tool, has shown significant improvement of MAGs refinement in many studies [243–245]. However, there is still room for improvement. For example, we still observed 21 out of 64 metagenome bins remained highly contaminated (>=10%) [99] after DASTool screening in our project. As we know, each bin represents an individual genome with single-taxon annotation in theory. However, recovered bins may contain mis-assigned contigs from other taxa, as close species always shared some conserved domains. Thus, it is possible to remove those contaminations by integrating the taxonomic classification result into MAG refinement. To avoid simply discarding these highly contaminated MAGs and improve the quality of MAGs, we developed a script to subset each MAG into collections of contigs from the same taxonomic classification. Our script is able to provide multiple contig subsets corresponding to different ranks for each bin, the user could then evaluate all refined subsets using MAG quality check tools like CheckM [98], and find best tradeoff between completeness and contamination. A detailed description of our MAG refinement workflow is described in **paper IV** and deployed at: https://github.com/yxxue/Recovery-and-refinement-of-MAGs-for-permafrost-metagenome.

Utilizing our refined workflow, we successfully reported 56 out of 64 MAGs with low contamination (<=10%). These 56 MAGs were from 13 phyla, including 8 high, 44 medium and 4 low quality draft according to MIMIG standards [99]. In total the analyzed MAGs constituted around 11.3% of the reads in each sample (min. 7.1%,

max. 13.4%). A detailed description of each MAG was published in **Paper III**. We found MAGs belonging to *Actinobacteria*, *Proteobacteria*, *Bacteroidetes*, *Acidobacteria* and *Chloroflexi* were dominant in Svalbard samples. In **Paper IV**, we further investigated microbial community composition based on changes in the MAG abundance.

Our results showed distinct composition differences between permafrost active layer (AL) and permafrost layer (PL) where predominant MAGs also changed with depth. In the AL, the most abundant phyla were *Acidobacteria* and Proteobacteria while PL MAGs were dominated by *Actinobacteria*, *Bacteroidetes*, *Chloroflexi* and *Proteobacteria*. Members of *Proteobacteria*, *Verrucomicrobia* and *Chloroflexi*, were ubiquitous in PL and had similar abundances in the upper PL (PL1 and PL2) than deep PL samples (PL3 and PL4). We also observed a declined trend in *Acidobacteria* and *Actinobacteria* abundances with increasing depth. Interestingly, more unique but highly represented MAGs were found in the deepest samples, such as *Chloroflexi* in PL3 and *Bacteroidetes* in PL4. Particularly, a previous marker gene analysis in the same core detected *Intrasporangiaceae* to be strongly dominant in the PL [208]. However, we could not detect similar trends in this dataset.

While several MAG refinement strategies are already deployed by IMG/M [143] and Anvi'o [148], our workflow provides a scalable and flexible approach where thousands of bins could be analyzed and screened. More importantly, previous bin refinement programs only consider contamination at phylum level, but it could be detected at all taxonomic levels based on our observation. Our script traces the hierarchical relationships using a user defined percentage threshold and subset bins spanning from phylum to species level. Still, some limitations remain. First, the performance of our bin refinement strategy heavily depends on the accuracy of the taxonomic annotation. In the current study we applied Kaiju [119] to annotate contigs against the NCBI nr database, as Kaiju is extreme fast for large contig sets and there is no specialized databases for permafrost. However, many tools are available to taxonomically annotate metagenomic data and we have not benchmarked them. Simon et al. evaluated the performance of 20 metagenomic classifiers, and pointed out that there was no single

best option: the choice of classifiers depends heavily on the scientific question, computational environment and target taxonomic domains, etc [117]. The rapid growth of reference databases will change the performance dramatically, and it is worth pointing out that removing contigs from a MAG may reduce completeness in some cases due to mis-assembly and annotation. Instead, we subset all possible cleaned MAGs and let users to decide the best tradeoff between completeness and contamination.

3.2.2 Coverage-based functional analysis in a MAG-centric view revealed key metabolic functions in Svalbard permafrost

There are two primary methods for metagenomic functional analysis: mapping the predicted genes against reference databases and then parsing the functional annotation result in either gene or pathway level [208,246]. However, both have some drawbacks. Gene-based approaches utilizes most dominant gene products, but they may overlook biological functions rely on multiple genes and focus on significantly abundant subset only. For another, pathway-level analysis can miss nuanced differences in functional variance as some pathways, especially core pathways, contain many shared sub-pathways or genes. To address above issues, we developed a novel comparative analysis strategy that utilizes KEGG Module (MO), a collection of manually defined functional units each encompassing a set of genes - represented by KO identifiers [247]. Compared to pathway or gene enriched analysis, module-based analysis directly links to specific metabolic capacity. Another issue is how to perform quantitative functional comparisons. Contig coverage is another important metagenomic characteristics [90,248], however, that is currently not used beyond binning [92,93,95,96]. In **paper IV** we utilized coverage patterns of presence/absence to split contigs into several pre-defined groups: each group represent a pattern across the samples with respect to their depth distribution, and further investigated the variance of key MO abundance across groups in a MAG-centric view. The source code and detailed description of our coverage-based functional analysis workflow is deployed at https://github.com/yxxue/Coverage-based-functional-analyisis-in-a-MAG-centric-view.

Here we only include contigs originating from refined MAGs. 20,573 contigs were then assigned to coverage-based classification groups (Table 3.1) according to their coverage distribution in the core. PL groups (PL_SUB, PL_ALL, PL_Pi) represented the largest portion by covering 60% of all contigs. Around 10% of them were shared in both layer (BO) while 13% were only found in AL. 18% was found in all PL while the majority was scattered in a smaller subset (PL_SUB, PL_Pi). We identified only a small fraction of contigs that had a strong correlation (0.9) with depth profile: 5% KI and 1% KD.

Groups	Definition	Criteria	Percentage
AL	Presence in AL	AL >= TH and ALL(PL) <= TH	13.1
BO	Presence Both in AL and PL	AL >= TH and ALL(PL) >= TH	10.2
LO	Absence Both in AL and PL	AL <= TH and ALL(PL) <= TH	14.1
PL_SUB	Presence in subset (2 or 3) PL	AL <= TH and SUB(PL) >= TH	26.4
PL_ALL	Presence in all PL	AL <= TH and ALL(PL) >= TH	18.8
PL_Pi	Presence in unique PL (P1, ... ,P4)	AL<=TH and UNIQUE(PL_Pi) >=TH	3.6 (P1), 2.3 (P2), 2.5 (P3), 5.5 (P4)
KI	Increasing trend in PL_ALL or PL_SUB	In (PL_ALL or PL_SUB) and CORR(PL, DEPTH) >= 0.9	4.9
KD	Decreasing trend in PL_ALL or PL_SUB	In (PL_ALL or PL_SUB) and CORR(PL, DEPTH) <= -0.9	0.9
UN	Unknown groups	Others	2.7

Table 3.1: Contig distribution across groups. AL: normalized coverage in Active layer, PL: normalized coverage in permafrost layer samples. TH: threshold (median of normalized coverage). DEPTH (cm under surface): 110, 122, 135, 170. CORR: Pearson correlation.

In total 451 out of 808 MO were detected in Svalbard MAGs. We manually selected 102 important MO with 8 pathways and observed distinct MO abundance among groups. We examined the trend in several key permafrost related pathways among different groups. MAGs belonging to group BO and PL showed a strong representation

of aerobic respiratory processes, such as F-type ATPase and NADH: quinone oxidoreductase (NQR). We further confirmed the aerobic respiration as the dominant carbon cycling pathway in this core, as dehydrogenases involved modules were in neither high abundance nor showed strong grouping trends. Polymer hydrolysis and CAZY functions were also found in abundance especially in core in PL groups (PL_ALL and PL_SUB). Several key metabolic functions of nitrogen and sulfur cycles were detected, whereas another key biogeochemical process methane metabolism was missed. Besides, we observed a suite of MO related with stress responses and antibiotic resistance – including KdpDE:potassium transport system, phosphate starvation response system (PhoR−PhoB), and redox response and chemotaxis, which may indicate how microbiota counter the extreme environmental stress in permafrost. Detailed result about variance of key metabolic functions was described in **paper IV**.

One of the most interesting findings in this study was that Svalbard MAGs were mostly aerobic and showed enriched functional potentials involved in ammonium, sulfur and phosphate metabolism. These results indicate that a substantial investment in energy production is required for permafrost microbiome. Our results are also in concurrence with previous activity measurements from the same location where Müller et al. [208] showed that permafrost thaw up to four times higher CO_2 respiration rate were observed under aerobic than anaerobic conditions through a series of incubations. Our analysis also indicated that one of the key permafrost microbiome traits is to obtain effective resources via various metabolic reactions.

In this study, we proposed a computational method that combine metagenomic and biological information in a MAG-centric view. Although we focused on Svalbard permafrost depth profile, our method enabled us to determine core functions and trends of multiple samples in a study. However, some limitations need to be discussed. First, it should be pointed out that the presence and abundance of MO within permafrost microbiota does not mean they are truly active or expressed, as our approach is based on genomics rather than transcriptomics. Therefore, these results only represent functional potential in the environment and provide supporting information of gene activities. But this limitation holds for all metagenomic datasets, even

metatranscriptomics is influenced by the time that sample is collected. In addition, metatranscriptomics is only applicable under some circumstances, it may fail due to the difficulty of extracting RNA from extreme environment like permafrost [213]. Second, compared with contig-based or gene-based functional profiling that utilized most of the reads, we focused on contigs that were able to be binned in a MAG-centric view. We excluded unbinned contigs as they can lead to erroneous conclusions or misinterpretations [249]. Focusing on MAGs enabled us to recover genomic information precisely and ensure pathway completion when calculating MO abundance. As described in many published metagenome studies, either read-based or assembly-based methods can annotate 20-30% of the data from the complex soil type environments. As a result, all these approaches bring their own biases in interpreting functional potential.

3.3 Reconstructing full-length rRNA sequences from total RNA metatranscriptomics

In **Paper V** we present the tool MetaRib for reconstructing rRNA genes from large scale total RNA metatranscriptomic data. Compared with existing tools, MetaRib is able to recover full-length rRNA contigs with a low false positive rate across multiple samples, even for very large datasets, together with accurate taxonomy-independent abundance estimation. We address the challenge posed by large complex datasets by integrating sub-assembly, dereplication and mapping in an iterative approach, with additional post-processing steps. Our approach utilized the uneven taxon-abundance distribution common for microbial communities [250], which makes it possible to reconstruct most abundant species with a small subset in a first few iterations, meanwhile reduces the redundancy the dataset iteratively helps to capture rarer species. To produce only full-length sequences, we applied overlap-based cluster to process fragmented contigs and duplicates. In addition, sample-based mapping statistics is not only used for estimating abundance, but also helps to filter false positive records. Although total RNA metatranscriptomics has not been widely used,

especially structural profiling, our approach opens up several new perspectives and enables a deeper understanding of how microbiota is structured and distributed.

To simulate the complexity of real microbiome communities, three simulated datasets were built, each dataset included 5 million reads, which was generated by 1000 randomly selected full-length rRNA contigs following a log-normal abundance distribution. In dataset a, all contigs were included in the corresponding reference. In dataset b, contigs were not identical but highly similar (between 95% to 99%) with the reference. Dataset c was the opposite case of dataset a that all contigs were removed from the reference.

In simulated datasets, MetaRib demonstrated significant speedup (60X) compared to EMIRGE using the same parameters, which could process 5 million reads in a few minutes instead of days. We further evaluated the performance of two tools in terms of Precision, Sensitivity and F1-score for all three simulated datasets. EMIRGE had a higher sensitivity compared to MetaRib, but the main issue was that it also produced a large number of 'false' contigs which lead to a quite low precision even in an ideal case (Dataset a). Conversely, MetaRib was able to recover almost all source contigs if they were represented in the reference while produced far fewer such 'false' sequences. In summary, MetaRib showed the best overall performance in all datasets with F1-score evaluation.

The AshBack project is a large-scale total RNA metatranscriptomic study – composed of 325 Gb rRNA sequences – to assess the impact of wood ash on agricultural and forest soil microbial communities [251]. Due to the lack of bioinformatic tools and computational constraints, previous rRNA analysis was performed on a small subset (1.5 million reads) of each sample [251]. We reanalyzed the full dataset using MetaRib. We observed more rRNA contigs but similar trends of richness and Shannon diversity as revealed by previous analyses. However, our MetaRib-based re-analysis indicated considerably less fluctuation of diversity in agricultural soil samples. MetaRib was also able to recover more rRNA sequences across all domains and capture more taxa than previous analysis. For example, we detected fungal

Mucoromycota, which appears to be dominant in high dose ash concentration at forest soil, although overlooked in the original analysis [251]. In addition, MetaRib could perform taxonomy-independent abundance estimation that were not possible when assembling reads sample-by-sample. We observed several interesting correlations between dominant contigs and metadata: As an example, *Proteobacteria* were ubiquitous in both soils and showed more variation in the forest. Further, *Acidobacteria* were dominant in the forest soil but dropped significantly at the highest ash concentration.

Still, some challenges remain. Both EMIRGE and consequently MetaRib perform best when reconstructing full-length contigs similar to sequences represented in the reference database. The contrasting results of simulated datasets indicate that MetaRib is largely applicable to relatively well-characterized environments, and emphasis the importance of reference database. Increasing the size of reference database would address the issue partially, but it will also result in longer execution time. Besides, since EMIRGE is limited to construct contigs with a maximum 97% similarity to each other [158], we recommend to use a non-redundant reference database with threshold similar to this. More recent tools such as MATAM [224] have been shown to perform better than EMIRGE in small datasets, and could be considered as a further improvement of MetaRib. An advantage of total RNA metatranscriptomics is to estimate relative abundances of rRNA sequences as proxies of microbial taxa without PCR bias. Notably, Blazwicz et al. pointed out that rRNA sequences were not an unbiased estimates of neither metabolic activity nor species growth in some cases, since rRNA gene copy number and patterns of ribosomal transcription and retention vary between species [217].

4. Concluding remarks

Advances in sequencing technology has greatly expanded our understanding of the fascinating microbial universe. The studies presented in this thesis address the extensive applications and challenges of NGS-based approaches, including marker gene, metagenomic and metatranscriptomic techniques, in microbiome research. The work of this thesis contributes novel insights in community composition and functional potential of various ecosystems, including human respiratory tract, permafrost, and wood ash by addressing bioinformatic challenges and suggesting solutions when working with such diverse environment.

The study on human airway microbiota explores the effect of contamination in COPD microbiome research, both experimentally and *in silico*. Our results underscore the importance of sampling methods and oropharyngeal contamination issues, and inform the use of protected specimen brushes in airway microbiota. This large-scale study examines the stability of airway microbiota over time utilizing optimized experimental protocols and bioinformatic workflows, demonstrating that the composition of airway microbiota changes overtime. Nonetheless, there seem to exist a core microbiome community. Future research should focus on identifying core microbiota and its correlation with COPD development with other omics technologies, as the current study is limited by the marker gene approach.

Predicting the metabolic response is one of the main challenges nowadays due to the enormous diversity and complexity in microbial communities. Soil represents one of the most complex environments and thus very challenging to study. Our analysis of Svalbard metagenomic samples provides novel methods and insights for understanding metabolic functionality in permafrost microbiota. The Metagenome Assembled Genomes (MAGs) centric strategy developed here enables us to focus on genomic contigs within biological context, and enables application of quantitative comparison of functional potential within pathway completion. Our results reveal novel metabolic potentials in Svalbard permafrost MAGs, especially aerobic response, including ammonium, sulfur and phosphate metabolism that have hitherto not been reported in a

permafrost environment. Further research could also focus on investigating aerobic metabolism in more detail, integrating with other omics, like metatranscriptomics, which could reveal functional activity and response in a more direct fashion.

rRNA sequences in metatranscriptomics provides novel insights about the structural information of active microbial communities. We have implemented MetaRib to enable reconstruction of full-length rRNA contigs from large-scale rRNA metatranscriptomic data, achieving a comparably efficiency and accuracy in benchmark performance, with the added taxonomy-independent abundance estimation. Further improvements could include the ability to handle novel rRNA contigs and unmapped reads, as current version of MetaRib is still a reference-based approach. Although total RNA analysis has not been widely used in microbiome research, we hope our approach will encourage researchers to make more use of the valuable rRNA sequences generated, which enables a deeper understanding of the structural information of microbial communities.

Conducting a robust bioinformatic analysis in microbiome research requires many steps and high degree of accuracy as the choice of proper approaches and parameters in each step can impact the final results and the biological interpretation significantly. In this thesis we explored the best practices to mining the valuable information from complex ecosystems, with a specific focus on poorly-characterized and low-biomass environments. We are still lacking standardized protocols right now to compare different studies, which is one of the most urgent things to be resolved. Multi-omics integration represents another approach in this field as this enable us to provide a more complete picture of the whole system beyond the capacity of any single approach, however there is no ***"golden standard"*** certain method at present due to the challenges of integrating different types of data. Despite these challenges, I am confident there will be coming more exciting discoveries from the microbiome research field in the future, given the rapid development of biological technologies and bioinformatics tools.

Bibliography

1. Hamarneh S. Measuring the Invisible World. The life and works of Antoni van Leeuwenhoek. A. Schierbeek. Abelard-Schuman, New York, 1959. 223 pp. $4. Science (80-). American Association for the Advancement of Science; 1960;132: 289–290. doi:10.1126/science.132.3422.289

2. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995;59: 143–169. doi:0146-0749/95/$04.00?0

3. Carraro L, Maifreni M, Bartolomeoli I, Martino ME, Novelli E, Frigo F, et al. Comparison of culture-dependent and -independent methods for bacterial community monitoring during Montasio cheese manufacturing. Res Microbiol. Elsevier Masson SAS; 2011;162: 231–239. doi:10.1016/j.resmic.2011.01.002

4. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci U S A. 1977;74: 5088–5090. doi:10.1073/pnas.74.11.5088

5. Liu WT, Marsh TL, Cheng H, Forney LJ. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl Environ Microbiol. 1997;63: 4516–4522.

6. Fischer SG, Lerman LS. Separation of random fragments of DNA according to properties of their sequences. Proc Natl Acad Sci. 1980;77: 4420–4424. doi:10.1073/pnas.77.8.4420

7. Kim DW. Real time quantitative PCR. Exp Mol Med. Cold Spring Harbor Laboratory Press; 2001;33: 101–109. doi:10.1006/meth.2001.1260

8. Langer-Safer PR, Levine M, Ward DC. Immunological methods for mapping genes on Drosophila polytene chromosomes. Proc Natl Acad Sci U S A. National Academy of Sciences; 1982;79: 4381–4385. doi:10.1073/pnas.79.14.4381

9. Loy A, Bodrossy L. Highly parallel microbial diagnostics using oligonucleotide microarrays. Clinica Chimica Acta. 2006. pp. 106–119. doi:10.1016/j.cccn.2005.05.041

10. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: The next culture-independent game changer. Front Microbiol. Frontiers; 2017;8: 1069. doi:10.3389/fmicb.2017.01069

11. Wilkins MHF, Stokes AR, Wilson HR. Molecular structure of deoxypentose nucleic acids. 50 Years DNA. Nature Publishing Group; 2016;171: 84–86. doi:10.1038/nature01396

12. Schneider GF, Dekker C. DNA sequencing with nanopores. Nat Biotechnol. 2012;30: 326–328. doi:10.1038/nbt.2181

13. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380. doi:10.1038/nature03959

14. Nyren P, Pettersson B, Uhlen M. Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. Anal Biochem. 1993;208: 171–175. doi:10.1006/abio.1993.1024

15. den Bakker HC, Moreno Switt AI, Govoni G, Cummings CA, Ranieri ML, Degoricija L, et al. Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of Salmonella enterica. BMC Genomics. BioMed Central; 2011;12: 245. doi:10.1186/1471-2164-12-245

16. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456: 53–59. doi:10.1038/nature07517

17. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: Past, present and future. Nature. Nature Publishing Group; 2017;550: 345–353. doi:10.1038/nature24286

18. Tan G, Opitz L, Schlapbach R, Rehrauer H. Long fragments achieve lower base quality in Illumina paired-end sequencing. Sci Rep. 2019;9: 2856. doi:10.1038/s41598-019-39076-7

19. Nakazato T, Ohta T, Bono H. Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read Archive. PLoS One. 2013;8. doi:10.1371/journal.pone.0077910

20. Levene HJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. Science (80-). 2003;299: 682–686. doi:10.1126/science.1079700

21. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics, Proteomics and Bioinformatics. 2015. pp. 278–289. doi:10.1016/j.gpb.2015.08.002

22. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. Nat Biotechnol. 2012;30: 344–348. doi:10.1038/nbt.2147

23. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: State of the art, applications and challenges. Brief Bioinform. 2017;19: 1256–1272. doi:10.1093/bib/bbx062

24. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chem Biol. 1998;5. doi:10.1016/S1074-5521(98)90108-9

25. Ren R, Sun Y, Zhao Y, Geiser D, Ma H, Zhou X. Phylogenetic Resolution of Deep Eukaryotic and Fungal Relationships Using Highly Conserved Low-Copy Nuclear Genes. Genome Biol Evol. 2016;8: 2683–2701. doi:10.1093/gbe/evw196

26. Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G. High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods. 2013;95: 401–414. doi:10.1016/j.mimet.2013.08.011

27. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, et al. Analysis of the Lung Microbiome in the "Healthy" Smoker and in COPD. Bereswill S, editor. PLoS One. 2011;6: e16384. doi:10.1371/journal.pone.0016384

28. Lanzén A, Jørgensen SL, Bengtsson MM, Jonassen I, Øvreås L, Urich T. Exploring the composition and diversity of microbial communities at the Jan Mayen hydrothermal vent field using RNA and DNA. FEMS Microbiol Ecol. 2011;77: 577–589. doi:10.1111/j.1574-6941.2011.01138.x

29. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, et al. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci Rep. 2017;7: 1–14. doi:10.1038/s41598-017-06665-3

30. Ziesemer KA, Mann AE, Sankaranarayanan K, Schroeder H, Ozga AT, Brandt BW, et al. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. Sci Rep. 2015;5: 1–19. doi:10.1038/srep16498

31. Gupta S, Mortensen MS, Schjørring S, Trivedi U, Vestergaard G, Stokholm J, et al. Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. Commun Biol. 2019;2: 291. doi:10.1038/s42003-019-0540-1

32. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12: 1–12. doi:10.1186/s12915-014-0087-z

33. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational meta'omics for microbial community studies. Mol Syst Biol. Nature Publishing Group; 2013;9: 1–15. doi:10.1038/msb.2013.22

34. Poinar HN, Schwarz C, Qi J, Shapiro B, MacPhee RDE, Buigues B, et al. Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA.

Science (80-). 2006;311: 392–394. doi:10.1126/science.1123360

35. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007. pp. 804–810. doi:10.1038/nature06244

36. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science (80-). 2015;348: 1261359–1261359. doi:10.1126/science.1261359

37. Loman NJ, Constantinidou C, Christner M, Chan JZM, Quick J, Weir JC, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic Escherichia coli O104:H4. JAMA - J Am Med Assoc. 2013;309: 1502–1510. doi:10.1001/jama.2013.3231

38. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science (80-). 2004;304: 66–74. doi:10.1126/science.1093857

39. Teeling H, Glockner FO. Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. Brief Bioinform. 2012;13: 728–742. doi:10.1093/bib/bbs039

40. Comtet-Marre S, Parisot N, Lepercq P, Chaucheyras-Durand F, Mosoni P, Peyretaillade E, et al. Metatranscriptomics Reveals the Active Bacterial and Eukaryotic Fibrolytic Communities in the Rumen of Dairy Cow Fed a Mixed Diet. Front Microbiol. 2017;8. doi:10.3389/fmicb.2017.00067

41. Bao G, Wang M, Doak TG, Ye Y. Strand-specific community RNA-seq reveals prevalent and dynamic antisense transcription in human gut microbiota. Front Microbiol. 2015;6. doi:10.3389/fmicb.2015.00896

42. Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC. Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome. Ward N, editor. PLoS One. 2008;3: e2527. doi:10.1371/journal.pone.0002527

43. Helbling DE, Ackermann M, Fenner K, Kohler HPE, Johnson DR. The activity level of a microbial community function can be predicted from its metatranscriptome. ISME J. 2012;6: 902–904. doi:10.1038/ismej.2011.158

44. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of metatranscriptomics in microbiome research. Bioinform Biol Insights. 2016;10: 19–25. doi:10.4137/BBI.S34610

45. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Corrigendum: Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35: 1211. doi:10.1038/nbt1217-1211b

46. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. Springer US; 2018;16: 410–422. doi:10.1038/s41579-018-0029-9

47. Reeder J, Knight R. The "rare biosphere": A reality check. Nat Methods. 2009;6: 636–637. doi:10.1038/nmeth0909-636

48. Andrews S, Krueger F, Seconds-Pichon A, Biggins F, Wingett S. FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics [Internet]. Babraham Institute. 2015. p. 1. doi:citeulike-article-id:11583827

49. Gordon A, Hannon GJ. Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpubl http//hannonlab cshl edu/fastx_ toolkit. 2010;

50. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing Noise From Pyrosequenced Amplicons. BMC Bioinformatics. 2011;12: 38. doi:10.1186/1471-2105-12-38

51. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27: 863–864. doi:10.1093/bioinformatics/btr026

52. Davis NM, Proctor DiM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. Microbiome; 2018;6: 1–14. doi:10.1186/s40168-018-0605-2

53. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods. 2011;8: 761–765. doi:10.1038/nmeth.1650

54. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27: 2194–2200. doi:10.1093/bioinformatics/btr381

55. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward D V., Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 2011;21: 494–504. doi:10.1101/gr.112730.110

56. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7: 335–336. doi:10.1038/nmeth.f.303

57. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75: 7537–7541. doi:10.1128/AEM.01541-09

58. Wright ES, Yilmaz LS, Noguera DR. DECIPHER, a Search-Based Approach to

Chimera Identification for 16S rRNA Sequences. Appl Environ Microbiol. 2012;78: 717–725. doi:10.1128/AEM.06516-11

59. Chen W, Zhang CK, Cheng Y, Zhang S, Zhao H. A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. Casiraghi M, editor. PLoS One. 2013;8: e70837. doi:10.1371/journal.pone.0070837

60. Caruso V, Song X, Asquith M, Karstens L. Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. Gibbons SM, editor. mSystems. 2019;4. doi:10.1128/mSystems.00163-18

61. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11: 2639–2643. doi:10.1038/ismej.2017.119

62. Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ. 2018;6: e5364. doi:10.7717/peerj.5364

63. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. Gilbert JA, editor. mSystems. 2017;2. doi:10.1128/msystems.00191-16

64. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13: 581–583. doi:10.1038/nmeth.3869

65. Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. Seo J-S, editor. PLoS One. 2020;15: e0227434. doi:10.1371/journal.pone.0227434

66. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm: Robust and fast clustering method for amplicon-based studies. PeerJ. 2014; doi:10.7717/peerj.593

67. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ. 2015;3: e1420. doi:10.7717/peerj.1420

68. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012;41: D590–D596. doi:10.1093/nar/gks1219

69. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72: 5069–5072. doi:10.1128/AEM.03006-05

70. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009;37. doi:10.1093/nar/gkn879

71. Balvočiute M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? BMC Genomics. 2017;18. doi:10.1186/s12864-017-3501-4

72. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Author Correction: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37: 1091–1091. doi:10.1038/s41587-019-0252-6

73. Matias Rodrigues JF, Schmidt TSB, Tackmann J, Von Mering C. MAPseq: Highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. Birol I, editor. Bioinformatics. 2017;33: 3808–3810. doi:10.1093/bioinformatics/btx517

74. Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. Gigascience. Oxford University Press; 2018;7: 1–10. doi:10.1093/gigascience/giy054

75. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 2017;18. doi:10.1186/s13059-017-1299-7

76. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048. doi:10.1093/bioinformatics/btw354

77. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research. 2018;7: 1338. doi:10.12688/f1000research.15931.2

78. Bushnell B. BBTools - DOE Joint Genome Institute [Internet]. 2014 [cited 3 Sep 2019]. Available: https://jgi.doe.gov/data-and-tools/bbtools/%0Ahttp://sourceforge.net/projects/bbmap/

79. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research. 2015;4: 900. doi:10.12688/f1000research.6924.1

80. Myers Jr EW. A history of DNA sequence assembly. it - Inf Technol. 2016;58. doi:10.1515/itit-2015-0047

81. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. Science (80-). 2000;287: 2196–2204. doi:10.1126/science.287.5461.2196

82. Myers EW. The fragment assembly string graph. Bioinformatics. 2005;21. doi:10.1093/bioinformatics/bti1114

83. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 2001;98: 9748–9753. doi:10.1073/pnas.171285098

84. Peng Y, Leung HCM, Yiu SM, Chin FYL. Meta-IDBA: A de Novo assembler for metagenomic data. Bioinformatics. 2011;27: i94–i101. doi:10.1093/bioinformatics/btr216

85. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28: 1420–1428. doi:10.1093/bioinformatics/bts174

86. Afiahayati, Sato K, Sakakibara Y. MetaVelvet-SL: An extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Res. 2015;22: 69–77. doi:10.1093/dnares/dsu041

87. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31: 1674–1676. doi:10.1093/bioinformatics/btv033

88. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: A new versatile metagenomic assembler. Genome Res. 2017;27: 824–834. doi:10.1101/gr.213959.116

89. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. Nat Methods. 2017;14: 1063–1071. doi:10.1038/nmeth.4458

90. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Corrigendum: Shotgun metagenomics, from sampling to analysis [Internet]. Nature biotechnology. Nature Publishing Group; 2017. p. 1211. doi:10.1038/nbt1217-1211b

91. Karlin S, Mrázek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol. 1997;179: 3899–3913. doi:10.1128/jb.179.12.3899-3913.1997

92. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;2015: e1165. doi:10.7717/peerj.1165

93. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32: 605–607. doi:10.1093/bioinformatics/btv638

94. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31: 533–538. doi:10.1038/nbt.2579

95. Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11: 1144–1146. doi:10.1038/nmeth.3103

96. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: An automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014;2014: e603. doi:10.7717/peerj.603

97. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol. Springer US; 2018;3: 836–843. doi:10.1038/s41564-018-0171-1

98. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25: 1043–1055. doi:10.1101/gr.186072.114

99. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35: 725–731. doi:10.1038/nbt.3893

100. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38: e191–e191. doi:10.1093/nar/gkq747

101. Hoff KJ. The effect of sequencing errors on metagenomic gene prediction. BMC Genomics. 2009; doi:10.1186/1471-2164-10-520

102. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res. 2012; doi:10.1093/nar/gkr1067

103. Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics. 2012;28: 2223–2230. doi:10.1093/bioinformatics/bts429

104. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010; doi:10.1093/nar/gkq275

105. Hoff KJ, Lingner T, Meinicke P, Tech M. Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res. 2009;37: W101–W105. doi:10.1093/nar/gkp327

106. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014; doi:10.1093/bioinformatics/btu153

107. Al-Ajlan A, El Allali A. Feature selection for gene prediction in metagenomic fragments. BioData Min. 2018; doi:10.1186/s13040-018-0170-z

108. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

109. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15. doi:10.1186/gb-2014-15-3-r46

110. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. bioRxiv. Cold Spring Harbor Laboratory; 2019; 762302. doi:10.1101/762302

111. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol. BioMed Central; 2018;19: 198. doi:10.1186/s13059-018-1568-0

112. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. PeerJ Inc.; 2017;3: e104. doi:10.7717/peerj-cs.104

113. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. BioMed Central; 2015;16: 236. doi:10.1186/s12864-015-1419-2

114. Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK- *S*. Bioinformatics. Narnia; 2016;32: 3823–3825. doi:10.1093/bioinformatics/btw542

115. Ainsworth D, Sternberg MJE, Raczy C, Butcher SA. k-SLAM: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. Nucleic Acids Res. 2016; gkw1248. doi:10.1093/nar/gkw1248

116. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: Rapid and sensitive classification of metagenomic sequences. Genome Res. 2016; doi:10.1101/gr.210641.116

117. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell. Elsevier Inc.; 2019;178: 779–794. doi:10.1016/j.cell.2019.07.010

118. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12: 59–60. doi:10.1038/nmeth.3176

119. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for

metagenomics with Kaiju. Nat Commun. 2016;7: 11257. doi:10.1038/ncomms11257

120. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. Eisen JA, editor. PLoS Comput Biol. 2012;8: e1002358. doi:10.1371/journal.pcbi.1002358

121. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. Springer US; 2018;15: 962–968. doi:10.1038/s41592-018-0176-y

122. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733–D745. doi:10.1093/nar/gkv1189

123. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23: 1282–1288. doi:10.1093/bioinformatics/btm098

124. Prestat E, David MM, Hultman J, Taş N, Lamendella R, Dvornik J, et al. FOAM (Functional Ontology Assignments for Metagenomes): a Hidden Markov Model (HMM) database with environmental focus. Nucleic Acids Res. 2014;42: e145–e145. doi:10.1093/nar/gku702

125. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016; doi:10.1093/nar/gkw387

126. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 2015;9: 207–216. doi:10.1038/ismej.2014.106

127. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47: W81–W87. doi:10.1093/nar/gkz310

128. Prakash T, Taylor TD. Functional assignment of metagenomic data: Challenges and applications. Brief Bioinform. 2012; doi:10.1093/bib/bbs033

129. Letunic I, Bork P. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 2018;46: D493–D496. doi:10.1093/nar/gkx922

130. Bateman A. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47: D506–D515. doi:10.1093/nar/gky1049

131. Fang G, Bhardwaj N, Robilotto R, Gerstein MB. Getting Started in Gene

Orthology and Functional Analysis. Troyanskaya O, editor. PLoS Comput Biol. 2010;6: e1000703. doi:10.1371/journal.pcbi.1000703

132. Tatusov RL. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28: 33–36. doi:10.1093/nar/28.1.33

133. Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, et al. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 2007;36: D250–D254. doi:10.1093/nar/gkm796

134. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28: 27–30. doi:10.1093/nar/28.1.27

135. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014;42: D206–D214. doi:10.1093/nar/gkt1226

136. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32: 258D – 261. doi:10.1093/nar/gkh036

137. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42: D222–D230. doi:10.1093/nar/gkt1223

138. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 2011;39: D225–D229. doi:10.1093/nar/gkq1189

139. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37: D211–D215. doi:10.1093/nar/gkn785

140. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, et al. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2017;45: D289–D295. doi:10.1093/nar/gkw1098

141. Finn RD, Clements J, Eddy SR. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. 2011; doi:10.1093/nar/gkr367

142. Keegan KP, Glass EM, Meyer F. MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods in Molecular Biology. 2016. doi:10.1007/978-1-4939-3369-3_13

143. Chen IMA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, et al. IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res. 2019;47: D666–D677. doi:10.1093/nar/gky901

144. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol. 2016; doi:10.1371/journal.pcbi.1004957

145. Hunter S, Corbett M, Denise H, Fraser M, Gonzalez-Beltran A, Hunter C, et al. EBI metagenomics - A new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res. 2014;42: D600–D606. doi:10.1093/nar/gkt961

146. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol. 2013;14: R2. doi:10.1186/gb-2013-14-1-r2

147. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: A metagenomic assembly, annotation and profiling framework. Bioinformatics. 2016;32: 2520–2523. doi:10.1093/bioinformatics/btw183

148. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3: e1319. doi:10.7717/peerj.1319

149. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase: The United States department of energy systems biology knowledgebase. Nature Biotechnology. 2018. pp. 566–569. doi:10.1038/nbt.4163

150. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, et al. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A. 2014;111: E2329–E2338. doi:10.1073/pnas.1319284111

151. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28: 3211–3217. doi:10.1093/bioinformatics/bts611

152. Leung HCM, Yiu SM, Chin FYL. IDBA-MTP: A hybrid metatranscriptomic assembler based on protein information. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014. doi:10.1007/978-3-319-05269-4_12

153. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29: 644–652. doi:10.1038/nbt.1883

154. Ye Y, Tang H. Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. Bioinformatics. 2016;32: 1001–1008. doi:10.1093/bioinformatics/btv510

155. Shakya M, Lo CC, Chain PSG. Advances and challenges in metatranscriptomic

analysis. Frontiers in Genetics. 2019. doi:10.3389/fgene.2019.00904

156. Gruber-Vodicka HR, Seah BK, Pruesse E. phyloFlash — Rapid SSU rRNA profiling and targeted assembly from metagenomes. bioRxiv. 2019; 521922. doi:10.1101/521922

157. Yuan C, Lei J, Cole J, Sun Y. Reconstructing 16S rRNA genes in metagenomic data. Bioinformatics. 2015;31: i35–i43. doi:10.1093/bioinformatics/btv231

158. Miller CS, Handley KM, Wrighton KC, Frischkorn KR, Thomas BC, Banfield JF. Short-Read Assembly of Full-Length 16S Amplicons Reveals Bacterial Diversity in Subsurface Sediments. Gilbert JA, editor. PLoS One. 2013;8: e56018. doi:10.1371/journal.pone.0056018

159. Wagner BD, Grunwald GK, Zerbe GO, Mikulich-Gilbertson SK, Robertson CE, Zemanick ET, et al. On the Use of Diversity Measures in Longitudinal Sequencing Studies of Microbial Communities. Front Microbiol. 2018;9. doi:10.3389/fmicb.2018.01037

160. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, et al. Conducting a Microbiome Study. Cell. 2014;158: 250–262. doi:10.1016/j.cell.2014.06.037

161. Shade A, Gregory Caporaso J, Handelsman J, Knight R, Fierer N. A meta-analysis of changes in bacterial and archaeal communities with time. ISME J. 2013; doi:10.1038/ismej.2013.54

162. Barwell LJ, Isaac NJB, Kunin WE. Measuring β-diversity with species abundance data. J Anim Ecol. 2015;84: 1112–1122. doi:10.1111/1365-2656.12362

163. Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol. 2007;62: 142–160. doi:10.1111/j.1574-6941.2007.00375.x

164. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26: 139–140. doi:10.1093/bioinformatics/btp616

165. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

166. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. McHardy AC, editor. PLoS Comput Biol. 2014;10: e1003531. doi:10.1371/journal.pcbi.1003531

167. Anderson MJ, Walsh DCI. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? Ecol Monogr. 2013;83: 557–574. doi:10.1890/12-2010.1

168. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001;26: 32–46. doi:10.1046/j.1442-9993.2001.01070.x

169. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Genes Dis. Elsevier Ltd; 2017;4: 138–148. doi:10.1016/j.gendis.2017.06.001

170. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 2016;10: 1669–1681. doi:10.1038/ismej.2015.235

171. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: And this is not optional. Front Microbiol. 2017;8: 1–6. doi:10.3389/fmicb.2017.02224

172. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML. Balances: a New Perspective for Microbiome Analysis. Lozupone C, editor. mSystems. 2018;3. doi:10.1128/msystems.00053-18

173. Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. PLoS Comput Biol. 2012;8. doi:10.1371/journal.pcbi.1002687

174. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. von Mering C, editor. PLoS Comput Biol. 2015;11: e1004226. doi:10.1371/journal.pcbi.1004226

175. Schwager E, Mallick H, Ventz S, Huttenhower C. A Bayesian method for detecting pairwise associations in compositional data. Blekhman R, editor. PLoS Comput Biol. 2017;13: e1005852. doi:10.1371/journal.pcbi.1005852

176. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Heal Dis. 2015;26. doi:10.3402/mehd.v26.27663

177. Yazdani M, Taylor BC, Debelius JW, Li W, Knight R, Smarr L. Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016. IEEE; 2016. pp. 1272–1280. doi:10.1109/BigData.2016.7840731

178. Teng F, Yang F, Huang S, Bo C, Xu ZZ, Amir A, et al. Prediction of Early Childhood Caries via Spatial-Temporal Variations of Oral Microbiota. Cell Host Microbe. 2015;18: 296–306. doi:10.1016/j.chom.2015.08.005

179. Chaudhary N, Sharma AK, Agarwal P, Gupta A, Sharma VK. 16S classifier: A tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. Dalby AR, editor. PLoS One. 2015;10: e0116106. doi:10.1371/journal.pone.0116106

180. Rasheed Z, Rangwala H. Metagenomic taxonomic classification using extreme learning machines. J Bioinform Comput Biol. 2012;10: 1250015. doi:10.1142/S0219720012500151

181. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol. 2016;12: 1–26. doi:10.1371/journal.pcbi.1004977

182. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 2018;6: 23. doi:10.1186/s40168-018-0401-z

183. Pflughoeft KJ, Versalovic J. Human Microbiome in Health and Disease. Annu Rev Pathol Mech Dis. 2012;7: 99–122. doi:10.1146/annurev-pathol-011811-132421

184. Gower JC. Generalized procrustes analysis. Psychometrika. 1975; doi:10.1007/BF02291478

185. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome. 2013;1: 17. doi:10.1186/2049-2618-1-17

186. DOLEDEC S, CHESSEL D. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshw Biol. 1994;31: 277–294. doi:10.1111/j.1365-2427.1994.tb01741.x

187. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 2016;10: 1669–1681. doi:10.1038/ismej.2015.235

188. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. Anal Chem. 2012;84: 5035–5039. doi:10.1021/ac300698c

189. Tuncbag N, Gosline SJC, Kedaigle A, Soltis AR, Gitter A, Fraenkel E. Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. Prlic A, editor. PLOS Comput Biol. 2016;12: e1004879. doi:10.1371/journal.pcbi.1004879

190. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2017;2: 16180. doi:10.1038/nmicrobiol.2016.180

191. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569: 655–662. doi:10.1038/s41586-019-1237-9

192. Vos T, Allen C, Arora M, Barber RM, Bhutta ZA, Brown A, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388: 1545–1602. doi:10.1016/S0140-6736(16)31678-6

193. Foreman MG, Campos M, Celedón JC. Genes and Chronic Obstructive Pulmonary Disease. Med Clin North Am. 2012;96: 699–711. doi:10.1016/j.mcna.2012.02.006

194. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Human gut microbes associated with obesity. Nature. 2006;444: 1022–1023. doi:10.1038/4441022a

195. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569: 655–662. doi:10.1038/s41586-019-1237-9

196. Dai Z, Coker OO, Nakatsu G, Wu WKK, Zhao L, Chen Z, et al. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome. 2018;6: 70. doi:10.1186/s40168-018-0451-2

197. Pragman AA, Kim HB, Reilly CS, Wendt C, Isaacson RE. The Lung Microbiome in Moderate and Severe Chronic Obstructive Pulmonary Disease. Taube C, editor. PLoS One. 2012;7: e47305. doi:10.1371/journal.pone.0047305

198. Huang YJ, Sethi S, Murphy T, Nariya S, Boushey HA, Lynch S V. Airway Microbiome Dynamics in Exacerbations of Chronic Obstructive Pulmonary Disease. J Clin Microbiol. 2014;52: 2813–2823. doi:10.1128/JCM.00035-14

199. Wang Z, Singh R, Miller BE, Tal-Singer R, Van Horn S, Tomsho L, et al. Sputum microbiome temporal variability and dysbiosis in chronic obstructive pulmonary disease exacerbations: An analysis of the COPDMAP study. Thorax. 2018; doi:10.1136/thoraxjnl-2017-210741

200. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, et al. Analysis of the lung microbiome in the "healthy" smoker and in COPD. PLoS One. 2011; doi:10.1371/journal.pone.0016384

201. Kiley JP, Caler E V. The Lung Microbiome. A New Frontier in Pulmonary Medicine. Ann Am Thorac Soc. 2014;11: S66–S70. doi:10.1513/AnnalsATS.201308-285MG

202. Mammen MJ, Sethi S. COPD and the microbiome. Respirology. 2016;21: 590–599. doi:10.1111/resp.12732

203. Grønseth R, Haaland I, Wiker HG, Martinsen EMH, Leiten EO, Husebø G, et al. The Bergen COPD microbiome study (MicroCOPD): rationale, design, and initial experiences. Eur Clin Respir J. 2014;1: 26196. doi:10.3402/ecrj.v1.26196

204. Alley RB, Barry R, Lemke P, Ren J, Alley RB, Allison I, et al. Observations: Changes in Snow, Ice and Frozen Ground. Climate Change 2007: The Physical Science Basis. Intergovernmental Panel on Climate Change; 2007. pp. 337–383.

