Computational and Structural Biotechnology Journal 18 (2020) 296–305

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Computational methods for 16S metabarcoding studies using Nanopore sequencing data

Andres Santos [a,b,c], Ronny van Aerle [c], Leticia Barrientos [a,b,c], Jaime Martinez-Urtaza [c,*]

[a] Applied and Molecular Biology Laboratory, Centre of Excellence in Translational Medicine, Universidad de La Frontera, Avenida Alemania 0458, 4810296 Temuco, Chile
[b] Scientific and Technological Bioresource Nucleus, Universidad de La Frontera, Avenida Francisco Salazar 01145, 481123 Temuco, Chile
[c] Centre for Environment, Fisheries and Aquaculture Science (Cefas), Barrack Road, Weymouth, Dorset DT4 8UB, UK

ARTICLE INFO

Article history:
Received 5 September 2019
Received in revised form 15 January 2020
Accepted 15 January 2020
Available online 31 January 2020

Keywords:
Third generation sequencing
MinION
Microbial diversity

ABSTRACT

Assessment of bacterial diversity through sequencing of 16S ribosomal RNA (16S rRNA) genes has been an approach widely used in environmental microbiology, particularly since the advent of high-throughput sequencing technologies. An additional innovation introduced by these technologies was the need of developing new strategies to manage and investigate the massive amount of sequencing data generated. This situation stimulated the rapid expansion of the field of bioinformatics with the release of new tools to be applied to the downstream analysis and interpretation of sequencing data mainly generated using Illumina technology. In recent years, a third generation of sequencing technologies has been developed and have been applied in parallel and complementarily to the former sequencing strategies. In particular, Oxford Nanopore Technologies (ONT) introduced nanopore sequencing which has become very popular among molecular ecologists. Nanopore technology offers a low price, portability and fast sequencing throughput. This powerful technology has been recently tested for 16S rRNA analyses showing promising results. However, compared with previous technologies, there is a scarcity of bioinformatic tools and protocols designed specifically for the analysis of Nanopore 16S sequences. Due its notable characteristics, researchers have recently started performing assessments regarding the suitability MinION on 16S rRNA sequencing studies, and have obtained remarkable results. Here we present a review of the state-of-the-art of MinION technology applied to microbiome studies, the current possible application and main challenges for its use on 16S rRNA metabarcoding.

Contents

* Corresponding author at: The Centre for Environment, Fisheries and Aquaculture Science (CEFAS), The Nothe, Barrack Road, Weymouth, Dorset DT4 8UB, UK.
E-mail address: jaime.martinez-urtaza@cefas.co.uk (J. Martinez-Urtaza).

https://doi.org/10.1016/j.csbj.2020.01.005

1. Introduction

Functionality, interaction, and dynamics of microbial communities are considered critical to the existence of ecological balance and life [1,2]. The fact that only less than 1% of microorganisms are cultivable under laboratory conditions [3] has presented historical constraints to providing a precise dimension of the microbial world, and to studying microbial diversity within a taxonomic context.

Since the foundations of molecular phylogeny were established in the 1960s and 70s, the 16S rRNA gene has been universally used for taxonomic studies of prokaryotic species [4,5]. 16S rRNA is part of the small ribosomal subunit (SSU) present in all prokaryotic cells and the gene encoding for this molecule possesses some distinctive characteristics that make it suitable for taxonomic profiling: 1) it is ubiquitous, being found in all prokaryotic and archaebacteria organisms [6]; 2) the relatively small size (~1500 bp) and high degree of functional conservation [5], 3) the presence of variable regions in the 16S rRNA gene as result of diverse rates of evolution among species, which can be used to distinguish different bacterial groups [7,8], and 4) the existence of highly conserved regions in the gene sequence, which can be used to design universal primers flanking different hypervariable regions (nine in total, V1-V9) identified in the gene [9]. On the other hand, the use of 16S rRNA for bacterial identification has some limitations, including the variable number of copies of these genes in bacterial genomes, the low taxonomic resolution at the species level for some bacterial groups, and the bias for taxonomic assignment of sequences depending on the variable region chosen for the analysis [10].

Until the late 1990s, the 16S rRNA gene was only applied in a taxonomic context to define species uniquely based on individual bacteria obtained from pure (mostly clinical) cultures [6,11]. However, in 1997, Pace et al. [12] described for the first time the composition of microbial communities without the need for cultivation in the laboratory by employing the sequence of the 16S rRNA gene using Sanger sequencing. This work led to the establishment of a universal approach to the study of microbial communities. Nowadays, sequence analysis of 16S rRNA continues to be the gold standard for studying microbial diversity, enabling accurate taxonomic profiling of the prokaryotic groups present in both clinical and environmental samples [11,12].

The introduction of Sanger sequencing technology in the investigation of microbial communities signified a revolution in the world of microbial ecology and entirely changed how microbial diversity was assessed. However, this approach required the analysis of individual sequences, implying that a cloning step was needed as a crucial prerequisite for the investigation of samples (Fig. 1a). As a result, sequences up to ~1000 can be generated.

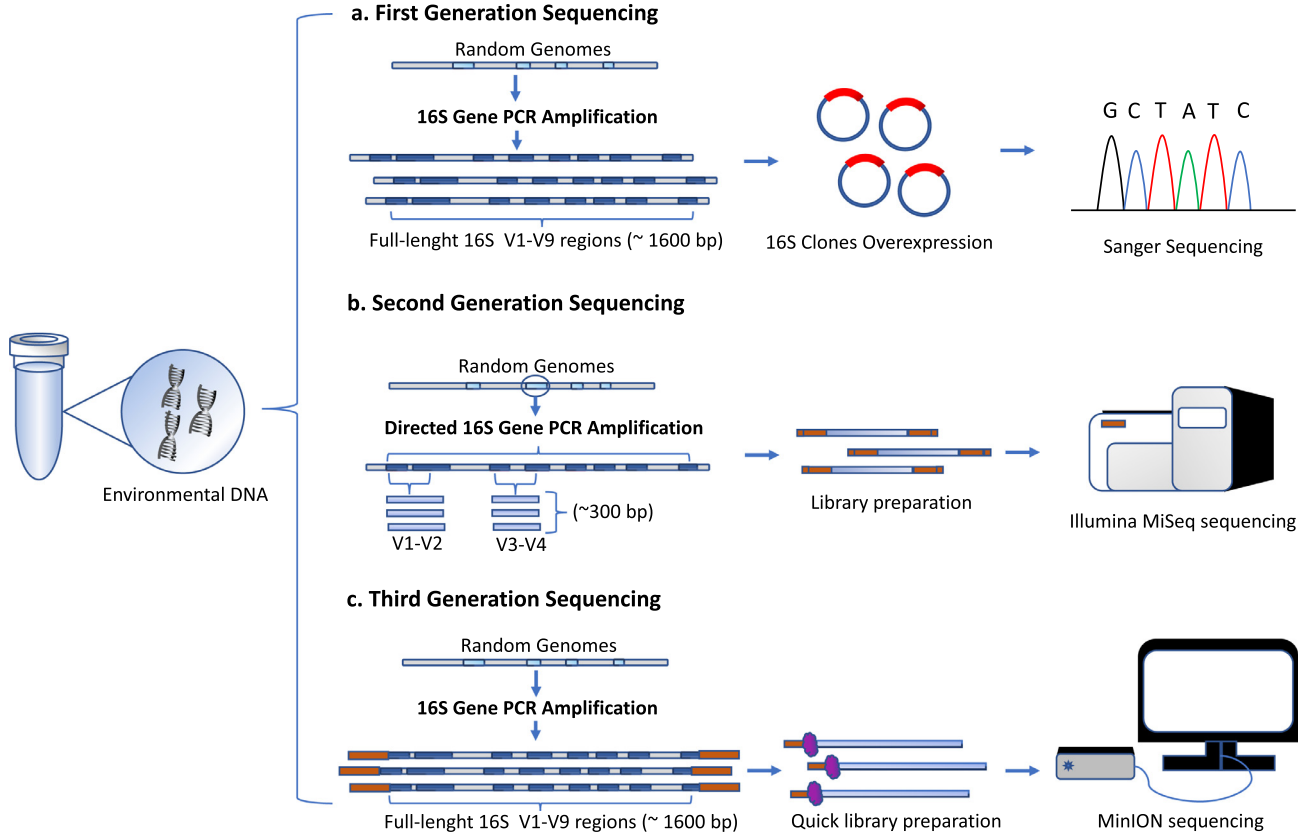

Fig. 1. Most common metabarcoding sequencing strategies for each sequencing technology generation. (a) First generation sequencing (Sanger). Under this approach, metabarcoding is classically performed by amplifying full-length 16S rRNA genes from an environmental DNA sample; once the amplicon has been obtained, the cloning of the 16S amplicons is performed, sequences are added into a vector and then transformed into a host; finally, plasmid extraction and purification are performed and the sequencing of 16S rRNA inserts is carried out by the Sanger method. (b) Second generation sequencing (Illumina). From environmental DNA samples, a PCR amplification of specific regions of de 16S rRNA gene is performed; depending on the scope of the study, one or two regions of the 16S gene can be amplified, with regions V1-V2 and V3-V4 being the most frequently used; by using these regions, a paired end library (the mix of DNA fragments with adapters attached to theirs ends and ready to be sequenced) preparation is often used for this purpose, adapters (exogenous nucleic acids that are ligated to a nucleic acid molecule to be sequenced) and index (unique DNA sequences ligated to fragments within a sequencing library, they allow the posterior sorting and identification of different samples sequenced on a same sequencing run) are added to 16S amplicon extremes and libraries of ~300 bp in length are finally sequenced on the Illumina MiSeq platform. (c) Third generation sequencing (Nanopore). This recently developed approach starts with the amplification of the full-length 16S rRNA gene from environmental DNA using universal primers; simultaneously, indexes for multiplexing are added to the amplicons in the same PCR reaction; once amplicons have been purified, the library preparation process is performed, consisting of the addition of a protein at a specific tagged region of the 16S amplicons (10 min for library preparation); finally direct sequencing of the samples is carried out on the MinION sequencer.

Table 1
Comparison of the available sequencing platforms for 16S metagenomic analysis using metabarcoding approach.

Sequencing Platform	Read Length (bp)	Accuracy	Output	Sequencing Chemistry	Run Time	Advantages in Metabarcoding approaches
Sanger	400–900	99.999%	1.9–84 Kb	Dideoxy chain termination	20 min −3 h	Long read length, high quality
Illumina MiSeq	75–300	99.9%	13.2–20 Gb	Sequencing by Synthesis	21–56 h	High Throughput, read quality
MinION	>200,000	~95%	~50 Gb	Single Sequencing real time-long reads	1–48 h	High Throughput, Long read length, portability
PacBio	10–15 Kb	99.999	5–10 Gb	Single Sequencing real time-long reads	4 h	Long read length and quality

However, the number of sequences that can be analyzed was limited due the output of Sanger platforms (Table 1). Therefore, a complete evaluation of bacterial diversity using Sanger sequencing became a serious challenge in terms of time and costs.

Globally, the advent of high-throughput sequencing, or Second Generation Sequencing (SGS) technologies, and its rapid and widespread application across laboratories in the early 2000s represented a paradigm shift in microbial ecology. The characteristic high output and data accuracy provided by these new technologies, along with the removal of tedious and time-consuming steps such as the cloning of DNA fragments and electrophoretic separation of sequencing products required for Sanger sequencing, makes possible the generation of massive sequencing data in short run processes. Among the different companies pioneering high-throughput sequencing, Illumina has achieved a leading position in the market, becoming the standard sequencing technology and the most frequently applied in microbial ecology studies [13,14]. The common elements in the sequences generated by this technology are the reduced length (from 50 bp to 300 bp), high throughput (from 2 Gb to 750 Gb), high accuracy, and reduced cost (starting from ~$40 USD per Gb approximately) [15] (Table 1).

Nevertheless, due to the differential characteristics of the Illumina and Sanger technologies in terms of sequence length, full-length sequences of the 16S rRNA gene are not achievable using Illumina sequencing alone. To overcome this limitation, 16S gene analysis with Illumina has been typically restricted to specific variable regions of the 16S rRNA, instead of the complete gene (Fig. 1b). However, the remarkable characteristics of Illumina sequencing in terms of outputs, accuracy and speed, have made this technology central in almost all of the most prominent studies based on 16S analysis carried out up to date, including the Human Microbiome Project [16], Earth Microbiome Project [17] and the Extreme Microbiome Project [18].

1.1. Current analytical approaches applied in 16S metagenomic studies

An additional innovation introduced by high-throughput sequencing technologies was the need for new strategies to manage and investigate the massive amount of sequencing data generated. From the user perspective, this change involved a transition from the application of basic computer programs accessible to general users in standard computers, to the need for sophisticated computational analysis requiring advanced bioinformatic skills. This situation stimulated the rapid expansion of the field of bioinformatics applied to microbial ecology studies, mainly with the release of new tools applied to the downstream analysis and interpretation of sequencing data. Nowadays, a large number of powerful tools are available which enable an efficient integration of different types of data [15–17].