205. Tarnocai C, Canadell JG, Schuur EAG, Kuhry P, Mazhitova G, Zimov S. Soil organic carbon pools in the northern circumpolar permafrost region. Global Biogeochem Cycles. 2009;23: n/a-n/a. doi:10.1029/2008GB003327

206. Jansson JK, Taş N. The microbial ecology of permafrost [Internet]. Nature Reviews Microbiology. 2014. pp. 414–425. doi:10.1038/nrmicro3262

207. Humlum O, Instanes A, Sollid JL. Permafrost in Svalbard: A review of research history, climatic background and engineering challenges. Polar Res. 2003;22: 191–215. doi:10.3402/polar.v22i2.6455

208. Müller O, Bang-Andreasen T, White RA, Elberling B, Taş N, Kneafsey T, et al. Disentangling the complexity of permafrost soil by using high resolution profiling of microbial community composition, key functions and respiration rates. Environ Microbiol. 2018;20: 4328–4342. doi:10.1111/1462-2920.14348

209. Wilhelm RC, Niederberger TD, Greer C, Whyte LG. Microbial diversity of active layer and permafrost in an acidic wetland from the Canadian high arctic. Can J Microbiol. 2011;57: 303–315. doi:10.1139/w11-004

210. Gittel A, Bárta J, Kohoutová I, Schnecker J, Wild B, Čapek P, et al. Site- and horizon-specific patterns of microbial community structure and enzyme activities in permafrost-affected soils of Greenland. Front Microbiol. 2014;5: 541. doi:10.3389/fmicb.2014.00541

211. Koyama A, Wallenstein MD, Simpson RT, Moore JC. Soil bacterial community composition altered by increased nutrient availability in Arctic tundra soils. Front Microbiol. 2014;5. doi:10.3389/fmicb.2014.00516

212. Deng J, Gu Y, Zhang J, Xue K, Qin Y, Yuan M, et al. Shifts of tundra bacterial and archaeal communities along a permafrost thaw gradient in Alaska. Mol Ecol. 2015;24: 222–234. doi:10.1111/mec.13015

213. Mackelprang R, Saleska SR, Jacobsen CS, Jansson JK, Taş N. Permafrost Meta-Omics and Climate Change. Annu Rev Earth Planet Sci. 2016;44: 439–462. doi:10.1146/annurev-earth-060614-105126

214. Shi Y, Tyson GW, Delong EF. Metatranscriptomics reveals unique microbial small RNAs in the oceans water column. Nature. 2009;459: 266–269. doi:10.1038/nature08055

215. Carvalhais LC, Dennis PG, Tyson GW, Schenk PM. Application of metatranscriptomics to soil environments. Journal of Microbiological Methods. 2012. pp. 246–251. doi:10.1016/j.mimet.2012.08.011

216. Jorth P, Turner KH, Gumus P, Nizam N, Buduneli N, Whiteley M. Metatranscriptomics of the human oral microbiome during health and disease. MBio. 2014;5. doi:10.1128/mBio.01012-14

217. Blazewicz SJ, Barnard RL, Daly RA, Firestone MK. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. ISME J. 2013;7: 2061–2068. doi:10.1038/ismej.2013.102

218. Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC. Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome. Ward N, editor. PLoS One. 2008;3: e2527. doi:10.1371/journal.pone.0002527

219. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome Biol. 2016;17: 260. doi:10.1186/s13059-016-1116-8

220. Westreich ST, Korf I, Mills DA, Lemay DG. SAMSA: a comprehensive metatranscriptome analysis pipeline. BMC Bioinformatics. 2016;17: 399. doi:10.1186/s12859-016-1270-8

221. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. SAMSA2: A standalone metatranscriptome analysis pipeline. BMC Bioinformatics. BMC Bioinformatics; 2018;19: 1–12. doi:10.1186/s12859-018-2189-z

222. Martinez X, Pozuelo M, Pascal V, Campos D, Gut I, Gut M, et al. MetaTrans: An open-source pipeline for metatranscriptomics. Sci Rep. Nature Publishing Group; 2016;6: 1–12. doi:10.1038/srep26447

223. Zeng F, Wang Z, Wang Y, Zhou J, Chen T. Large-scale 16S gene assembly using metagenomics shotgun sequences. Bioinformatics. 2017;33: 1447–1456. doi:10.1093/bioinformatics/btx018

224. Pericard P, Dufresne Y, Couderc L, Blanquart S, Touzet H. MATAM: Reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes. Bioinformatics. 2018;34: 585–591. doi:10.1093/bioinformatics/btx644

225. Papi A, Bellettato CM, Braccioni F, Romagnoli M, Casolari P, Caramori G, et al. Infections and Airway Inflammation in Chronic Obstructive Pulmonary Disease Severe Exacerbations. Am J Respir Crit Care Med. 2006;173: 1114–1121. doi:10.1164/rccm.200506-859OC

226. Bafadhel M, McKenna S, Terry S, Mistry V, Reid C, Haldar P, et al. Acute Exacerbations of Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med. 2011;184: 662–671. doi:10.1164/rccm.201104-0597OC

227. Wang Z, Bafadhel M, Haldar K, Spivak A, Mayhew D, Miller BE, et al. Lung

microbiome dynamics in COPD exacerbations. Eur Respir J. 2016;47: 1082–1092. doi:10.1183/13993003.01406-2015

228. Cabrera-Rubio R, Garcia-Nunez M, Seto L, Anto JM, Moya A, Monso E, et al. Microbiome Diversity in the Bronchial Tracts of Patients with Chronic Obstructive Pulmonary Disease. J Clin Microbiol. 2012;50: 3562–3568. doi:10.1128/JCM.00767-12

229. Dickson RP, Erb-Downward JR, Martinez FJ, Huffnagle GB. The Microbiome and the Respiratory Tract. Annu Rev Physiol. 2016;78: 481–504. doi:10.1146/annurev-physiol-021115-105238

230. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018;6: 90. doi:10.1186/s40168-018-0470-z

231. Kioroglou D, Mas A, Portillo M del C. Evaluating the Effect of QIIME Balanced Default Parameters on Metataxonomic Analysis Workflows With a Mock Community. Front Microbiol. 2019;10. doi:10.3389/fmicb.2019.01084

232. Dickson RP, Huffnagle GB. The Lung Microbiome: New Principles for Respiratory Bacteriology in Health and Disease. Goldman WE, editor. PLOS Pathog. 2015;11: e1004923. doi:10.1371/journal.ppat.1004923

233. Huffnagle GB, Dickson RP, Lukacs NW. The respiratory tract microbiome and lung inflammation: a two-way street. Mucosal Immunol. 2017;10: 299–306. doi:10.1038/mi.2016.108

234. Sinha R, Weissenburger-Moser LA, Clarke JL, Smith LM, Heires AJ, Romberger DJ, et al. Short term dynamics of the sputum microbiome among COPD patients. Chotirmall SH, editor. PLoS One. 2018;13: e0191499. doi:10.1371/journal.pone.0191499

235. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4: e2584. doi:10.7717/peerj.2584

236. Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, et al. Advancing Our Understanding of the Human Microbiome Using QIIME. Methods in Enzymology. 2013. pp. 371–444. doi:10.1016/B978-0-12-407863-5.00019-8

237. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. Xu J, editor. mSystems. 2018;3. doi:10.1128/mSystems.00187-18

238. Yue JC, Clayton MK. A Similarity Measure Based on Species Proportions.

Commun Stat - Theory Methods. 2005;34: 2123–2131. doi:10.1080/STA-200066418

239. Einarsson GG, Comer DM, McIlreavey L, Parkhill J, Ennis M, Tunney MM, et al. Community dynamics and the lower airway microbiota in stable chronic obstructive pulmonary disease, smokers and healthy non-smokers. Thorax. 2016;71: 795–803. doi:10.1136/thoraxjnl-2015-207235

240. Drengenes C, Wiker HG, Kalananthan T, Nordeide E, Eagan TML, Nielsen R. Laboratory contamination in airway microbiome studies. BMC Microbiol. 2019;19: 187. doi:10.1186/s12866-019-1560-1

241. Daniel R. The metagenomics of soil. Nat Rev Microbiol. 2005;3: 470–478. doi:10.1038/nrmicro1160

242. Jansson JK, Taş N. The microbial ecology of permafrost. Nat Rev Microbiol. 2014;12: 414–425. doi:10.1038/nrmicro3262

243. Danczak RE, Johnston MD, Kenah C, Slattery M, Wilkins MJ. Capability for arsenic mobilization in groundwater is distributed across broad phylogenetic lineages. Pereira IAC, editor. PLoS One. 2019;14: e0221694. doi:10.1371/journal.pone.0221694

244. Imperato V, Kowalkowski L, Portillo-Estrada M, Gawronski SW, Vangronsveld J, Thijs S. Characterisation of the Carpinus betulus L. Phyllomicrobiome in Urban and Forest Areas. Front Microbiol. 2019;10. doi:10.3389/fmicb.2019.01110

245. Seitz KW, Dombrowski N, Eme L, Spang A, Lombard J, Sieber JR, et al. Asgard archaea capable of anaerobic hydrocarbon cycling. Nat Commun. 2019;10: 1822. doi:10.1038/s41467-019-09364-x

246. Mackelprang R, Burkert A, Haw M, Mahendrarajah T, Conaway CH, Douglas TA, et al. Microbial survival strategies in ancient permafrost: Insights from metagenomics. ISME J. Nature Publishing Group; 2017;11: 2305–2318. doi:10.1038/ismej.2017.93

247. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40: D109–D114. doi:10.1093/nar/gkr988

248. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013;23: 111–120. doi:10.1101/gr.142315.112

249. Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and Complete Genomes from Metagenomes. bioRxiv. Cold Spring Harbor Laboratory; 2019; 808410. doi:10.1101/808410

250. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc Natl Acad Sci. 2006;103: 12115–12120. doi:10.1073/pnas.0605127103

251. Bang-Andreasen T, Anwar MZ, Lanzén A, Kjøller R, Rønn R, Ekelund F, et al. Total RNA sequencing reveals multilevel microbial community changes and functional responses to wood ash application in agricultural and forest soil. FEMS Microbiol Ecol. 2019; doi:10.1093/femsec/fiaa016

Appendices

Appendices

Paper I

Protected sampling is preferable in bronchoscopic studies of the airway microbiome

Grønseth, R.*, Drengenes, C., Wiker, H. G., Tangedal, S., **Xue, Y.**, Husebø, G. R., Svanes, Ø., Lehmann, S., Aardal, M., Hoang, T., Kalananthan, T., Hjellestad Martinsen, E. M., Orvedal Leiten, E., Aanerud, M., Nordeide, E., Haaland, I., Jonassen, I., Bakke, P., & Eagan, T.

ERJ open research, 3(3), 00019-2017. (2017).

https://doi.org/10.1183/23120541.00019-2017.

ERJ open research

ORIGINAL ARTICLE
AIRWAY MICROBIOME

Protected sampling is preferable in bronchoscopic studies of the airway microbiome

Rune Grønseth[1], Christine Drengenes[1,2], Harald G. Wiker[2,3], Solveig Tangedal[1,2], Yaxin Xue[4], Gunnar Reksten Husebø[1,2], Øistein Svanes[1,2], Sverre Lehmann[1,2], Marit Aardal[1], Tuyen Hoang[2], Tharmini Kalananthan[1], Einar Marius Hjellestad Martinsen[2], Elise Orvedal Leiten[2], Marianne Aanerud[1], Eli Nordeide[1,2], Ingvild Haaland[1,2], Inge Jonassen[4], Per Bakke[2] and Tomas Eagan[1,2]

Affiliations: [1]Dept of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway. [2]Dept of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway. [3]Dept of Microbiology, Haukeland University Hospital, Bergen, Norway. [4]Computational Biology Unit, Dept of Informatics, University of Bergen, Bergen, Norway.

Correspondence: Rune Grønseth, Dept of Thoracic Medicine, Haukeland University Hospital, Jonas Lies vei, Bergen 5021, Norway. E-mail: nielsenrune@me.com

ABSTRACT The aim was to evaluate susceptibility of oropharyngeal contamination with various bronchoscopic sampling techniques.

67 patients with obstructive lung disease and 58 control subjects underwent bronchoscopy with small-volume lavage (SVL) through the working channel, protected bronchoalveolar lavage (PBAL) and bilateral protected specimen brush (PSB) sampling. Subjects also provided an oral wash (OW) sample, and negative control samples were gathered for each bronchoscopy procedure. DNA encoding bacterial 16S ribosomal RNA was sequenced and bioinformatically processed to cluster into operational taxonomic units (OTU), assign taxonomy and obtain measures of diversity.

The proportion of Proteobacteria increased, whereas Firmicutes diminished in the order OW, SVL, PBAL, PSB ($p<0.01$). The alpha-diversity decreased in the same order ($p<0.01$). Also, beta-diversity varied by sampling method ($p<0.01$), and visualisation of principal coordinates analyses indicated that differences in diversity were smaller between OW and SVL and OW and PBAL samples than for OW and the PSB samples. The order of sampling (left *versus* right first) did not influence alpha- or beta-diversity for PSB samples.

Studies of the airway microbiota need to address the potential for oropharyngeal contamination, and protected sampling might represent an acceptable measure to minimise this problem.

@ERSpublications
Protected bronchoscopic sampling is most suitable for identification of a distinct airway microbiome http://ow.ly/qIIy30eqB9M

Cite this article as: Grønseth R, Drengenes C, Wiker HG, *et al*. Protected sampling is preferable in bronchoscopic studies of the airway microbiome. *ERJ Open Res* 2017; 3: 00019-2017 [https://doi.org/10.1183/23120541.00019-2017].

This article has supplementary material available from openres.ersjournals.com

Received: Feb 16 2017 | Accepted after revision: June 21 2017

Support statement: The current work has been funded through unrestricted grants and fellowships from Helse Vest, GlaxoSmithKline, the Endowment of Timber Merchant A. Delphin and Wife through the Norwegian Medical Association and Bergen Medical Research Foundation. Funding information for this article has been deposited with the Crossref Funder Registry.

Conflict of interest: Disclosures can be found alongside this article at openres.ersjournals.com

Introduction

High-throughput sequencing has opened up a new window in microbial ecology, enabling the characterisation of microbial communities in biological compartments thought to be completely sterile only a few years ago. The implications for health and disease are widely unexplored, but are likely to be significant [1]. Recent studies have found compelling evidence for the lungs to have a distinct microbiome [2], providing a bacterial presence with which our immune system interacts [3, 4]. As almost all pulmonary diseases have a local inflammatory component, there is a possibility of a disrupted microbiome being integral to disease pathogenesis.

Thus, there is a current push to characterise the pulmonary microbiome, and its relation to different pulmonary diseases. However, sampling the pulmonary microbiome is difficult. Sputum is fraught with significant contamination from the oral cavity, and percutaneous sampling is unpractical with a high risk of complications like pneumothorax or bleeding. The emerging gold standard for sampling is bronchoscopy. But bronchoscopy also has its technical challenges, besides issues of discomfort, cost and sedation. The bronchoscope must pass through either the oral or nasal cavity in addition to the pharyngeal cavity, and might carry contaminants from the upper airways to the lower biomass compartment of the lower airways. Samples are collected through the same bronchoscope working channel through which fluid is suctioned up and out. The different modes of sampling (bronchoalveolar lavage (BAL) brushings, biopsies) might be carried through catheters, which may or may not have a wax-sealed tip to ensure sterility. Added to this is the conundrum caused by the constant influx of microbiota by microaspiration and inhalation that probably is responsible for maintenance and creation of a large fraction of the lung microbiome [5].

In 25 studies of the human lung microbiome sampling the airway microbiome by bronchoscopy of healthy subjects [2–4, 6–9] and patients with chronic obstructive pulmonary disease (COPD) [10–14], asthma [15, 16], interstitial lung disease [17, 18], cystic fibrosis (CF) [19], HIV [20–23] and lung-transplanted subjects [24–27]; only five used protected sterile brushes (PSB) to avoid contamination from the working channel [7, 8, 16, 19, 22]. Some authors reported that suction was not used prior to entering the trachea [2–4, 6–10, 20, 22], and three studies used separate bronchoscopes for anaesthesia and sampling of some or all participants [3, 4, 7]. No study performed bronchoalveolar lavage (BAL) through a protected catheter (protected BAL), and no study with more than 20 sampled subjects has compared protected with unprotected sampling methods.

In preparation for the analyses of a large, ongoing COPD microbiome study [28], we sought to reduce contamination as well as assess the performance of different sampling techniques. In the current paper we present analyses to examine the degree of oropharyngeal influence on the airway microbiome applying protected bronchoscopic sampling techniques. In addition we present an analysis on the effect of sampling the left or right lung first.

Material and methods

The design of the entire MicroCOPD study has been published previously [28]. The current analysis includes 58 control subjects, 64 subjects with COPD and three subjects with asthma. All participants were at least 35 years old and were recruited from previous longitudinal case–control studies in addition to a few volunteers [29]. Subjects had neither acute respiratory symptoms nor any reported use of antibiotics or oral corticorticosteroids within the last 14 days prior to bronchoscopy. Other inclusion/exclusion criteria are listed in the supplementary material.

The Regional Committee for Medical and Health Research Ethics approved the study (REK Nord, project number 2011/1307). All participants provided written informed consent.

All participants received at least 0.4 mg of salbutamol through a spacer before the bronchoscopy procedure. Flexible video-bronchoscopy was performed *via* the oral route in supine position. No suction was used prior to having entered the trachea. All subjects received local anaesthesia with lidocaine both before and during the procedure. All but 18 subjects received mild sedation (alfentanil) parenterally. Participants were monitored according to current guidelines, and were observed for at least 2 h after the procedure [30]. Six procedural samples, of which five were obtained during bronchoscopy, were analysed for each participant: oral wash (OW); three protected specimen brushes (PSBs) from the right lower lobe (right PSB) and three from the left upper lobe (left PSB); two 50-mL fractions of protected bronchoalveolar lavage of the right middle lobe (PBAL1 and PBAL2); and small-volume lavage (SVL) in the left upper lobe. In addition, we included negative control samples (NCSs) from the same bottle of phosphate-buffered saline that was used for the procedure of the corresponding individual. For 49 subjects, we examined the left lung before the right lung. BAL and SVL were always collected after obtaining PSB samples. Protected specimen brushes and protected bronchoalveolar lavage are illustrated in supplementary figures S1 and S2.

Bacterial DNA was extracted using a combination of enzymatic lysis with lysozyme, mutanolysin and lysostaphin, and mechanical lysis methods using the FastPrep-24 as described by the manufacturers of the FastDNA Spin Kit (MP Biomedicals, LLC, Solon, OH, USA).

Library preparation and sequencing of the V3-V4 region of the 16S rRNA gene was carried out according to the Illumina 16S Metagenomic Sequencing Library Preparation guide (Part no. 15044223 Rev. B). The V3-V4 region was PCR amplified (45 cycles) and prepared for a subsequent index PCR step using primers adapted from Kozich *et al.* [31] as follows. 16S amplicon PCR forward primer (overhang adaptor sequences are underlined): 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGG NGGCWGCAG. 16S amplicon PCR reverse primer (overhang adaptor sequences are underlined): 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC. The samples were pooled and prepared for 2×300 cycles of paired-end sequencing on the Illumina Miseq sequencing platform using reagents from the Miseq reagent kit v3 (Illumina Inc., San Diego, CA, USA).

The chosen bioinformatic pipeline was Quantitative Insights Into Microbial Ecology (QIIME, http://qiime.org) v1.9.1. After creating a library of joined reads, operational taxonomic units (OTUs) were picked at a 97% similarity threshold, small OTUs and OTUs seen in negative control samples were removed, taxonomy was assigned to the OTUs and a phylogenetic tree was constructed after alignment. We used the GreenGenes version 13.8 as reference database [32]. Further details on the bioinformatic procedures can be found in the supplementary material.

Differences in relative abundance of taxa were evaluated by applying a beta distribution and non-parametric trend tests. Alpha-diversity was evaluated using Faith's phylogenetic diversity (PD), or "PD wholetree". Beta-diversity was estimated with unweighted UniFrac and visualised by principal coordinates analyses (PCoA) [33]. Diversity analyses require a similar number of sequences in each sample, which was ensured by rarefaction. Statistical significance for alpha-diversity and beta-diversity between sampling methods was evaluated by Bonferroni-corrected Wilcoxon matched-pairs test in Stata version 13.2 (Statacorp, Texas, USA) and permutational ANOVA (permanova) tests in QIIME, respectively.

Results

Only three subjects had asthma: two men and one woman. The 64 COPD subjects were slightly older, included more men and had a larger tobacco-smoking burden than the 58 control subjects (table 1).

For each of the 125 participants, seven samples were sequenced (negative control sample, OW, right PSB, PBAL1, PBAL2, left PSB, SVL). A total of 12.5 million sequences were obtained from the six procedural samples after bioinformatics clean-up, as described in the methods section. For alpha- and beta-diversity, we rarefied our data at 1000 sequences.

Taxonomy

Figure 1 shows the taxonomic classification by sampling method at the phylum level. As the degree of protection from influence of oral environment increased, the proportion of Proteobacteria increased, whereas Firmicutes diminished ($p<0.01$). At the genus level all sample types where dominated by streptococci, but the mean proportion of the largest *Streptococcus* OTU showed the same declining pattern by sample type (OW 14.5%, SVL 13.6%, PBAL1 11.8%, PBAL2 11.3%, right PSB 8,6% and left PSB 5.4%; non-parametric trend test $p<0.001$).

TABLE 1 Characteristics of 125 subjects of the MicroCOPD study

	COPD	Asthma	Control
Subjects	64	3	58
Males	34 (53.1%)	2 (67.7%)	34 (58.6%)
Current smokers	15 (23.4%)	0	16 (27.6%)
Ex-smokers	48 (75.0%)	2 (67.7%)	35 (60.3%)
Never-smokers	1 (1.6%)	1 (33.3%)	7 (12.1%)
Smoking exposure pack-years	28.49±16.08	20.88±24.22	22.83±18.55
FEV1 % predicted	56.83±16.30	88.31±11.37	100.71±11.00
Age years	68.73±7.23	64.41±9.1	64.89±8.43
Use of inhaled corticosteroids	44 (68.8%)	1 (33.3%)	1 (1.7%)

Data are presented as mean±SD unless otherwise stated. COPD: chronic obstructive pulmonary disease; FEV1: forced expiratory volume in 1 s.

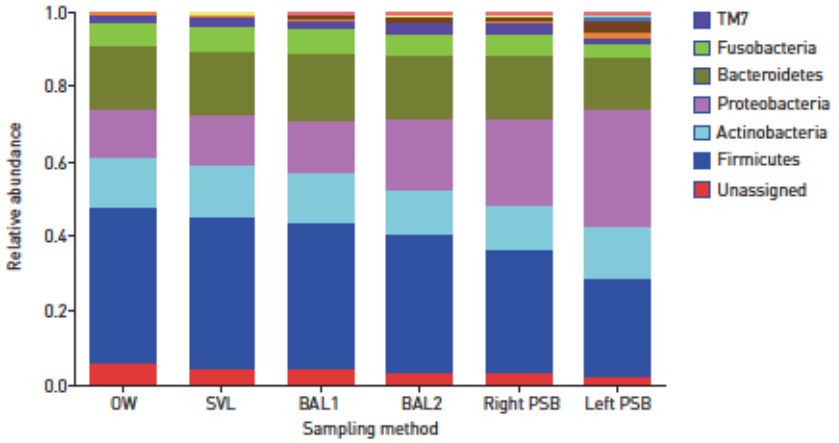


FIGURE 1 Mean taxonomic distribution at the phylum level, by sampling method, for all 125 individuals (unrarefied). OW: oral wash; SVL: small-volume lavage in the left upper lobe; BAL1: first fraction of protected bronchoalveolar lavage (BAL) from right middle lobe; BAL2: second fraction of protected BAL from right middle lobe; PSB: protected specimen brush from right lower lobe and left upper lobe. No legend for smallest phylae.

Alpha-diversity

Figure 2 shows a boxplot of the alpha-diversity metric, Faith's phylogenetic diversity, by sampling method and by disease category, excluding the three asthma subjects. The phylogenetic diversity within a sample is an indication of richness as the diversity increases both when a higher number of different OTUs are present, and when the phylogenetic distance is larger within the phylogenetic tree (less genetically similar). Bonferroni-corrected Wilcoxon matched-pairs signed-ranks tests showed that the oral wash samples were more alpha-diverse than all other sampling methods ($p<0.001$). The diversity was lower in COPD patients than controls, for most all sample types (figure 2). Importantly, the diversity decreased as the samples

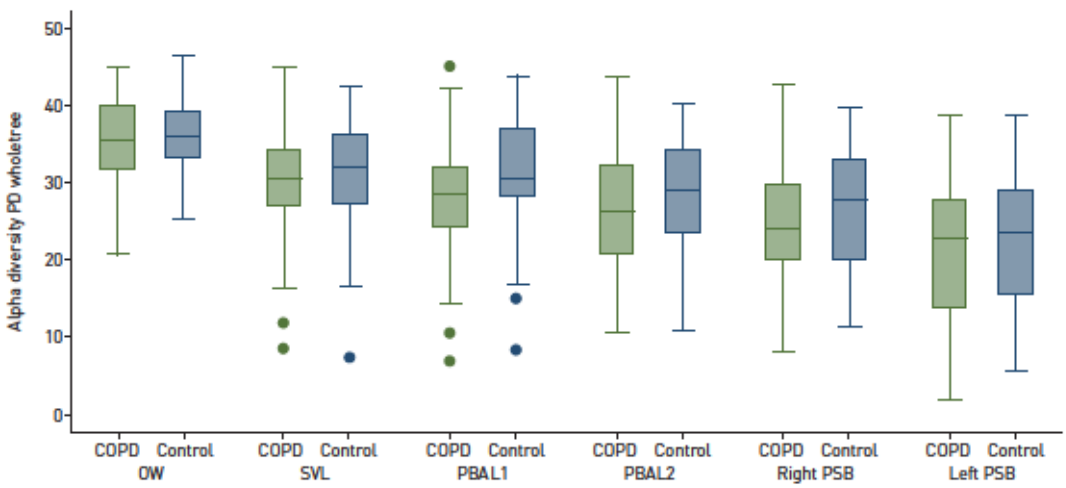


FIGURE 2 Box-plot of alpha-diversity measured by wholetree phylogenetic differences grouped according to sampling method and chronic obstructive pulmonary disease (COPD) status. Rarefied at 1000 sequences. OW: oral wash sample; SVL: small-volume lavage from left upper lobe; PBAL1: first fraction of protected bronchoalveolar lavage (BAL); PBAL2: second fraction of protected BAL; right PSB: protected specimen brush from right lower lobe; left PSB: protected specimen brush from left upper lobe.

were less exposed to potential oral and bronchoscope contamination (OW>SVL>PBAL1>PBAL2>rightPSB>leftPSB, non-parametric trend test $p<0.01$).

Beta-diversity

To compare between sample compositions (beta-diversity), we constructed principal coordinates analysis (PCoA) plots of unweighted UniFrac distances including all procedural samples. Figure 3 shows the PCoA plots for the oral wash *versus* each of the other sampling methods. Each dot represents a diversity measurement for one sample, and the OW sample is always shown in green. As can be seen, most respiratory tract samples clustered differently from the OW samples, but the visual impression is that the differences in diversity were smaller between OW and SVL and OW and PBAL samples than for OW and the PSB samples. Another way of comparing the beta-diversity was employed using a permanova test; estimating the beta-diversity between OW samples and each of the other sampling methods. This method tests to which degree the variation in a matrix of UniFrac distances can be explained by an imposed categorisation (*i.e.* sampling method). Overall permanova test confirmed that the beta-diversity differed by sampling method (pseudo F 8.73, $p=0.001$, 999 permutations). When the distance matrix was split according to the comparisons in figure 3, all were significant ($p<0.01$, permanova, corrected for multiple comparison), with the permanova pseudo F-statistic gradually increasing for the comparison of OW with SVL, PBAL1, PBAL2, right PSB and left PSB respectively, again indicating that PSB samples were more clearly separated from OW samples than SVL and PBAL.

Finally we investigated whether the order of sampling (left *versus* right lung first) influenced alpha- and beta-diversity in PSB samples. We found no significant difference in alpha- or beta-diversity for the right or the left PSBs as judged by phylogenetic diversity and unweighted UniFrac (supplementary figures S3 and S4).

Discussion

We have shown that protected BAL and protected brush samples differed more from oral wash samples than unprotected lavage through the bronchoscope working channel. Thus, unprotected sampling of the airway microbiome might convey an image of a microbiome that is more similar to the oral microbiome, than it would have been with protected sampling.

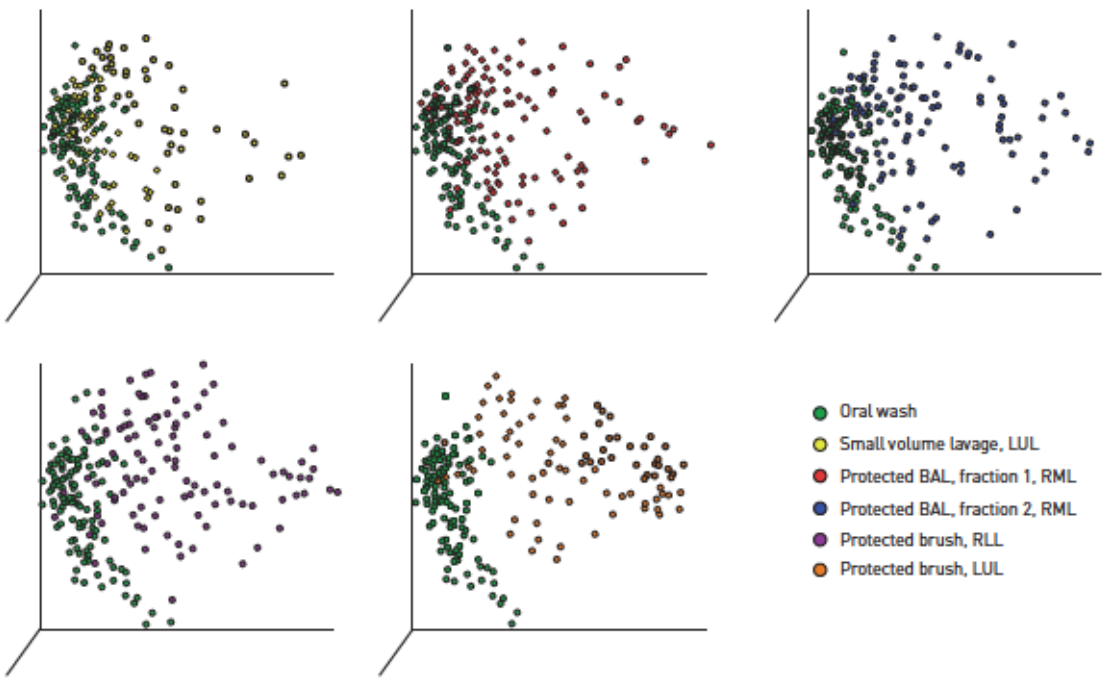


FIGURE 3 Principal coordinates analyses on unweighted UniFrac distance matrix comparing sampling methods in the MicroCOPD to oral wash samples. Rarefied at 1000 sequences. LUL: left upper lobe; BAL: bronchoalveolar lavage; RML: right middle lobe; RLL: right lower lobe.

To our best knowledge this is the first study that presents both protected brush *and* protected lavage sampling as compared with both the oral microbiome and unprotected sampling. With more than 120 examined subjects it is by today the largest single site bronchoscopy study of the lung microbiome.

As other authors we find evidence of a lung microbiome separated from the oral microbiome by a larger fraction of Proteobacteria and a proportionately lower fraction of Firmicutes [2, 8, 15, 20]. However, SEGAL and associates [3, 4] mainly found that the airway microbiome was characterised by enrichment from supraglottic areas of the respiratory tract, and in particular by *Prevotella* and *Veillonella* OTUs, which are Bacteriodetes and Firmicutes, respectively. They examined 49 subjects, with supraglottic brushes and BAL through the working channel, and observed that two clusters dominated airway samples: one dominated by OTUs present in negative control samples, and one dominated by OTUs present in supraglottic brushes. One interpretation might be that these two clusters represent two different modalities of contamination, the first one from laboratory procedures and the second from bronchoscopic carryover. SEGAL *et al.* argue that if it was bronchoscopic carry-over, they would have observed a dilutional effect when they compared a first BAL of the lingula, with the second BAL of the right middle lobe. However, this comparison was done for only 15 individuals, and anatomically one might expect lower biomass in the lingula than the right middle lobe.

Other authors have also investigated the possibility of bronchoscopic carryover. BASSIS *et al.* examined oral wash samples of 12 subjects and compared them with a first BAL of the lingula and a second BAL of the right middle lobe [6]. They did not find any difference in quantitative PCR between the first and second BAL, and no difference in beta-diversity when comparing the OW with the two BALs. Their interpretation was that if there was significant carryover, there should have been observed some sort of dilutional effect. Nevertheless, the two sampled sites are separated by the carina, and the bronchoscope must be repositioned between sampling, and these two sites are indeed in different communication with the outside world, possibly leading to an *a priori* larger biomass in the right lung. Also, DICKSON *et al.* compared supraglottic brushes with PSB and BAL through the working channel [8]. In principal component analyses of beta-diversity they found no clustering by sample type, except that the supraglottic samples differed from the intrapulmonary sample communities. However, by performing unprotected BAL before PSB, residual BAL fluid might have affected the brush areas making them more similar to the BAL sample sites. Finally, 15 sampled subjects might not be sufficient to detect the differences we observed in the current study with more than 100 participants.

It is quite plausible that microbes migrate from the oropharyngeal cavity to the airways, generating a normal overlap between the oropharyngeal and airway microbiomes [5]. But as we have shown, co-existing sample contamination likely also is an issue. The oropharyngeal microbiome has a known large biomass, with a high diversity. By passing through this cavity, contamination to the outside of the bronchoscope including its tip is inevitable. Use of suction will contaminate the working channel [7]. Since the oral biomass is much greater than the airway biomass, even a small contamination will have a disproportionate effect on the supposed airway microbiome if the unprotected measurements are performed through the working channel. Using the working channel for unprotected lavage repeatedly at different lobes will lead to contamination from one lobe to another. Using larger volume lavage may negate this effect to some degree, but not eliminate the problem.

Results from the current study suggest that protected sheet sampling is the superior sampling methodology. Comparing unprotected SVL and PSB both taken from the upper left lobe in our study, SVL was most similar to the oral sample by visual assessment of the 10 most abundant taxa, and likewise both by alpha- and beta-diversity. A direct comparison of protected and unprotected lavage from the same lobe is impossible, as any washing will impact the contents of later washings. However, the diversity of PBAL from the right middle lobe was intermediate between that found in OW and that found in the PSB.

Besides the above-mentioned study by DICKSON and colleagues [8], only two other studies have compared PSBs to other sampling methods [7, 19]. CHARLSON *et al.* [7] sampled laboratory reagents, the bronchoscope itself during various parts of the procedure, and the oropharyngeal microbiome in addition to BAL through the working channel and PSBs. They concluded that the microbiome from the lower respiratory tract was indiscriminate from the oropharyngeal microbiome irrespective of sampling method. However, the study included only one PSB per sampling, had lower sequencing depth than the current study, included only six healthy individuals and there were no adjustments made for OTUs seen in the negative control samples [7]. HOGAN *et al.* compared PSB, and SVL samples of nine CF patients [19]. For eight CF patients who had PSB and SVL taken from the same lobe, diversity was consistently higher in the PSB samples [19], the opposite of our findings. HOGAN *et al.* employed the PSB only at visible mucus plugs, and the airways of adult CF patients are perhaps no longer representing a low biomass environment. In addition the number of study subjects was limited.

The main strength of our study was comprehensive sampling of a large, heterogeneous sample of subjects with and without COPD, while taking precautions to avoid excessive influence from laboratory and bronchoscopic contamination. However, some potential weaknesses should be acknowledged. First, we have not performed quantitative PCR, and thus cannot conclude regarding the amount of 16S rRNA gene copies in the samples before amplification. Second, our analyses do not include a mock community, and we are therefore not able to provide sequencing error rates for the current study. We could also have spiked our samples with bacteria that would have indicated the efficiency of our DNA extraction. Third, pre-bronchoscopy all participants received 0.4 mg salbutamol. This was done for obtaining pre-bronchoscopy post-bronchodilator lung function values, but had the added benefit of protecting against procedural bronchospasm. Salbutamol was given as an aerosol through large volume spacers that are cleaned daily, and we are not aware of reports on contamination through metered dose inhalers. Furthermore, since both patients and controls received salbutamol, our conclusions should not be affected. Fourth, some results are difficult to compare with those of other authors because of differences in DNA extraction, PCR amplification, sequencing and bioinformatic approach. This is the result of a field where standards for 16S rRNA gene amplicon studies of microbial communities currently do not exist. To facilitate reproducibility we have used well-documented analytic approaches and mostly default settings for our bioinformatic pipeline (QIIME), in addition to using primers and PCR recommendations from a major next-generation sequencing provider (Illumina). Regardless of this, we cannot rule out that some of our findings only pertain to the current set of methodological choices such as the choice of sequencing hypervariable region V3V4 [34]. To minimise the influence of small/spurious OTUs we have excluded singletons by using default settings in our OTU picking, and removed OTUs that constituted less than 0.005% of the total number of sequences.

Insights concerning the airway microbiome in disease and health might provide vital understanding of disease mechanisms and provide new targets for treating lung diseases such as COPD, asthma, cystic fibrosis and interstitial lung diseases. However, to date only a minority of studies have performed protected sampling, and might have been affected by exposure to exposure to microbiota encountered before reaching the sampled sites. We have shown that unprotected sampling is likely to be affected by this phenomenon, and we encourage the use of protected specimen brushes when sampling the airway microbiota.

Acknowledgements

The MicroCOPD study is a large undertaking with many contributing partners. The authors wish to thank Lise Monsen (Dept of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway), Hildegunn Fleten (Dept of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway), Randi Sandvik (Dept of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway), Tove Folkestad (Dept of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway), Ane Aamli (Dept of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway) and Kristina Apalseth (Dept of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway) for help with data collection and aspects of laboratory handling.

R. Grønseth was the guarantor of the study and all authors had full access to all of the data in the study and take full responsibility for the integrity of the data and the accuracy of the data analysis. R. Grønseth, H.G. Wiker, M. Aanerud, P. Bakke and T. Eagan designed the study. R. Grønseth, H.G. Wiker, G.R. Husebø, Ø. Svanes, S. Lehmann, M. Aardal, T. Kalananthan, E.M. Hjellestad Martinsen, E. Orvedal Leiten, E. Nordeide, I. Haaland, I. Jonassen, P. Bakke and T. Eagan took part in the data collection. T. Hoang, C. Drengenes, H.G. Wiker and T. Kalananthan performed DNA extraction and high-throughput sequencing analyses. R. Grønseth, Y. Xue, S. Tangedal, T. Eagan and I. Jonassen. performed statistical and bioinformatic analyses. Data were interpreted by R. Grønseth, C. Drengenes, H.G. Wiker, S. Tangedal, Y. Xue, I. Haaland, I. Jonassen and T. Eagan. R. Grønseth, C. Drengenes, S. Tangedal and T. Eagan drafted the paper. All authors revised the draft and approved the version to be published.

References

1 Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet* 2012; 13: 260–270.
2 Morris A, Beck JM, Schloss PD, *et al.* Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *Am J Respir Crit Care Med* 2013; 187: 1067–1075.
3 Segal LN, Alekseyenko AV, Clemente JC, *et al.* Enrichment of lung microbiome with supraglottic taxa is associated with increased pulmonary inflammation. *Microbiome* 2013; 1: 19.
4 Segal LN, Clemente JC, Tsay JC, *et al.* Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype. *Nat Microbiol* 2016; 1: 16031.
5 Dickson RP, Erb-Downward JR, Martinez FJ, *et al.* The microbiome and the respiratory tract. *Annu Rev Physiol* 2016; 78: 481–504.
6 Bassis CM, Erb-Downward JR, Dickson RP, *et al.* Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals. *MBio* 2015; 6: e00037–15.
7 Charlson ES, Bittinger K, Haas AR, *et al.* Topographical continuity of bacterial populations in the healthy human respiratory tract. *Am J Respir Crit Care Med* 2011; 184: 957–963.
8 Dickson RP, Erb-Downward JR, Freeman CM, *et al.* Spatial variation in the healthy human lung microbiome and the adapted island model of lung biogeography. *Ann Am Thorac Soc* 2015; 12: 821–830.

9 Venkataraman A, Bassis CM, Beck JM, *et al.* Application of a neutral community model to assess structuring of the human lung microbiome. *MBio* 2015; 6: e02284-14.
10 Cabrera-Rubio R, Garcia-Nunez M, Seto L, *et al.* Microbiome diversity in the bronchial tracts of patients with chronic obstructive pulmonary disease. *J Clin Microbiol* 2012; 50: 3562–3568.
11 Einarsson GG, Comer DM, McIlreavey L, *et al.* Community dynamics and the lower airway microbiota in stable chronic obstructive pulmonary disease, smokers and healthy non-smokers. *Thorax* 2016; 71: 795–803.
12 Erb-Downward JR, Thompson DL, Han MK, *et al.* Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PLoS One* 2011; 6: e16384.
13 Pragman AA, Kim HB, Reilly CS, *et al.* The lung microbiome in moderate and severe chronic obstructive pulmonary disease. *PLoS One* 2012; 7: e47305.
14 Zakharkina T, Heinzel E, Koczulla RA, *et al.* Analysis of the airway microbiota of healthy individuals and patients with chronic obstructive pulmonary disease by T-RFLP and clone sequencing. *PLoS One* 2013; 8: e68302.
15 Hilty M, Burke C, Pedro H, *et al.* Disordered microbial communities in asthmatic airways. *PLoS One* 2010; 5: e8578.
16 Huang YJ, Nelson CE, Brodie EL, *et al.* Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. *J Allergy Clin Immunol* 2011; 127: 372–381.
17 Garzoni C, Brugger SD, Qi W, *et al.* Microbial communities in the respiratory tract of patients with interstitial lung disease. *Thorax* 2013; 68: 1150–1156.
18 Molyneaux PL, Cox MJ, Willis-Owen SA, *et al.* The role of bacteria in the pathogenesis and progression of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2014; 190: 906–913.
19 Hogan DA, Willger SD, Dolben EL, *et al.* Analysis of lung microbiota in bronchoalveolar lavage, protected brush and sputum samples from subjects with mild-to-moderate cystic fibrosis lung disease. *PLoS One* 2016; 11: e0149998.
20 Beck JM, Schloss PD, Venkataraman A, *et al.* Multicenter comparison of lung and oral microbiomes of HIV-infected and HIV-uninfected individuals. *Am J Respir Crit Care Med* 2015; 192: 1335–1344.
21 Cribbs SK, Uppal K, Li S, *et al.* Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection. *Microbiome* 2016; 4: 3.
22 Lozupone C, Cota-Gomez A, Palmer BE, *et al.* Widespread colonization of the lung by *Tropheryma whipplei* in HIV infection. *Am J Respir Crit Care Med* 2013; 187: 1110–1117.
23 Twigg HL, Knox KS, Zhou J, *et al.* Effect of advanced HIV Infection on the respiratory microbiome. *Am J Respir Crit Care Med* 2016; 194: 226–235.
24 Borewicz K, Pragman AA, Kim HB, *et al.* Longitudinal analysis of the lung microbiome in lung transplantation. *FEMS Microbiol Lett* 2013; 339: 57–65.
25 Dickson RP, Erb-Downward JR, Freeman CM, *et al.* Changes in the lung microbiome following lung transplantation include the emergence of two distinct *Pseudomonas* species with distinct clinical associations. *PLoS One* 2014; 9: e97214.
26 Dickson RP, Erb-Downward JR, Prescott HC, *et al.* Cell-associated bacteria in the human lung microbiome. *Microbiome* 2014; 2: 28.
27 Willner DL, Hugenholtz P, Yerkovich ST, *et al.* Reestablishment of recipient-associated microbiota in the lung allograft is linked to reduced risk of bronchiolitis obliterans syndrome. *Am J Respir Crit Care Med* 2013; 187: 640–647.
28 Grønseth R, Haaland I, Wiker HG, *et al.* The Bergen COPD microbiome study (MicroCOPD): rationale, design, and initial experiences. *Eur Clin Respir J* 2014; 1: 26196.
29 Eagan TM, Ueland T, Wagner PD, *et al.* Systemic inflammatory markers in COPD: results from the Bergen COPD Cohort Study. *Eur Respir J* 2010; 35: 540–548.
30 Du Rand IA, Blaikley J, Booton R, *et al.* British Thoracic Society guideline for diagnostic flexible bronchoscopy in adults: accredited by NICE. *Thorax* 2013; 68: Suppl. 1, i1–i44.
31 Klindworth A, Pruesse E, Schweer T, *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 2013; 41: e1.
32 McDonald D, Price MN, Goodrich J, *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 2012; 6: 610–618.
33 Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005; 71: 8228–8235.
34 Hiergeist A, Reischl U, Priority PIMCQAP, *et al.* Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. *Int J Med Microbiol* 2016; 306: 334–342.

Online supplement for

Protected sampling is preferable in bronchoscopic studies of the airways microbiome

MATERIAL AND METHODS

Study design

The design of the MicroCOPD study has been published previously (1). The current analysis includes 58 control subjects, 64 subjects with COPD, and 3 subjects with asthma, all examined between 11th of April, 2013 and 14th of April, 2015 at the outpatient clinic of the Department of Thoracic Medicine at Haukeland University Hospital. All participants were at least 35 years old, and were recruited from previous longitudinal case-control studies (2), and a few volunteers. Participation was postponed for subjects who had an ongoing respiratory symptom exacerbation or had used antibiotics or oral corticosteroids within the last 14 days. Subjects using anti-coagulants or double platelet inhibition, subjects with unstable coronary heart disease, hypoxemia (SpO_2 < 90% when receiving 3 liters of oxygen/minute through a nasal canula), hypercapnia at rest (pCO_2 > 6.65 kPa), or with known allergies against lidocaine or alfentanil were not included.

Control subjects had a post-bronchodilator (BD) FEV_1/FVC ratio ≥0.70 and no clinical diagnosis of obstructive lung disease as evaluated by the study physician. COPD cases had a post-BD FEV_1/FVC < 0.70, and a clinical diagnosis of COPD. Moderate, severe and very severe COPD was defined as post-BD FEV_1 between 50 and 80% of predicted, < 50% of predicted and < 30% of predicted by Norwegian pre-BD reference values,

respectively (3). The Regional Committee for Medical and Health Research Ethics approved the study (REK Nord, project number 2011/1307). All participants provided written informed consent.

Sample collection

All participants received at least 0.4 mg of salbutamol through a spacer before the bronchoscopy procedure. Sterile phosphate-buffered saline (PBS) in bottles of 500 mL were unsealed maximum 24 hours before the procedure. Immediately before bronchoscopy all participants delivered an oral wash (OW) sample by gargling 10 mL PBS. 1 mL of PBS from the same bottle was used as a negative control sample, and all PBS fluid used for samples for one subject came from the same bottle.

Flexible video-bronchoscopy was performed via the oral route in supine position. No suction was used prior to having entered the trachea. All subjects received local anesthesia with lidocaine both before and during the procedure. All but 18 subjects received mild sedation alfentanil parenterally. Participants were monitored according to current guidelines, and were observed for at least 2 hours after the procedure (4).

The following samples were taken in the same consecutive order during bronchoscopy:

1. Three wax-plug protected specimen brushes (PSB) from the right lower lobe (Conmed, Utica, NY, USA). The three brushes were cut off with sterile scissors, and placed together in an Eppendorf tube with 1 ml PBS.
2. Protected bronchoalveolar lavage (PBAL) of the right middle lobe by instilling two fractions each of 50 mL PBS (PBAL1 and PBAL2) using a wax-plug protected catheter (Plastimed Combicath, Le Plessis Bouchard, France).

3. Three wax-plug protected specimen brushes (PSB) from the left upper lobe, treated as the right lobe PSBs.

4. Small-volume lavage (SVL) of 20 mL PBS in the left upper lobe, from the same segment as the left PSB was taken. This lavage was sampled using the suction from the broncoscope´s working channel, thus mimicking the way BAL is most often sampled.

For 49 participants we examined the left side before the right (i.e according to the numbering above we performed sample 3., 4., 1., 2.).

DNA Extraction

1800 µl of OW, PBAL and SVL samples and 450 µl PSB and PBS NC samples were used for DNA extraction. An equal volume of Sputasol (Oxoid) was added to the samples followed by a 15 minute incubation in a thermomixer (1000 rpm) at 37 °C. The bacterial cells were then pelleted by centrifugation at 15700 g for 8 minutes. The supernatant was discarded and the bacterial cells were resuspended in 250 µl PBS.

Bacterial DNA was then extracted from the cells using enzymatic and mechanical lysis methods. As mechanical lysis methods tend to result in the shearing of free DNA, the samples were first treated with an enzyme cocktail consisting of 25 µl lysozyme (10 mg/mL, Sigma-Aldrich), 3 µl mutanolysin (25 KU/mL, Sigma-Aldrich), 1.5 µl lysostaphin (4000 U/mL, Sigma-Aldrich) and 20.5 µl TE5 buffer (10 mM Tris-HCl, 5mM EDTA, pH 8) and incubated at 37°C for 1 hour in a thermomixer (350 rpm). Before proceeding with mechanical cell lysis, the samples were centrifuged at 15700 g for 15 minutes to pellet any bacterial cells not sufficiently lysed by the enzymes. The supernatant containing the extracted DNA was transferred to a new eppendorf tube and stored at 4 °C while further processing of the bacterial cell pellet. The pellet was resuspended in 800 µl CLS-TC lysis

buffer from the FastDNA Spin Kit (MP Biomedicals, LLC, Solon, OH, USA) and transferred to a Lysing Matrix A tube (FastDNA Spin Kit). The sample was then subjected to mechanical lysis using the FastPrep-24 instrument (MP Biomedicals) at a speed setting of 6.0 m/s for 40 seconds. The lysate was then pooled with the supernatant from the enzyme lysis step and DNA further purified as described by the manufacturers for the FastDNA Spin Kit.

16S rRNA library preparation and high-throughput sequencing

Library preparation and sequencing of the V3-V4 region of the 16S rRNA gene was carried out according to the Illumina 16S Metagenomic Sequencing Library Preparation guide (Part # 15044223 Rev. B). The V3-V4 region was PCR amplified and prepared for a subsequent index PCR step using primers adapted from Klindworth et al. (5):

- 16S Amplicon PCR Forward Primer (overhang adaptor sequences underlined) = 5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG
- 16S Amplicon PCR Reverse Primer (overhang adaptor sequences underlined) = 5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC

The PCR reaction was carried out using the following cycling conditions: an initial cycle at 95°C for 3 minutes followed by 45 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds and a final extension cycle at 72 °C for 5 minutes. The incorporation of dual indexes to the library in the subsequent 8 cycles of Index PCR step was carried out using primers from the Nextera XT Index Kit (Illumina Inc., San Diego. CA, USA) and enabled the sequencing of 96 samples in each setup. The samples were DNA quantified using the Qubit dsDNA HS Assay Kit (Life Technologies) and normalized to 4nM. The samples were pooled and prepared for 2 x 300 cycles of paired-end sequencing on the

Illumina Miseq sequencing platform using reagents from the Miseq reagent kit v3 (Illumina).

Statistics and bioinformatics

The chosen bioinformatic pipeline was Quantitative Insights Into Microbial Ecology (QIIME - http://qiime.org) version 1.9.1, run on Macintosh OSX using the MacQIIME package. Two files per sample from the Illumina MiSeq, one forward read, and one with reverse read was first joined with at least an overlap of 100 base pairs. The resulting files were merged to one library and sequences of poor quality were discarded, demaning a base quality score (phred score) of 19 or higher. Operational taxonomic units (OTUs) were picked using the open reference based approach in QIIME using Uclust with a 97% sequence similarity threshold (6) and GreenGenes version 13.8 as the reference database (7). All OTUs that constituted less than 0.0005% of the total sequence number were removed (8). The GreenGenes database (v.13.8) was also used for taxonomic assignment of OTUs (7) with the Ribosomal Database Project (RDP) classifier (9). All sequences from OTUs seen in corresponding negative control samples were deleted for the downstream analyses (10). Phylogenetic tree construction was performed with FastTree (11), after alignment using PyNAST (12).

In order to assess similiarity between samples obtained by different bronchoscopic sampling techniques and the oropharyngeal microbiome, the taxonomic distribution and diversity of OTUs from the OW samples were compared to all other sample types. Alpha-diversity was evaluated using Faith's phylogenetic diversity (PD), or "PD wholetree". Beta-diversity was estimated with unweighted UniFrac, as well as visualization of

taxonomic distribution of OTUs and beta-diversity with principal coordinates analyses (PCoA) of UniFrac distance matrices for the entire data-set (13). Diversity analyses require a similar number of sequences in each sample, which was ensured by setting rarefaction levels. Samples with fewer sequences than the rarefaction level were excluded, whereas a number of sequences equal to the rarefaction level was chosen at random from the remaining samples. Due to the previous removal of a large number of sequences (the negative control sample OTUs), the rarefaction levels were relatively low. The proportion of taxa by sample type was tested using the *betafit* command in Stata as well as non-parametric trend tests. Statistical significance for alpha-diversity and beta-diversity between sampling methods was evaluated by Bonferroni-corrected Wilcoxon matched-pairs test in Stata version 13.2 (Statacorp, Texas, USA) and Bonferroni-corrected permutational ANOVA (permanova) with 1000 permutations in QIIME, respectively.

REFERENCES

1. Grønseth R, Haaland I, Wiker HG, Martinsen EMH, Leiten EO, Husebø G, Svanes Ø, Bakke PS, Eagan TM. The Bergen COPD microbiome study (MicroCOPD): rationale, design, and initial experiences. European Clinical Respiratory Journal. 2014;1:26196.
2. Eagan TM, Ueland T, Wagner PD, Hardie JA, Mollnes TE, Damas JK, Aukrust P, Bakke PS. Systemic inflammatory markers in COPD: results from the Bergen COPD Cohort Study. Eur Respir J. 2010;35:540-548.
3. Gulsvik A, Tosteson T, Bakke P, Humerfelt S, Weiss ST, Speizer FE. Expiratory and inspiratory forced vital capacity and one-second forced volume in asymptomatic never-smokers in Norway. Clin Physiol. 2001;21:648-660.
4. Du Rand IA, Blaikley J, Booton R, Chaudhuri N, Gupta V, Khalid S, Mandal S, Martin J, Mills J, Navani N, Rahman NM, Wrightson JM, Munavvar M. British Thoracic Society guideline for diagnostic flexible bronchoscopy in adults: accredited by NICE. Thorax. 2013;68 Suppl 1:i1-i44.
5. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013;41:e1.
6. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460-2461.
7. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6:610-618.
8. Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, Ursell LK, Lauber C, Zhou H, Song SJ, Huntley J, Ackermann GL, Berg-Lyons

D, Holmes S, Caporaso JG, Knight R. Advancing our understanding of the human microbiome using QIIME. Methods Enzymol. 2013;531:371-444.

9. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73:5261-5267.
10. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.
11. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26:1641-1650.
12. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 2010;26:266-267.
13. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71:8228-8235.

FIGURE LEGENDS

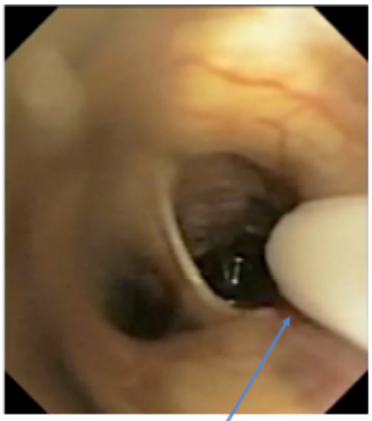
Figure S1: Illustration of protected bronchoalveolar lavage (PBAL) with phosphate buffered saline (PBS).

Figure S2: Illustration of protected specimen brush (PSB) sampling.

Figure S3: Box-plot of alpha-diversity measured by wholetree phylogenetic differences by which lung that was sampled first and right/left protected specimen brush (PSB). Rarefied at 1000 sequences.

Figure S4: Principal coordinates analyses of unweighted UniFrac distances by which lung that was sampled first (red dots – left side first, blue dots – right side first) in right protected specimen brushes (PSB) from the right lower lobe and left PSBs from the left upper lobe. Rarefied at 1000 sequences.

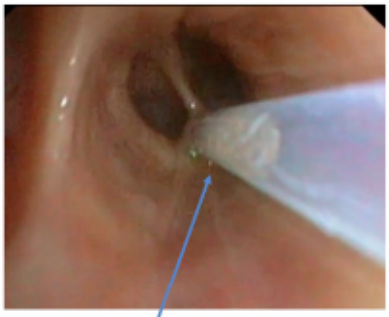
Protected sterile lavage



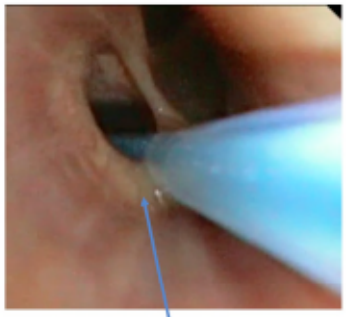
Sterile inner catheter after wax tip has been released

Lavage collected with manual suction in the same sterile 50 mL syringes through which PBS buffered saline was inserted

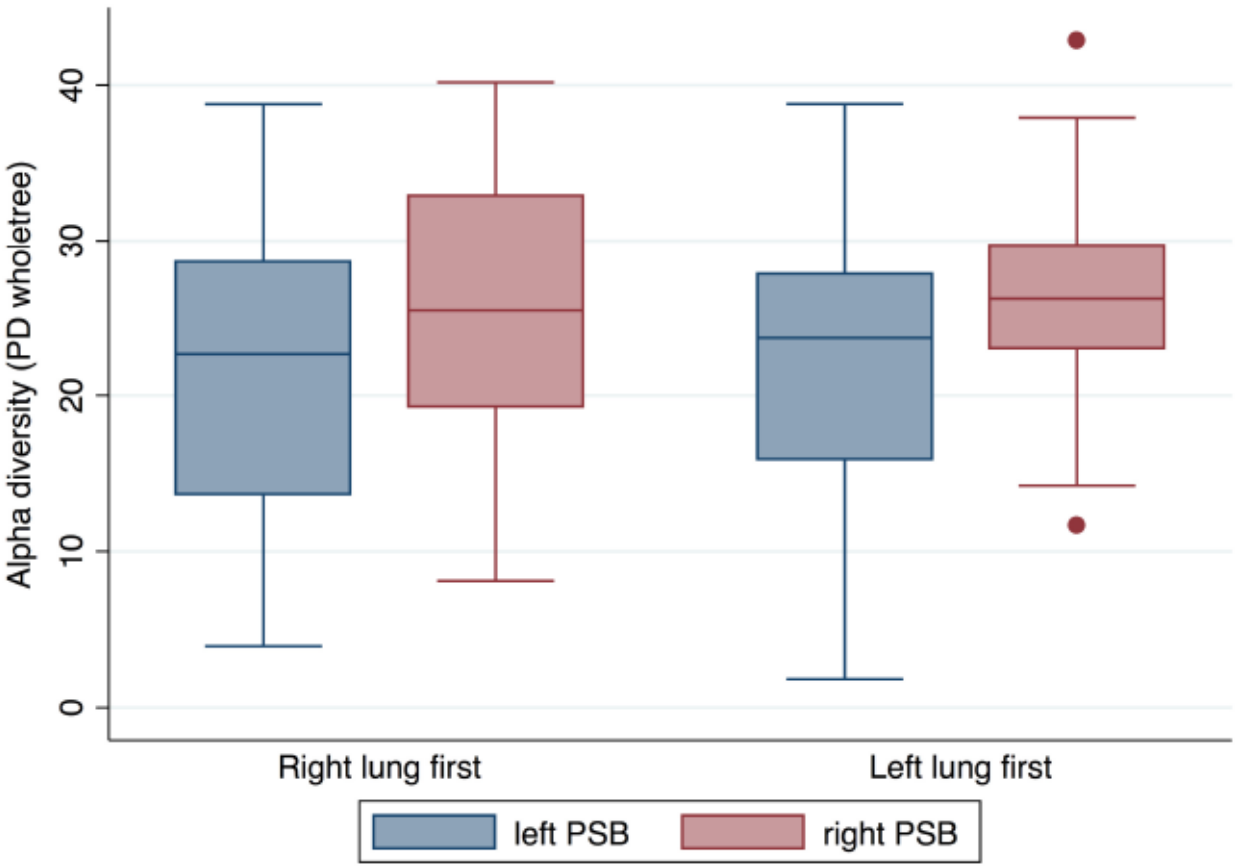
Protected sterile brush

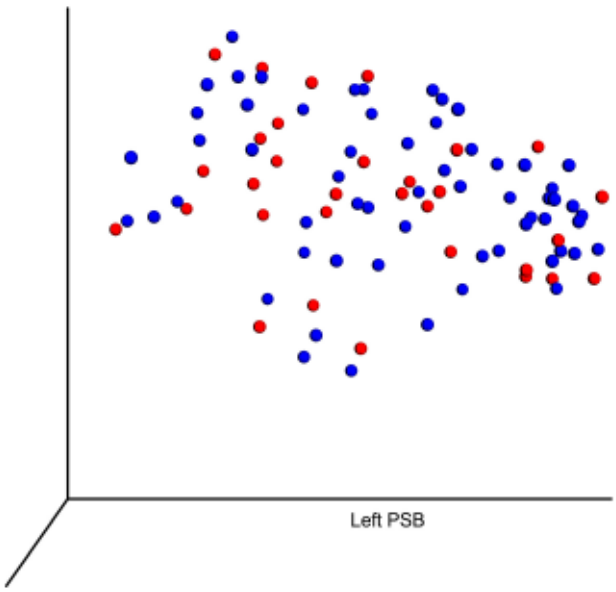


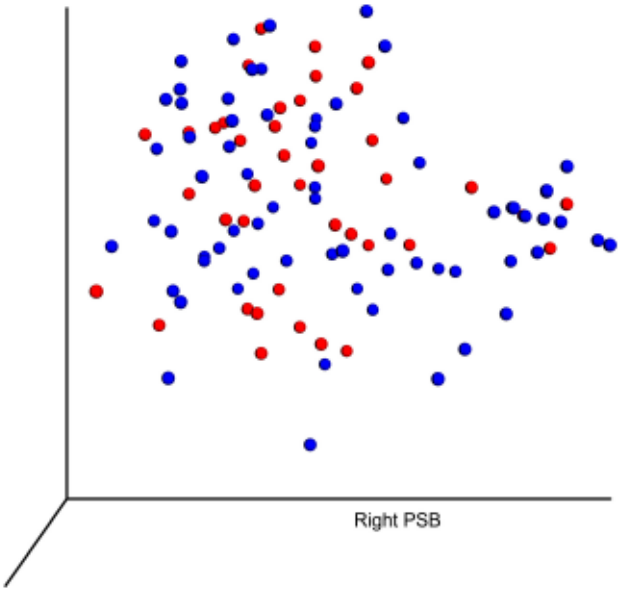
Wax tip before release



Released and inserted sterile brush


Alpha diversity (PD wholetree)
0
10
20
30
40
Right lung first
Left lung first
left PSB
right PSB


Left PSB


Right PSB

Paper II

Repeated bronchoscopy in health and obstructive lung disease: Is the airway microbiome stable?

Grønseth, R.*, **Xue, Y.***, Jonassen, I., Haaland, I., Kommedal, Kommedal O., Wiker, H. G., Drengenes, C., Bakke, P., & Eagan, T.

(submitted)

Paper III

Bacterial and Archaeal Metagenome-Assembled Genome Sequences from Svalbard Permafrost.

Xue, Y.*, Jonassen, I., Øvreås, L., & Taş, N.

***Microbiology resource announcements**, 8(27), e00516-19*. (2020)

https://doi.org/10.1128/MRA.00516-19.

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

OMICS DATA SETS

Bacterial and Archaeal Metagenome-Assembled Genome Sequences from Svalbard Permafrost

Yaxin Xue,[a] Inge Jonassen,[a] Lise Øvreås,[b,c] Neslihan Taş[d,e]

[a]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway
[b]Department of Biological Sciences, University of Bergen, Bergen, Norway
[c]University Center in Svalbard, UNIS, Longyearbyen, Norway
[d]Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[e]Biosciences Area, Lawrence Berkeley National Laboratory, Berkeley, California, USA

ABSTRACT Permafrost contains one of the least known soil microbiomes, where microbial populations reside in an ice-locked environment. Here, 56 prokaryotic metagenome-assembled genome (MAG) sequences from 13 phyla are reported. These MAGs will provide information on metabolic pathways that could mediate biogeochemical cycles in Svalbard permafrost.

Permafrost covers over 25% of the exposed land surface of the Northern Hemisphere and hosts a diversity of microbes proposed to be unique to cold habitats (1). These frozen soils contain a large reservoir of soil organic matter (SOM) that can have a significant impact on global climate upon thawing (2). The permafrost thaw may stimulate microbial activity and thus enable SOM decomposition. Previous studies have shown differences in microbial diversity between active layer (seasonally thawed and refrozen topsoil) and permafrost microbial communities (1–5). Although permafrost microbiomes are known to be highly diverse (1), they are largely underrepresented in global surveys. In this study, we investigated the microbial communities through a depth profile from Svalbard, and we report the binned metagenomic coassembly of five metagenome samples (6) and 56 metagenome-assembled genome (MAG) sequences.

Soil samples were obtained from an ice-wedge polygon site in the Adventdalen Valley in Svalbard, Norway (78.186N, 15.9248E). The site soil geochemistry was described previously (6). Five depth segments, namely, one active layer mineral horizon and four permafrost layers, were collected at the following depths: 0 to 14, 101 to 118, 118 to 126, 126 to 144, and 161 to 181 cm below the soil surface. Total community genomic DNA was extracted using a PowerSoil DNA isolation kit, and sequencing libraries were prepared using a TruSeq DNA library kit. An Illumina HiSeq 2500 instrument was used to acquire paired-end 150-bp metagenomic sequences, generating 20 Gb of raw reads per sample (7). The microbial community diversity and composition were reported elsewhere (6).

After adapter and low-quality reads were trimmed using MOCAT2 v2.0.0 (7), all cleaned reads were merged and then coassembled with MEGAHIT v1.1.3 (8), resulting in 566,254 contigs of ≥1 kb. We binned the contigs with MaxBin2 v2.2.5 (9) and MetaBAT2 v2.12.1 (10) and then dereplicated and aggregated them into MAGs using DAS Tool v1.1.0 (11), which resulted in 64 MAGs. We used CheckM v1.0.11 (12) to determine the completeness and contamination of these MAGs. We further examined the taxonomic distribution of contigs within each MAG based on Kaiju v1.6.2 (13) annotations and removed contaminating contigs. This process resulted in a total of 56 MAGs with contamination less than 10%. Default parameters were used with all software. We recovered 8 high-, 44 medium-, and 4 low-quality draft MAGs in accordance with minimum information about metagenome-assembled genome (MIMAG) standards (14). The MAGs

Citation Xue Y, Jonassen I, Øvreås L, Taş N. 2019. Bacterial and archaeal metagenome-assembled genome sequences from Svalbard permafrost. Microbiol Resour Announc 8:e00516-19. https://doi.org/10.1128/MRA.00516-19.

Editor Kenneth M. Stedman, Portland State University

Address correspondence to Yaxin Xue, xue.ethan@gmail.com, or Lise Øvreås, Lise.Ovreas@uib.no.

Received 6 May 2019
Accepted 6 June 2019
Published 3 July 2019

TABLE 1 Detailed completeness and contamination results, genome size, GC content, MIMAG status, taxonomy, and ENA accession information of MAGs

MAG alias	Completeness (%)	Contamination (%)	Genome size (bp)	GC content (%)	MIMAG classification	Taxonomy[a]	ENA accession no.
Maxbin2.039_sub	98.2	9.2	3,147,504	55.5	Medium	*Acidobacteria* sp.	ERZ870056
Metabat.113	96.8	0.9	2,959,789	67.9	High	*Actinobacteria* sp.	ERZ870109
Metabat.158	96.6	2.4	4,406,707	63.9	High	*Alphaproteobacteria* sp.	ERZ870094
Metabat.151	96.4	5.1	4,482,786	36.9	Medium	*Bacteroidetes* sp.	ERZ870080
Metabat.89	96.3	2.2	2,753,811	53.6	High	*Verrucomicrobia* sp.	ERZ870097
Metabat.179	95.3	0.9	2,724,314	69.2	High	*Chloroflexi* sp.	ERZ870110
Metabat.143	94.4	2.0	2,442,640	66.0	High	*Chloroflexi* sp.	ERZ870099
Metabat.177_sub	94.3	6.7	4,572,140	59.2	Medium	*Proteobacteria*	ERZ870074
Metabat.40	93.6	3.4	3,692,750	65.4	High	*Betaproteobacteria*	ERZ870086
Metabat.123_sub	93.2	9.6	4,243,256	68.2	Medium	*Actinobacteria* sp.	ERZ870064
Metabat.14	92.6	3.9	2,553,466	66.1	High	*Chloroflexi* sp.	ERZ870083
Metabat.133	91.6	1.9	2,305,255	67.3	High	Candidate *Dormibacteraeota* sp.	ERZ870101
Metabat.147	91.3	7.9	4,040,741	55.9	Medium	*Verrucomicrobia* sp.	ERZ870070
Metabat.67	89.9	5.5	1,906,190	68.3	Medium	*Actinobacteria* sp.	ERZ870077
Maxbin2.041	89.7	2.2	3,901,541	59.3	Medium	*Acidobacteria* sp.	ERZ870096
Metabat.164_sub	89.4	4.7	2,849,413	64.2	Medium	*Chloroflexi* sp.	ERZ870081
Maxbin2.071_sub	86.5	8.3	3,144,416	70.1	Medium	*Actinobacteria* sp.	ERZ870067
Metabat.51	85.9	8.2	2,827,458	60.8	Medium	*Gemmatimonadetes* sp.	ERZ870069
Maxbin2.021_sub	85.7	6.8	2,132,093	70.0	Medium	*Chloroflexi* sp.	ERZ870073
Metabat.154	84.8	2.5	2,330,430	69.6	Medium	*Actinobacteria* sp.	ERZ870091
Metabat.156	84.7	1.5	2,372,385	35.6	Medium	*Bacteroidetes* sp.	ERZ870107
Maxbin2.102_sub	84.6	1.8	2,720,713	64.2	Medium	*Acidobacteriaceae* sp.	ERZ870102
Metabat.138	84.4	2.0	2,813,002	55.1	Medium	*Verrucomicrobia* sp.	ERZ870098
Metabat.172_sub	83.2	2.4	2,237,822	65.1	Medium	*Rhizobiales* sp.	ERZ870093
Maxbin2.128	82.1	9.8	2,270,224	51.7	Medium	*Alphaproteobacteria* sp.	ERZ870062
Maxbin2.086_sub	81.9	9.7	3,605,629	57.9	Medium	*Acidobacteria* sp.	ERZ870063
Metabat.159_sub	81.7	8.7	2,099,345	55.9	Medium	*Verrucomicrobia* sp.	ERZ870066
Metabat.121	80.2	3.5	2,452,147	35.8	Medium	*Bacteroidetes* sp.	ERZ870085
Metabat.122	77.3	8.3	2,004,053	67.9	Medium	*Actinobacteria* sp.	ERZ870068
Metabat.163_sub	73.8	2.0	2,166,091	71.1	Medium	*Solirubrobacterales* sp.	ERZ870100
Metabat.72	72.9	3.2	3,967,186	40.8	Medium	*Bacteroidetes* sp.	ERZ870087
Metabat.167	72.5	2.3	2,102,822	70.7	Medium	*Actinobacteria* sp.	ERZ870095
Metabat.115	72.1	5.1	1,795,856	70.2	Medium	*Actinobacteria* sp.	ERZ870079
Metabat.174	71.6	2.5	2,317,750	35.4	Medium	*Bacteroidetes* sp.	ERZ870092
Metabat.53	71.3	8.2	5,534,727	37.1	Medium	*Bacteroidetes* sp.	ERZ879091
Metabat.100	69.8	0.9	2,344,086	68.8	Medium	*Solirubrobacterales* sp.	ERZ870111
Metabat.26	67.9	0.8	2,094,082	68.3	Medium	*Actinobacteria* sp.	ERZ870112
Metabat.119	67.2	0.0	731,988	47.4	Medium	*Saccharibacteria* sp.	ERZ870115
Metabat.140	67.1	6.0	1,381,010	69.0	Medium	*Chloroflexi* sp.	ERZ870075
Metabat.16	66.2	1.5	844,132	41.3	Medium	*Thaumarchaeota* sp.	ERZ870108
Maxbin2.015	65.5	4.0	2,138,105	49.3	Medium	*Proteobacteria* sp.	ERZ870082
Maxbin2.090	64.2	5.9	2,561,445	65.2	Medium	*Gemmatimonadetes* sp.	ERZ870076
Metabat.48	63.6	1.7	741,844	38.9	Medium	Candidate *Levybacteria* sp.	ERZ870104
Metabat.28	63.5	2.6	2,845,538	67.0	Medium	*Burkholderiales* sp.	ERZ870090
Metabat.166	63.3	0.2	739,124	45.6	Medium	*Saccharibacteria* sp.	ERZ870114
Maxbin2.012	63.2	6.9	2,750,113	55.1	Medium	*Proteobacteria* sp.	ERZ870072
Metabat.12	63.0	1.6	2,221,067	39.3	Medium	*Bacteroidetes* sp.	ERZ870106
Metabat.155_sub	58.6	2.9	1,479,786	56.8	Medium	*Nitrosomonadales* sp.	ERZ870089
Metabat.94	58.2	3.1	3,546,342	59.7	Medium	*Acidobacteria* sp.	ERZ870088
Maxbin2.095_sub	53.4	8.8	2,850,869	56.6	Medium	*Nitrospirae* sp.	ERZ870065
Metabat.1	51.9	0.6	1,114,730	51.0	Medium	*Nitrosospira* sp.	ERZ870113
Metabat.170	51.4	3.6	3,578,256	59.6	Medium	*Acidobacteria* sp.	ERZ870084
Metabat.175	48.3	1.6	1,833,825	41.8	Low	*Bacteroidetes* sp.	ERZ870105
Maxbin2.011	42.4	5.2	2,493,859	62.5	Low	*Rhizobiales* sp.	ERZ870078
Maxbin2.064_sub	40.9	7.4	1,652,927	43.4	Low	*Firmicutes* sp.	ERZ870071
Maxbin2.096_sub	31.1	1.8	1,233,990	54.5	Low	*Acidobacteria* sp.	ERZ870103

[a]Uncultured isolates were used.

were distributed across the following phyla: *Actinobacteria*, 11; *Proteobacteria*, 11; *Bacteroidetes*, 8; *Acidobacteria*, 7; *Chloroflexi*, 6; *Verrucomicrobia*, 4; *Saccharibacteria*, 2; *Gemmatimonadetes*, 2; candidate phylum *Dormibacteraeota* (AD3), 1; candidate phylum *Levybacteria*, 1; *Firmicutes*, 1; *Nitrospirae*, 1; and *Thaumarchaeota*, 1 (Table 1). Here, we report MAGs with 31.07 to 98.20% estimated completeness, and therefore the MAG sizes range from 731,988

to 5,534,727 bp. The MAGs will be used to investigate metabolic pathways that could impact SOM decomposition in permafrost soils. Results from the comparative genomic analyses of these MAGs will be published elsewhere.