Within this context, several bioinformatics programs and tools for processing amplicon sequencing data are presently available, most of them designed to work with V3 and V4 variable regions of the 16S rRNA gene. The most popular packages for 16S amplicon analysis are QIIME [20], MOTHUR [21] and Phyloseq [22]. In particular for 16S metagenomic studies, standard analysis packages and pipelines typically include a workflow comprising demultiplexing and quality control steps, followed by the generation of Operational Taxonomic Units (OTU picking) and/or "Amplicon Sequence Variants analysis" (ASV) analysis, which allows the taxonomic assignment of representative sequences and diversity analysis of the sample (Fig. 2). Consequently, taxonomic assignment of sequences is a critical step and the most informative element for microbial diversity analyses.

A detailed pipeline of the most conventional workflows for 16S rRNA Illumina sequences are presented in Fig. 2. Despite the differences between the different packages, the principal components in the workflow are analog and shared a common process, which includes: quality control of sequences, clustering or ASV analyses, taxonomic assignment and diversity analyses (Fig. 3).

2. Third generation of sequencing technologies

In recent years, a third generation of sequencing (TSG) technologies has been developed and have been used in parallel and complementarily to the former sequencing strategies. These new technologies interrogate a single molecule of DNA in real time and produce very long reads (from 1 to 100 kb). In 2011, Pacific Biosciences introduced the first TSG technology, which was termed single-molecule real-time sequencing [19,20]. Recent releases of a new sequencer, in particular the Sequel, has improved the output by increasing read length and throughput per run by 10- and 100-fold respectively. However, despite that this new platform is two-fold cheaper than the previous versions, it is still less cost-effective than Illumina and therefore the applications of this platform to 16S metagenomic studies remain scarce. In addition, the error rate falls in the same range as the first PacBio version (~13%) [25] and the output is still lower than Illumina. Therefore, price and limited output has restricted the application of the PacBio system in microbial community studies [22–24] (Table 1).

In 2014, Oxford Nanopore Technologies (ONT) introduced nanopore sequencing [28]. Nanopore sequencing was developed at the end of the 1980s [29], although the first successful use of this sequencing technology was reported in 2012 [30]. This sequencing technology directly detects the nucleotides without active DNA synthesis, since a long stretch of single stranded DNA passes through a protein nanopore that is stabilized in an electrically resistant polymer membrane [25–27]. Specifically, nucleotide detection is based on setting a voltage across this membrane, which is composed by sensors that are able to detect the ionic current changes shifted by nucleotides occupying the pore in real time while the DNA molecule passes through.

Applying this technology, ONT released the MinION platform in 2014, with some remarkable advantages such as low price, portability, and fast sequencing chemistry [33]. MinION is basically a base to grip a flowcell responsible for the direct sequencing of indi-

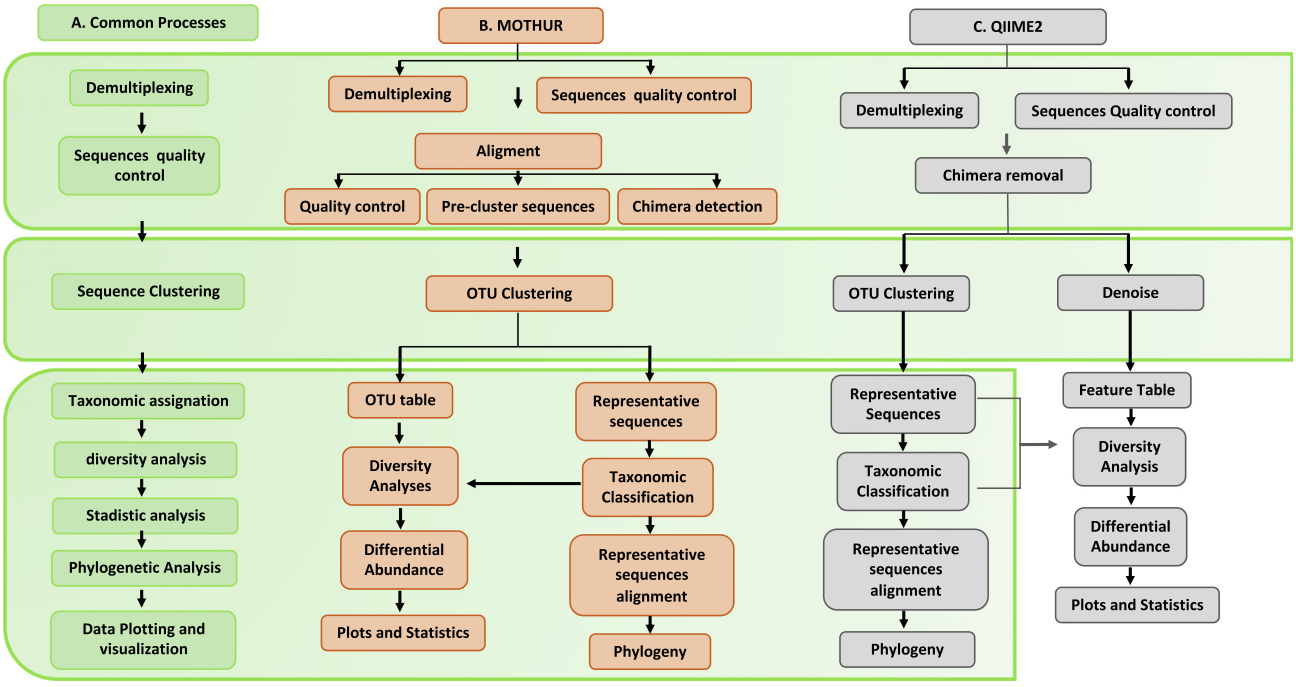

Fig. 2. Classic pipelines MOTHUR [21] and QIIME2 [20] and their complete workflow for 16S rRNA amplicons analyses, the "common processes" flow contains all common steps in both pipelines.