Data availability. The shotgun sequence data were deposited in the European Nucleotide Archive (ENA) database under the study number PRJEB30872 with the accession numbers ERR3078909 to ERR3078913. The MAGs are publicly available in the ENA under the analysis accession numbers ERZ870056, ERZ870062 to ERZ870115, and ERZ879091.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Research School in Bioinformatics, Biostatistics, and Systems Biology (NORBIS) to Yaxin Xue. Funding for this work was provided to Neslihan Taş by the Office of Biological and Environmental Research in the DOE Office of Science—Early Career Research Program. This study is part of the project "Microorganisms in the Arctic: major drivers of biogeochemical cycles and climate change" (RCN 227062), funded by the Norwegian Research Council (principal investigator [PI], Lise Øvreås). Lise Øvreås was awarded the Fulbright Arctic Chair 2012 to 2013 (Fulbright Foundation).

REFERENCES

1. Jansson JK, Taş N. 2014. The microbial ecology of permafrost. Nat Rev Microbiol 12:414–425. https://doi.org/10.1038/nrmicro3262.
2. Mackelprang R, Saleska SR, Jacobsen CS, Jansson JK, Taş N. 2016. Permafrost meta-omics and climate change. Annu Rev Earth Planet Sci 44:439–462. https://doi.org/10.1146/annurev-earth-060614-105126.
3. Gittel A, Bárta J, Kohoutová I, Schnecker J, Wild B, Capek P, Kaiser C, Torsvik VL, Richter A, Schleper C, Urich T. 2014. Site- and horizon-specific patterns of microbial community structure and enzyme activities in permafrost-affected soils of Greenland. Front Microbiol 5:541. https://doi.org/10.3389/fmicb.2014.00541.
4. Deng J, Gu Y, Zhang J, Xue K, Qin Y, Yuan M, Yin H, He Z, Wu L, Schuur EAG, Tiedje JM, Zhou J. 2015. Shifts of tundra bacterial and archaeal communities along a permafrost thaw gradient in Alaska. Mol Ecol 24:222–234. https://doi.org/10.1111/mec.13015.
5. Taş N, Prestat E, Wang S, Wu Y, Ulrich C, Kneafsey T, Tringe SG, Torn MS, Hubbard SS, Jansson JK. 2018. Landscape topography structures the soil microbiome in arctic polygonal tundra. Nat Commun 9:777. https://doi.org/10.1038/s41467-018-03089-z.
6. Müller O, Bang-Andreasen T, White RA, Elberling B, Taş N, Kneafsey T, Jansson JK, Øvreås L. 2018. Disentangling the complexity of permafrost soil by using high resolution profiling of microbial community composition, key functions and respiration rates. Environ Microbiol 20:4328–4342. https://doi.org/10.1111/1462-2920.14348.
7. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, Voigt AY, Zeller G, Sunagawa S, Bork P. 2016. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics 32:2520–2523. https://doi.org/10.1093/bioinformatics/btw183.
8. Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033.
9. Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607. https://doi.org/10.1093/bioinformatics/btv638.
10. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165. https://doi.org/10.7717/peerj.1165.
11. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 3:836–843. https://doi.org/10.1038/s41564-018-0171-1.
12. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.
13. Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 7:11257. https://doi.org/10.1038/ncomms11257.
14. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731. https://doi.org/10.1038/nbt.3893.

Paper IV

Metagenome-assembled Genome Distribution and Key Functionality Highlight Importance of Aerobic Metabolism in Svalbard Permafrost.

Xue, Y.*, Jonassen, I., Øvreås, L., & Taş, N.

FEMS microbiology ecology, 96(5), fiaa057. (2020)

https://doi.org/10.1093/femsec/fiaa057.

FEMS Microbiology Ecology, 96, 2020, fiaa057

doi: 10.1093/femsec/fiaa057
Advance Access Publication Date: 17 April 2020
Research Article

RESEARCH ARTICLE

Metagenome-assembled genome distribution and key functionality highlight importance of aerobic metabolism in Svalbard permafrost

Yaxin Xue[1], Inge Jonassen[1], Lise Øvreås[2,3] and Neslihan Taş[4,5,*,†]

[1]Computational Biology Unit, Department of Informatics, University of Bergen,Thormøhlensgt 55 N-5008, Bergen, Norway, [2]Department of Biological Sciences, University of Bergen, Thormøhlensgt 53 N-5020, Bergen, Norway, [3]University Center in Svalbard, UNIS, N-9171, Longyearbyen, Norway, [4]Ecology Department, Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA and [5]Environmental Genomics and Systems Biology, Biosciences Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

*Corresponding author: Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 70A-3317 Berkeley, CA 94720, USA. Tel: +15104865538; E-mail: ntas@lbl.gov

One sentence summary: Two novel computational methods were developed to conduct a deep metagenomic analysis of Svalbard permafrost samples, resulting in previously unreported trends in permafrost, especially the importance of aerobic metabolisms.

Editor: Max Häggblom

†Neslihan Taş, http://orcid.org/0000-0001-7525-2331

ABSTRACT

Permafrost underlies a large portion of the land in the Northern Hemisphere. It is proposed to be an extreme habitat and home for cold-adaptive microbial communities. Upon thaw permafrost is predicted to exacerbate increasing global temperature trend, where awakening microbes decompose millennia old carbon stocks. Yet our knowledge on composition, functional potential and variance of permafrost microbiome remains limited. In this study, we conducted a deep comparative metagenomic analysis through a 2 m permafrost core from Svalbard, Norway to determine key permafrost microbiome in this climate sensitive island ecosystem. To do so, we developed comparative metagenomics methods on metagenomic-assembled genomes (MAG). We found that community composition in Svalbard soil horizons shifted markedly with depth: the dominant phylum switched from *Acidobacteria* and *Proteobacteria* in top soils (active layer) to *Actinobacteria*, *Bacteroidetes*, *Chloroflexi* and *Proteobacteria* in permafrost layers. Key metabolic potential propagated through permafrost depths revealed aerobic respiration and soil organic matter decomposition as key metabolic traits. We also found that Svalbard MAGs were enriched in genes involved in regulation of ammonium, sulfur and phosphate. Here, we provide a new perspective on how permafrost microbiome is shaped to acquire resources in competitive and limited resource conditions of deep Svalbard soils.

Keywords: Svalbard; permafrost; microbiome; metagenome-assembled genomes; aerobic metabolism

Received: 31 October 2019; **Accepted:** 9 April 2020

INTRODUCTION

Permafrost covers nearly one quarter of Earth's terrestrial surface and stores an estimated amount of 20%–50% of global soil organic matter (SOM) (Schuur *et al.* 2008; Tarnocai *et al.* 2009). In the Northern Hemisphere as much of 24% of the soil is permanently frozen (Alley *et al.* 2007). These ecosystems are proposed to provide a unique environment for cold-adapted microorganisms and shown to contain highly diverse microbial communities (Jansson and Taş 2014). Global warming is expected to have its largest impact through thawing of permafrost and the scale of this impact depends strongly on the amount and vertical distribution of ground ice (Kokelj *et al.* 2017). During the past decade, with steadily rising temperatures, permafrost thaw has accelerated across the Arctic areas (Hayes *et al.* 2014). The effect of large-scale permafrost thaw becomes a serious concern as it may increase the microbial activity leading to SOM degradation and release of more greenhouse gases (GHGs) – such as carbon dioxide (CO_2) and methane (CH_4) – hence contributing to further global warming (Jansson and Taş 2014). Therefore, it is highly relevant to characterize the bacterial community residing in the permafrost in terms of species composition and its metabolic and functional potential. Advances in next-generation sequencing (NGS) has expanded our ability to characterize the microbiome and investigate potential metabolisms from permafrost samples. For example, metagenomics was critical to identify substantial functional and compositional differences between active layer (AL: experiences seasonal thaw-refreeze) and permafrost layer (PL: constantly frozen for more than two consecutive years), which showed that transition from frozen to thaw state stimulates SOM-degrading microbes (Mackelprang *et al.* 2011). While metagenomics continues to transform our understanding of microbial functions upon thaw (Jansson and Taş 2014; Hultman *et al.* 2015; Woodcroft *et al.* 2018) most of our current knowledge is still based on studies that are focused on 16S rRNA gene-sequencing analysis (Wilhelm *et al.* 2011; Gittel *et al.* 2014; Koyama *et al.* 2014; Deng *et al.* 2015; Mackelprang *et al.* 2016a). These studies are informative for describing species or groups of species in a community permafrost microbiome but is less suited for exploring functional potential and novel species distribution (Knight *et al.* 2018).

The Svalbard archipelago is a unique permafrost environment located at Arctic–Atlantic Ocean border. About 60% of the land is covered by glaciers but remainder periglacial environment contains the largest permafrost area in Europe outside of Russia. In contrast to other regions with extensive permafrost areas, such as Siberia and Northern Alaska, permafrost in Svalbard is presumably of young age (i.e. Holocene) specifically at low altitude areas around central Spitsbergen. However, high altitude permafrost in Svalbard may represent an exception to this (Humlum, Instanes and Sollid 2003). The North Atlantic Current dampens polar influence in Svalbard where especially winter temperatures could be up to 20°C higher than similar latitudes in Russia and Canada (Humlum, Instanes and Sollid 2003). As a result, permafrost in Svalbard is proposed to be more sensitive to changes in temperature and soil thickness (Humlum, Instanes and Sollid 2003). Research in Svalbard provides an opportunity to study the immediate effects climate change and permafrost thaw. Svalbard had been a focal point of studying glacial, subglacial (recently deglaciated), cryoconite sediments (Kastovská *et al.* 2005; Edwards *et al.* 2011) and tundra microbiomes (Tveit *et al.* 2013; Schostag *et al.* 2015; Bang-Andreasen *et al.* 2017). The Arctic tundra in Svalbard contains diverse microorganisms which are active throughout the winter despite the freezing conditions (Schostag *et al.* 2015). Peatlands of Svalbard are shown to be inhabited by microbes governing biogeochemical cycles through hydrolysis of plant polymers, fermentation, methanogenesis and methanotrophy where *Actinobacteria* was identified as a key phylum carrying out SOM degradation (Tveit *et al.* 2013). However, in comparison with other soils, our knowledge of the Svalbard permafrost microbiome is limited. In a previous publication from Adventdalen Valley permafrost, we showed that PL were significantly different from the AL, where microbial community structure changed strongly with depth and *Actinobacteria* were identified as the dominant microbial phylum of PL via 16S rRNA gene sequencing (Müller *et al.* 2018). However, others also showed that *Actinobacteria*, *Bacteroidetes*, *Firmicutes* and *Proteobacteria* are major parts of the microbiome (Bang-Andreasen *et al.* 2017) of near surface permafrost at this location suggesting that Adventdalen Valley permafrost is likely to have a highly heterogeneous composition.

In this study, we investigated the microbial composition and functional potential through a permafrost core from Svalbard's Adventdalen Valley in order to determine key microbial functional potential. Although metagenomics provides holistic view to microbial functions from largely unculturable permafrost microbiome (Mackelprang *et al.* 2016b), several aspects of bioinformatic analysis remain challenging. For example, we are still lacking an effective and robust workflow for recovering quality metagenome-assembled genomes (MAGs) from the permafrost communities due to the large complexity and heterogeneity present in these soils. More importantly tools enabling systematic comparison among metagenomes by taking full advantage of data and maximize the information driven from samples, are urgently needed. To address these issues, we developed computational tools to aid high-quality MAG recovery and to identify key functions through comparative functional analysis. We aimed to capture the variances in microbial composition and trends in functional potential throughout the depth profile (AL to PL). We hypothesized that (i) phylogenetically related MAGs resides in PL where (ii) SOM-degradation pathways in key permafrost microbiome are represented by mix of aerobic and anaerobic processes.

MATERIALS AND METHODS

Sample collection

Soil samples were obtained from an ice-wedge polygon site in the Adventdalen Valley in Svalbard, Norway (78.186 N, 15.9248E) in 2011. Adventdalen represents a classic high-arctic fjord-valley, which are sediment filled paleo fjords characteristic to formerly glaciated mountain coastal areas. Detailed description and procedures for core collection and characterizing soil samples were described previously (Müller *et al.* 2018). In short, the permafrost core was collected in by automated drilling in April 2011 in Adventdalen, Svalbard. The total length of the core was 198 cm, and the core was immediately frozen at −20°C, until further processing. The entire core was scanned by X-ray computed tomography (CT) imaging, and cut into 1–2 cm slices using saw blades sterilized with ethanol. To remove potential surface contaminants (Bang-Andreasen *et al.* 2017) from the core fragments the outermost 2 cm were cut off using sterile blades. Based on the results from the temperature loggers, CT scanning and water content of the permafrost core (Müller *et al.* 2018) active and PL depths were decided. Five fragments, one from AL and four PL,

with different depths AL1 (7 cm), PL1 (110 cm), PL2 (122 cm), PL3 (135 cm), PL4 (170 cm) below the soil surface were subjected to metagenomics analyses. Both AL and PL soils were acidic (pH: 4.6 AL; pH: 4.5–5.0 PL) and contained 1.3%–1.7% C gr soil (Müller et al. 2018).

Metagenomic sequencing, recovery and refinement of MAGs

DNA was extracted and libraries prepared using procedures described previously (Xue et al. 2019). Metagenome sequencing was performed using the Illumina HiSeq 2500 instrument to acquire 150 bp paired-end sequences, generating around 20Gbp per sample after quality control (trim and discard low-quality sequences) with MOCAT2 v2.0.0 (Kultima et al. 2016). The analysis workflow used here organizes several bioinformatic scripts to recover and refine MAGs (Fig. 1A). Firstly, all quality controlled reads were co-assembled with MEGAHIT v1.1.3 (Li et al. 2015). Two binning tools, MaxBin2 v2.2.5 (Wu, Simmons and Singer 2016) and MetaBAT2 v2.12.1 (Kang et al. 2015), were used and output bins were further dereplicated and aggregated with DASTool v1.1.10 (Sieber et al. 2018). The checkM v1.0.11 (Parks et al. 2015) was used to determine completeness and contamination of MAGs. We observed that a large portion of bins had a high contamination percentage even after using DASTool. To improve the quality of MAGs, we developed a script, called 'Decon_MAG_by_taxa.py', that will subset each bin into collections of contigs from the same taxonomic classification. In theory each bin represents an individual genome with single-taxon annotation. However, in practice bins contain contigs from other taxa due to the complexity of microbial communities. Yet it is possible to remove those contaminations by parsing their taxonomic classification. First, each bin was annotated with Kaiju v1.6.2 (Menzel, Ng and Krogh 2016) using default parameters utilizing the NCBI nr database to classify each contig into a taxonomic rank, from phylum to species. Then script extracts contigs with the same taxonomic classification at each rank and generates multiple subsets of fasta files corresponding to each rank.

By default, Kaiju will return a 'NA' if it cannot find a taxonomic classification at certain ranks, which results in many 'NA's at lower rank and loss of hierarchical taxonomic structure while contamination may happen in any rank. To maximum utilize the taxonomic annotation, here we considered 'NA' in Kaiju annotation as a special taxonomic rank, and sustained the hierarchical structure under the following rules: (i) when 'NA' observed in a non-phylum level, a label is generated via combining higher taxonomic rank information with '_NA_' denotation as a rank identifier (P: Phylum, C: Class, O: Order, F: Family, G: Genus, S: Species), (ii) if 'NA' appeared at the phylum level a label is generated as 'P_NA'. For example, if a contig is annotated as: 'C1; Proteobacteria; Alphaproteobacteria; Rhizobiales; NA; NA; Unknown species', then it will be converted to: 'C1; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiales_NA_F; Rhizobiales_NA_F_NA_G; Unknown species'. Later, the script calculates the percentage of every taxa label in each rank and keeps labels whose percentage were higher than a user-defined threshold (default = 0.5). As the script provides multiple subsets of fasta corresponding to different ranks for each bin, the user can run CheckM with all of these subsets and evaluate the best tradeoff between completeness and contamination. More detailed description of our MAG refinement method is available at: https://github.com/yxxue/Recovery-and-refinement-of-MAGs-for-permafrost-metagenome.

MAGs were annotated to a taxonomic rank based on Kaiju and GTDB-Tk v0.3.3 (Parks et al. 2018) annotation. For each sample, we aligned sequence data against all refined MAGs using BBMAP v37.36 (https://sourceforge.net/projects/bbmap/) with default parameters. The relative abundance of each MAG was calculated by aggregating the mapping ratio of contigs pertained to this MAG. RAST annotations for the MAGs are publicly available at KBase narrative (Arkin et al. 2018): https://narrative.kbase.us/narrative/ws.50152.obj.370 (KBase account required).

Coverage-based functional analysis in a MAG-centric view

Normalization coverage

To perform quantitative comparative analysis, we utilized a normalization strategy – TPM (Transcripts Per Kilobase Million) – which is commonly used in normalizing gene expression in RNA-seq analysis (Wagner, Kin and Lynch 2012). Our normalization method consists of three steps. Firstly, we considered coverage of contigs as RPK value of contigs, as coverage represent the number of mapped reads divided by the length of the contig, which is analogous to be the concept of PRK value. Second, we calculated the 'per million' scale factor by dividing total mapped read counts with 1 million in each sample. For example, the mapped reads count in AL1 was 9 171 534, thus the scaling factor in AL1 would be 9.171534 (9171,534/1000,000). Finally, coverage was normalized by dividing corresponding scaling factor, respectively.

Definition of groups

We pre-defined several groups combining the coverage patterns with geographical significance (Table 1). To capture the distinct variation in terms of coverage profiles among contigs, we chose median of the normalized coverage as a global threshold to classify contigs and removed low coverage contigs (LO). Active layer (AL) was simple case in our data sets since there was only one sample representing the active layer while we found that coverage distribution in PL were more complicated and needed to be considered separately: some contigs were only present in specific samples, while others appeared in full or in part in all PL samples. Therefore, we defined three groups for PL samples: PL_Pi(only present in specific samples), PL_SUB (present in some of the samples) and PL_ALL(present in all samples). Besides, we derived contigs that had a strong correlation (0.9) between depth and coverage from PL samples, namely KI and KD. Group BO represented the ubiquitous contigs in Svalbard AL and PL, remaining contigs were assigned to UN (unknown).

Calculating KEGG Module abundance of MAGs

We considered each MAG as an independent unit and normalized coverage was used to represent KEGG Orthology (KO) abundance. An illustration of our strategy is shown in Fig. 1B. First, we used Prodigal v2.6.3 (Hyatt et al. 2012) with meta procedure to predict genes for all MAGs. Predicted protein file was then uploaded to perform KO annotation using GhostKOALA (Kanehisa, Sato and Morishima 2016). Later, we converted the gene-based KO annotation to a MAG-centric hierarchical structure and calculated KEGG module abundance. KEGG Module (MO) is a collection of KOs, which represents tight functional components with a clearer biological significance comparing with KO identifiers. In each MAG, abundance of KEGG Modules (MOs) was calculated by summing the average existing KO and then dividing

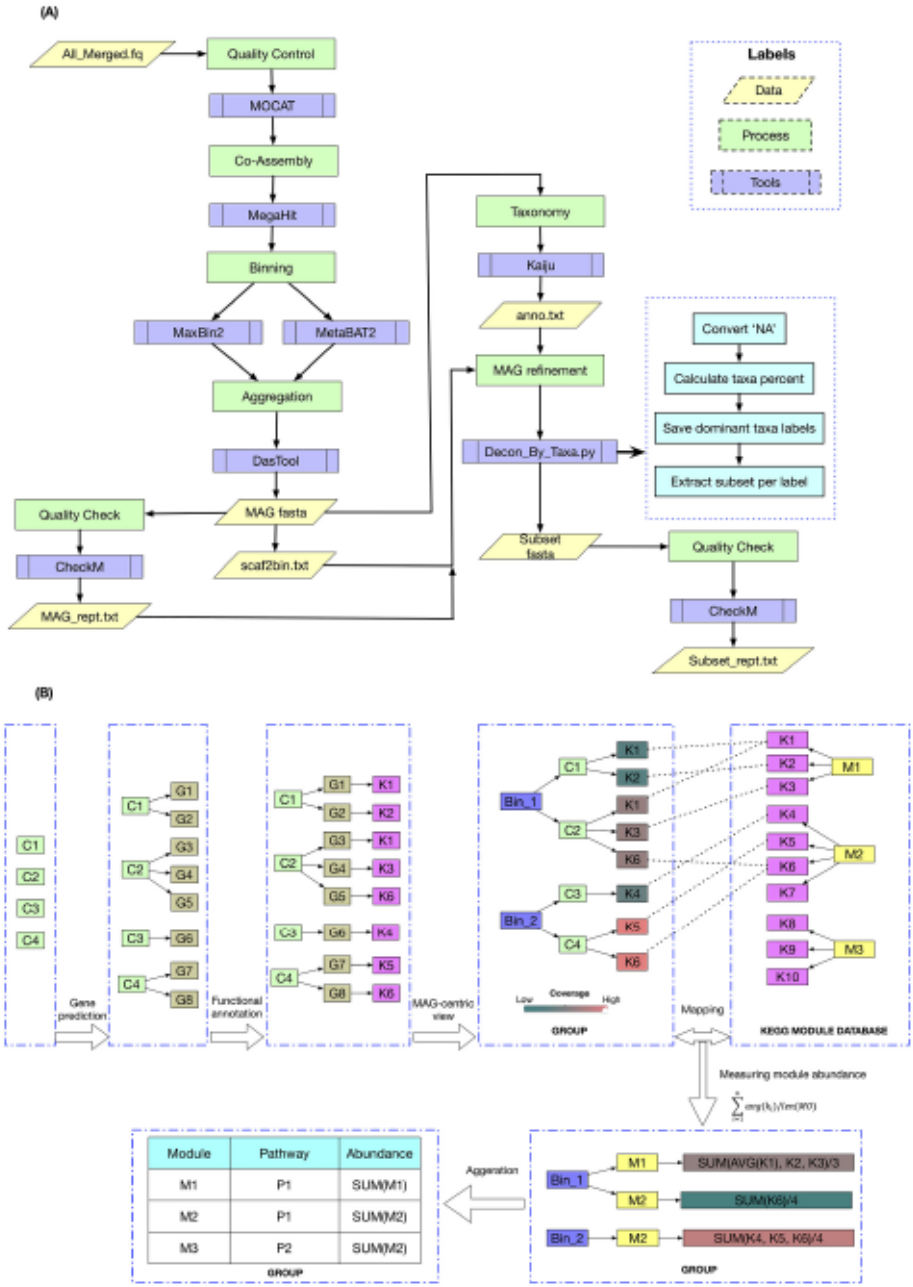


Figure 1. Overview of Svalbard permafrost bioinformatic strategies. (A), An improved workflow for MAG recovery and refinement. The entire workflow includes several steps and tools, including quality control, co-assembly, binning, aggregation, quality check and MAG refinement. See details in methods. (B), A schematic illustration for coverage-based functional analysis in a MAG-centric view. Each contig (C), contains multiple genes (G) that can be annotated with KEGG Orthology (K) and linked with KEGG Module (M) database. Coverage can be used as a quantitative measure for each KEGG Module hence allowing analysis of trends of increasing or decreasing representation between and within sample set (Table 1). SUM: Summation. AVG: Average.

Table 1. Definition of sample groups. AL: normalized coverage in active layer, PL: normalized coverage in permafrost layer samples. TH: threshold (median of normalized coverage). DEPTH (cm under surface): 110, 122, 135, 170. CORR: Pearson correlation.

Groups	Definition	Criteria
AL	Presence in AL	AL > = TH and ALL(PL) < = TH
BO	Presence Both in AL and PL	AL > = TH and ALL(PL) > = TH
LO	Absence Both in AL and PL	AL < = TH and ALL(PL) < = TH
PL_SUB	Presence in subset (2 or 3) PL	AL < = TH and SUB(PL) > = TH
PL_ALL	Presence in all PL	AL < = TH and ALL(PL) > = TH
PL_Pi	Presence in unique PL (P1, ..., P4)	AL< = TH and UNIQUE(PL_Pi) > = TH
KI	Increasing trend in PL_ALL or PL_SUB	In (PL_ALL or PL_SUB) and CORR(PL, DEPTH) > = 0.9
KD	Decreasing trend in PL_ALL or PL_SUB	In (PL_ALL or PL_SUB) and CORR(PL, DEPTH) < = −0.9
UN	Unknown groups	Others

by total number of KO identifiers in this module. MO abundance in each group was measured by aggregating MO abundance of all MAGs presented at each group, respectively. As the demonstration shown in Fig. 1B, M1 consists of 3 KO (K1-K3) and M2 of 4 KO (K4-K6). Bin_1 includes two weighted (normalized coverage) contigs with 5 KO: C1 (w_1K1, w_1K2) and C2 (w_2K1, w_2K3, w_2K6). Based on the definition of MAG, we suppose that contigs in the same MAG are able to share their KO: we further use average if there are multiple hits for identical KO in the same MAG. Therefore, M1 abundance in Bin_1 is: $SUM(AVG(w_1K1, w_2K1), w_1K2, w_2K3) / 3$. Similarly, only K4 in M2 is detected in Bin_1 while M2 consists of 4 KO, thus M2 abundance in Bin_1 is: $SUM(w_2K6) / 4$. Finally, M1 abundance in this group is simply aggregating all M1 abundance of each MAG. A detailed demonstration of performing our coverage-based analysis and source code are available at https://github.com/yxxue/Coverage-based-functional-analyisis-in-a-MAG-centric-view.

RESULTS

Unique MAGs become abundant with depth in Svalbard permafrost

We reconstructed 56 MAGs from 13 phyla, including 8 high, 44 medium and 4 low-quality draft in accordance with MIMIG standards (Bowers *et al.* 2017). In total, the analyzed MAGs constituted on average 11.3% of the reads obtained for each sample (min. 7.1% and max. 13.4%). In this location, we found several MAGs belonging to *Actinobacteria*, *Proteobacteria*, *Bacteroidetes*, *Acidobacteria* and *Chloroflexi* to be most abundant (Fig. 2). Additionally, MAGs belonging to *Verrucomicrobia*, *Saccharibacteria*, *Gemmatimonadetes*, *Firmicutes*, *Nitrospirae*, *Thaumarchaeota*, *candidate phylum Dormibacteraeota* (AD3) and *candidate phylum Levybacteria* were found in lower abundance. We did not recover any methanogenic archaea in this location. Detailed description of MAGs were published previously (Xue *et al.* 2019). MAGs showed low similarity to publicly available genomes (Table S1, Supporting Information) suggesting that they represent novel species. We also compared these MAGs to microbiomes of recent stable isotope probing showing activity at subzero conditions (Tuorto *et al.* 2014; Gadkari *et al.* 2019). Svalbard MAGs were distantly related to these novel populations and showed 75%–88% similarity on 16S rRNA genes (Table S2).

Microbial community composition based on changes in MAG abundance showed distinct differences between AL and PL where predominant MAG also changed with depth (Fig. 2, Fig. S1, Supporting Information). In the AL, the most abundant phyla were *Acidobacteria* and *Proteobacteria* while PL MAGs were dominated by *Actinobacteria*, *Bacteroidetes*, *Chloroflexi* and *Proteobacteria*. The most dominant MAGs in AL – maxbin2.039_sub (*Acidobacteria*), metabat.158 (*Proteobacteria*), metabat.89 (*Verrucomicrobia*) – declined to nearly undetectable levels in the PL. Members of *Proteobacteria*, *Verrucomicrobia* and *Chloroflexi*, were ubiquitous in PL and had similar abundances in the upper PL (PL1 and PL2) than deep PL samples (PL3 and PL4). We observed a decline in *Acidobacteria* and some *Actinobacteria* MAG abundances with depth. Previous 16S rRNA based analysis detected a single Actinobacteria family – *Intrasporangiaceae* – to be strongly dominant throughout the PL (Müller *et al.* 2018). However, we could not detect similar populations in this data set. We further examined both assembled contigs and un-assembled raw reads by Kaiju annotation and BBMAP alignment and found that *Intrasporangiaceae* constituted a relatively small portion of the contigs in assembled reads (1.2%) and in general of metagenomes as represented by raw reads (total of 3.3% in all metagenomes). More unique but highly represented MAGs were found in the deepest samples, like metabat.179 (*Chloroflexi*) in PL3 and metabat.151 (*Bacteroidetes*) in PL4. Likewise *Saccharibacteria*, candidate phylum *Dormibacteraeota* (AD3) and candidate phylum *Levybacteria* had their highest abundance in deep permafrost.

Determining the complexities of the Svalbard permafrost by coverage-based groups

Many permafrost studies are focused on sample specific comparative analysis (Yergeau *et al.* 2010; Mackelprang *et al.* 2017; Müller *et al.* 2018), however, sample-based analysis is not able to reflect the complexity of microbial spatial arrangement directly. Moreover, we observed that there were some regular patterns in coverage distribution across multiple samples. To utilize to the maximum the information and enable a deeper understanding of permafrost microbial universe at a high-resolution, we developed a comparative strategy to investigate the variance of functional potential combing the genomic (coverage) and functional (KEGG) information in a MAG-centric view. Only contigs from MAGs were included in this analysis. 20,573 contigs originating from refined MAGs were assigned to classification groups (Table 1). PL group represented the largest portion of the data by covering 60% of the contigs (Fig. 3). About 10% of the contigs were shared between both AL and PL and ubiquitous at all samples (BO) while 13% of the contigs were only found in AL. After filtering 14% of low abundance contigs (LO), only 3% could not be assigned to any of the above groups (UN). Within PL 26% of the contigs fell under subset of PL (PL_SUB) category, 19% of the contigs was found in all 4 PL (PL_ALL) and represent the key functions in Svalbard permafrost. Only a small portion of the contigs were specific to each depth (a quarter of contigs were exclusively observed in only one sample (PL_P1, PL_P2, PL_P3, PL_P4) covering 2%–6% of the total contigs. We identified only a small fraction of contigs in PL_ALL and PL_SUB that had a strong correlation with depth profile: about 5% of contigs decreased (KD) and

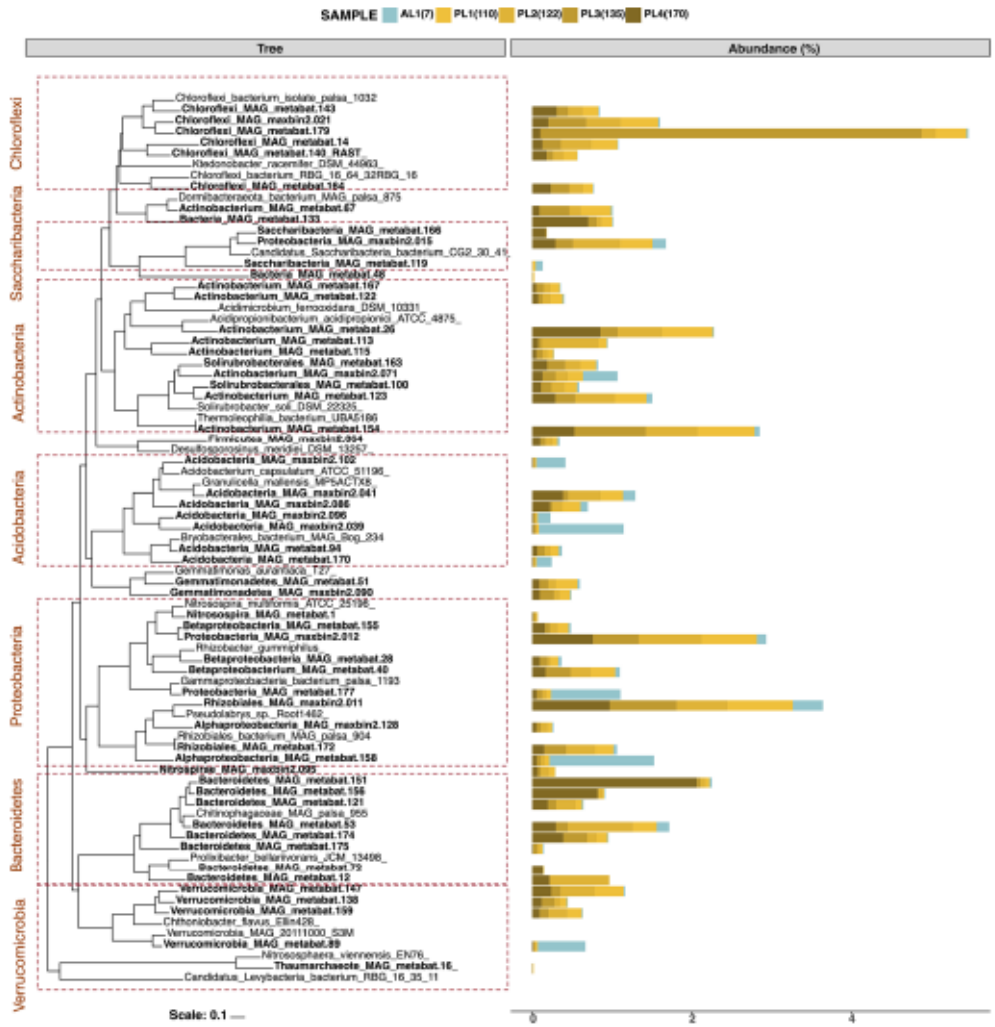


Figure 2. The relative abundance of MAGs shifts between samples. Percent MAG abundance in five soil layers, one active layer (AL, blue color) and four permafrost layers (PL, yellow to brown color), with different depths below the soil surface are shown: AL1 (7 cm), PL1 (110 cm), PL2 (122 cm), PL3 (135 cm) and PL4 (170 cm). Maximum likelihood phylogenetic tree was constructed by using 49 highly conserved COG families from publicly available genomes.

1% function represented in contigs increased (KI) with depth. Group-based abundance distribution showed a clear distinct difference of dominant phylum among groups (Fig. S1, Supporting Information): *Acidobacteria* and *Proteobacteria* in AL; *Proteobacteria* in BO; *Actinobacteria*, *Bacteroidetes*, *Chloroflexi* and *Proteobacteria* in PL.

Key metabolic functions governing carbon and nutrient cycles in Svalbard permafrost

About 451 out of 808 MO in the database were detected in Svalbard MAGs, several pivotal MO were selected and assigned into corresponding metabolic pathways manually, finally 8 pathways with 102 MO were retained (Table S3). Here we report MO of different pathways showed distinct abundance among groups (Fig. 4, Fig. S2).

Carbon cycling and energy production

We examined the trends in carbon cycle and energy production genes among different groups by focusing on hydrolysis of polymers, carbohydrate active enzymes (CAZY), sugar utilization, fatty acid oxidation, oxidative phosphorylation and energy production categories. One of the most abundant MO was F-type ATPase (F-ATPase), which was present in both BO and PL_ALL. This process is important because in Bacteria most ATP is produced by F-ATPase in the cytoplasmic membrane under aerobic conditions (otherwise by glycolysis and fermentation under anaerobic conditions) (Kühlbrandt 2019). MAGs belonging

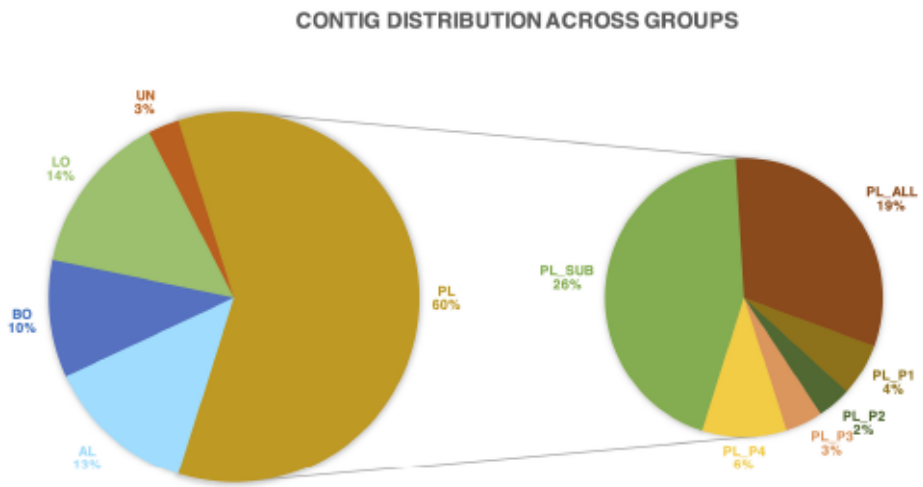


Figure 3. Contig distribution across groups. In total, 20573 contigs from all MAGs were assigned to each group based on pre-defined criteria (Table 1). KI (1%) and KD (5%) were not presented in the pie chart.

to group BO and PL_ALL also included a large number of aerobic respiratory chain complex modules, such as NADH: quinone oxidoreductase (NQR). Most living systems prefer to use conserved energy currencies, including proton motive force (PMF), NADH and ATP. NQR connects these energy currencies by using NADH produced during nutrient breakdown to generate a PMF, which is subsequently used for ATP synthesis (Barquera 2014). Collectively these trends show strong representation of aerobic respiratory processes in Svalbard permafrost, however, we also observed a decreasing trend in their abundance with depth (KD>KI, Fig. S2, Supporting Information). We further investigated dehydrogenases involved in fermentation, however, these were neither in high abundance nor showed strong grouping trends hence confirming the aerobic respiration as the dominant carbon cycling pathway in this location (Fig. S3, Supporting Information).