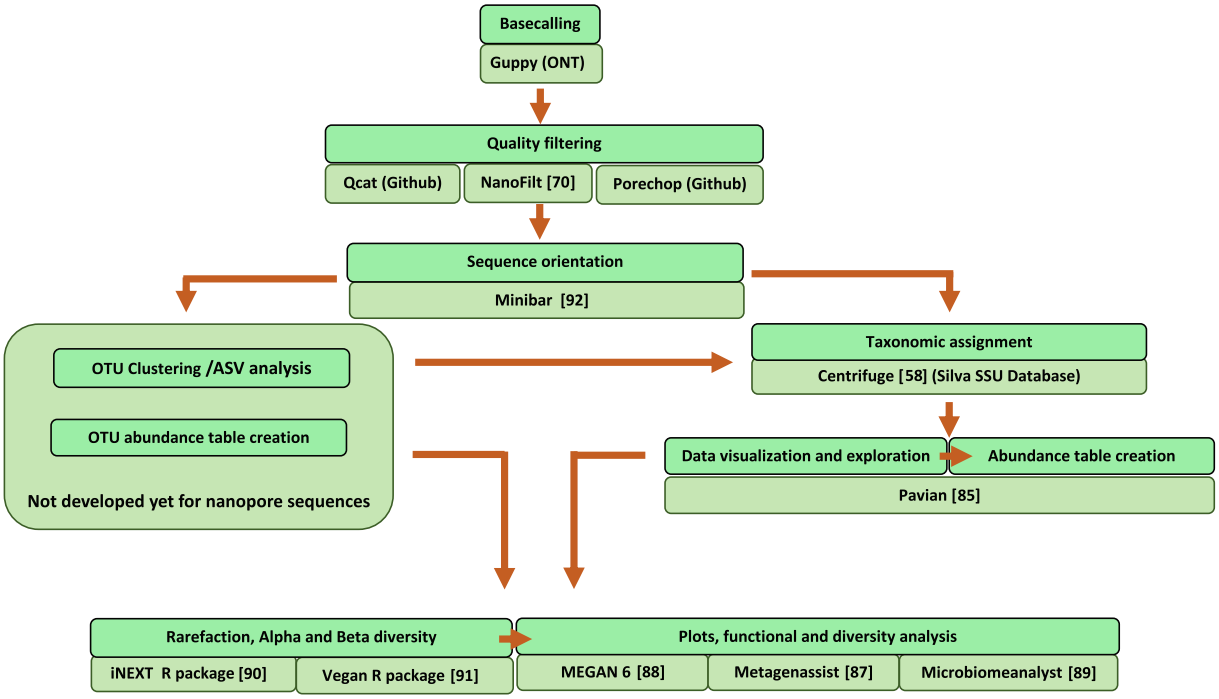

Fig. 3. Recommended MinION 16S rRNA amplicons pipeline for bacterial diversity analysis. (See above-mentioned references for further information.)

vidual DNA strands that translocate nanoscale the pores in the semiconductor membrane [34]. The most remarkable characteristic of the MinION Nanopore sequencer is the length of the sequences generated by the flowcell and the amount of data that can be produced per run. Moreover, MinION is a miniaturized sequencing device and the smallest available today in the market, with dimensions of 10 × 3 × 2 cm and weight of 87 g. One particular feature is that the sequencing process does not utilize a secondary signal such as light or pH, as with Illumina and PacBio [35]. According to the manufacturer, the most recent chemistry used in the R9.4.5 version of the flowcell provides an accuracy of ~95% with an output of ~20 Gb. However, the quality of the reads generated by the R9.4.5 flowcell is still lower than those of Illumina, which possess an accuracy of 99.9% (Table 1). Typical problems in Nanopore reads are the frequent presence of insertions and deletions artificially generated in the sequences that may introduce some obstacles to correctly analyze and interpret data from MinION [32].

Another remarkable characteristic of ONT platforms is that data analysis can be performed from the beginning of the sequencing run, which could considerably reduce the time of analysis compared to Illumina platforms. In addition, costs associated with the analyses performed by MinION are much lower compared with other sequencing platforms currently applied for 16S metagenomic studies (Table 1). All these characteristics make the MinION an accessible technology for many laboratories, which has generated a rapid expansion of the use of this technology across the scientific community. Within this context, a remarkable and original feature that ONT have developed is the "nanopore community," which is part of the ONT website. This "community" provides a common space where users can get help and feedback on device performance, methodologies, and bioinformatic analysis. It is important to note that there are other ONT platforms that can produce larger quantities of sequencing data than the MinION platform, with the same characteristics, such as GridION (100 Gb) and PromethION (6 Tb) [30]

3. The potential of the Nanopore sequencing for 16S rRNA studies

Nanopore sequencing brings to 16S rRNA metabarcoding studies the benefits of both first and second-generation sequencing. ONT platforms generate long reads, allowing cover the full-length sequence of 16S rRNA gene (V1-V9 regions) through a fast, cheap, and high throughput process. One of the most relevant advantages of the full-length 16S rRNA sequences is that they offer a higher level of taxonomic and phylogenetic resolution for bacterial identification since all the informative sites of 16S rRNA genes are considered in the analysis [36]. With Illumina sequencing, the conventional strategy for sequencing the 16S rRNA uses the hypervariable regions V1-V2 and/or V3-V4 [37], and taxonomy is assigned based only on these short variable regions of the 16S rRNA gene of approximately ~300 bp. The analysis of these short regions provides a limited taxonomic resolution in most cases, failing to reliably discriminate sequences beyond genus level [31,32]. Moreover, the choice of these regions will produce a direct effect on the specificity of the taxonomic assignment. For example, V4 regions better represent the whole bacterial diversity in host-associated studies, while V1-V2 are more specific for skin microbiota studies. In addition, taxonomic resolution varies for different groups of bacteria when using different portions of the 16S rRNA gene [40]. By contrast, the resolution obtained with Nanopore sequencing is only comparable to levels provided by Sanger 16S rRNA sequencing, with the potential for providing better discrimination among taxa, a deeper phylogenetic signal, and a more accurate taxonomic placement of 16S rRNA nanopore sequences [34,31,30]. Another advantage of ONT, is that data can be generated in a short runtime (1–48 h) and at an affordable price (~ $50 USD per sample) Table 1.

As previously mentioned, MinION is one of the most popular ONT platforms today and has been used extensively in genomics and transcriptomics studies [35–40], and over the last two years is rapidly growing in studies on microbial diversity. However, despite the evident benefits of the use of ONT technology in microbial ecology studies, there are still several factors limiting the implementation of these new approaches in the routine analysis of microbial diversity. The scarcity of tools specifically designed to work with full sequences of the 16S gene have made it extremely challenging to carry out a specialized taxonomic analysis of Nanopore sequences. Moreover, the limited quality of Nanopore 16S sequences has represented a serious constraint to apply exiting tools designed for other technologies (mostly Illumina) to analyze these sequences.

3.1. Nanopore 16S metagenomic studies

Studies applying Nanopore sequencing to describe microbial diversity have conventionally applied a similar approach than previous studies, which were mostly Illumina-based, regardless of the fact that Nanopore generates full-length 16S sequences. With Nanopore, the full length 16S rRNA gene is amplified by PCR using universal primers (27F and 1493R). The library is prepared by the addition of adapters in the amplicon sequences, and samples are sequenced directly with a flowcell gripped on the MinION device (Fig. 1 c).