Polymer hydrolysis and CAZY functions were also found in abundance especially in core in PL groups (PL_ALL and PL_SUB). We found that galactose could be utilized to glucose (via Leloir) or to pyruvate (via De Ley) as both pathways were well represented in permafrost MAGs. Though a known bottleneck in Leloir is galactose transportation from outside of the cell, we also observed an over-representation of ABC transporters in PL group (Pathway: Transporters), which demonstrated the genetic potential of permafrost microbiomes to degrade galactose in carbohydrate metabolism. MAGs also showed potential to degrade more complex carbon sources all the way to CO_2 (Figs S4 and S5, Supporting Information). For example, the most abundant MAG in this set *Chloroflexi* MAG metabat.179 (Genus: UBA5189) had xylulose kinase and xylose transporters (Table S4, Supporting Information), but lacked genes encoding xylose isomerase, the first enzyme of the isomerase pathway of xylose metabolism. Therefore, it was likely that only xylulose could be utilized. MAG metabat.179 also had three copies of GH3 family beta-hexosaminidase (chitinolytic) and related N-acetyl-D-glucosamine (GlcNAc) transporters. These enzymes can cleave monomers of GlcNAc from the non-reducing end of chitin oligomers. Additionally, this MAG contained a CO dehydrogenase and could use organic acids (L-Lactate dehydrogenase and Aconitate hydratase) hence showing the potential to utilize a range of polymeric carbon to CO_2. Trehalose biosynthesis, a known carbon source and cryoprotectant, was also highly represented in PL (PL_ALL and PL_SUB). Pyruvate oxidation genes were found in both BO and PL indicating its importance for both AL and PL. We observed a decreasing trend (KD, Fig. S2, Supporting Information) in almost all polymer hydrolysis and CAZY functions except trehalose biosynthesis and pyruvate oxidation.

Nitrogen, methane and sulfur metabolisms

Within Svalbard MAGs nitrogen cycle was restricted to denitrification and dissimilatory nitrate reduction to ammonia. Both pathways were abundant in both BO and PL yet in comparison with other MOs, nitrogen cycling genes constituted a small portion of the genetic potential. Even so, some MAGs, like Bacteroidetes MAG metabat.151, showed a potential of full denitrification (Fig. S6, Supporting Information). We did not detect MO and genes involved in nitrification. At least one copy of glutamine synthetase (EC 6.3.1.2), glutamate synthase (EC 1.4.1.13) and ammonium transporters (Amt) were found in most abundant MAGs and were also well represented in both AL, BO and PL groups. All together, these results show the potential to use organic nitrogen and available ammonia in the environment through the depth profile in Svalbard soils. In this set only *Firmicutes* MAG maxbin2.064_sub (Genus: *Desulfosporosinus*) was found to be capable of nitrogen fixation, whereas another key biogeochemical process methane metabolism was not found in Svalbard MAGs.

Genes for dissimilatory sulfite reduction, the sulfur oxidation (SOX) gene complexes mediating thiosulfate oxidation and assimilatory sulfite reductase MOs were present in Svalbard MAGs. These MOs were in low abundance, but internal comparison among the groups revealed distinct trends. For example, assimilatory sulfate reduction was abundant in all groups while dissimilatory sulfate reduction had its strongest trend in PL.

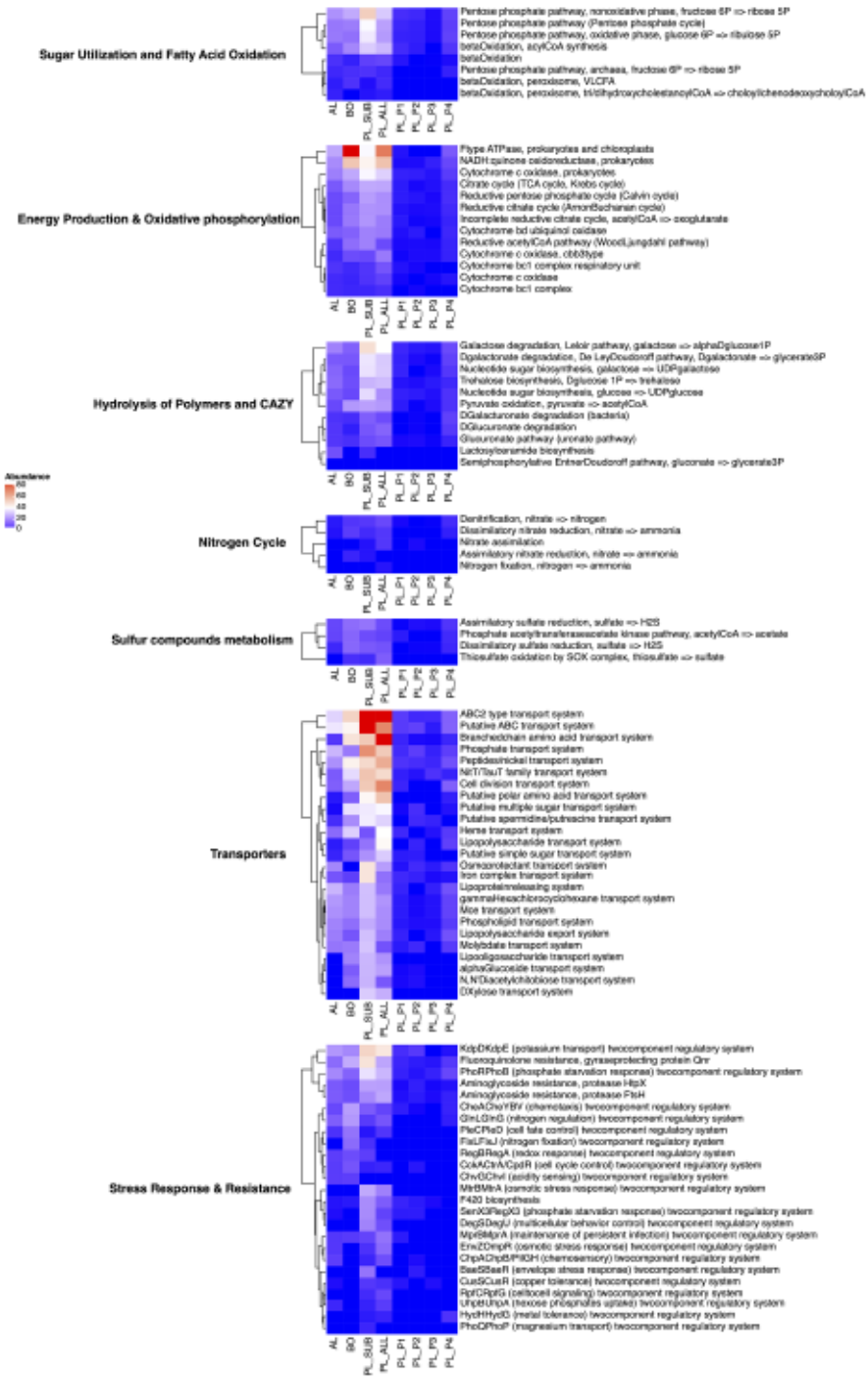


Figure 4. Trends in KEGG MO abundance in each group. The abundance of MO was calculated with normalized coverage in a MAG-centric strategy (see Methods).

However, we also detected co-occurrence of these pathways. For example, one of the most abundant MAGs, *Proteobacteria* MAG maxbin2.012 (Genus: *Gallionella*) contained genes involved both in assimilatory and dissimilatory sulfate reduction (Figs S7 and S8, Supporting Information). Additionally, thiosulfate oxidation by SOX complex was found mainly dominant in PL_ALL. This complex has been shown to produce either sulfate (complete pathway) or elemental sulfur (incomplete pathway) in diverse organisms (Houghton *et al.* 2016). We detected a decreasing trend (KD, Fig. S2, Supporting Information) in assimilatory sulfate reduction with increasing depth but not with dissimilatory sulfite reduction. These findings underlined the importance of ability to metabolize sulfur in Svalbard MAG lifecycle.

Stress responses and antibiotic resistance

Permafrost microorganisms have reportedly been shown to contain a suite of systems to deal with environmental stressors, such as cold-shock proteins and osmotic stress proteins, to counter the extreme physical and chemical stresses, including freezing temperatures, oligotrophic conditions and high salinity (Mackelprang *et al.* 2016a). We observed enrichment of KdpDE: potassium transport system in PL (PL_ALL and PL_SUB), which is required for maintaining the intracellular pH by buffering the negative charge of amino acids and used in many bacteria as a compatible solute to counteract osmotic stress (Gundlach, Commichau and Stulke 2018). Additionally, we found several two component regulatory transport systems involved in cell processes and cycle control, redox response and chemotaxis in high abundance in PL (PL_ALL and PL_SUB). Another major stress response MO was phosphate starvation response system (PhoR–PhoB), which was highly abundant in PL (PL_ALL and PL_SUB) groups, especially in PL4. Concomitantly, phosphate transport systems were among highly abundant transporters shared between AL and PL groups. These findings indicate that regulation intracellular pH and phosphorus availability are pivotal for Svalbard MAGs.

Besides MO managing environmental stressors, several antibiotic resistance genes acting against aminoglycosides and fluoroquinolones were highly abundant in PL. The aminoglycosides are natural antibiotics produced by soil bacteria where broad-spectrum bactericidal activity is achieved by interference with protein synthesis, including corruption of the genetic code via bind to rRNA and proteins within the 30S subunit of the ribosome (Cox *et al.* 2015). Fluoroquinolones are another class of broad-spectrum antibiotics that target the type II topoisomerases (DNA gyrase and topoisomerase IV) involved in the maintenance of DNA topology (Rutgersson *et al.* 2014). In a previous work, Qnr has been found as a novel mechanism of natural fluoroquinolones resistance in bacteria (Chen *et al.* 2013).

DISCUSSION

Complexity and unmatched diversity in soil metagenomes provide many challenges to data analysis; especially to those seeking to recover high-quality MAGs. DASTool (Sieber *et al.* 2018), a recently published bin refinement tool, aims to recover more near-complete genomes by aggregating and integrating bins generated from established binning algorithms (Kang *et al.* 2015; Wu, Simmons and Singer 2016). Applications of DASTool (Danczak *et al.* 2019; Imperato *et al.* 2019; Seitz *et al.* 2019) showed significantly improved MAG refinement and recovery. Yet when reconstructing permafrost MAGs these efforts might still not be sufficient. For example, in this study we observed that 21 out of 64 metagenome bins remained highly contaminated (> = 10%) even after using DASTool. We developed a script to recover bins that would be otherwise discarded (Fig. 1A). While several bin refinement strategies are deployed by IMG/M (Chen *et al.* 2019) and Anvi'o (Eren *et al.* 2015) our workflow provides a scalable and flexible alternative where thousands of bins could be analyzed systematically. We picked Kaiju as taxonomic classifier due to its extensibility as it provides fast and sensitive annotations of large contig sets. With our script, the user can choose different taxonomic reference databases – such as RefSeq, NCBI nr database or local – depending on their research goals. More importantly, contaminated contigs could be detected at all taxonomic levels and bins could be refined up to species level. Our script traces the hierarchical relationships using a user defined percentage threshold and subset contaminated bins for all ranks from phylum to species level. Removing possible contaminated contigs from a MAG may reduce completeness in some cases due to the inaccuracy in the taxonomic assignments. With our improved workflow for MAG refinement, we successfully reported 56 out of 64 MAGs with low contamination (< = 10%).

Here, we also developed a new comparative strategy for investigating functional potential based on coverage with a MAG-centric view (Fig. 1B). Generally, metagenomic functional analysis was achieved by mapping short reads or assembled contigs with predicted genes against reference databases followed by parsing the result in gene or pathway level approaches (Mackelprang *et al.* 2017; Müller *et al.* 2018). Gene-by-gene approaches utilizes most dominant gene products while overlooking the fact that biological functions rely on multiple genes while only a subset of them may be significantly abundant. For another, pathway-level analysis can miss nuanced differences in functional variance as a key pathway could contain many shared sub-pathways or genes. Motivated by this, we deployed a comparative analysis strategy that utilizes KEGG Module, a collection of manually defined functional units each encompassing a set of genes – represented by KO identifiers (Kanehisa *et al.* 2012). Comparing with pathway or gene enriched analysis, module-based analysis directly links to specific metabolic capacity (Kanehisa *et al.* 2014). Coverage is another important metagenomic characteristic (Albertsen *et al.* 2013; Sharon *et al.* 2013; Quince *et al.* 2017) that is currently not used beyond binning assembled contigs into MAGs (Alneberg *et al.* 2014; Imelfort *et al.* 2014; Kang *et al.* 2015; Wu, Simmons and Singer 2016). Our approach takes into account coverage and patterns of presence/absence and changes in coverage between samples through defining profiles or groups (Table 1) and analyzing KEGG Module-based functional information across these groups. In Svalbard permafrost this approached allowed identification of functions linked with depth in addition to aiding capture of new trends distinguishing AL and PL (Fig. 4). Although we have focused on permafrost metagenomics in this work, strategies similar to those applied here are applicable to other metagenomic studies, especially for well-characterized environments such as human gut with more accurate taxonomic classification and available MAGs as well as additional information on samples.

Svalbard soil and PLs were previously described via 16S rRNA gene amplicon sequencing up to a depth of 2 m where microbial communities in PL were dominated by the *Actinobacteria* (family *Intrasporangiaceae*). *Intrasporangiaceae* 16S rRNA gene was found in an average abundance of 70% in PL; however, we only found this group to account for 3.3% of the all raw reads and 1.2% of assembled contigs. This could be caused by differences in biases between the two sequencing methodologies. Currently

sequenced *Intrasporangiaceae* genomes (JGI IMG/M) contain 1–5 copies of 16S rRNA gene which could cause an overestimation when analyzed via amplicon sequencing. Another reason for this mismatch can originate from under-sampling of *Intrasporangiaceae* populations during metagenome sequencing. *Intrasporangiaceae* genomes are really high-GC content populations (68%–74% of GC range 63 genomes in JGI IMG/M), hence such high-GC rich fragments can be under-sampled during metagenomic library preparation, fail to pass quality checks during base calling and have difficulties during assembly (Bowers *et al.* 2015).

The grouping approach proposed here enabled us to determine key functions and trends in different cell and biochemical cycles propagated by each MAG through a permafrost depth profile. The most strikingly abundant microbial metabolism in this set of MAGs was aerobic. Vertical soil profiles are often depicted as aerobic zones transitioning neatly into anaerobic zones where terminal electron accepting processes and fermentation govern carbon decomposition (Mackelprang *et al.* 2016b). Yet soil systems, especially permafrost, are shown to be more complex. In permafrost aerobic microsites can exist within ice where low-to-freezing temperatures enable oxygen transfer into water (Jansson and Taş 2014). Via use of ^{14}C-acetate and ^{14}C-glucose microbial communities in permafrost from Canadian high Arctic were shown to be active at near ambient subzero temperatures (−5°C to −15°C) (Steven *et al.* 2008). More recently activity of both tundra and permafrost microbes at subzero temperatures were shown via stable isotope probing (Tuorto *et al.* 2014; Gadkari *et al.* 2019). Carbon degradation pathways identified in cold soils and permafrost show abundance and activity of various aerobic and anaerobic pathways at different locations. Genes involved in starch, lignocellulose, chitin, cellulose and trehalose degradation in both the active layer and permafrost (Yergeau *et al.* 2010; Mackelprang *et al.* 2011; Gadkari *et al.* 2019) and anaerobic metabolism was identified as a common microbial trait in permafrost metagenomes (Lipson *et al.* 2013; Hultman *et al.* 2015; Woodcroft *et al.* 2018). Our current knowledge of intact and thawing permafrost points to a large variance in metabolic potential and its utilization among different geographical locations (Mackelprang *et al.* 2016b). In Svalbard permafrost, we found aerobic processes as the key metabolism (Fig. 4) of recovered MAGs which showed previously unreported metabolic potential in permafrost. Besides genes involved SOM degradation (Fig. S3, Supporting Information), we found that in permafrost MAGs for aerobic processes dominating cellular metabolism. These results indicate that a substantial investment by permafrost MAGs in energy production is required to maintain reactions in order to survive at low temperatures. These results are also in concurrence with previous activity measurements from the same location where through a series of incubations Müller *et al.* (Müller *et al.* 2018) showed upon permafrost thaw up to four times higher CO_2 respiration rate were observed under aerobic than anaerobic conditions. Additionally, permafrost samples emitted similar quantities of CO_2 to active layer soils suggesting that Svalbard permafrost microbiome can stimulate its aerobic metabolism upon thaw. CH_4 is an important component of soil GHG fluxes in the Arctic which is shown to be released upon permafrost thaw as a result of significant changes in microbial populations and their interactions (Singleton *et al.* 2018; Woodcroft *et al.* 2018). In this study; however, we did not find any methanogenic MAGs or methane oxidation potential genes and anaerobic incubation experiments yielded no CH_4 production (Müller *et al.* 2018).

Arctic soils and permafrost are nitrogen limited where importance of nitrogen fixation for permafrost microbiome was highlighted by earlier metagenomics efforts (Yergeau *et al.* 2010; Mackelprang *et al.* 2011). It was hypothesized that the frozen conditions in permafrost sequester biologically available nitrogen, making nitrogen fixation necessary to contain metabolic activity. Hultman *et al.* (Hultman *et al.* 2015) showed that the permafrost microbiome was poised to assimilate nitrogen where genes encoding both *glutamine-* and *glutamate synthases* were transcribed and translated in permafrost. These pan-arctic observations were also paralleled in Svalbard active layer soils where Schostag et al (Schostag *et al.* 2015) detected high abundance of nitrogen-fixing bacteria via 16S rRNA gene sequencing. Svalbard permafrost MAGs showed similar trends to these previous findings where throughout the depth profile most abundant MAGs had *glutamine synthetase*, *glutamate synthase* and ammonium transporters to assimilate nitrogen. Earlier research showed that 450–550 μg/L ammonia could be found in Svalbard permafrost layers (Müller *et al.* 2018). In contrast, nitrogen fixation potential was limited, which collectively suggest nitrogen limitation as an important constraint to cellular activity in intact and thawed permafrost.

Sulfur metabolism have been shown to be widely present in permafrost microbes (Hansen *et al.* 2007; Vatsurina *et al.* 2008; Lipson *et al.* 2013; Chauhan *et al.* 2014). While sulfite reduction and sulfur oxidation were found in permafrost at different depths (Jansson and Taş 2014; Hultman *et al.* 2015), sulfate reduction rates were only high in bog samples while almost negligible in intact permafrost (Hultman *et al.* 2015). Current knowledge from metagenome data suggest that redox conditions become favorable for sulfate reduction after permafrost thaw. Svalbard MAGs provide a new perspective to sulfur metabolism in permafrost where abundant MAGs to contained genes involved both in assimilatory and dissimilatory sulfate reduction (Figs S7 and S8, Supporting Information). Genomic evidence suggests that *Gallionella* (one of the main sulfur cycle MAGs: maxbin2.012) are adapted to extremely low oxygen levels, it is possible that they are capable of growth at dissolved O_2 concentrations below the oxygen detection limits to occupy a narrow niche between O_2 and redox gradients (Emerson *et al.* 2013; Berg *et al.* 2019). We hypothesize that Svalbard MAGs retain flexibility in their sulfur metabolism in order to fully utilize limited resources propagated by ice and formation or microsites.

Genes involved in stress responses, resistance and resilience are shown to be crucial part of not only permafrost microorganisms but also psychrophiles in general (Ayala-Del-Río *et al.* 2010; Mykytczuk *et al.* 2013). Microbial survival in permafrost is challenging: proteins are less flexible and are prone to denaturation (Mykytczuk *et al.* 2013), cell membranes often susceptible to lose their fluidity (Ayala-Del-Río *et al.* 2010), water retention can be challenging and nutrient transport can be constrained. As a result, efficient anion and cation transporters is beneficial for cell survival. We observed an enrichment of potassium transport regulatory system in abundant permafrost MAGs (Fig. 4). The presence of potassium transporter protein in permafrost was also confirmed by a previous metaproteomics study (Hultman *et al.* 2015). As these transporters serve an important role in maintaining the intracellular pH, counteract osmotic stress and also required as cofactors for many enzymes. Finally, potassium is essential for the activity of many enzymes and protein complexes including the ribosome as well as for the regulation of gene expression. Their enrichment in MAGs shows high capability in regulating cellular functions and potential activity in frozen soils. Hultman et al. (Hultman *et al.* 2015) found high numbers of cold-shock proteins in permafrost. Though present in Svalbard MAGs cold-shock proteins were not highly abundant

in MAGs; instead cell fate and cycle control, redox response and chemotaxis regulatory systems were of high abundance. Transmembrane receptors are ubiquitously used by prokaryotes in environmental sensing (Bi, Jin and Sourjik 2018). As a result, it can be expected that cellular functions controlling these systems are retained and maybe enriched in permafrost. Surprisingly we did not identify spore forming potential as a key functional potential of Svalbard MAGs. This in line with the previous assessment that spores are not the best survival strategy for freezing conditions (Mondav et al. 2014). Besides environmental stressors, several antibiotic resistance genes acting against aminoglycosides and fluoroquinolones were among key functions shared among Svalbard permafrost MAGs. Antibiotic resistant bacteria were found both among the Arctic and Antarctic isolates (Mindlin and Petrova 2017) where about one third of the isolated permafrost strains were resistant to more than one antibiotic. Aminoglycosides were observed in ancient permafrost samples as well (Dcosta et al. 2011; Kashuba et al. 2017). Resistance against fluoroquinolones, which directly inhibit DNA synthesis, is a widespread microbial survival strategy (Rutgersson et al. 2014). Antibiotic resistance is an inherent property of permafrost microbiome however we are yet to understand the importance of these mechanisms on permafrost microbial diversity and biochemical cycles beyond their apparent role in survival.

Svalbard MAGs carry signatures of metabolic pathways that provide tight control of growth and resources. Almost all living cells sophisticatedly regulate their phosphate uptake that enables survival under phosphate-limiting conditions (Marzan and Shimizu 2011). In particular, regulation of phosphate may play an important role when nitrogen is also limiting. We found that metabolism involved in recycling and acquisition of ammonium was concomitant with strong representation phosphate regulation (i.e. starvation response and related transporters). Especially in phosphate depleted soils efficient phosphorus transporters are pivotal, as they allow microorganisms to compete for bioavailable phosphorus. Here, we hypothesize that microbial growth, survival and diverse metabolism including energy and central carbon cycling in Svalbard permafrost is facilitated by coupled regulation of ammonium, sulfur and phosphate metabolism. Even though we are not able to tie this hypothesis to availability of nutrients or gene expression that regulates these metabolisms, it is tempting to speculate that under freezing conditions Svalbard microbial populations regulate extra- and intra-cellular nutrient stoichiometry and availability closely to survive and utilize a wide range of carbon resources.

CONCLUSIONS

Predicting metabolic functionality and responses to changing environmental conditions from metagenomic data are among the greatest challenges in microbial ecology today (Myrold, Zeglin and Jansson 2014). Still metagenomics can be used to generate novel hypotheses about microbial metabolism and lifestyle. Permafrost in Svalbard is predicted to be more sensitive to increases in soil temperature and active layer thickness than the permafrost of extensive lowlands in Siberia, northern Canada and Alaska. In addition, Svalbard is an archipelago located near the northern most branches of the North Atlantic Current and the southern limit of the polar icepack. Even small variations in these important phenomena will induce rapid climatic variations with potential effects on the local Svalbard climate and permafrost. In this study, we provide an in-depth analysis of key permafrost microbial functions in Svalbard via a MAG-centric analysis. Svalbard MAGs were mostly aerobic and showed enrichment in functions regulating ammonium, sulfur and phosphate metabolism. Among different permafrost depths we repeatedly observed these metabolic pathways. Their perseverance point to their potential importance to life in permafrost. Our analysis also identified effective resource acquisition from the environment in potentially competitive and limited resource conditions as a key permafrost microbiome property. Collectively our results showed that Svalbard MAGs contain previously unreported metabolic functions in a permafrost environment.

DATA AND CODE AVAILABILITY

The shotgun sequence data and recovered MAGs were deposited in the European Nucleotide Archive (ENA) database under the study number PRJEB30872.

An instruction of refining MAGs and source code is available at https://github.com/yxxue/Recovery-and-refinement-of-MAGs-for-permafrost-metagenome.

A demonstration of comparative functional analysis by coverage in Svalbard metagenome and related source code are available at https://github.com/yxxue/Coverage-based-functional-analyisis-in-a-MAG-centric-view.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSEC online.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Research School in Bioinformatics, Biostatistics and Systems Biology (NORBIS) to Yaxin Xue. Funding for this work was provided to Neslihan Taş by the Office of Biological and Environmental Research in the DOE Office of Science – Early Career Research Program. This study is part of the project 'Microorganisms in the Arctic: major drivers of biogeochemical cycles and climate change' (RCN 227062), funded by the Norwegian Research Council (principal investigator [PI], Lise Øvreås). Lise Øvreås was awarded the Fulbright Arctic Chair 2012–2013 (Fulbright Foundation).

Conflict of interest. None declared.

REFERENCES

Albertsen M, Hugenholtz P, Skarshewski A *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 2013;31:533–8.

Alley RB, Barry R, Lemke P *et al.* Observations: changes in snow, ice and frozen ground. *Clim Chang 2007 Phys Sci Basis* 2007;4:337–83.

Alneberg J, Bjarnason BS, De Bruijn I *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014;11:1144–6.

Arkin AP, Cottingham RW, Henry CS *et al.* KBase: the United States department of energy systems biology knowledgebase. *Nat Biotechnol* 2018;36:566–9.

Ayala-Del-Río HL, Chain PS, Grzymski JJ *et al.* The genome sequence of psychrobacter arcticus 273–4, a psychroactive siberian permafrost bacterium, reveals mechanisms for

adaptation to low-temperature growth. *Appl Environ Microbiol* 2010;76:2304–12.

Bang-Andreasen T, Schostag M, Priemé A *et al.* Potential microbial contamination during sampling of permafrost soil assessed by tracers. *Sci Rep* 2017;7:43338.

Barquera B. The sodium pumping NADH:quinone oxidoreductase (Na+-NQR), a unique redox-driven ion pump. *J Bioenerg Biomembr* 2014;46:289–98.

Berg JS, Jézéquel D, Duverger A *et al.* Microbial diversity involved in iron and cryptic sulfur cycling in the ferruginous, low-sulfate waters of Lake Pavin. *PLoS One* 2019;14, doi:10.1371/journal.pone.0212787.

Bi S, Jin F, Sourjik V. Inverted signaling by bacterial chemotaxis receptors. *Nat Commun* 2018;9:2927.

Bowers RM, Clum A, Tice H *et al.* Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* 2015;16:856.

Bowers RM, Kyrpides NC, Stepanauskas R *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;35:725–31.

Chauhan A, Layton AC, Vishnivetskaya TA *et al.* Metagenomes from thawing low-soil-organic-carbon mineral cryosols and permafrost of the canadian high arctic. *Genome Announc* 2014;2, doi:10.1128/genomeA.01217-14.

Chen B, Yang Y, Liang X *et al.* Metagenomic profiles of antibiotic resistance genes (ARGs) between human impacted estuary and deep ocean sediments. *Environ Sci Technol* 2013;47:12753–60.

Chen IMA, Chu K, Palaniappan K *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* 2019;47:D666–77.

Cox G, Stogios PJ, Savchenko A *et al.* Structural and molecular basis for resistance to aminoglycoside antibiotics by the adenylyltransferase ANT(2″)-Ia. Bush K (ed.). *MBio* 2015;6, doi:10.1128/mBio.02180-14.

Danczak RE, Johnston MD, Kenah C *et al.* Capability for arsenic mobilization in groundwater is distributed across broad phylogenetic lineages. Pereira IAC (ed.). *PLoS One* 2019;14:e0221694.

Dcosta VM, King CE, Kalan L *et al.* Antibiotic resistance is ancient. *Nature* 2011;477:457–61.

Deng J, Gu Y, Zhang J *et al.* Shifts of tundra bacterial and archaeal communities along a permafrost thaw gradient in Alaska. *Mol Ecol* 2015;24:222–34.

Edwards A, Anesio AM, Rassner SM *et al.* Possible interactions between bacterial diversity, microbial activity and supraglacial hydrology of cryoconite holes in Svalbard. *ISME J* 2011;5:150–60.

Emerson D, Field EK, Chertkov O *et al.* Comparative genomics of freshwater Fe-oxidizing bacteria: Implications for physiology, ecology, and systematics. *Front Microbiol* 2013;4, doi:10.3389/fmicb.2013.00254.

Eren AM, Esen ÖC, Quince C *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 2015;3:e1319.

Gadkari PS, McGuinness LR, Männistö MK *et al.* Arctic tundra soil bacterial communities active at subzero temperatures detected by stable isotope probing. *FEMS Microbiol Ecol* 2019, doi:10.1093/femsec/fiz192.

Gittel A, Bárta J, Kohoutová I *et al.* Site- and horizon-specific patterns of microbial community structure and enzyme activities in permafrost-affected soils of Greenland. *Front Microbiol* 2014;5:541.

Gundlach J, Commichau FM, Stülke J. Perspective of ions and messengers: an intricate link between potassium, glutamate, and cyclic di-AMP. *Curr Genet* 2018;64:191–5.

Hansen AA, Herbert RA, Mikkelsen K *et al.* Viability, diversity and composition of the bacterial community in a high Arctic permafrost soil from Spitsbergen, Northern Norway. *Environ Microbiol* 2007;9:2870–84.

Hayes DJ, Kicklighter DW, McGuire AD *et al.* The impacts of recent permafrost thaw on land–atmosphere greenhouse gas exchange. *Environ Res Lett* 2014;9:045005.

Houghton JL, Foustoukos DI, Flynn TM *et al.* Thiosulfate oxidation by Thiomicrospira thermophila: metabolic flexibility in response to ambient geochemistry. *Environ Microbiol* 2016;18:3057–72.

Hultman J, Waldrop MP, Mackelprang R *et al.* Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* 2015;521:208–12.

Humlum O, Instanes A, Sollid JL. Permafrost in Svalbard: A review of research history, climatic background and engineering challenges. *Polar Res* 2003;22:191–215.

Hyatt D, Locascio PF, Hauser LJ *et al.* Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* 2012;28:2223–30.

Imelfort M, Parks D, Woodcroft BJ *et al.* GroopM: An automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2014;2014:e603.

Imperato V, Kowalkowski L, Portillo-Estrada M *et al.* Characterisation of the Carpinus betulus L. Phyllomicrobiome in Urban and Forest Areas. *Front Microbiol* 2019;10, doi:10.3389/fmicb.2019.01110.

Jansson JK, Taş N. The microbial ecology of permafrost. *Nat Rev Microbiol* 2014;12:414–25.

Kanehisa M, Goto S, Sato Y *et al.* KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;40:D109–14.

Kanehisa M, Goto S, Sato Y *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;42:D199–205.

Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* 2016;428:726–31.

Kang DD, Froula J, Egan R *et al.* MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 2015;2015:e1165.

Kashuba E, Dmitriev AA, Kamal SM *et al.* Ancient permafrost Staphylococci carry antibiotic resistance genes. *Microb Ecol Health Dis* 2017;28:1345574.

Kastovská K, Elster J, Stibal M *et al.* Microbial assemblages in soil microbial succession after glacial retreat in Svalbard (high arctic). *Microb Ecol* 2005;50:396–407.

Knight R, Vrbanac A, Taylor BC *et al.* Best practices for analysing microbiomes. *Nat Rev Microbiol* 2018;16:410–22.

Kokelj SV, Lantz TC, Tunnicliffe J *et al.* Climate-driven thaw of permafrost preserved glacial landscapes, northwestern Canada. *Geology* 2017;45:371–4.

Koyama A, Wallenstein MD, Simpson RT *et al.* Soil bacterial community composition altered by increased nutrient availability in Arctic tundra soils. *Front Microbiol* 2014;5, doi:10.3389/fmicb.2014.00516.

Kühlbrandt W. Structure and mechanisms of F-type ATP synthases. *Annu Rev Biochem* 2019;88:515–49.
Kultima JR, Coelho LP, Forslund K *et al.* MOCAT2: A metagenomic assembly, annotation and profiling framework. *Bioinformatics* 2016;32:2520–3.
Li D, Liu CM, Luo R *et al.* MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;31:1674–6.
Lipson DA, Haggerty JM, Srinivas A *et al.* Metagenomic insights into anaerobic metabolism along an arctic peat soil profile. Zhou Z (ed.). *PLoS One* 2013;8:e64659.
Mackelprang R, Burkert A, Haw M *et al.* Microbial survival strategies in ancient permafrost: insights from metagenomics. *ISME J* 2017;11:2305–18.
Mackelprang R, Saleska SR, Jacobsen CS *et al.* Permafrost meta-omics and climate change. *Annu Rev Earth Planet Sci* 2016a;44:439–62.
Mackelprang R, Saleska SR, Jacobsen CS *et al.* Permafrost meta-omics and climate change. *Annu Rev Earth Planet Sci* 2016b;44:439–62.
Mackelprang R, Waldrop MP, DeAngelis KM *et al.* Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 2011;480:368–71.
Marzan L, Shimizu K. Metabolic regulation of Escherichia coli and its phoB and phoR genes knockout mutants under phosphate and nitrogen limitations as well as at acidic condition. *Microb Cell Fact* 2011;10:39.
Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* 2016;7:11257.
Mindlin SZ, Petrova MA. On the Origin and Distribution of Antibiotic Resistance: Permafrost Bacteria Studies. *Mol Genet Microbiol Virol* 2017;32:169–79.
Mondav R, Woodcroft BJ, Kim E-H *et al.* Discovery of a novel methanogen prevalent in thawing permafrost. *Nat Commun* 2014;5:3212.
Müller O, Bang-Andreasen T, White RA *et al.* Disentangling the complexity of permafrost soil by using high resolution profiling of microbial community composition, key functions and respiration rates. *Environ Microbiol* 2018;20:4328–42.
Mykytczuk NCS, Foote SJ, Omelon CR *et al.* Bacterial growth at −15°C; molecular insights from the permafrost bacterium Planococcus halocryophilus Or1. *ISME J* 2013;7:1211–26.
Myrold DD, Zeglin LH, Jansson JK. The potential of metagenomic approaches for understanding soil microbial processes. *Soil Sci Soc Am J.* 2014;78:3–10.
Parks DH, Chuvochina M, Waite DW *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;36:996.
Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–55.
Quince C, Walker AW, Simpson JT *et al.* Corrigendum: Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017;35:1211.
Rutgersson C, Fick J, Marathe N *et al.* Fluoroquinolones and qnr genes in sediment, water, soil, and human fecal flora in an environment polluted by manufacturing discharges. *Environ Sci Technol* 2014;48:7825–32.
Schostag M, Stibal M, Jacobsen CS *et al.* Distinct summer and winter bacterial communities in the active layer of Svalbard permafrost revealed by DNA- and RNA-based analyses. *Front Microbiol* 2015;6, doi:10.3389/fmicb.2015.00399.
Schuur EAG, Bockheim J, Canadell JG *et al.* Vulnerability of permafrost carbon to climate change: implications for the global carbon cycle. *Bioscience* 2008;58:701–14.
Seitz KW, Dombrowski N, Eme L *et al.* Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat Commun* 2019;10:1822.
Sharon I, Morowitz MJ, Thomas BC *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* 2013;23:111–20.
Sieber CMK, Probst AJ, Sharrar A *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018;3:836–43.
Singleton CM, McCalley CK, Woodcroft BJ *et al.* Methanotrophy across a natural permafrost thaw environment. *ISME J* 2018;12:2544–58.
Steven B, Pollard WH, Greer CW *et al.* Microbial diversity and activity through a permafrost/ground ice core profile from the Canadian high Arctic. *Environ Microbiol* 2008;10:3388–403.
Tarnocai C, Canadell JG, Schuur EAG *et al.* Soil organic carbon pools in the northern circumpolar permafrost region. *Glob Biogeochem Cycles* 2009;23:2023–2034.
Tuorto SJ, Darias P, McGuinness LR *et al.* Bacterial genome replication at subzero temperatures in permafrost. *ISME J* 2014;8:139–49.
Tveit A, Schwacke R, Svenning MM *et al.* Organic carbon transformations in high-Arctic peat soils: Key functions and microorganisms. *ISME J* 2013;7:299–311.
Vatsurina A, Badrutdinova D, Schumann P *et al.* Desulfosporosinus hippei sp. nov., a mesophilic sulfate-reducing bacterium isolated from permafrost. *Int J Syst Evol Microbiol* 2008;58:1228–32.
Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* 2012;131:281–5.
Wilhelm RC, Niederberger TD, Greer C *et al.* Microbial diversity of active layer and permafrost in an acidic wetland from the Canadian high arctic. *Can J Microbiol* 2011;57:303–15.
Woodcroft BJ, Singleton CM, Boyd JA *et al.* Genome-centric view of carbon processing in thawing permafrost. *Nature* 2018;560:49–54.
Wu YW, Simmons BA, Singer SW. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;32:605–7.
Xue Y, Jonassen I, Øvreås L *et al.* Bacterial and archaeal metagenome-assembled genome sequences from Svalbard Permafrost. Stedman KM (ed). *Microbiol Resour Announc* 2019;8:e00516–19.
Yergeau E, Hogues H, Whyte LG *et al.* The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. *ISME J* 2010;4:1206–14.

Figure S1: The relative abundance of MAGs represented among main groups. Three dominant sample groups are presented: AL, BO, and PL.

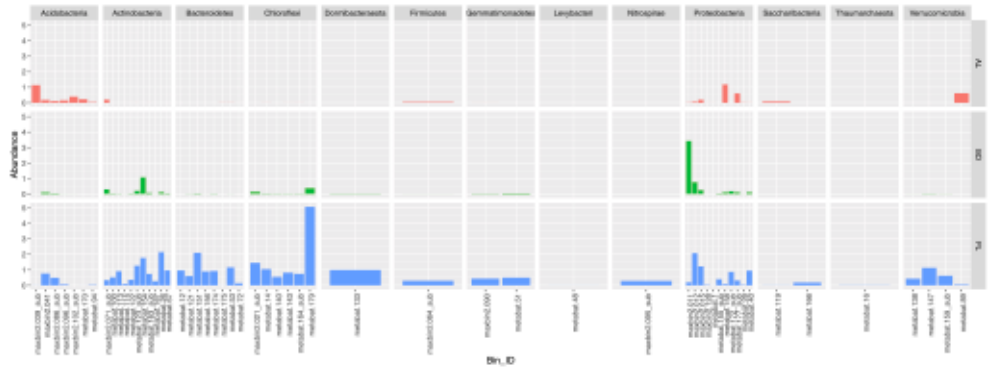


Figure S2: Heatmap shows the abundance of selected KEGG MO between KI (Increasing trend in PL_ALL and PL_SUB) and KD (Decreasing trend in PL_ALL and PL_SUB). KI and KD represent contigs with strong correlations (KI:>= 0.9, KD<= -0.9) between depth (cm: 110, 122, 135, 170) and normalized coverage in PL samples.