Authors have tried to standardize a different 16S-based amplicon barcoding protocol by using a two PCR step-based protocol, with the first process to amplify the 16S rRNA gene and a second one for the addition of adapters for the 16S amplicons sequencing [48,49]. Another strategy has been based on the use of an ONT 1D2 chemistry library preparation where both DNA strands are sequenced (similar to the paired-end sequencing of Illumina), improving the quality of the reads by sequencing both strands of the target DNA [50]. Although different strategies have been applied in published studies using Nanopore sequencing for 16S rRNA metabarcoding, the 16S barcoding Kit of Oxford Nanopore Technologies has been predominantly used with satisfactory results [41–44].

Similar to sample preparation, methodologies introduced to analyze Nanopore 16S amplicons have included a broad range of bioinformatic tools. Nevertheless, despite the different tools, the central process in all the published studies is the application of a strategy based on taxonomic assignment [43–45,47].

3.2. Taxonomic assignment using Nanopore 16S sequences

Compared with Illumina, there is a scarcity of bioinformatic tools and protocols designed specifically for the analysis of Nanopore 16S sequences. The most extensively used tool is the cloud-based data analysis service EPI2ME (ONT), which provides a number of workflows for end-to-end analysis of nanopore 16S data: 16S taxonomic classification, a barcoding protocol, and quality filter of reads. For taxonomic assignment, FASTQ files are uploaded on the FASTQ 16S protocol of the EPI2ME platform, reads are filtered by quality and then taxonomy is assigned using BLAST to the NCBI database, with a minimum horizontal coverage of 30% and a minimum accuracy of 77% as default parameters (ONT). However, this tool is not publicly available and only ONT customers can gain access to this tool through a web platform. Moreover, quality filters, adapter trimming, or setting of alignment parameters such as identity and coverage of sequences, are already configured by default and the user cannot modify more than the initial parameters of the quality of reads. Furthermore, the format of the final output with the taxonomic assignment results is not compatible with other tools for performing downstream analyses such as diversity and taxonomic differential abundance. To overcome these limitations of EPI2ME software, it is necessary to define a different analytical pipeline that considers other bioinformatic tools available.

Cusco [48] applied a mapping approach for taxonomic assignment using the tool Minimap, and was able to determine the taxonomic composition at the genus and species level for bacterial isolates, mock communities, and complex skin samples. However, the study suggested the need for a more accurate bioinformatic protocol to achieve more reliable results. Another important result of this research is that taxonomic accuracy can be improved by analyzing sequences longer than 16S rRNA gene, such as the rrn operon (16S rRNA-ITS-23S rRNA; 4500 bp). Using Minimap2 [54], Kai et al. [52] reported a species-level bacteria identification with more than 90% of reads correctly assigned to each species. A

subsequent study carried out by Hardegen et al. [49] used a BLAST-based classification and concluded that their pipeline can be suitable for taxonomic assignment of 16S rRNA reads from Nanopore sequencing. Edwards et al. [51] used VSEARCH [55] for taxonomic assignment and reached a confidence level of ~75% at the phylum and family level. A different approach was performed by Ma et al. [50], who carried out taxonomic classification using RDP classifier [56], and reported in pure-culture an average annotation accuracy of 93.8% and 82.0% at the phyla and genus level, respectively. Mitsuhashi et al. [57] analyzed a mock community of pleural effusion from a patient with empyema using Centrifuge [58] and BLAST for taxonomic analysis, successfully identifying all the species presents in the mock community applying Centrifuge [58]. Turner et al. [53] described the microbiome of a new invasive nemertean species using Centrifuge [58] for taxonomic assignment, identifying 2054 species associated with the microbiome.

Considering all of the aforementioned studies, Centrifuge [58] and Minimap [54] have been the most frequently used taxonomic classifiers for Nanopore datasets [50,41,44,43,45]. Regarding the characteristics of both bioinformatic tools, Centrifuge [58] is capable of accurately identifying reads when using databases containing multiple highly similar reference genomes, such as different strains of a bacterial species. Moreover, Centrifuge works by building a database of genomes in which unique segments of these genomes are identified to build an FM-index (a compressed data structure for full-text pattern searching). This FM-index can be used for efficient searches of sequenced reads against genome segments in a database. On the other hand, Minimap2 [54] is a general-purpose alignment program that maps long DNA sequences against reference genomes such as Human, fungal, bacterial, or viral genomes. Minimap2 is >30 times faster than long-read mapping tools or cDNA mapping tools and also possesses higher accuracy, surpassing most aligners specialized in a single type of alignment. Although both tools have been applied with success to the analysis of Nanopore data, Minimap was specifically developed for mapping long reads while Centrifuge was conceived for a more general purpose (mapping against full genomes databases) in metagenomic analyses. However, in terms of parameter setting and configuration, Centrifuge offers more variety of modules and versatility, which could result in a more reliable taxonomic assignment.

Other tools such as BLASTN, MEGABLAST and LASTZ [52,50] have also applied for taxonomic assignment in metabarcoding studies using Illumina sequencing. Nevertheless, it is important to highlight that due to the differences between Nanopore and Illumina reads in terms of longer and poorer quality resulting from the presence of insertions and deletions on sequences, many of these standardized bioinformatics tools and pipelines are not suitable to be used with Nanopore data. In this context, Magi et al [60,61] have made an assessment of alignment and mapping tools and concluded that mapping or aligning Nanopore reads against a database is particularly challenging due to the size, high number and non-uniform error profiles of these long sequences. This study also found that mapping and alignment tools such as LAST, BWA, BLASR, and MarginAlign, were inefficient to process Nanopore data and the outcomes of these analyses were deeply influenced by the sequence lengths, since longer sequences contained more errors [53,54,14,46]. Moreover, Centrifuge has been included as part of the pipeline for the analysis of nanopore sequences in the new tool MINDS [62]. Based on these studies, Centrifuge and Minimap2 have proven to be the most suitable tools to work with Nanopore data, and they could be considered the best choices at present.

In addition, a second critical aspect to consider in taxonomic assignment is the composition of the database, which generally has a strong influence on the percentage of sequences correctly assigned to different taxonomic levels [63,64]. To date, there are few curated databases available for microbial identification—the most frequently used for 16S studies SILVA [65], Greengenes [66], RDP [56], and NCBI [67]. SILVA database contains taxonomic information for the domains of Bacteria, Archaea, and Eukarya. It is based primarily on phylogenies for small subunit rRNAs (16S for prokaryotes and 18S for Eukarya) [64]. Their taxonomic hierarchy and rank are constructed according to Bergey's Taxonomic Outlines, List of Prokaryotic Names with Standing in Nomenclature (LPSN), and manual curation [68]. Greengenes is the most popular and widely used database, since it is the default database in the QIIME pipeline (http://qiime.org/index.html). It provides Bacterial and Archaeal taxonomy based on phylogenetic trees inferred from chimera-free, consistent multiple sequence alignments, but it has not been updated since May 2013. The NCBI taxonomy contains the names of all organisms associated with submissions to the NCBI sequence data bases. It is manually curated based on current systematic literature, and uses over 150 sources. It contains some duplicate names that represent different organisms. Each NCBI database node has a scientific name and may have some synonyms assigned to it. Is important to note that this has been the most used database in articles of MinION 16S sequences classification [57,51,59,53,52]. The RDP database is based on 16S rRNA sequences from Bacteria, Archaea, and Fungi (Eukarya). It contains16S rRNA sequences available from the International Nucleotide Sequence Database Collaboration (INSDC) database. Another new database is EzBiocloud, which is a species level resolution database made of 61 700 species/phylotypes, including 13 132 species/phylotypes with validly published names, and 62 362 whole-genome assemblies that were identified taxonomically at the genus, species, and subspecies levels [69].