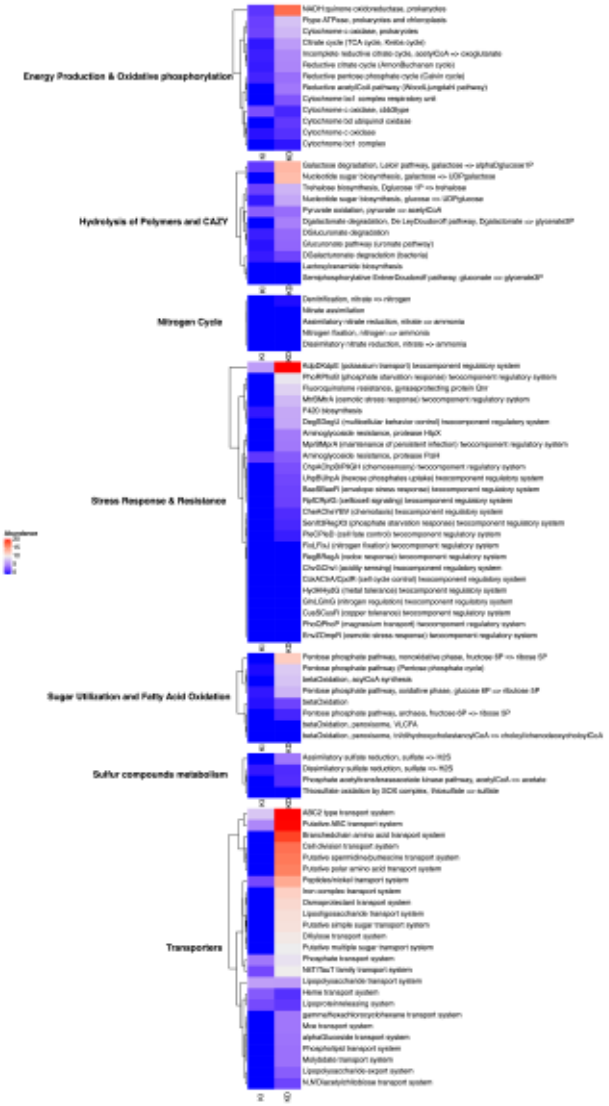


Figure S3: Alcohol Dehydrogenase at Svalbard MAGs among sample groups. The bar chart shows abundance distribution of KEGG Orthology (KO) related with Alcohol Dehydrogenase metabolism among groups in a MAG-centric view.

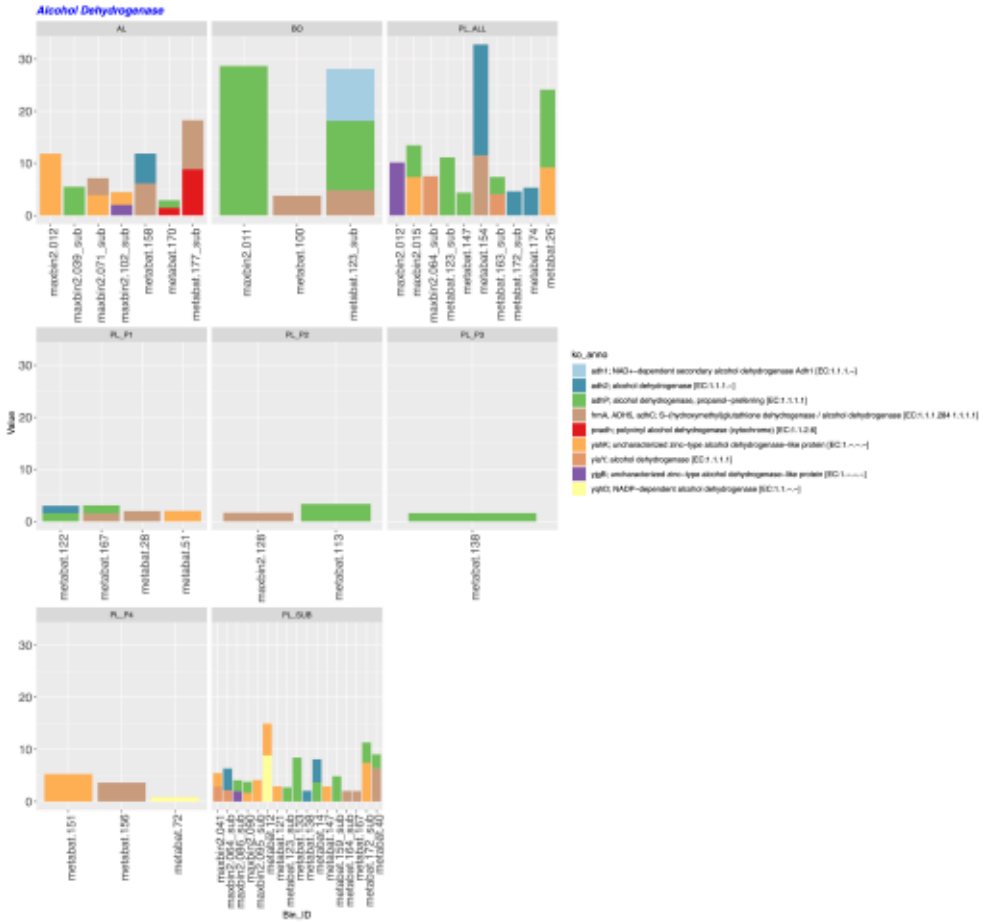


Figure S4: Cellulotic Enzymes at Svalbard MAGs among sample groups. The bar chart shows abundance distribution of KEGG Orthology (KO) related with Cellulotic Enzymes metabolism among groups in a MAG-centric view.

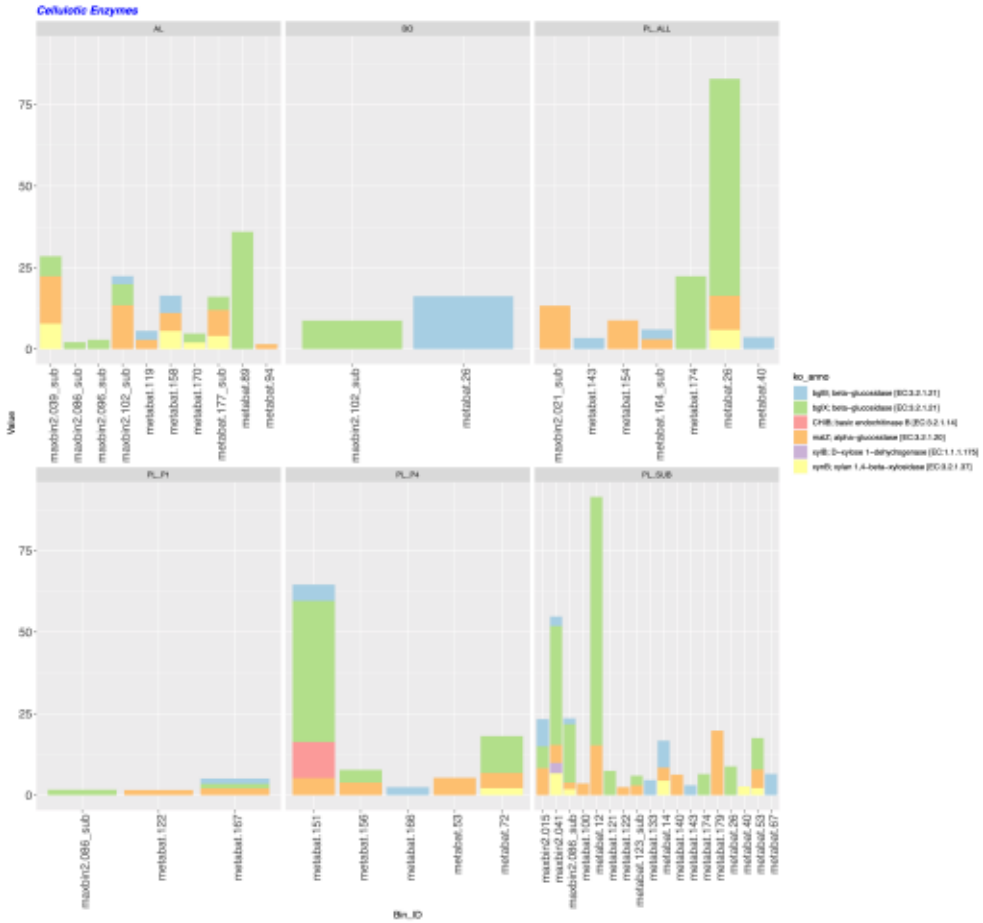


Figure S5: CO Dehydrogenase at Svalbard MAGs among sample groups. The bar chart shows abundance distribution of KEGG Orthology (KO) related with CO Dehydrogenase metabolism among groups in a MAG-centric view.

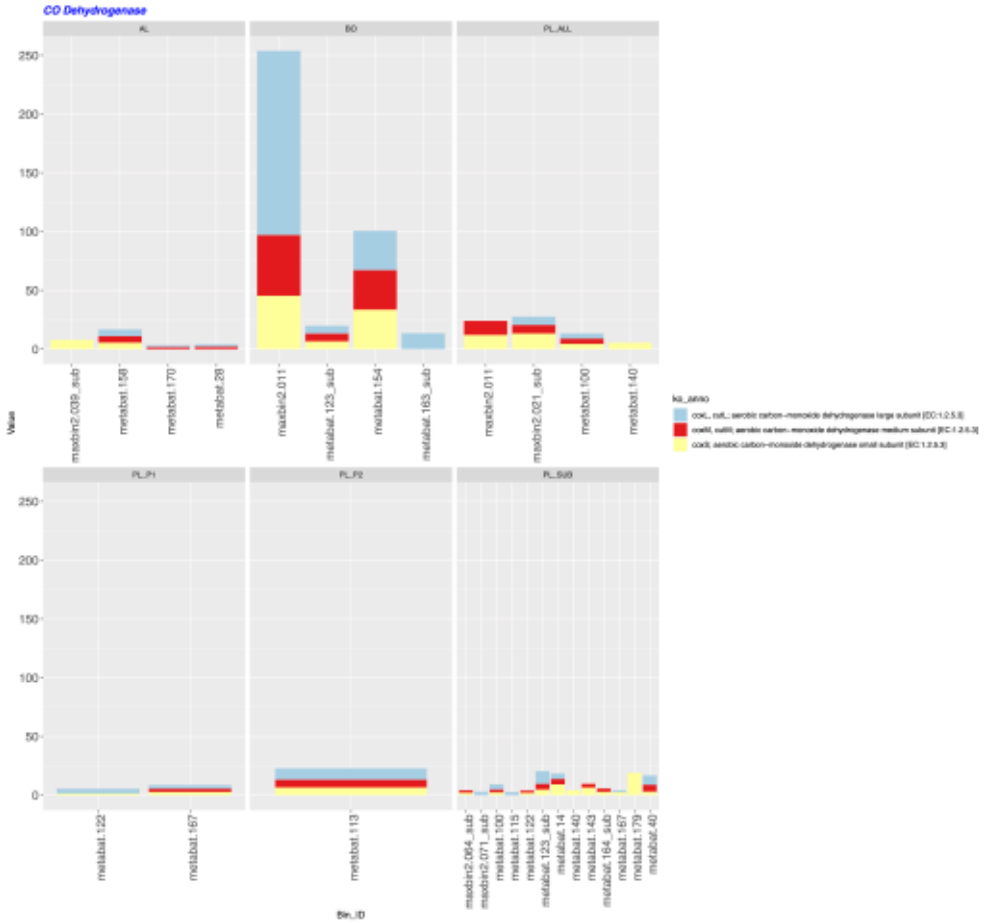


Figure S6: Nitrogen cycle at Svalbard MAGs among groups. The bar chart shows abundance distribution of KEGG Orthology (KO) related with Nitrogen cycle metabolism among groups in a MAG-centric view.

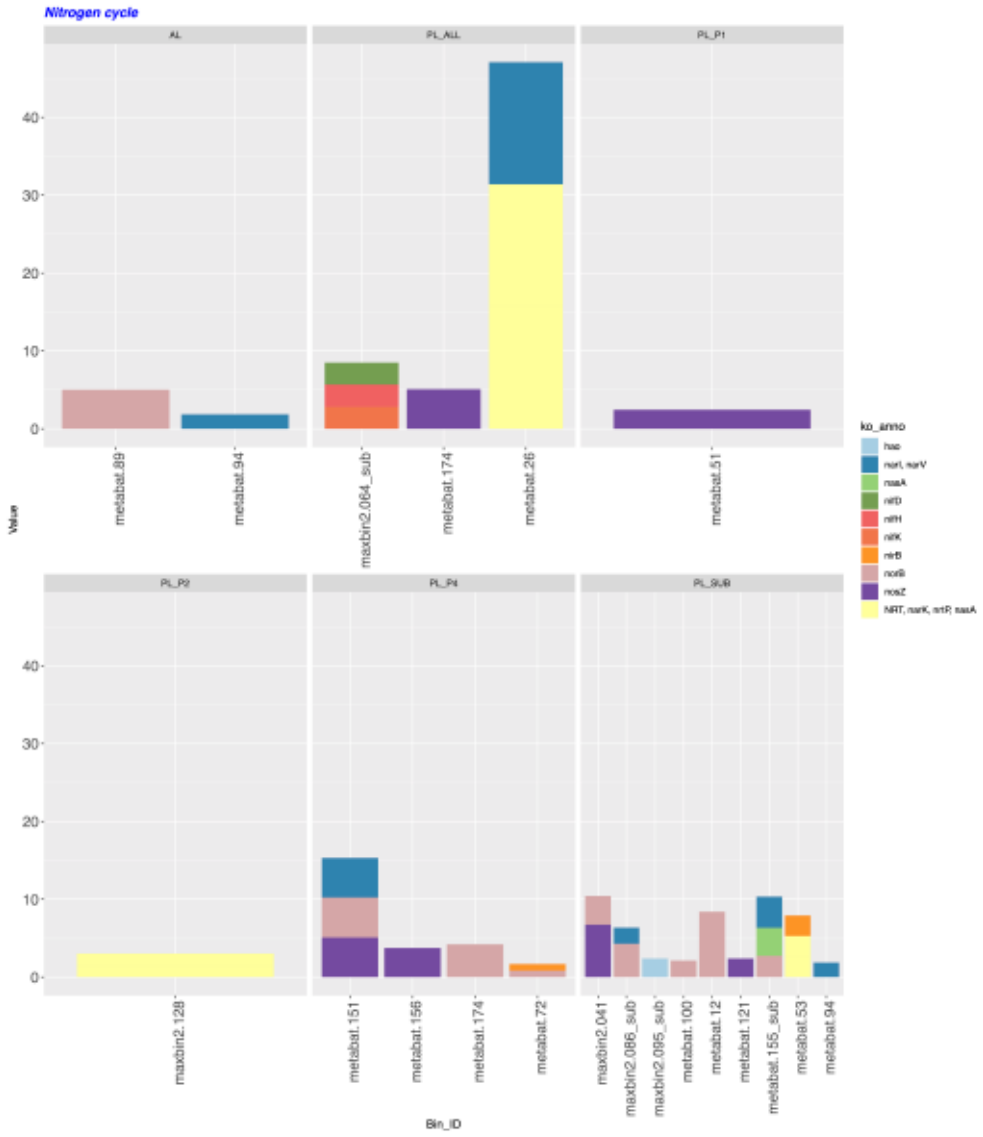


Figure S7: Dissimilatory sulfate reduction at Svalbard MAGs among groups. The bar chart shows abundance distribution of KEGG Orthology (KO) related with Dissimilatory sulfate reduction metabolism among groups in a MAG-centric view.

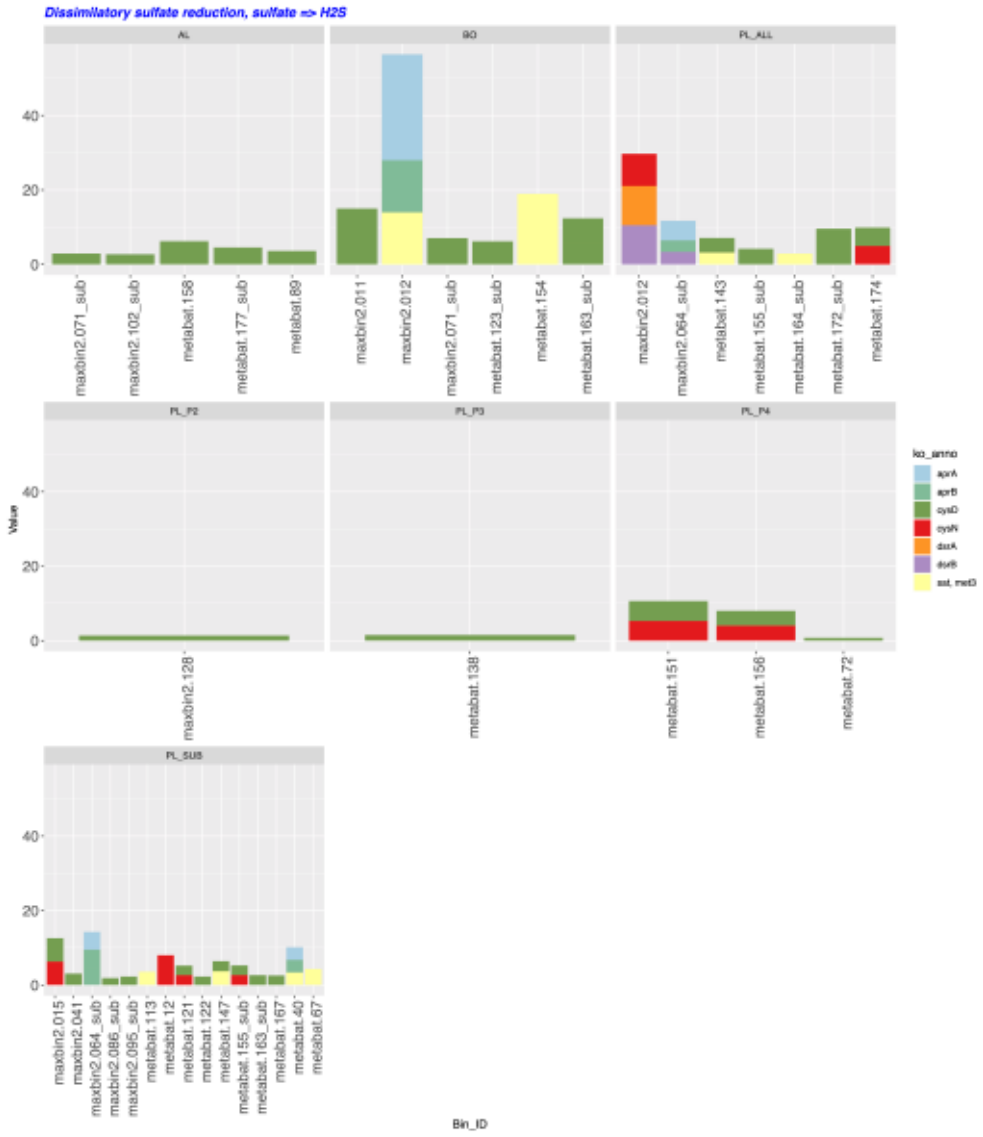


Figure S8: Assimilatory sulfate reduction at Svalbard MAGs among groups. The bar chart shows abundance distribution of KEGG Orthology (KO) related with Assimilatory sulfate reduction metabolism among groups in a MAG-centric view.

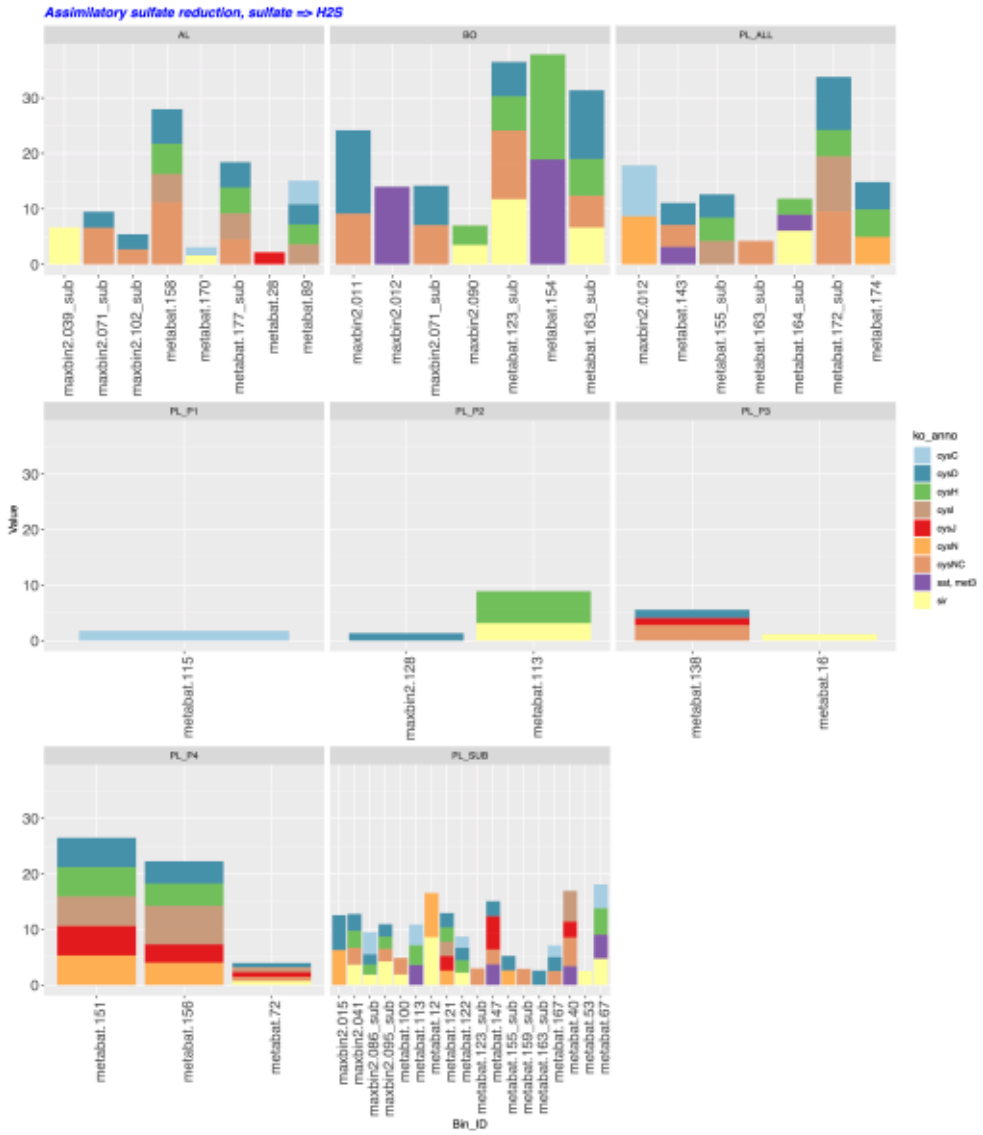


Table S1: Taxonomic classification of MAGs.

User Genome	GTDB Classification	FastANI Reference	FastANI Reference Radius	FastANI Taxonomy	FastANI ANI	FastANI Alignment Fraction	Closest Placement Reference	Closest Placement Taxonomy	Closest Placement Pla	Closest Placement Classification Method	Note	Other R AA Percent	RED Value
Acidobacteriota_MAG_metabin3.086	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__Acidobacteriales;f__Acidobacteriaceae;g__;s__									Placement	taxonomic novelty determined using RED	74.84	0.763675734
Acidobacteriota_MAG_metabin3.089	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__UBA7541;f__UBA7541;g__UBA7541;s__						GCA_000181195.1	d__Bacteria;p__Acidobacteriota;c__Aci	74.28	0.14 Placement	taxonomic classification fully defined by lca; GCA_00	88.16	0.803844755
Acidobacteriota_MAG_metabin3.041	d__Bacteria;p__Acidobacteriota;c__Acidobact GCA_002413085.1	95	d__Bacteria;p__Acido	97.33	0.69	GCA_002413085.1	d__Bacteria;p__Acidobacteriota;c__Aci	97.33	0.69 ANI Placement	topological placement and ANI have congru GCA_00	84.15	
Bacteroidetes_MAG_metabat.53	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Chitinophagaceae;g__Pelka-165;s__						GCA_000132765.1	d__Bacteria;p__Bacteroidota;c__Bacte	77.31	0.2 Placement	taxonomic classification fully defined by lca; GCA_00	87.5	0.830512038
Acidobacteriota_MAG_metabat.84	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__Bryobacteriales;f__Bryobacteraceae;g__BOG-224;s__						GCA_000136575.1	d__Bacteria;p__Acidobacteriota;c__Aci	77.77	0.29 Placement	taxonomic classification fully defined by lca; GCA_00	58.63	0.880035529
Verrucomicrobia_MAG_metabat.89	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Pedosphaerales;f__Pedosphaeraceae;g__UBA11358;s__						GCA_000133815.1	d__Bacteria;p__Verrucomicrobiota;c__\	81.47	0.44 Placement	taxonomic classification fully defined by lca; GCA_00	86.88	0.84663259
Bacteroidetes_MAG_metabat.72	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Prolixibacteraceae;g__UBA6934;s__						GCA_000429585.1	d__Bacteria;p__Bacteroidota;c__Bacte	93.19	0.81 Placement	taxonomic classification fully defined by lca; GCA_00	76.35	0.863765491
Gemmatimonadetes_MAG_metabat.51	d__Bacteria;p__Gemmatimonadota;c__Gemm GCA_002136875.1	95	d__Bacteria;p__Gemr	98.27	0.68	GCA_002136875.1	d__Bacteria;p__Gemmatimonadota;c__(	98.27	0.66 ANI Placement	topological placement and ANI have congru GCA_00	74.62	
Betaproteobacterium_MAG_metabat.40	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Pelsa-1006;g__PALSA-1003;s__						GCA_000150475.1	d__Bacteria;p__Proteobacteria;c__Gan	77.15	0.29 Placement	taxonomic novelty determined using RED	81.33	0.867566835
Actinobacterium_MAG_metabat.67	d__Bacteria;p__Dormibacteraeota;c__Dormibacteria;o__UBA6260;f__UBA6260;g__Pelka-879;s__						GCA_000150125.1	d__Bacteria;p__Dormibacteraeota;c__Dor	81.3	0.64 Placement	taxonomic novelty determined using RED	88.75	0.899030987
Actinobacterium_MAG_metabat.28	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Propionibacteriales;f__Propionibacteriaceae;g__Pen-668;s__									Placement	taxonomic novelty determined using RED	68.45	0.872394575
Bacteria_MAG_metabat.48	d__Bacteria;p__Patescibacteria;c__Microgenomatia;o__Levybacteriales;f__GWA1-39-11;g__;s__									Placement	taxonomic classification fully defined by topology	44.29	0.892163975
Bacteroidetes_MAG_metabat.114	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Chitinophagaceae;g__Pelka-165;s__						GCA_000500465.1	d__Bacteria;p__Bacteroidota;c__Bacte	80.58	0.44 Placement	taxonomic classification fully defined by lca; GCA_00	62.28	0.848911527
Bacteroidetes_MAG_metabat.115	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__f__;g__;s__									Placement	taxonomic novelty determined using RED	39.76	0.964855493
Acidobacteriota_MAG_metabat.170	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__Bryobacteriales;f__Bryobacteraceae;g__Pelka-85;s__									Placement	taxonomic classification fully defined by lca; GCA_00	45.85	0.941897297
Actinobacterium_MAG_metabat.167	d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__IMCC26256;f__PALSA-555;g__PALSA-555;s__						GCA_000140175.1	d__Bacteria;p__Actinobacteriota;c__Ac	76.38	0.57 Placement	taxonomic novelty determined using RED	80.12	0.877780017
Rhizobiales_MAG_metabat.172	d__Bacteria;p__Proteobacteria;c__Alphaprote GCA_002135455.1	95	d__Bacteria;p__Prote	97.34	0.67	GCA_002135455.1	d__Bacteria;p__Proteobacteria;c__Alph	97.34	0.67 ANI Placement	topological placement and ANI have congru GCA_00	83.03	
Solirubrobacterales_MAG_metabat.163	d__Bacteria;p__Actinobacteriota;c__Thermoleophilia;o__Solirubrobacterales;f__Solirubrobacteraceae;g__QHBN01;s__						GCA_002244205.1	d__Bacteria;p__Actinobacteriota;c__Th	77.6	0.34 Placement	taxonomic novelty determined using RED	88.12	0.854551218
Alphaproteobacteria_MAG_metabat.158	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__UBA1301;f__UBA1301;g__UBA6038;s__						GCA_001866865.1	d__Bacteria;p__Proteobacteria;c__Alph	78	0.38 Placement	taxonomic novelty determined using RED	84.11	0.866899819
Chloroflexi_MAG_metabat.184	d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__SH23-28-21_RBG-16-64-32;g__;s__									Placement	taxonomic novelty determined using RED	79.98	0.798955528
Saccharibacteria_MAG_metabat.188	d__Bacteria;p__Patescibacteria;c__Saccharimonadia;o__Saccharimonadales;f__Saccharimonadaceae;g__UBA1647;s__									Placement	taxonomic classification fully defined by lca; GCA_00	76.65	0.826771572
Bacteroidetes_MAG_metabat.156	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Chitinophagaceae;g__Pelka-165;s__						GCA_000132765.1	d__Bacteria;p__Bacteroidota;c__Bacte	76.75	0.13 Placement	taxonomic classification fully defined by lca; GCA_00	84.03	0.819626148
Chloroflexi_MAG_metabat.114	d__Bacteria;p__Dormibacteraeota;c__Dormibacteria;o__f__;g__;s__									Placement	taxonomic novelty determined using RED	91.41	0.558888742
Chloroflexi_MAG_metabat.143	d__Bacteria;p__Chloroflexota;A;c__Ellin6529;o__CSP1-4;f__CSP1-4;g__Pelka-1532;s__						GCA_000169726.1	d__Bacteria;p__Chloroflexota_A;c__Ell	77.8	0.26 Placement	taxonomic novelty determined using RED	84.29	0.885648058
Chloroflexi_MAG_metabat.149	d__Bacteria;p__Chloroflexota;A;c__Ellin6529;o__QHBD01;f__QHBD01;g__;s__						GCA_000243665.1	d__Bacteria;p__Chloroflexota_A;c__Ell	78.56	0.37 Placement	taxonomic novelty determined using RED	82.34	0.84322864
Verrucomicrobia_MAG_metabat.139	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__UBA1240;g__WA40;s__									Placement	taxonomic novelty determined using RED	76.33	0.839525487
Bacteroidetes_MAG_metabat.131	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Chitinophagaceae;g__Pelka-165;s__						GCA_000132765.1	d__Bacteria;p__Bacteroidota;c__Bacte	76.15	0.13 Placement	taxonomic classification fully defined by lca; GCA_00	75.58	0.81864729
Bacteroidetes_MAG_metabat.151	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Chitinophagaceae;g__Pelka-165;s__						GCA_000132765.1	d__Bacteria;p__Bacteroidota;c__Bacte	77.27	0.13 Placement	taxonomic classification fully defined by lca; GCA_00	83.41	0.812883095
Verrucomicrobia_MAG_metabat.147	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__UBA1240;g__;s__									Placement	taxonomic classification fully defined by topology	81.19	0.858381584
Actinobacterium_MAG_metabat.154	d__Bacteria;p__Actinobacteriota;c__Thermole GCA_002413085.1	95	d__Bacteria;p__Actino	99.09	0.65	GCA_002413085.1	d__Bacteria;p__Actinobacteriota;c__Th	99.09	0.65 ANI Placement	topological placement and ANI have congruent spec	77.58	
Actinobacterium_MAG_metabat.123	d__Bacteria;p__Actinobacteriota;c__Thermoleophilia;o__Solirubrobacterales;f__Solirubrobacteraceae;g__Pelka-485;s__									Placement	taxonomic novelty determined using RED	80.47	0.893693141
Bacteria_MAG_metabat.133	d__Bacteria;p__Dormibacteraeota;c__Dormibacteria;o__UBA6260;f__UBA6260;g__;s__									Placement	taxonomic novelty determined using RED	89.18	0.851485088
Thaumarchaeota_MAG_metabat.16	d__Archaea;p__Crenarchaeota;c__Nitrososphi GCA_002501855.1	95	d__Archaea;p__Crena	99.16	0.79	GCA_002501855.1	d__Archaea;p__Crenarchaeota;c__Nitro	99.16	0.79 ANI Placement	topological placement and ANI have congru GCA_00	51.58	
Verrucomicrobia_MAG_metabat.159	d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Chthoniobacterales;f__UBA1240;g__;s__						GCA_000169975.1	d__Bacteria;p__Verrucomicrobiota;c__\	77.18	0.27 Placement	taxonomic novelty determined using RED	73.95	0.836783788
Betaproteobacteria_MAG_metabat.155	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Gallionellaceae;g__Gallionella;s__						GCA_002429965.1	d__Bacteria;p__Proteobacteria;c__Gan	93.74	0.63 Placement	taxonomic novelty determined using RED	52.68	0.888031791
Bacteroidetes_MAG_metabat.13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__FEN-979;g__;s__									Placement	taxonomic novelty determined using RED	76.58	0.79271245
Actinobacterium_MAG_metabat.122	d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__IMCC26256;f__PALSA-555;g__;s__						GCA_000140175.1	d__Bacteria;p__Actinobacteriota;c__Ac	76.64	0.27 Placement	taxonomic novelty determined using RED	68.89	0.762540351
Actinobacterium_MAG_metabat.113	d__Bacteria;p__Actinobacteriota;c__UBA4738;o__UBA4738;f__HRB-N12;g__;s__						GCA_002996915.1	d__Bacteria;p__Actinobacteriota;c__UB	76.6	0.22 Placement	taxonomic novelty determined using RED	84.27	0.775008554
Solirubrobacterales_MAG_metabat.169	d__Bacteria;p__Actinobacteriota;c__Thermoleophilia;o__Solirubrobacterales;f__Solirubrobacteraceae;g__Pelka-485;s__						GCA_000137375.1	d__Bacteria;p__Actinobacteriota;c__Th	76.63	0.16 Placement	taxonomic classification fully defined by lca; GCA_00	84.33	0.802185978
Actinobacterium_MAG_metabat.115	d__Bacteria;p__Actinobacteriota;c__UBA4738;o__UBA4738;f__UBA4738;g__;s__									Placement	taxonomic classification fully defined by lca; GCA_00	88.17	0.77123829
Alphaproteobacteria_MAG_maxbin2.128	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Ambami ellaceae;g__PALSA-937;s__						GCA_000151375.1	d__Bacteria;p__Proteobacteria;c__Alph	85.69	0.77 Placement	taxonomic novelty determined using RED	74.44	0.888158258
Acidobacteria_MAG_maxbin2.102	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__Acidobacteriales;f__Acidobacteriaceae;g__KBS-60;s__						GCF_000381585.1	d__Bacteria;p__Acidobacteriota;c__Aci	78.73	0.47 Placement	taxonomic novelty determined using RED	73.08	0.847662684
Saccharibacteria_MAG_metabat.119	d__Bacteria;p__Patescibacteria;c__Saccharimonadia;o__Saccharimonadales;f__UBA1312;g__;s__									Placement	taxonomic classification fully defined by topology	71.98	0.837934596
Nitrosospira_MAG_metabat.1	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Nitrosomonadaceae;g__Nitrosospira;s__						GCF_900103495.1	d__Bacteria;p__Proteobacteria;c__Gan	76.31	0.56 Placement	taxonomic novelty determined using RED	GCF_90	45.3	0.916867575
Acidobacteria_MAG_metabin3.096	d__Bacteria;p__Acidobacteriota;c__Acidobacteriae;o__UBA7541;f__;g__;s__									Placement	taxonomic novelty determined using RED	39.23	0.885753847
Nitrospinae_MAG_maxbin2.095	d__Bacteria;p__Nitrospinota;c__Nitrospinia;o__Nitrospinales;f__Nitrospinaceae;g__Pelka-1315;s__						GCA_002135435.1	d__Bacteria;p__Nitrospinota;c__Nitrospi	92.69	0.6 Placement	taxonomic classification fully defined by lca; GCA_00	49.98	0.899282482
Gemmatimonadetes_MAG_maxbin2.096	d__Bacteria;p__Gemmatimonadota;c__Gemmatimonadetes;o__Gemmatimonadales;f__GWC2-71-9;g__;s__									Placement	taxonomic classification fully defined by topology	60.83	0.791186217
Actinobacterium_MAG_maxbin2.071	d__Bacteria;p__Actinobacteriota;c__Thermoleophilia;o__Solirubrobacterales;f__Solirubrobacteraceae;g__Pelka-74;s__						GCA_000136675.1	d__Bacteria;p__Actinobacteriota;c__Th	80.01	0.59 Placement	taxonomic classification fully defined by lca; GCA_00	74.03	0.840844212
Proteobacteria_MAG_maxbin2.015	d__Bacteria;p__Patescibacteria;c__Saccharimonadia;o__Saccharimonadales;f__Saccharimonadaceae;g__UBA1647;s__						GCA_002429075.1	d__Bacteria;p__Patescibacteria;c__Sac	94.9	0.8 Placement	taxonomic classification fully defined by lca; GCA_00	75.95	0.857277217
Chloroflexi_MAG_maxbin2.021	d__Bacteria;p__Chloroflexota;A;c__Ellin6529;o__CSP1-4;f__CSP1-4;g__;s__									Placement	taxonomic classification fully defined by lca; GCA_00	72.06	0.830007363
Rhizobiales_MAG_maxbin2.017	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Pseudolabrys;s__						GCA_000152455.1	d__Bacteria;p__Proteobacteria;c__Alph	93.32	0.74 Placement	taxonomic classification fully defined by lca; GCA_00	38.29	0.899503388
Firmicutes_MAG_maxbin2.064	d__Bacteria;p__Firmicutes_D;c__Desulfitobacteriia;o__Desulfitobacteriales;f__Desulfitobacteriaceae;g__Desulfosporosinus;s__						GCA_002413075.1	d__Bacteria;p__Firmicutes_D;c__Desulf	97.2	0.63 Placement	taxonomic classification fully defined by lca; GCA_00	41.09	0.899334862
Proteobacteria_MAG_maxbin2.012	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Gallionellaceae;g__Gallionella;s__						GCA_002429965.1	d__Bacteria;p__Proteobacteria;c__Gan	97.06	0.5 Placement	taxonomic classification fully defined by lca; GCA_00	61.17	0.898364578
Proteobacteria_MAG_metabat.177	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Steroidobacterales;f__Steroidobacteraceae;g__Bsp-1198;s__						GCA_000165495.1	d__Bacteria;p__Proteobacteria;c__Gan	78.58	0.37 Placement	taxonomic classification fully defined by lca; GCA_00	86.66	0.869129081
Betaproteobacteria_MAS_metabat.29	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Rhizobacter;s__						GCA_000152565.1	d__Bacteria;p__Proteobacteria;c__Gan	80.17	0.56 Placement	taxonomic classification fully defined by lca; GCF_00	57.5	0.856396252
Chloroflexi_MAS_metabat.179	d__Bacteria;p__Chloroflexota;A;c__Ellin6529;o__CSP1-4;f__CSP1-4;g__UBA6189;s__									Placement	taxonomic classification fully defined by lca; GCA_00	89.94	0.899675718

Table S2: 16S rRNA comparative analysis of Svalbard MAGs with recent stable isotope probing studies.