Some authors have evaluated the differences in taxonomic assignment using these databases, [64] and showed that NCBI is the bigger one in terms of number of sequences, followed by SILVA, RDP and Greengenes, respectively. In addition, they found that Silva shares the most taxonomic units with NCBI, and that green genes is the less diverse data base. Moreover, only green genes and NCBI could get taxonomic assignment to the species level rank, while SILVA allows only genus as the lowest rank. Importantly, NCBI database is not curated for all the groups of microorganisms and may contain duplicated copies of 16S sequences, which can lead to a bias in taxonomic assignment by an overestimation because of the high number of some bacterial groups. An example of this is the high number of available sequences belonging to pathogenic bacterial groups given by the NCBI repository. Contrasting with clinical strains, sequences belonging to extreme environments still remain scarce in the NCBI database and may be underrepresented when a taxonomic assignment is carried out. More detailed guidelines for the selection of the database is provided by Park & Won 2018 [68].

A final consideration for the selection of tools is the format for output data, since they cannot be compatible with other bioinformatics tools applied for downstream analysis. This particularly relates to those tools performing statistical tests, and generating plots and comparative analyses of taxonomic profiles identified in samples. A detailed description of the different options and applications of the available tools for 16S metagenomic studies using Nanopore data are summarized in Table 2.

3.3. Constraints to move beyond taxonomic assignment with Nanopore sequencing data

Since most of the analytical tools for taxonomic assignment have been developed to be applied to Illumina data and cannot be used for Nanopore sequences, the potential benefits of using full-length 16S rRNA sequences has not been systematically explored. The deeper taxonomic resolution provided by the full

Table 2
Different tools used to analyze Nanopore 16S data in metabarcoding studies.

Analysis approach	Data processes included	Tools used for analysis	Taxonomic Data Base	Reference
Profiling of bacterial communities	Basecalling, Demultiplexing, adapters and barcode trimming, chimera removal, taxonomic assignment	Albacore V2.3.1, Porechop, Yacrd 0.3, Minimap, EPI2ME	NCBI and rrn database	[48]
In field metagenome bacterial community analysis	Basecalling, Demultiplexing, Taxonomic assignment, diversity analysis	Albacore v1.10, SilNTAX, usearch v10.0.240	Ribosomal Database Project	[51]
Rapid bacterial pathogens identification	Basecalling, human reads removal, bacterial reads taxonomic assignment	Albacore 2.2.4, TanTan v13, Minimap2, R	GenomeSync database, NCBI database	[52]
Monitoring microbial of an anaerobic digestion system	Basecalling, Demultiplexing, adapter trimming, Taxonomic assignment	Metrichor, EPI2ME, poRe, Porechop, QIIME, BLAST,	GreenGenes database	[49]
Microbiome characterization	Basecalling, OTU picking, taxonomy assignment.	Metrichor v2.42.2, Poretools, QIIME 1.9. RDP classifier, BLASTn	GreenGenes database	[50]
Microbiome amplicon sequencing workflow	Bassecalling, alignment, re-orientation of reads, de-novo clustering, chimera removal,	Fast5-to-fastq, seqtk, INC-Seq, blastn, Graphmap, POA, chopSeq, nanoClust, R	No taxonomic assignment	[81]

16S gene sequence can reach the genus and species level with higher specificity than other approaches, [68–70]. This methodology has been applied with success in clinical, forensic and quality control of industrial processes where many of the microorganisms to be identified are well represented in databases due to their medical/human relevance [29,61].

However, taxonomic assignment is not always the best approach in other ecological contexts where the microbial community has not been previously studied. In these circumstances, the most representative microorganisms living in these habitats may remain unexplored and consequently their genomic data are not present in databases, which makes the taxonomic identification for many of the reads impossible. This situation is probably even more critical working with Nanopore data, since databases are predominantly composed by fragments of the 16S rRNA gene and presence of full-length sequences is frequently the exception and not the rule, limiting a reliable taxonomic identification based on the full sequence of the gene. On the other hand, the presence of a large number of reads without taxonomic assignment has a direct impact in providing a realistic measure of the biological diversity in the sample, leading to an underestimation of the real number of species. In this context, and as described in section 2, to overcome these limitations and the bias induced by a direct taxonomic assignment of reads, approaches such as Operational

Table 3
Bioinformatic tools for 16S rRNA metabarcoding Nanopore data.

Process	Tool	Input file	Programming languages	Available from	Reference
Basecalling	Albacore	Fast5	Python	https://nanoporetech.com/	ONT
	Guppy	Fast5	Python	https://nanoporetech.com/	ONT
	Deep Nano	fast5	Python	https://bitbucket.org/vboza/deepnano	[77]
	Chiron	Fast5	Python	https://github.com/haotianteng/Chiron	[78]
Sequencing report	NanoPlot	fastq, fasta, sequencing_summary (Albacore or guppy basecaller)	Python	https://github.com/wdecoster/NanoPlot	[82]
	pOre	fastq, fasta	R	https://sourceforge.net/projects/rpore/files/	[83]
	pauvre	fastq		https://github.com/conchoecia/pauvre	Github
	poretools	fastq, fast5	Python	https://github.com/arq5x/poretools	[84]
Demultiplexing	Albacore	Fast5	Python	https://nanoporetech.com/	ONT
	qcat	fastq	Python	https://github.com/nanoporetech/qcat	Github
	porechop	fastq, fasta	C++, Python	https://github.com/rrwick/Porechop	Github
Filtering and trimming	NanoFilt	fastq	Python	https://github.com/wdecoster/nanofilt	[82]
	Filtlong	fastq	C++, Python	https://github.com/rrwick/Filtlong	Github
	Porechop	fastq	C++, Python	https://github.com/rrwick/Porechop	Github
Taxonomic assignment	Minimap2	fastq, fasta	C++, Python	https://github.com/lh3/minimap2	[54]
	Wimp	fastq	Cloud-based	https://nanoporetech.com/	ONT
	Centrifuge	fastq, fasta	g++	https://ccb.jhu.edu/software/centrifuge	[58]
	LASTZ	fastq, fasta	g++, python	https://github.com/lastz/lastz	Github
Clustering	NanoClust	USEARCH/VSEARCH format	Python	https://github.com/umerijaz/nanopore/blob/master/nanoCLUST.py	[81]
	CARNAC-LR	paf	C++, Python	https://github.com/kamimrcht/CARNAC-LR	[80]
Data exploration	Pavian	Kraken and MetaPhlan formats	R	https://github.com/fbreitwieser/pavian	[85]
	PHINCH	biom	Cloud-based	https://github.com/PitchInteractiveInc/Phinch	[86]
	Krona	Krona format	–	https://github.com/marbl/Krona/wiki	[87]
	MEGAN6	OTU table	–	http://ab.inf.uni-tuebingen.de/software/megan6/	[88]
	Microbiome Analyst	OTU table, taxonomy table	Cloud-based	https://www.microbiomeanalyst.ca/	[89]