SvalbardMAG_aln_Gadkari

qseqid	Svalbard_MAG	Svalbard_Taxa	sseqid	SIP_Gadkari_2020_Taxa	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore
k127_268166	reasbin2.015	Proteobacteria	SeqRun8_OTU135	Saccharibacteria	76.471	340	54	23	135	460	98	425	6.45E-93	281
k127_458188	reasbin2.015	Proteobacteria	SeqRun8_OTU69	Saccharibacteria	85.613	424	57	4	717	1138	425	4	1.94E-127	441
k127_171559	reasbin2.041	Acidobacteria	SeqRun8_OTU37	Saccharibacteria	77.108	352	65	10	845	1170	98	424	3.22E-49	181
k127_3181191	reatabat.14	Chloroflexi	SeqRun8_OTU16	Verrucomicrobia	76.899	316	66	7	6622	6933	130	442	1.65E-45	171

Tourto_aln_SvalbardMAG

qseqid	SIP_Tourto_2014_Taxa	sseqid	Svalbard_MAG	Svalbard_Taxa	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore
DB59814.1	Gemmatimonadetes	k127_171559	reasbin2.041	Acidobacteria	87.568	915	88	18	1	900	173	1295	0	1046
DB59815.1	Gemmatimonadetes	k127_171559	reasbin2.041	Acidobacteria	86.652	929	91	20	1	902	173	1295	0	998
DB59816.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	84.834	1107	127	29	1	1892	6154	7244	0	1085
DB59817.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	79.78	1093	172	25	1	1879	6154	7211	0	740
DB59818.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	82.955	854	120	11	238	1077	6383	7211	0	725
DB59819.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	81.101	1106	167	33	1	1879	6154	7244	0	866
DB59820.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	80.888	1104	173	31	1	1879	6154	7244	0	835
DB59821.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	81.112	863	129	18	238	1080	6383	7221	0	660
DB59822.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	81.208	894	128	23	238	1098	6378	7244	0	684
DB59823.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	80.848	875	136	19	217	1080	6378	7227	1.92E-179	630
DB59824.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	81.299	859	129	20	217	1097	6378	7244	0	680
DB59825.1	Actinobacteria	k127_3301391	reatabat.14	Chloroflexi	81.187	859	130	20	235	1095	6378	7244	0	684
DB59826.1	Chloroflexi	k127_1845442	reatabat.358	Alphaproteobacteria	74.721	1078	228	34	1	1856	57018	58089	3.40E-121	440
DB59827.1	Chloroflexi	k127_3301391	reatabat.14	Chloroflexi	81.671	1112	166	29	1	1898	6154	7244	0	876
DB59828.1	Chloroflexi	k127_3301391	reatabat.14	Chloroflexi	81.579	1102	168	27	1	1880	6154	7243	0	880
DB59829.1	Chloroflexi	k127_3301391	reatabat.14	Chloroflexi	82.45	1094	178	19	1	1875	6154	7244	0	937
DB59830.1	Chloroflexi	k127_3301391	reatabat.14	Chloroflexi	82.835	1100	159	25	1	1879	6154	7244	0	957
DB59831.1	Chloroflexi	k127_3301391	reatabat.14	Chloroflexi	84.407	877	113	19	286	1100	6378	7244	0	892
DB59832.1	Chloroflexi	k127_3301391	reatabat.14	Chloroflexi	82.849	1103	168	33	1	1877	6154	7244	0	905
DB59833.1	Chloroflexi	k127_3301391	reatabat.14	Chloroflexi	82.871	1104	158	31	1	1879	6154	7244	0	963
DB59834.1	Chloroflexi	k127_3301391	reatabat.14	Chloroflexi	82.486	1102	157	30	1	1877	6154	7244	0	933
DB59835.1	Chloroflexi	k127_3301391	reatabat.14	Chloroflexi	82.996	1102	158	32	1	1877	6154	7244	0	928
DB59836.1	Acidobacteria	k127_3301391	reatabat.14	Chloroflexi	82.634	857	111	24	280	1080	6378	7282	0	722
DB59837.1	Acidobacteria	k127_1355597	reasbin2.815	Proteobacteria	82.864	762	100	18	346	1123	3370	3619	0	671
DB59838.1	Acidobacteria	k127_1355597	reasbin2.815	Proteobacteria	82.182	740	99	17	345	1080	3370	3680	1.17E-171	604
DB59839.1	Acidobacteria	k127_1845442	reatabat.358	Alphaproteobacteria	80.332	1083	166	36	1	1867	57018	58089	0	779
DB59840.1	Acidobacteria	k127_1355597	reasbin2.815	Proteobacteria	86.243	756	84	20	316	1060	3370	3624	0	802
DB59841.1	Proteobacteria	k127_3301391	reatabat.14	Chloroflexi	78.838	1143	172	41	1	1125	6154	7244	0	706
DB59842.1	Proteobacteria	k127_1845442	reatabat.358	Alphaproteobacteria	84.725	1078	143	19	1	1871	57018	58089	0	1007
DB59843.1	Proteobacteria	k127_1845442	reatabat.358	Alphaproteobacteria	87.492	1052	128	8	1	1852	57018	58089	0	1208
DB59844.1	Proteobacteria	k127_1845442	reatabat.358	Alphaproteobacteria	85.3	1068	129	25	1	1856	57018	58089	0	1077
DB59845.1	Proteobacteria	k127_3301391	reatabat.14	Chloroflexi	79.854	1097	188	26	1	1875	6154	7238	0	771
DB59846.1	Proteobacteria	k127_1845442	reatabat.358	Alphaproteobacteria	81.895	870	152	12	234	1059	57222	58089	0	708
DB59847.1	Proteobacteria	k127_3301391	reatabat.14	Chloroflexi	77.571	1118	191	43	1	1109	6154	7244	9.15E-178	635
DB59848.1	Proteobacteria	k127_1845442	reatabat.358	Alphaproteobacteria	82.828	891	101	29	226	1104	57219	58089	0	750
DB59849.1	Proteobacteria	k127_1845442	reatabat.358	Alphaproteobacteria	80.814	860	125	18	227	1080	57219	58044	0	638
DB59850.1	Proteobacteria	k127_1845442	reatabat.358	Alphaproteobacteria	82.15	865	131	17	235	1109	57218	58089	0	676
DB59851.1	Proteobacteria	k127_1845442	reatabat.358	Alphaproteobacteria	81.297	859	122	18	235	1080	57218	58083	0	658

Table S3: Selected KEGG Modules (MO) and their corresponding pathways.

Module_ID	Module_Anno	Pathway
M00009	Citrate cycle (TCA cycle, Krebs cycle)	Energy Production & Oxidative phosphorylation
M00152	Cytochrome bc1 complex	Energy Production & Oxidative phosphorylation
M00151	Cytochrome bc1 complex respiratory unit	Energy Production & Oxidative phosphorylation
M00153	Cytochrome bd ubiquinol oxidase	Energy Production & Oxidative phosphorylation
M00154	Cytochrome c oxidase	Energy Production & Oxidative phosphorylation
M00156	Cytochrome c oxidase, cbb3-type	Energy Production & Oxidative phosphorylation
M00155	Cytochrome c oxidase, prokaryotes	Energy Production & Oxidative phosphorylation
M00157	F-type ATPase, prokaryotes and chloroplasts	Energy Production & Oxidative phosphorylation
M00620	Incomplete reductive citrate cycle, acetyl-CoA => oxoglutarate	Energy Production & Oxidative phosphorylation
M00144	NAD(P)-quinone oxidoreductase, prokaryotes	Energy Production & Oxidative phosphorylation
M00377	Reductive acetyl-CoA pathway (Wood-Ljungdahl pathway)	Energy Production & Oxidative phosphorylation
M00173	Reductive citrate cycle (Arnon-Buchanan cycle)	Energy Production & Oxidative phosphorylation
M00165	Reductive pentose phosphate cycle (Calvin cycle)	Energy Production & Oxidative phosphorylation
M00552	D-galactonate degradation, De Ley-Doudoroff pathway, D-galactonate => glycerate-3P	Hydrolysis of Polymers and CAZY
M00631	D-Galacturonate degradation (bacteria)	Hydrolysis of Polymers and CAZY
M00061	D-Glucuronate degradation	Hydrolysis of Polymers and CAZY
M00632	Galactose degradation, Leloir pathway, galactose => alpha-D-glucose-1P	Hydrolysis of Polymers and CAZY
M00014	Glucuronate pathway (uronate pathway)	Hydrolysis of Polymers and CAZY
M00066	Lactosylceramide biosynthesis	Hydrolysis of Polymers and CAZY
M00554	Nucleotide sugar biosynthesis, galactose => UDP-galactose	Hydrolysis of Polymers and CAZY
M00549	Nucleotide sugar biosynthesis, glucose => UDP-glucose	Hydrolysis of Polymers and CAZY
M00307	Pyruvate oxidation, pyruvate => acetyl-CoA	Hydrolysis of Polymers and CAZY
M00633	Semi-phosphorylative Entner-Doudoroff pathway, gluconate => glycerate-3P	Hydrolysis of Polymers and CAZY
M00565	Trehalose biosynthesis, D-glucose 1P => trehalose	Hydrolysis of Polymers and CAZY
M00531	Assimilatory nitrate reduction, nitrate => ammonia	Nitrogen Cycle
M00529	Denitrification, nitrate => nitrogen	Nitrogen Cycle
M00530	Dissimilatory nitrate reduction, nitrate => ammonia	Nitrogen Cycle
M00615	Nitrate assimilation	Nitrogen Cycle
M00175	Nitrogen fixation, nitrogen => ammonia	Nitrogen Cycle
M00049	Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	Nucleic Acid Metabolism
M00260	DNA polymerase III complex, bacteria	Nucleic Acid Metabolism
M00050	Guanine ribonucleotide biosynthesis IMP => GDP,GTP	Nucleic Acid Metabolism
M00048	Inosine monophosphate biosynthesis, PRPP + glutamine => IMP	Nucleic Acid Metabolism
M00005	PRPP biosynthesis, ribose 5P => PRPP	Nucleic Acid Metabolism
M00546	Purine degradation, xanthine => urea	Nucleic Acid Metabolism
M00046	Pyrimidine degradation, uracil => beta-alanine, thymine => 3-aminoisobutanoate	Nucleic Acid Metabolism
M00053	Pyrimidine deoxyribonucleotide biosynthesis, CDP/CTP => dCDP/dCTP,dTDP/dTTP	Nucleic Acid Metabolism
M00052	Pyrimidine ribonucleotide biosynthesis, UMP => UDP/UTP,CDP/CTP	Nucleic Acid Metabolism
M00183	RNA polymerase, bacteria	Nucleic Acid Metabolism

M00051	Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP	Nucleic Acid Metabolism
M00742	Aminoglycoside resistance, protease FtsH	Stress Response & Resistance
M00743	Aminoglycoside resistance, protease HtpX	Stress Response & Resistance
M00450	BaeS-BaeR (envelope stress response) two-component regulatory system	Stress Response & Resistance
M00512	CckA-CtrA/CpdR (cell cycle control) two-component regulatory system	Stress Response & Resistance
M00506	CheA-CheYBV (chemotaxis) two-component regulatory system	Stress Response & Resistance
M00507	ChpA-ChpB/PilGH (chemosensory) two-component regulatory system	Stress Response & Resistance
M00520	ChvG-ChvI (acidity sensing) two-component regulatory system	Stress Response & Resistance
M00452	CusS-CusR (copper tolerance) two-component regulatory system	Stress Response & Resistance
M00478	DegS-DegU (multicellular behavior control) two-component regulatory system	Stress Response & Resistance
M00445	EnvZ-OmpR (osmotic stress response) two-component regulatory system	Stress Response & Resistance
M00378	F420 biosynthesis	Stress Response & Resistance
M00524	FixL-FixJ (nitrogen fixation) two-component regulatory system	Stress Response & Resistance
M00729	Fluoroquinolone resistance, gyrase-protecting protein Qnr	Stress Response & Resistance
M00497	GlnL-GlnG (nitrogen regulation) two-component regulatory system	Stress Response & Resistance
M00499	HydH-HydG (metal tolerance) two-component regulatory system	Stress Response & Resistance
M00454	KdpD-KdpE (potassium transport) two-component regulatory system	Stress Response & Resistance
M00460	MprB-MprA (maintenance of persistent infection) two-component regulatory system	Stress Response & Resistance
M00461	MtrB-MtrA (osmotic stress response) two-component regulatory system	Stress Response & Resistance
M00444	PhoQ-PhoP (magnesium transport) two-component regulatory system	Stress Response & Resistance
M00434	PhoR-PhoB (phosphate starvation response) two-component regulatory system	Stress Response & Resistance
M00511	PleC-PleD (cell fate control) two-component regulatory system	Stress Response & Resistance
M00523	RegB-RegA (redox response) two-component regulatory system	Stress Response & Resistance
M00517	RpfC-RpfG (cell-to-cell signaling) two-component regulatory system	Stress Response & Resistance
M00443	SenX3-RegX3 (phosphate starvation response) two-component regulatory system	Stress Response & Resistance
M00473	UhpB-UhpA (hexose phosphates uptake) two-component regulatory system	Stress Response & Resistance
M00087	beta-Oxidation	Sugar Utilization and Fatty Acid Oxidation
M00086	beta-Oxidation, acyl-CoA synthesis	Sugar Utilization and Fatty Acid Oxidation
M00862	beta-Oxidation, peroxisome, tri/dihydroxycholestanoyl-CoA => choloyl/chenodeoxycholoyl-CoA	Sugar Utilization and Fatty Acid Oxidation
M00861	beta-Oxidation, peroxisome, VLCFA	Sugar Utilization and Fatty Acid Oxidation
M00004	Pentose phosphate pathway (Pentose phosphate cycle)	Sugar Utilization and Fatty Acid Oxidation
M00580	Pentose phosphate pathway, archaea, fructose 6P => ribose 5P	Sugar Utilization and Fatty Acid Oxidation
M00007	Pentose phosphate pathway, non-oxidative phase, fructose 6P => ribose 5P	Sugar Utilization and Fatty Acid Oxidation
M00006	Pentose phosphate pathway, oxidative phase, glucose 6P => ribulose 5P	Sugar Utilization and Fatty Acid Oxidation
M00176	Assimilatory sulfate reduction, sulfate => H2S	Sulfur compounds metabolism
M00596	Dissimilatory sulfate reduction, sulfate => H2S	Sulfur compounds metabolism
M00579	Phosphate acetyltransferase-acetate kinase pathway, acetyl-CoA => acetate	Sulfur compounds metabolism
M00595	Thiosulfate oxidation by SOX complex, thiosulfate => sulfate	Sulfur compounds metabolism
M00254	ABC-2 type transport system	Transporters
M00201	alpha-Glucoside transport system	Transporters
M00237	Branched-chain amino acid transport system	Transporters
M00256	Cell division transport system	Transporters
M00215	D-Xylose transport system	Transporters
M00669	gamma-Hexachlorocyclohexane transport system	Transporters
M00259	Heme transport system	Transporters
M00240	Iron complex transport system	Transporters
M00252	Lipooligosaccharide transport system	Transporters
M00320	Lipopolysaccharide export system	Transporters
M00250	Lipopolysaccharide transport system	Transporters
M00255	Lipoprotein-releasing system	Transporters
M00670	Mce transport system	Transporters
M00189	Molybdate transport system	Transporters
M00606	N,N'-Diacetylchitobiose transport system	Transporters
M00188	NitT/TauT family transport system	Transporters
M00209	Osmoprotectant transport system	Transporters
M00239	Peptides/nickel transport system	Transporters
M00222	Phosphate transport system	Transporters
M00210	Phospholipid transport system	Transporters
M00258	Putative ABC transport system	Transporters
M00207	Putative multiple sugar transport system	Transporters
M00236	Putative polar amino acid transport system	Transporters
M00221	Putative simple sugar transport system	Transporters
M00193	Putative spermidine/putrescine transport system	Transporters

Table S4: Key SOM degradation genes in Chloroflexi MAG metabat.179.

Feature ID	Type	Function	Length	Contig Name
Chloroflexi_MAG_metabat.179.RAST.CDS.700	gene	Xylulose kinase (EC 2.7.1.17)	1,449	k127_713037
Chloroflexi_MAG_metabat.179.RAST.CDS.1980	gene	D-xylose transport ATP-binding protein XylG	816	k127_2612503
Chloroflexi_MAG_metabat.179.RAST.CDS.1981	gene	Xylose ABC transporter, permease protein XylH	1,320	k127_2612503
Chloroflexi_MAG_metabat.179.RAST.CDS.1982	gene	Xylose ABC transporter, periplasmic xylose-binding protein XylF	1,110	k127_2612503
Chloroflexi_MAG_metabat.179.RAST.CDS.1296	gene	Beta-hexosaminidase (EC 3.2.1.52)	1,242	k127_1303888
Chloroflexi_MAG_metabat.179.RAST.CDS.1376	gene	Beta-hexosaminidase (EC 3.2.1.52)	1,593	k127_1761644
Chloroflexi_MAG_metabat.179.RAST.CDS.782	gene	Beta-hexosaminidase (EC 3.2.1.52)	1,842	k127_854051
Chloroflexi_MAG_metabat.179.RAST.CDS.1252	gene	N-Acetyl-D-glucosamine ABC transport system, permease protein 2	1,083	k127_1054772
Chloroflexi_MAG_metabat.179.RAST.CDS.1253	gene	N-Acetyl-D-glucosamine ABC transport system, permease protein 1	963	k127_1054772
Chloroflexi_MAG_metabat.179.RAST.CDS.1861	gene	N-Acetyl-D-glucosamine ABC transport system, permease protein 2	939	k127_2270236
Chloroflexi_MAG_metabat.179.RAST.CDS.1862	gene	N-Acetyl-D-glucosamine ABC transport system, permease protein 1	945	k127_2270236
Chloroflexi_MAG_metabat.179.RAST.CDS.1631	gene	carbon monoxide dehydrogenase E protein	1401	k127_2237853
Chloroflexi_MAG_metabat.179.RAST.CDS.1245	gene	L-lactate dehydrogenase (EC 1.1.2.3)	1098	k127_1054772
Chloroflexi_MAG_metabat.179.RAST.CDS.1191	gene	Aconitate hydratase (EC 4.2.1.3)	2865	k127_1054772

Paper V

Reconstructing Ribosomal Genes From Large Scale Total RNA Meta-Transcriptomic Data.

Xue, Y.*, Lanzén, A., & Jonassen, I.

***Bioinformatics (Oxford, England)**, 36(11), 3365–3371.* (2020)

https://doi.org/10.1093/bioinformatics/btaa177.

Bioinformatics, 36(11), 2020, 3365–3371
doi: 10.1093/bioinformatics/btaa177
Advance Access Publication Date: 13 March 2020
Original Paper

Sequence analysis

Reconstructing ribosomal genes from large scale total RNA meta-transcriptomic data

Yaxin Xue [1,*], **Anders Lanzén** [2,3] **and Inge Jonassen** [1,*]

[1]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway, [2]AZTI-Tecnalia, Herrera Kaia, 20110 Pasaia, Spain and [3]Ikerbasque, Basque Foundation for Science, 48011 Bilbao, Spain

*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received on November 17, 2019; revised on February 1, 2020; editorial decision on March 8, 2020; accepted on March 10, 2020

Abstract

Motivation: Technological advances in meta-transcriptomics have enabled a deeper understanding of the structure and function of microbial communities. 'Total RNA' meta-transcriptomics, sequencing of total reverse transcribed RNA, provides a unique opportunity to investigate both the structure and function of active microbial communities from all three domains of life simultaneously. A major step of this approach is the reconstruction of full-length taxonomic marker genes such as the small subunit ribosomal RNA. However, current tools for this purpose are mainly targeted towards analysis of amplicon and metagenomic data and thus lack the ability to handle the massive and complex datasets typically resulting from total RNA experiments.

Results: In this work, we introduce MetaRib, a new tool for reconstructing ribosomal gene sequences from total RNA meta-transcriptomic data. MetaRib is based on the popular rRNA assembly program EMIRGE, together with several improvements. We address the challenge posed by large complex datasets by integrating sub-assembly, dereplication and mapping in an iterative approach, with additional post-processing steps. We applied the method to both simulated and real-world datasets. Our results show that MetaRib can deal with larger datasets and recover more rRNA genes, which achieve around 60 times speedup and higher F1 score compared to EMIRGE in simulated datasets. In the real-world dataset, it shows similar trends but recovers more contigs compared with a previous analysis based on random sub-sampling, while enabling the comparison of individual contig abundances across samples for the first time.

Availability and implementation: The source code of MetaRib is freely available at https://github.com/yxxue/MetaRib.

Contact: yaxin.xue@uib.no or Inge.Jonassen@uib.no

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advances in next-generation sequencing have boosted the study of microbial communities in many ecosystems. Meta-transcriptomics, the direct sequencing and analysis of all RNA in a microbial community, has been widely used in investing microbial universe from various environments (Carvalhais *et al.*, 2012; Jorth *et al.*, 2014; Shi *et al.*, 2009). It provides an informative perspective about the current state of functional output, as it can elucidate which members and functions of a community are active in certain circumstances, rather than only the genomic contents (Franzosa *et al.*, 2015). Meta-transcriptomics is considered to be more efficient in observing rapid regulatory responses than meta-proteomics (Carvalhais *et al.*, 2012). Moreover, it could capture the information missing in DNA-based metagenomics, such as RNA viruses (Culley, 2006; Zhang *et al.*, 2006). The whole microbial RNA pool is dominated by rRNA and tRNA (95–99%), while only small fractions are mRNA (1–5%) (Carvalhais *et al.*, 2012). To date, most meta-transcriptomic studies have focused on function (mRNA) rather than structure, depleting rRNA both experimentally and *in silico*.

'Total RNA meta-transcriptomics' involves the isolation and sequencing of reverse transcribed total RNA pools—including mRNA (gene expression), rRNA (abundance), RNA viruses, tRNA and other non-coding RNA—from samples without any PCR or cloning step. In contrast to normal meta-transcriptomics, this approach enables us to obtain both structural and functional information simultaneously in a microbial community (Urich *et al.*, 2008). It answers two fundamental questions in microbial research—'who is there?' and 'what are they doing?'—with a few advantages. In terms of structural investigation, total RNA meta-transcriptomics assesses

taxonomic diversity in all three domains of life, meanwhile avoiding amplification bias, compared to PCR-based amplicon surveys. Ribosomal RNA is also essential for protein synthesis in all organisms. Thus, its relative abundance across taxa generally reflects the overall structural activity in a community. For functional profiling, it provides novel insights into current gene activity status with corresponding structural profiling simultaneously in one experiment.

Several tools are available for meta-transcriptomics, e.g. IMP (Narayanasamy *et al.*, 2016), SAMSA (Westreich *et al.*, 2016), MetaTrans (Martinez *et al.*, 2016), but they are geared mainly for studying the functional profiling. Though typically disregarded in meta-transcriptomics, rRNA and its corresponding gene is widely used as a genetic marker to study bacterial phylogeny and taxonomy, as it is present in all domains and has both highly conserved regions and regions that vary between species. Currently, most structural rRNA profiling relies on amplicon sequencing (meta-barcoding) using 'universal' primers to target and amplify hypervariable regions of rRNA or other taxonomic markers as broadly as possible (Rosselli *et al.*, 2016). Although amplicon sequencing represents a fundamentally important method for studying microbial and other biological communities, it is susceptible to biases depending on the specificity/universality of the primers used and other PCR conditions. Thus, it may lead to an incomplete or biased profile of the true biodiversity present in a given sample (Lanzén *et al.*, 2011; Shakya *et al.*, 2013). By using total RNA meta-transcriptomics for structural profiling, such biases can be avoided. Furthermore, it allows for the reconstruction of full-length rRNA sequences, enabling a higher resolution for taxonomy profiling. This is typically not feasible in meta-barcoding; using short-read sequencing technologies results in amplicons with insufficient phylogenetic signal, while long-read sequencing allow for longer amplicons but is currently restricted by higher error rates. Existing *de novo* assembly tools for shotgun sequence reads are designed primarily for genomic or metagenomic data and do not perform well on rRNA genes (Yuan *et al.*, 2015). Instead, there are several tools developed specifically for rRNA recovery and assembly, such as EMIRGE (Miller *et al.*, 2013), REAGO (Yuan *et al.*, 2015), RAMBL (Zeng *et al.*, 2017) and MATAM (Pericard *et al.*, 2018). However, these tools were designed for analysis of smaller datasets and cannot be used directly to analyze total RNA meta-transcriptomics studies.

Here, we present MetaRib, a novel tool for constructing full-length ribosomal gene sequences optimized for total RNA meta-transcriptomic data. Firstly, its dereplication process enables us to identify both existing species and novel species, while minimizing false positives. Furthermore, it significantly reduces the running time and memory usage by an iterative sampling approach, making it possible to assemble rRNA sequences from very large datasets: combining several samples also allows for reconstructing rRNA from less abundant species. Thus, MetaRib allows us to study the distributions of assembled rRNA sequences across multiple samples, independent of taxonomical classification. This is done by mapping reads to the resulting assembled small subunit ribosomal RNA (SSU rRNA) sequences, which we consider as operational taxonomic units (OTUs).

Our approach exploits the uneven taxon-abundance distribution common for microbial communities, with few dominating taxa and a long tail of rarer ones, often referred to as the 'rare biosphere' (Sogin *et al.*, 2006).

In practice, this leads to high redundancy in total RNA meta-transcriptomic data, with many sequences originating from the most abundant species. Our assumption is that rRNA of highly abundant species can be reconstructed from a relatively small subsample of the sequences. Subsequently, all rRNA sequences in the whole dataset related with the same species could be removed from further analysis, enabling reconstruction of less abundant species iteratively. Merging reads from several samples or datasets can also help to reconstruct rarer species, below the assembly threshold in smaller datasets. We evaluated our tool using three simulated total RNA datasets (limited to prokaryotic rRNA with special design to access different scenarios) and benchmarked its performance. Moreover, a real-world dataset from a large-scale soil total RNA experiment consisting of three billion SSU rRNA reads was analyzed, showing that MetaRib could recover more information than what was possible in the previous study of the same data.

2 Materials and methods

2.1 Metarib workflow

The MetaRib algorithm consists of three major modules: (i) initialization, (ii) iterative reconstruction and (iii) post-processing, summarized in Figure 1.

2.1.1 Initialization

A configuration file is needed to initiate the workflow, which first controls the availability of data and standalone software tools (dependencies). A case-specific workflow script is then generated and executed. A full description of the input configuration file and data structure is found in the GitHub repository (https://github.com/yxxue/MetaRib).

2.1.2 Iterative reconstruction

MetaRib uses an iterative process to reconstruct rRNA contigs. The workflow is initiated on a randomly picked subset of the total reads, which are assembled, and used to filter remaining reads by removing those that can be mapped perfectly to the resulting contigs. This process (random selection, assembly and filtering) is then repeated until a pre-defined termination criterion is reached. This module is composed of five steps:

Step 1: Subsampling reads

The first step is initial subsampling of sequencing data from the remaining unmapped reads. In each iteration, a subset of n reads (provided in the configuration file, by default $n = 100\,000$) is randomly picked from the total unmapped reads U, of size N (initially containing all reads). MetaRib will change the seed number automatically at each iteration to avoid repetitive sampling of reads.

Step 2: Assembly of subset

The randomly picked subset of size n is used as input to EMIRGE (Miller *et al.*, 2013) for reference-assisted assembly into rRNA contigs. The EMIRGE assembly parameters, including the reference sequence database used, can be specified in the configuration file. Considering that the community structure is relative uneven, for most natural communities, contigs corresponding to highly abundant species are more likely to be assembled in the first several iterations even when $n \ll N$.

Step 3: Dereplication of contigs

When the assembly is completed, contigs resulting from Step 2 of the current iteration are compared with the existing assembled rRNA contig set C (initially empty). New contig sequences are first concatenated to existing ones, then sorted by sequence length and

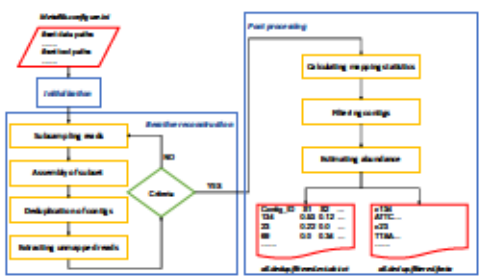


Fig. 1. Schematic overview of MetaRib workflow. Dash rectangles depict main modules, solid line rectangles represent major steps in each module and red elements denote input and output files

renamed with unique IDs. Overlap-based clustering is then performed to eliminate duplicated and keep longest contig sequences for each cluster using a stringent threshold, considering the high similarity of rRNA contigs.

Step 4: Mapping of remaining unmapped reads

All unmapped reads, U (i.e. all reads in the first iteration) are aligned against the dereplicated contig set C using stringent parameters, considering the presence of highly conserved regions in the rRNA gene. Reads that align to the contigs are removed from U, leaving only unmapped reads for subsequent iterations. Since sequences from highly abundant taxa are more likely to be assembled in the first iterations, a large proportion of raw reads are likely to be removed, which facilitates assembly of remaining reads. This is the key approach of MetaRib to reduce the complexity, memory and time requirement when assembling large datasets from typical, uneven biological communities.

Step 5: Terminating criteria

The iterative process will be terminated under three circumstances: (i) it reaches a maximum of 11 iterations; (ii) the remaining unmapped reads is less than n; or (iii) the last iteration produced a sufficiently small number of novel contigs (<1% of the current contig set). The last situation may indicate that assembly from a subsample of size n is difficult due to poor coverage of all taxa present. To counteract this, a final extra iteration is carried out using a subsample of size $2n$.

2.1.3 Post-processing

Once any of the criteria for halting the iteration have been met, a final non-redundant contig set is generated. MetaRib will then start post-processing to filter out low-quality contigs and estimate their relative abundance across individual samples.

Step 1: Calculating mapping statistics

Raw reads from each sample are aligned to the contig set C to generate several mapping statistics by BBMAP, including the mapping rate (%), coverage and covered percentage of each particular contig in C.

Step 2: Filtering contigs

Low-quality contigs are filtered by parsing mapping statistics report from Step 1. We consider a particular contig in C is a false positive record if either its average coverage or percent of bases covered are below a pre-defined threshold (by default 2 and 80%, respectively).

Step 3: Estimating abundance

The mapping rate is used to represent relative abundance of contigs in each sample. As rRNA genes contain both conserved and variable regions, we choose to include both 'unambiguous' mapping (where a merged read is aligned to only one contig) and 'ambiguous' mapping (where a read can be aligned to more than one contig).

Finally, MetaRib will generate two files: one containing the high-quality contig sequences (in FASTA format) and one matrix ('OTU table') that summarizes the abundance information across samples, which each row representing a contig and each column representing a sample. These numbers are assumed to approximate the abundances of taxa corresponding to the reconstructed contigs. An exception is species with considerable intra-specific rRNA sequence variation, for which total abundance instead can be obtained by identifying and adding the relative abundances for their contigs.

2.2 Implementation

MetaRib is developed with Python2.7 and is distributed under the GNU GPL v3.0 license. MetaRib is freely available on https://github.com/yxxue/MetaRib. Dependencies include the Python libraries Pandas (used for data analysis).

MetaRib also requires EMIRGE for rRNA assembly. EMIRGE was chosen by default as it is one the most widely used for reconstructing full-length rRNA genes and has shown better performance than other methods.

The BBtools suit (https://jgi.doe.gov/data-and-tools/bbtools/) is also required for MetaRib and utilized for several tasks including read mapping and dereplication. BBtools/reformat.sh is used for format conversion and subsampling. BBtools/dedupe.sh is an overlap-based dereplication tool allowing a specified number of substitutions or edit distance, applied in MetaRib's dereplication step. Default dedupe.sh parameters are maximum five indels and minimum 99% similarity (*fo = t ow = t c = t mcs = 1 e = 5 mid = 99*). BBtools/bbmap.sh is used to map (align) reads to contigs. BBMap has a few advantages for our implementation, including output of unmapped reads immediately (bypassing SAM/BAM format output), which accelerates the iteration process. Furthermore, it performs global rather than local alignment that can avoid excluding excessive reads due to highly conserved regions of rRNA genes. In addition, it returns detailed mapping statistics, used in post-processing. Default parameters for BBMAP is *minid = 0.96 maxindel = 1 minbits = 2 idfilter = 0.98* and users can modify those parameters in the configuration file.

The BBtools suits and EMIRGE need to be installed before MetaRib, and their parameters defined in the configuration file.

2.3 Evaluation with simulated datasets

2.3.1 Generation of simulated datasets

To simulate the complexity of real microbiome communities, three *in silico* simulated datasets were built. As a full-length rRNA reference dataset, we used the SILVA SSU rRNA reference database (Quast *et al.*, 2012) (release 123). To simulate sequence reads for dataset a, one thousand full-length sequences were randomly picked from a version of the reference database clustered at 94% identity using maximum linkage. These reference sequences were used to simulate 5 million Illumina pair-end sequencing reads following a log-normal abundance distribution, using ART (Huang *et al.*, 2012). For Dataset b, we randomly selected 1000 sequences from the full non-redundant version of Silva v123 (i.e. not clustered using 94%). Only full-length sequences with a similarity between 95% and 99% to the clustered reference database were retained and used to generate 5 million sequence read pairs with ART following the same distribution. Finally, Dataset c was similar with a, but all full-length sequences used to generate them were removed from the reference database used by EMIRGE during assembly. An overview of simulated datasets is shown in Figure 2. The intra-dataset sequence similarity was evaluated by performing global all-against-all alignment for each dataset (exclude self-alignment) with minimum pairwise identity 90 (Supplementary Fig. S1). All simulated datasets and corresponding EMIRGE references are deposited at NIRD research data archive (https://archive.sigma2.no/pages/public/datasetDetail.jsf?id=10.11582/2019.00040).

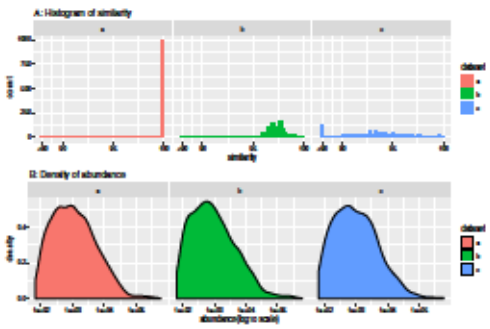


Fig. 2. Overview of simulated datasets. (A) The global pairwise similarity distribution of picked contigs aligning to the reference is shown. (B) Contig abundance distribution for simulated reads in the three test datasets is shown

2.3.2 Evaluating performance

Simulated datasets were used to compare the performance of MetaRib with EMIRGE (run non-iteratively). All programs were tested on the same computer cluster using 40 cores (in-house compute server, 80 cores, 1 TB RAM). For the running time comparison benchmark, we used the GNU 'time' command to capture both real (elapsed), system and user time.

For the simulated datasets, we could assess the correctness of the reconstructed contigs, i.e. how similar each reconstructed contig was to the 'source contig', recording for each reconstructed contig the similarity of the closest source sequence, and vice versa; for each source sequence, the similarity of the closest reconstructed contig. For this analysis, we used Vsearch (Rognes *et al.*, 2016) for performing pair-wise global alignment with 90% minimum identity.

For a range of similarity thresholds, we then counted statistical measures of the performance of two methods. *True positives (TP)* correspond to the number of 'correctly' reconstructed contigs (having a reconstructed contig with similarity above the threshold) and *false positives (FP)*—the number of reconstructed incorrect contigs (below the similarity threshold). *False negatives (FN)* correspond to the number of un-reconstructed sequences in the source contig. Finally, we calculate *Precision*, *Sensitivity* and *F1-score* based on the number of *TP*, *FP* and *FN*.

To evaluate the accuracy of abundance estimation, we then performed Pearson's correlation test between the real abundance of source contigs with the abundance output of the closest reconstructed contigs.

2.4 Real-world dataset

In order to evaluate the performance of MetaRib on real-world total RNA sequence data, we utilized the data 3 billion sequence reads generated as part of the AshBack project (Bang-Andreasen *et al.*, 2020). Bang-Andreasen *et al.* (2020) conducted a large-scale total RNA meta-transcriptomic study to access the impact of wood ash on agricultural and forest soil microbial communities and functional expression simultaneously applying four doses of wood ash concentration: 0, 3, 12 and 90 t ha^{-1} (Conc: 0, 3, 12, 90). Each dose was applied to two soil types: agricultural and forest soil and total community RNA extracted and sequenced after 0, 10, 30 and 100 days of incubation (D0, D3, D30, D100). The large-scale and complexity made it an ideal case to apply MetaRib.

A total of 325 Gb rRNA sequences were collected from the wood ash dataset (PRJNA512608). Due to the lack of bioinformatic tools and computational constraints, previous rRNA analysis was performed on a small subset (1.5 million randomly selected sequences) of each sample, using EMIRGE (Bang-Andreasen *et al.*, 2020). We reanalyzed the complete dataset using MetaRib with default parameters, and, in a repeated analysis with $n = 1\,000\,000$ considering the larger size of the dataset. Downstream analysis was performed with Phyloseq (McMurdie and Holmes, 2013) and DADA2 (Callahan *et al.*, 2016), figures were generated using ggplot2 (Ginestet, 2011) and ComplexHeatmap (Gu *et al.*, 2016). Since all samples were analyzed together in MetaRib, we could also detect the presence (here defined as a relative abundance $\geq 1e{-}5$) of contigs across samples.

Table 1. Comparison of programs running time

	User (s)		Elapsed (HH:MM:SS)		Iterations
	EMIRGE	MetaRib	EMIRGE	MetaRib	MetaRib
a	2224330	19278	37:02:26	0:28:27	5
b	3913996	59468	57:36:44	1:07:30	5
c	2499052	45901	37:45:27	0:47:58	7

Note: User is the amount of CPU time spent; elapsed is the time from start to finish the program. Iteration is the iteration number in MetaRib for each dataset.

3 Results

3.1 Run time comparison

Table 1 shows statistics of time usage when analyzing the three simulated datasets using EMIRGE non-iteratively and with MetaRib (otherwise using the same parameters). MetaRib could assemble simulated datasets (5 million sequences each) in a few minutes while EMIRGE needs days to run, representing around 60X speedup compared to using EMIRGE out of the box with the same parameters.