Taxonomic units (OTU) picking and/or denoising pipelines are commonly used for 16S Illumina data analysis [71–73] Both OTU picking and ASV analyses reduce the duplication and error of representative sequences and allow the analysis of bacterial groups without a database limitation, which allows for a more reliable taxonomic assignment resulting in a more robust definition of microbial communities (Table 3).

These analyses need to be performed in order to execute a taxonomic assignment and diversity analysis (Fig. 3). As described previously, tools such as DADA2 and Deblur are the most commonly applied in Illumina sequencing pipelines. However, because of the particular characteristics of Nanopore 16S reads (length and quality), the use of DADA2 and Deblur or any other algorithm based on ASV detection, has not as of yet been viable for Nanopore data. The number of errors—mainly insertions/deletions—typically introduced through the Nanopore sequencing, represent an extraordinary limitation to finding similarity between reads. Furthermore, the artificial divergence in sequences caused by the poor quality of reads, even when they come from a single organism, can produce the effect that each read is identified as a single sequence variant, leading to an overestimation of bacterial diversity [71]. As a consequence, the analysis of Nanopore reads with inappropriate OTU clustering tools or using an ASV approach could provide a completely incorrect picture of the microbial diversity of the sample showing a dataset with very divergent sequences.

Therefore, although the ASV approach is the most complete way to assess bacterial diversity, it is impracticable for Nanopore data analysis, with the only option available being the application of an OTUs-based clustering approach. However, similar limitations to the ones identified using ASV can be found when the most popular clustering algorithms are applied [74], such as UCLUST [75], VSEARCH [55] or CDHIT [76]. The use of the popular pipeline QIIME to analyze Nanopore 16S sequences was assessed in a recent study [50], indicating that the tool failed at the step of OTU picking, which corroborates the aforementioned issue of applying tools designed for Illumina to Nanopore data. By performing a close or open reference OTU clustering, only a small fraction of the data would be clustered and the main proportion of a dataset will be composed of singletons, which cause an erroneous overestimation of the bacterial diversity in the samples.

As previously mentioned, read quality is one of the most important constraints for nanopore data analysis. Basecalling is the most determinant process for the improvement of sequence quality. Nanopore sequencing is based on the detection of changes in electric currents produced by the passing of DNA strands through a nanopore. Each base ideally should have a specific current variation, called an event. Each event is summarized by the mean and variance of the current and by the event duration [77,51]. Translation of this event into a DNA sequence is known as the basecalling process. Original basecallers of ONT used Hidden Markov Models (HMM), however nowadays new strategies based on the use of machine learning are applied in all modern nanopore sequences basecallers, such as Guppy, DeepNano, and Chiron [77,78]. This machine learning-based basecallers use neural networks that can be trained with real sequencing data. The use of machine learning approaches has been shown to be effective for improving the quality of nanopore sequencing data and limiting the impact of base modifications, insertions, and deletions commonly present in raw data [79]. Therefore, the use of these new approach of machine learning on nanopore data has been crucial for the sequence quality improvement and in the short term will probably allow the necessary improvement of nanopore sequences to go beyond the taxonomic assignment of 16S sequences.

A final and important point to be considered is the difference in the orientations of reads produced by Illumina and Nanopore sequencing technologies. With Illumina, read orientation is defined from the beginning of sequencing and therefore sequences are all in the same orientation, which greatly facilitates bioinformatic data analysis. This homogeneity in the sequencing data is essential for alignment and clustering because reads can be compared more easily. On the other hand, with the 1D sequencing chemistry of Nanopore, adapters can be ligated to one or both ends of the DNA template [71] and DNA strands are sequenced in random orientations. Consequently, after the basecalling process the dataset is composed by a mix of forward and reverse sequences that are not complementary to each other. Hence, it may be critical to incorporate an additional step to evaluate the orientation of reads prior to the analysis of Nanopore data in order to reach consistent results.

According to the points discussed in previous sections relating to the availability of tools and their applications for working with Nanopore sequences, a workflow for 16S rRNA data analysis is proposed in the Fig. 3.

4. Summary and outlook

With the advent of modern technologies for sequencing, microbial ecology studies based on the analysis of the microbial 16S rRNA gene have become one of the most popular techniques in metabarcoding studies. Most of the studies conducted to date using Nanopore sequences report pipelines applied with a narrow scope, typically using a specific bioinformatic protocol to detect a particular pathogen or a target bacterial group or taxon, without considering the analysis of the whole microbial community present in the sample. However, most of the current aligners, clustering algorithms, and tools cannot process Nanopore data [74] and this remains a challenge to performing a more comprehensive analysis of Nanopore 16S rRNA data.

Due to the potential bias introduced by taxonomic assignment, OTU clustering may represent a more convenient alternative. In this regard, the new tools developed for transcriptomic de-novo clustering could represent an alternative to explore in the future [66,67]. As several transcriptomic based studies have been carried out with Nanopore, a possible alternative would be to apply these varieties of tools for *de-novo* clustering of all the transcripts originating from a single gene, and apply the same strategy to group all the variants of the 16S gene in a sample. Moreover, some of these tools have been developed to deal with the particular features of the Nanopore sequences and, therefore, can be used as a first approach to implement a specific clustering tool for 16S sequences from Nanopore.