3.2 Correctness

The relative performance of two tools is shown in terms of *Precision*, *Sensitivity* and *F1-score* for all three simulated datasets representing different scenario. MetaRib shows the best overall performance in all datasets with *F1-score* evaluation (Fig. 3 and Supplementary Table S1). EMIRGE recovers almost all source sequences if they are represented in the reference (Dataset a). For b and c, where source sequences are less similar to the reference database, EMIRGE has a higher sensitivity compared to MetaRib. However, as shown in Figure 3, EMIRGE is also producing a large number 'false' contigs, which leads to a quite low precision and F1-score even in an ideal case (Dataset a). Conversely, MetaRib is producing far fewer such 'false' sequences. We also test the performance of the 'contig filtering' step done as part of the post-processing.

Our results demonstrate that filtering low-quality contigs using mapping statistics (MetaRib_F) improves the performance compared with the unfiltered result (MetaRib_R; see 2.1.3 step 3). More detailed results—like statistical metrics of contigs length in each iteration (Supplementary Table S2) and comparison of contigs length distribution between tools (Supplementary Fig. S2)—can be found in the Supplementary Data.

3.3 Abundance estimation

Figure 4 shows the scatter plot of comparation of relative abundance between source contigs(src_ab) with the closest reconstructed contigs (est_ab). MetaRib could estimate the relative abundance accurately when the nearly full-length contigs are reconstructed (sim $\geq$ 97.5), even for very low-abundant records (src_ab $\leq$ 1e−2). As we expected, it has the best performance in an ideal scenario (Dataset a); however, it comes up with over-estimation problem at low-abundant records caused by 'ambiguous' mapping of conserved region in rRNA sequences which are distinct from the reference (Dataset b).

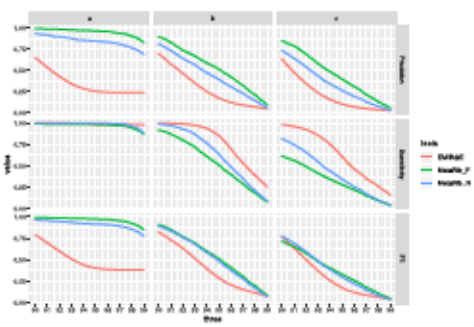


Fig. 3. Overview evaluation of correctness. The *X*-axis represents different similarity thresholds used to determine if a reconstructed contig is correct. The *Y*-axis represents the value of measurements (precision, sensitivity and F1-score). MetaRib_F represents contigs filtered with low-quality records, while MetaRib_R is the original output

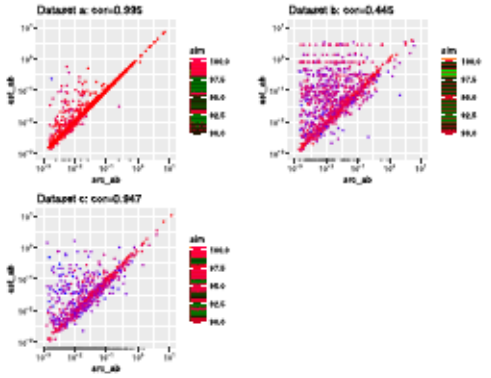


Fig. 4. Evaluation of abundance estimation. Values for real abundance (src_ab) and the closest estimated abundance (est_ab) displayed on log–log coordinates and colored with the similarity score (sim). Pearson correlation is calculated between src_ab and est_ab

Table 2. Comparison of MetaRib running time with different sampling reads number

Sampling_num (*n*)	User (s)	Elapsed (HH:MM:SS)
100 000 (100 K)	7 927 023	38:00:48
1 000 000 (1 M)	12 330 130	62:16:55

Note: The program is performed with 80 cores.

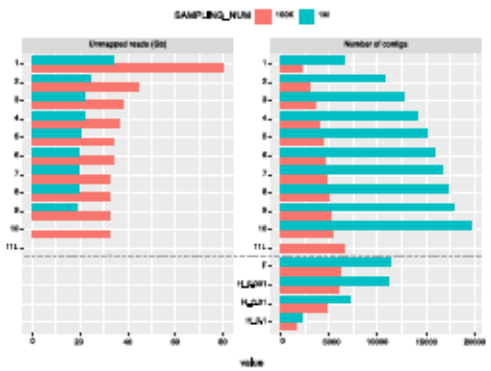


Fig. 5. MetaRib performance for ASHBACK dataset per iteration with two different read subsampling numbers *n*. 1–11: iterations. F: filtered contigs. H: contigs which relative abundance is higher than certain thresholds (0.001%: H_0001, 0.01%: H_0.01, 0.1%: H_0.1) in at least one sample

3.4 Real-world dataset

MetaRib could complete the analysis of 320 Gb (3 billion reads) in approximately 1–2 days using default parameters with 80 cores. However, the CPU and run time is nearly doubled when using a larger reads sampling number ($n = 1\,000\,000$; Table 2).

The read subsampling number *n* also effects the performance of MetaRib, both for the iteration process and final result resulting in 11 iterations for the default value ($n = 100k$) and 9 iterations for $n = 1M$ (see Fig. 5). As we expected, the size of U decreases significantly in the first few iterations and thus becomes stabilized; while smaller values of *n* need more iterations to converge and result in more remaining unmapped reads after the last iteration. However, larger *n* values also result in more potential false positives. For example, the size of U ceases to decrease after five iterations, whereas the number of C maintains a continuous increase. Particularly, nearly half of the C fail to pass the filter step (F) using $n = 1M$. We further check the number of contigs which relative abundance is higher than certain thresholds (0.001%: H_0001, 0.01%: H_0.01, 0.1%: H_0.1) in at least one sample according to their estimated abundance. We find that the number of 'dominant' (high abundance) contigs using the default value ($n = 100k$) gives closer results to $n = 1M$ with a higher threshold, which indicates that the smaller, default value of *n* was sufficient to reconstruct the majority rRNA contigs in a complex community. Results obtained using $n = 1M$ were thus excluded from further analysis.

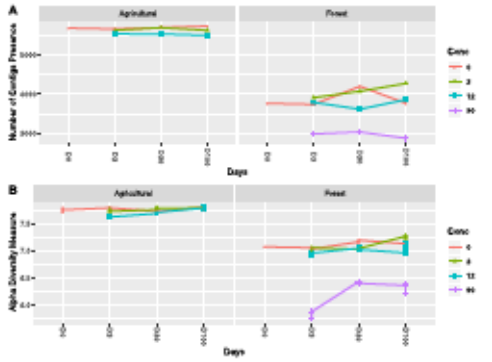


Fig. 6. Number of contigs and Shannon diversity across the two soils at increasing wood ash amendment and incubation times ($n = 100$ K). The presence of contig is determined by the average abundance within each measure and soil (≥1e−5). Shannon diversity is estimated based on relative abundance table

We observe more rRNA contigs in both sites and similar trends of richness and Shannon diversity across treatments in forest soil as those revealed by previous analysis (Bang-Andreasen *et al.*, 2020), except considerably less fluctuation of diversity across treatments and time in agricultural soil (Fig. 6).

MetaRib is able to recover more rRNA contigs across all domains and captures more taxa than before. For example, the fungal division *Mucoromycota* appears to be dominant in both with an abundance of approximately 3.5% in Forest at the highest ash concentration, while missing in the previous analysis (Bang-Andreasen *et al.*, 2020) (Fig. 7). MetaRib also allowed us to carry out taxonomy-independent statistics that were not possible when assembling reads sample-by-sample. Thus, we observed several interesting abundance patterns among the top 100 dominant contigs, illustrated as a heatmap in Figure 7. For example, while *Proteobacteria* were ubiquitous in both soils, different contigs dominated and showed more fluctuations in the forest. Contigs affiliated to the *Acidobacteria* were dominant in the forest soil and most of their abundances were positively correlated with concentration; however, they dropped significantly at the highest ash concentration. Besides, one *Firmicutes* affiliated contig was only presented in agricultural soil, while other *Firmicutes* contigs were only abundant in the highest dose in forest soil. *Verrucomicrobia* associated contigs showed the opposite trend.

4 Discussion

Here, we present the tool MetaRib for reconstructing rRNA genes from large scale total RNA meta-transcriptomic data. Its main advantage compared to existing methods is to quickly and reliably assemble rRNA contigs across multiple samples, even in very large datasets, with a low false positive rate and a taxonomy-independent relative abundance estimation.

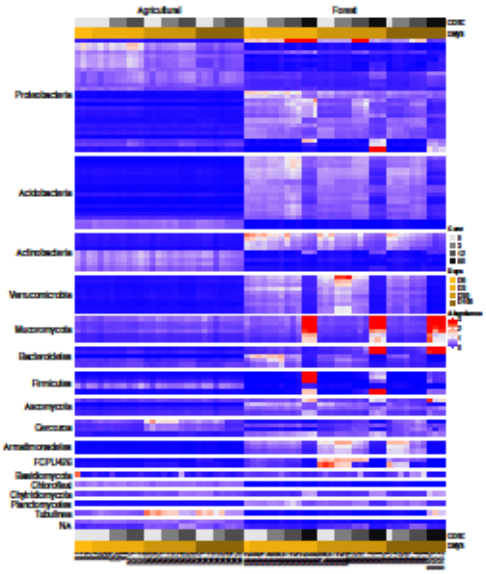


Fig. 7. Heatmap of abundance distribution for top 100 most dominant contigs among samples ($n = 100$ K). Samples are ordered at increasing incubation times and wood ash amendment

Using simulated datasets, we show that MetaRib performs similarly to EMIRGE (representing the current state-of-the-art) in terms of recovering the underlying full-length true sequences, at the same time avoiding generating as many unreliable sequences (false positives) with a significant speedup. Besides, it provides an opportunity to have an overview of the abundance distribution across multiple samples, which could indicate important functions or patterns when combined with biological information.

Still, some challenges remain. Both EMIRGE and MetaRib are reference-based approaches, which could have issues in recovering novel and similar contigs when there is lacking information in the reference database (Datasets b and c): only partial sequences could be reconstructed in such extreme scenario. The contrasting results of simulated datasets indicate that MetaRib is able to capture most information in relatively well-characterized environments while it is more likely to generate false positives and partial sequences for poorly characterized environments. It also illustrates that the reference database is crucial for performance. While the most recent release of Silva includes over 9 million SSU sequences, our simulations used a less inclusive, earlier version, clustered at 94% sequence identity. It is likely that a more recent version will result in higher similarity for rRNA sequences, but it also result in longer execution times. At any rate, a non-redundant reference database is recommended, since EMIRGE is limited to reconstructing sequences with maximum 97% similarity to each other (Miller *et al.*, 2013). Other recent tools for rRNA assembly such as MATAM (Pericard *et al.*, 2018) have been shown to perform better than EMIRGE on small datasets, and future work could include using MATAM within the MetaRib tool.

An advantage of total RNA meta-transcriptomics is the ability to estimate relative abundances of rRNA sequences as proxies of microbial taxa, without PCR bias. Similarly, applications of third-generation sequencing like Oxford nanopore also have this advantage together with extreme long sequencing reads and real-time identification, which has shown great potential in microbial research (Jain *et al.*, 2016; Shin *et al.*, 2016).

However, it is important to point out that the number of rRNA reads does not represent an unbiased estimate of neither the metabolic activity nor the abundance (biomass or cell numbers) of the taxa as such, since rRNA gene copy number and patterns of ribosomal transcription and retention vary between organisms (Blazewicz *et al.*, 2013). In addition, so far it seems to be no commercial kit from Oxford Nanopore for sequencing of prokaryotic or total RNA, only eukaryotic, poly A-tagged mRNA sequencing.

Several parameter settings will also impact the performance of MetaRib, especially for large scale datasets, as illustrated here using a real-world dataset. In particular, the trade-off between execution time and the quality of the final results needs to be considered carefully. For example, increasing the read subsampling number will lead to longer execution times, but generate more low abundance contigs from rare organisms, thus recovering more of the diversity. However, it also leads to more false positives in terms of incorrectly assembled contigs.

In the current implementation, MetaRib discards any remaining unmapped reads after the iteration process is finished. However, taxonomy-independent rRNA assembly tools like REAGO could be considered as a further step to assemble discarded reads in order to maximize the information recovered from total RNA datasets.

Our approach opens up several new perspectives for total RNA meta-transcriptomics. First of all, it simplifies the analysis of the large and redundant datasets generated, via iterative reconstruction. In doing so, it also reduces false positives and allows for taxonomy-independent comparisons of contig abundances across samples. In spite of its advantages, total RNA has not been widely used compared to other environmental genomics techniques We hope that MetaRib will enable researchers to make more use of this technique and the valuable rRNA sequence data generated, with full-length sequences free of primer bias. Ultimately, this enables a deeper understanding of how natural microbial communities are structured, as well their function.

Financial Support

none declared.

Conflict of Interest

none declared.

References

Bang-Andreasen,T. *et al.* (2020) Total RNA-sequencing reveals multi-level microbial community changes and functional responses to wood ash application in agricultural and forest soil. *FEMS Microbiol Ecol.*, 96. doi: 10.1093/femsec/fiaa016.

Blazewicz,S.J. *et al.* (2013) Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *ISME J*, 7, 2061–2068.

Callahan,B.J. *et al.* (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods*, 13, 581–583.

Carvalhais,L.C. *et al.* (2012) Application of metatranscriptomics to soil environments. *J. Microbiol. Methods*, 91, 246–251.

Culley,A.I. (2006) Metagenomic analysis of coastal RNA virus communities. *Science*, 312, 1795–1798.

Franzosa,E.A. *et al.* (2015) Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.*, 13, 360–372.

Ginestet,C. (2011) ggplot2: elegant graphics for data analysis. *J. R. Stat. Soc.*, 174, 245–246.

Gu,Z. *et al.* (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32, 2847–2849.

Huang,W. *et al.* (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, 28, 593–594.

Jain,M. *et al.* (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.*, 17, 239.

Jorth,P. *et al.* (2014) Metatranscriptomics of the human oral microbiome during health and disease. *MBio*, 5, e01012–e01014.

Lanzén,A. *et al.* (2011) Exploring the composition and diversity of microbial communities at the Jan Mayen hydrothermal vent field using RNA and DNA. *FEMS Microbiol. Ecol.*, 77, 577–589.

Martinez,X. *et al.* (2016) MetaTrans: an open-source pipeline for metatranscriptomics. *Sci. Rep.*, 6, 1–12.

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.

Miller,C.S. *et al.* (2013) Short-read assembly of full-length 16S Amplicons reveals bacterial diversity in subsurface sediments. *PLoS One*, 8, e56018.

Narayanasamy,S. *et al.* (2016) IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.*, 17, 260.

Pericard,P. *et al.* (2018) MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes. *Bioinformatics*, 34, 585–591.

Quast,C. *et al.* (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, 41, D590–D596.

Rognes,T. *et al.* (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 2016, e2584.

Rosselli,R. *et al.* (2016) Direct 16S rRNA-seq from bacterial communities: a PCR-independent approach to simultaneously assess microbial diversity and functional activity potential of each taxon. *Sci. Rep.*, 6, 32165.

Shakya,M. *et al.* (2013) Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.*, 15, 1882–1899.

Shi,Y. *et al.* (2009) Metatranscriptomics reveals unique microbial small RNAs in the oceans water column. *Nature*, 459, 266–269.

Shin,J. *et al.* (2016) Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Sci. Rep.*, 6, 29681.

Sogin,M.L. *et al.* (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl. Acad. Sci. USA*, 103, 12115–12120.

Urich,T. *et al.* (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One*, 3, e2527.

Westreich,S.T. *et al.* (2016) SAMSA: a comprehensive metatranscriptome analysis pipeline. *BMC Bioinformatics*, 17, 399.

Yuan,C. *et al.* (2015) Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics*, 31, i35–i43.

Zeng,F. *et al.* (2017) Large-scale 16S gene assembly using metagenomics shotgun sequences. *Bioinformatics*, 33, 1447–1456.

Zhang,T. *et al.* (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.*, 4, 0108–0118.

Supplementary information

Fig. S1. Histogram of global pair wise similarity distribution of each dataset.

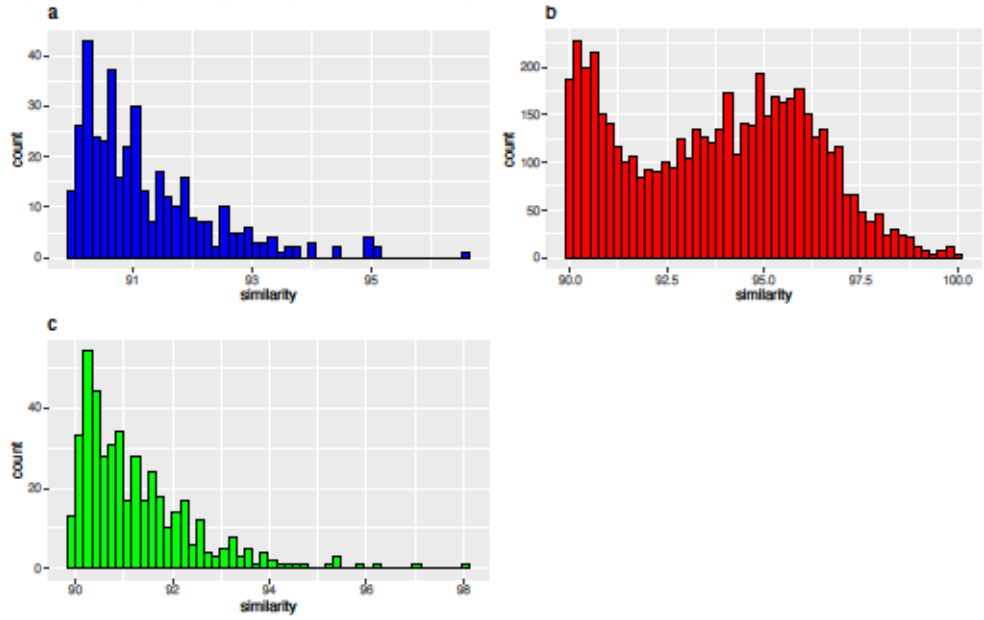


Fig. S2. Histogram of contig length distribution between source (SRC), MetaRib and EMIRGE using log10 transformation for both axes.

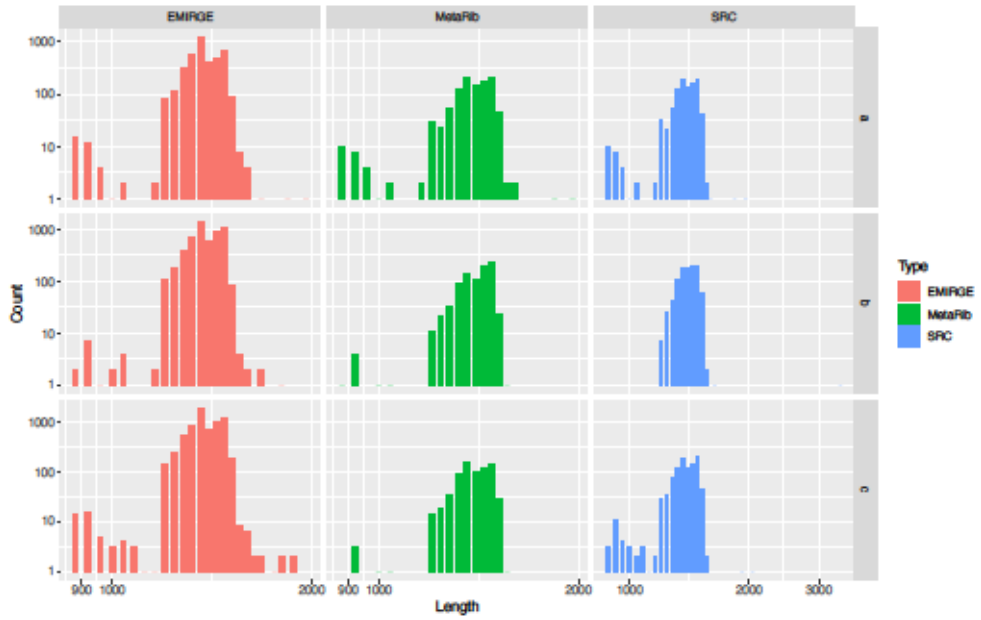


Table S1: The detailed performance of all tools in all simulated datasets.

datasets	tools	thres	TP	FP	FN	Precision	Sensitivity	F1
a	EMIRGE	90	2690	1444	0	0.6507015	1	0.7883939
a	EMIRGE	90.5	2445	1689	0	0.5914369	1	0.7432741
a	EMIRGE	91	2194	1940	0	0.5307209	1	0.693426
a	EMIRGE	91.5	1983	2151	2	0.4796807	0.9989924	0.6481451
a	EMIRGE	92	1778	2356	4	0.4300919	0.9977553	0.6010818
a	EMIRGE	92.5	1576	2558	5	0.3812288	0.9968374	0.5515311
a	EMIRGE	93	1415	2719	5	0.3422835	0.9964789	0.5095427
a	EMIRGE	93.5	1279	2855	6	0.3093856	0.9953307	0.4720428
a	EMIRGE	94	1175	2959	10	0.2842284	0.9915612	0.4418124
a	EMIRGE	94.5	1095	3039	11	0.2648766	0.9900543	0.4179389
a	EMIRGE	95	1052	3082	12	0.2544751	0.9887218	0.4047711
a	EMIRGE	95.5	1020	3114	14	0.2467344	0.9864603	0.3947368
a	EMIRGE	96	999	3135	14	0.2416546	0.9861797	0.3881873
a	EMIRGE	96.5	990	3144	15	0.2394775	0.9850746	0.385289
a	EMIRGE	97	986	3148	16	0.2385099	0.9840319	0.3839564
a	EMIRGE	97.5	984	3150	17	0.2380261	0.983017	0.3832522
a	EMIRGE	98	984	3150	17	0.2380261	0.983017	0.3832522
a	EMIRGE	98.5	982	3152	18	0.2375423	0.982	0.3825477
a	EMIRGE	99	982	3152	18	0.2375423	0.982	0.3825477
a	MetaRib_R	90	1196	85	0	0.9336456	1	0.9656843
a	MetaRib_R	90.5	1183	98	0	0.9234973	1	0.9602273
a	MetaRib_R	91	1176	105	0	0.9180328	1	0.957265
a	MetaRib_R	91.5	1156	125	0	0.90242	1	0.9487074
a	MetaRib_R	92	1147	134	0	0.8953942	1	0.9448105
a	MetaRib_R	92.5	1137	144	1	0.8875878	0.9991213	0.9400579
a	MetaRib_R	93	1123	158	1	0.8766589	0.9991103	0.9338877
a	MetaRib_R	93.5	1103	178	1	0.8610461	0.9990942	0.9249476
a	MetaRib_R	94	1096	185	1	0.8555816	0.9990884	0.921783
a	MetaRib_R	94.5	1089	192	2	0.8501171	0.9981668	0.9182125
a	MetaRib_R	95	1082	199	3	0.8446526	0.997235	0.9146238
a	MetaRib_R	95.5	1076	205	4	0.8399688	0.9962963	0.9114782
a	MetaRib_R	96	1071	210	4	0.8360656	0.9962791	0.9091681
a	MetaRib_R	96.5	1056	225	5	0.824356	0.9952875	0.9017933
a	MetaRib_R	97	1034	247	10	0.8071819	0.9904215	0.8894624
a	MetaRib_R	97.5	1015	266	15	0.7923497	0.9854369	0.8784076
a	MetaRib_R	98	982	299	36	0.7665886	0.9646365	0.8542845
a	MetaRib_R	98.5	944	337	62	0.7369243	0.9383698	0.8255356
a	MetaRib_R	99	886	395	115	0.6916472	0.8851149	0.7765118
a	MetaRib_F	90	1040	8	9	0.9923664	0.9914204	0.9918932
a	MetaRib_F	90.5	1040	8	10	0.9923664	0.9904762	0.9914204
a	MetaRib_F	91	1038	10	11	0.990458	0.9895138	0.9899857
a	MetaRib_F	91.5	1034	14	12	0.9866412	0.9885277	0.9875836
a	MetaRib_F	92	1033	15	12	0.985687	0.9885167	0.9870999
a	MetaRib_F	92.5	1031	17	12	0.9837786	0.9884947	0.986131
a	MetaRib_F	93	1024	24	14	0.9770992	0.9865125	0.9817833
a	MetaRib_F	93.5	1019	29	14	0.9723282	0.9864472	0.9793369
a	MetaRib_F	94	1019	29	14	0.9723282	0.9864472	0.9793369
a	MetaRib_F	94.5	1016	32	15	0.9694656	0.985451	0.977393
a	MetaRib_F	95	1011	37	17	0.9646947	0.983463	0.9739884

51	a	MetaRib_F	95.5	1008	40	19	0.9618321	0.9814995	0.9715663
52	a	MetaRib_F	96	1006	42	19	0.9599237	0.9814634	0.970574
53	a	MetaRib_F	96.5	993	55	24	0.9475191	0.9764012	0.9617433
54	a	MetaRib_F	97	982	66	30	0.9370229	0.9703557	0.9533981
55	a	MetaRib_F	97.5	973	75	34	0.9284351	0.9662363	0.9469586
56	a	MetaRib_F	98	948	100	56	0.9045802	0.9442231	0.9239766
57	a	MetaRib_F	98.5	920	128	81	0.8778626	0.9190809	0.897999
58	a	MetaRib_F	99	872	176	129	0.8320611	0.8711289	0.8511469
59	b	EMIRGE	90	3858	1668	4	0.6981542	0.9989643	0.8219003
60	b	EMIRGE	90.5	3528	1998	5	0.6384365	0.9985848	0.7788939
61	b	EMIRGE	91	3183	2343	10	0.5760043	0.9968681	0.7301296
62	b	EMIRGE	91.5	2874	2652	15	0.5200869	0.9948079	0.683066
63	b	EMIRGE	92	2579	2947	30	0.4667029	0.9885013	0.6340504
64	b	EMIRGE	92.5	2251	3275	37	0.4073471	0.9838287	0.5761454
65	b	EMIRGE	93	1949	3577	51	0.3526963	0.9745	0.5179378
66	b	EMIRGE	93.5	1641	3885	72	0.2969598	0.9579685	0.4533775
67	b	EMIRGE	94	1386	4140	93	0.2508143	0.9371197	0.3957173
68	b	EMIRGE	94.5	1157	4369	127	0.2093739	0.9010903	0.3397944
69	b	EMIRGE	95	956	4570	165	0.1730004	0.85281	0.2876486
70	b	EMIRGE	95.5	793	4733	220	0.1435034	0.7828233	0.2425447
71	b	EMIRGE	96	677	4849	290	0.1225118	0.7001034	0.2085323
72	b	EMIRGE	96.5	572	4954	367	0.1035107	0.6091587	0.1769528
73	b	EMIRGE	97	503	5023	437	0.0910242	0.5351064	0.1555831
74	b	EMIRGE	97.5	434	5092	517	0.0785378	0.4563617	0.1340127
75	b	EMIRGE	98	377	5149	592	0.0682229	0.3890609	0.1160893
76	b	EMIRGE	98.5	319	5207	668	0.0577271	0.3232016	0.0979579
77	b	EMIRGE	99	256	5270	742	0.0463265	0.256513	0.0784795
78	b	MetaRib_R	90	1188	275	11	0.8120301	0.9908257	0.892562
79	b	MetaRib_R	90.5	1133	330	21	0.7744361	0.9818024	0.865877
80	b	MetaRib_R	91	1075	388	32	0.7347915	0.971093	0.8365759
81	b	MetaRib_R	91.5	1006	457	47	0.6876282	0.9553656	0.799682
82	b	MetaRib_R	92	937	526	72	0.6404648	0.9286422	0.7580906
83	b	MetaRib_R	92.5	883	580	90	0.6035543	0.9075026	0.7249589
84	b	MetaRib_R	93	830	633	122	0.5673274	0.8718487	0.6873706
85	b	MetaRib_R	93.5	751	712	170	0.5133288	0.815418	0.6300336
86	b	MetaRib_R	94	691	772	212	0.4723172	0.765227	0.5841082
87	b	MetaRib_R	94.5	634	829	270	0.4333561	0.7013274	0.5356992
88	b	MetaRib_R	95	575	888	325	0.393028	0.6388889	0.4866695
89	b	MetaRib_R	95.5	508	955	403	0.3472317	0.557629	0.4279697
90	b	MetaRib_R	96	451	1012	469	0.3082707	0.4902174	0.3785145
91	b	MetaRib_R	96.5	387	1076	545	0.2645249	0.4152361	0.3231733
92	b	MetaRib_R	97	338	1125	614	0.2310321	0.355042	0.2799172
93	b	MetaRib_R	97.5	266	1197	711	0.1818182	0.272262	0.2180328
94	b	MetaRib_R	98	210	1253	781	0.1435407	0.2119072	0.1711491
95	b	MetaRib_R	98.5	146	1317	852	0.0997949	0.1462926	0.118651
96	b	MetaRib_R	99	90	1373	911	0.0615174	0.0899101	0.0730519
97	b	MetaRib_F	90	767	88	69	0.897076	0.9174641	0.9071555
98	b	MetaRib_F	90.5	745	110	82	0.871345	0.9008464	0.8858502
99	b	MetaRib_F	91	712	143	103	0.8327485	0.8736196	0.8526946
100	b	MetaRib_F	91.5	671	184	130	0.7847953	0.8377029	0.8103865
101	b	MetaRib_F	92	630	225	158	0.7368421	0.7994924	0.7668898

102	b	MetaRib_F	92.5	602	253	186	0.7040936	0.7639594	0.7328058
103	b	MetaRib_F	93	574	281	222	0.671345	0.7211055	0.6953362
104	b	MetaRib_F	93.5	535	320	275	0.625731	0.6604938	0.6426426
105	b	MetaRib_F	94	502	353	322	0.5871345	0.6092233	0.597975
106	b	MetaRib_F	94.5	472	383	370	0.5520468	0.5605701	0.5562758
107	b	MetaRib_F	95	440	415	420	0.5146199	0.5116279	0.5131195
108	b	MetaRib_F	95.5	398	457	488	0.4654971	0.4492099	0.4572085
109	b	MetaRib_F	96	362	493	544	0.4233918	0.3995585	0.41113
110	b	MetaRib_F	96.5	316	539	611	0.3695906	0.3408846	0.3546577
111	b	MetaRib_F	97	277	578	670	0.3239766	0.2925026	0.3074362
112	b	MetaRib_F	97.5	222	633	753	0.2596491	0.2276923	0.242623
113	b	MetaRib_F	98	181	674	808	0.2116959	0.1830131	0.1963124
114	b	MetaRib_F	98.5	129	726	868	0.1508772	0.1293882	0.1393089
115	b	MetaRib_F	99	82	773	918	0.0959064	0.082	0.0884097
116	c	EMIRGE	90	4475	2580	81	0.6343019	0.9822212	0.7708208
117	c	EMIRGE	90.5	3896	3159	104	0.5522325	0.974	0.7048394
118	c	EMIRGE	91	3369	3686	140	0.4775337	0.9601026	0.6378266
119	c	EMIRGE	91.5	2872	4183	167	0.4070872	0.9450477	0.5690509
120	c	EMIRGE	92	2432	4623	201	0.3447201	0.9236612	0.5020644
121	c	EMIRGE	92.5	2034	5021	242	0.2883062	0.8936731	0.4359661
122	c	EMIRGE	93	1662	5393	275	0.2355776	0.8580279	0.3696619
123	c	EMIRGE	93.5	1354	5701	319	0.1919206	0.8093246	0.3102658
124	c	EMIRGE	94	1101	5954	364	0.1560595	0.7515358	0.2584507
125	c	EMIRGE	94.5	894	6161	410	0.1267186	0.6855828	0.2139012
126	c	EMIRGE	95	718	6337	464	0.1017718	0.607445	0.1743353
127	c	EMIRGE	95.5	584	6471	516	0.0827782	0.5309091	0.143225
128	c	EMIRGE	96	484	6571	564	0.0686038	0.4618321	0.1194619
129	c	EMIRGE	96.5	418	6637	601	0.0592488	0.4102061	0.1035422
130	c	EMIRGE	97	357	6698	651	0.0506024	0.3541667	0.0885526
131	c	EMIRGE	97.5	310	6745	694	0.0439405	0.3087649	0.0769326
132	c	EMIRGE	98	263	6792	739	0.0372785	0.2624751	0.0652848
133	c	EMIRGE	98.5	212	6843	789	0.0300496	0.2117882	0.0526316
134	c	EMIRGE	99	159	6896	841	0.0225372	0.159	0.0394786
135	c	MetaRib_R	90	982	356	218	0.7339312	0.8183333	0.7738377
136	c	MetaRib_R	90.5	924	414	254	0.690583	0.7843803	0.7344992
137	c	MetaRib_R	91	862	476	291	0.6442451	0.7476149	0.6920915
138	c	MetaRib_R	91.5	795	543	331	0.5941704	0.7060391	0.6452922
139	c	MetaRib_R	92	725	613	374	0.5418535	0.6596906	0.5949938
140	c	MetaRib_R	92.5	650	688	426	0.4857997	0.6040892	0.5385253
141	c	MetaRib_R	93	583	755	480	0.435725	0.5484478	0.485631
142	c	MetaRib_R	93.5	507	831	537	0.3789238	0.4856322	0.4256927
143	c	MetaRib_R	94	456	882	578	0.3408072	0.4410058	0.3844857
144	c	MetaRib_R	94.5	404	934	625	0.3019432	0.3926142	0.3413604
145	c	MetaRib_R	95	351	987	670	0.2623318	0.3437806	0.2975837
146	c	MetaRib_R	95.5	303	1035	709	0.2264574	0.2994071	0.2578723
147	c	MetaRib_R	96	256	1082	750	0.1913303	0.2544732	0.21843
148	c	MetaRib_R	96.5	222	1116	780	0.1659193	0.2215569	0.1897436
149	c	MetaRib_R	97	171	1167	829	0.1278027	0.171	0.1462789
150	c	MetaRib_R	97.5	135	1203	865	0.1008969	0.135	0.1154833
151	c	MetaRib_R	98	98	1240	902	0.0732436	0.098	0.0838323
152	c	MetaRib_R	98.5	69	1269	931	0.0515695	0.069	0.0590248

153	c	MetaRib_R	99	41	1297	959	0.0306428	0.041	0.0350727
154	c	MetaRib_F	90	611	106	380	0.8521618	0.6165489	0.7154567
155	c	MetaRib_F	90.5	585	132	413	0.8158996	0.5861723	0.6822157
156	c	MetaRib_F	91	565	152	439	0.7880056	0.562749	0.656595
157	c	MetaRib_F	91.5	533	184	476	0.7433752	0.5282458	0.617613
158	c	MetaRib_F	92	495	222	507	0.6903766	0.494012	0.5759162
159	c	MetaRib_F	92.5	456	261	549	0.6359833	0.4537313	0.5296167
160	c	MetaRib_F	93	417	300	592	0.58159	0.4132805	0.4831981
161	c	MetaRib_F	93.5	378	339	629	0.5271967	0.3753724	0.4385151
162	c	MetaRib_F	94	347	370	662	0.4839609	0.3439049	0.4020857
163	c	MetaRib_F	94.5	316	401	693	0.4407252	0.3131814	0.3661645
164	c	MetaRib_F	95	283	434	725	0.3947001	0.280754	0.3281159
165	c	MetaRib_F	95.5	252	465	753	0.3514644	0.2507463	0.2926829
166	c	MetaRib_F	96	221	496	784	0.3082287	0.2199005	0.2566783
167	c	MetaRib_F	96.5	191	526	811	0.2663877	0.1906188	0.2222222
168	c	MetaRib_F	97	152	565	848	0.2119944	0.152	0.177053
169	c	MetaRib_F	97.5	125	592	875	0.1743375	0.125	0.1456028
170	c	MetaRib_F	98	93	624	907	0.1297071	0.093	0.1083285
171	c	MetaRib_F	98.5	67	650	933	0.0934449	0.067	0.0780431
172	c	MetaRib_F	99	40	677	960	0.055788	0.04	0.0465929

Table S2: statistical metrics of contigs length in each iteration and abundance estimation.

# Iterative reconstruction							
file	format	type	num_seqs	sum_len	min_len	avg_len	max_len
./dataset_a/MetaRib/Iteration/iter_1/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	975	1,360,997	912	1,395.90	1,975
./dataset_a/MetaRib/Iteration/iter_2/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,277	1,785,886	901	1,398.50	1,975
./dataset_a/MetaRib/Iteration/iter_3/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,280	1,790,078	901	1,398.50	1,975
./dataset_a/MetaRib/Iteration/iter_4/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,281	1,791,532	901	1,398.50	1,975
./dataset_b/MetaRib/Iteration/iter_1/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	981	1,378,854	904	1,405.60	1,588
./dataset_b/MetaRib/Iteration/iter_2/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,418	1,984,759	902	1,399.70	1,593
./dataset_b/MetaRib/Iteration/iter_3/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,439	2,014,399	902	1,399.90	1,593
./dataset_b/MetaRib/Iteration/iter_4/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,455	2,036,567	902	1,399.70	1,593
./dataset_b/MetaRib/Iteration/iter_5_L/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,463	2,047,874	902	1,399.80	1,593
./dataset_c/MetaRib/Iteration/iter_1/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	767	1,068,275	904	1,392.80	1,595
./dataset_c/MetaRib/Iteration/iter_2/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,169	1,626,992	904	1,391.80	1,595
./dataset_c/MetaRib/Iteration/iter_3/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,236	1,720,808	904	1,392.20	1,595
./dataset_c/MetaRib/Iteration/iter_4/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,272	1,770,652	904	1,392	1,595
./dataset_c/MetaRib/Iteration/iter_5/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,295	1,802,927	904	1,392.20	1,597
./dataset_c/MetaRib/Iteration/iter_6/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,308	1,821,199	904	1,392.40	1,597
./dataset_c/MetaRib/Iteration/iter_7_L/emirge_amp/iter.20/all.dedup.fasta	FASTA	DNA	1,338	1,862,851	904	1,392.30	1,597
# Abundance							
file	format	type	num_seqs	sum_len	min_len	avg_len	max_len
./dataset_a/MetaRib/Abundance/all.dedup.fasta	FASTA	DNA	1,281	1,791,532	901	1,398.50	1,975
./dataset_a/MetaRib/Abundance/all.dedup.filtered.fasta	FASTA	DNA	1,048	1,471,133	901	1,403.80	1,975
./dataset_b/MetaRib/Abundance/all.dedup.fasta	FASTA	DNA	1,463	2,047,874	902	1,399.80	1,593
./dataset_b/MetaRib/Abundance/all.dedup.filtered.fasta	FASTA	DNA	855	1,216,191	905	1,422.40	1,588
./dataset_c/MetaRib/Abundance/all.dedup.fasta	FASTA	DNA	1,338	1,862,851	904	1,392.30	1,597
./dataset_c/MetaRib/Abundance/all.dedup.filtered.fasta	FASTA	DNA	717	1,011,562	928	1,410.80	1,595

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

UNIVERSITAS BERGENSIS

uib.no

ISBN: 9788230844229 (print)
9788230862773 (PDF)