Finally, many challenges for data analysis have surfaced since the development of the new sequencing technologies. The correct use of available tools has contributed to extending the use of 16S data from Nanopore for a first evaluation of the microbial composition. For Nanopore, efforts have been primarily focused on designing tools for basecalling, demultiplexing, and taxonomic assignment, according to the demand of consumers and end-users of this technology. Certainly, we are still in the first stages of the genomic revolution and the future will bring new possibilities for the expansion of these technologies and development of a new generation of powerful bioinformatic tools. The best parameters concerning the identity, alignment, and database choice must also be evaluated for each dataset in particular if the identification at the species level is required. The 2019 release by ONT of the new version (R10) of the flowcell with a new chemistry, will offer a substantial improvement in quality and quantity of data, with a consensus accuracy reaching 99% and an output of 50 Gb. All these developments in Nanopore outputs will generate new challenges for bioinformatic analysis, but will also bring new opportunities to revolutionize microbial ecology studies.

CRediT authorship contribution statement

Andres Santos: Writing - original draft. **Ronny van Aerle:** Writing - review & editing. **Leticia Barrientos:** Writing - review & editing. **Jaime Martinez-Urtaza:** Writing - review & editing.

Acknowledgements

Andres Santos work was supported by the grants CONICYT-Doctorado Nacional-2017-21171392; Universidad de La Frontera CD-FRO1204; Network for Extreme Environments Research (NXR17-0003); DI 19-0079 and Cefas Seedcorn.

We thank Judith Hoffman from Northern Light Translations for English revision of this review.

References

[1] Zhou J, He Z, Yang Y, Deng Y, Tringe SG, Alvarez-Cohen L. High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. MBio 2015;6. https://doi.org/10.1128/mBio.02288-14.
[2] Levin SA. Fundamental questions in biology. PLoS Biol 2006;4:e300.
[3] Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: Lessons learned from the uncultivated majority. Curr Opin Microbiol 2016. https://doi.org/10.1016/j.mib.2016.04.020.
[4] Dubnau D, Smith I, Morell P, Marmur J. Gene conservation in Bacillus species. I. Conserved genetic and nucleic acid base sequence homologies. Proc Natl Acad Sci 1965. https://doi.org/10.1073/pnas.54.2.491.
[5] Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 1977.
[6] Amit Roy SR. Molecular Markers in Phylogenetic Studies-A Review. J Phylogenetics Evol Biol 2014. https://doi.org/10.4172/2329-9002.1000131.
[7] Gutell RR, Gray MW, Schnare MN. A compilation of large subunit (23S and 23S-like) ribosomal RNA structures: 1993. Nucleic Acids Res 1993. https://doi.org/10.1093/nar/21.13.3055.
[8] Clarridge JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 2004. https://doi.org/10.1128/CMR.17.4.840-862.2004.
[9] Gray MW, Sankoff D, Cedergren RJ. On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. Nucleic Acids Res 1984. https://doi.org/10.1093/nar/12.14.5837.
[10] Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS ONE 2014. https://doi.org/10.1371/journal.pone.0093827.
[11] Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 2007. https://doi.org/10.1128/JCM.01228-07.
[12] Pace NR. A molecular view of microbial diversity and the biosphere. Science 1997;276:734–40.
[13] Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI. The effect of 16s rRNA region choice on bacterial community metabarcoding results. Sci Data 2019. https://doi.org/10.1038/sdata.2019.7.
[14] Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol 2014. https://doi.org/10.1111/1462-2920.12250.
[15] Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet 2016. https://doi.org/10.1038/nrg.2016.49.
[16] A framework for human microbiome research. Nature 2012;486:215–21. doi:10.1038/nature11209.
[17] Gilbert JA, Jansson JK, Knight R. Earth microbiome project and global systems biology. MSystems 2018;3:e00217–e317. https://doi.org/10.1128/mSystems.00217-17.
[18] Tighe S, Afshinnekoo E, Rock TM, McGrath K, Alexander N, McIntyre A, et al. Genomic methods and microbiological technologies for profiling novel and extreme environments for the extreme microbiome project (XMP). J Biomol Tech 2017;28:31–9. https://doi.org/10.7171/jbt.17-2801-004.
[19] Simon C, Daniel R. Metagenomic analyses: past and future trends. Appl Environ Microbiol 2011. https://doi.org/10.1128/aem.02345-10.
[20] 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;7:335–6. https://doi.org/10.1038/nmeth.f.303.
[21] 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. https://doi.org/10.1128/AEM.01541-09.
[22] McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 2013;8:. https://doi.org/10.1371/journal.pone.0061217e61217.
[23] Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. Nat Biotechnol 2009. https://doi.org/10.1038/nbt.1561.
[24] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science (80-) 2009. https://doi.org/10.1126/science.1162986.
[25] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genom 2012. https://doi.org/10.1186/1471-2164-13-341.
[26] Mosher JJ, Bowman B, Bernberg EL, Shevchenko O, Kan J, Korlach J, et al. Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. J Microbiol Methods 2014. https://doi.org/10.1016/j.mimet.2014.06.012.
[27] Myer PR, Kim MS, Freetly HC, Smith TPL. Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers. J Microbiol Methods 2016. https://doi.org/10.1016/j.mimet.2016.06.004.
[28] Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. Genome Biol 2016. https://doi.org/10.1186/s13059-016-1103-0l.
[29] Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol 2016. https://doi.org/10.1038/nbt.3423.
[30] van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet 2018. https://doi.org/10.1016/j.tig.2018.05.008.
[31] Feng Y, Zhang Y, Ying C, Wang D, Du C. Nanopore-based fourth-generation DNA sequencing technology. Genom Proteom Bioinf 2015. https://doi.org/10.1016/j.gpb.2015.01.009.
[32] Kono N, Arakawa K. Nanopore sequencing: review of potential applications in functional genomics. Dev Growth Differ 2019. https://doi.org/10.1111/dgd.12608.
[33] Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nat Methods 2015. https://doi.org/10.1038/nmeth.3290.
[34] Schneider GF, Dekker C. DNA sequencing with nanopores. Nat Biotechnol 2012;30:326.
[35] Plesivkova D, Richards R, Harbison S. A review of the potential of the MinION™ single-molecule sequencing system for forensic applications. Wiley Interdiscip Rev Forensic Sci 2019;1:. https://doi.org/10.1002/wfs2.1323e1323.
[36] Bahram M, Anslan S, Hildebrand F, Bork P, Tedersoo L. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment. Environ Microbiol Rep 2018. https://doi.org/10.1111/1758-2229.12684.
[37] Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R. PrimerProspector: De novo design and taxonomic analysis of barcoded polymerase chain reaction primers. Bioinformatics 2011. https://doi.org/10.1093/bioinformatics/btr087.
[38] Kerkhof LJ, Dillon KP, Häggblom MM, McGuinness LR. Profiling bacterial communities by MinION sequencing of ribosomal operons. Microbiome 2017;5:116. https://doi.org/10.1186/s40168-017-0336-9.
[39] Pollock J, Glendinning L, Wisedchanwet T, Watson M. The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. Appl Environ Microbiol 2018;84. https://doi.org/10.1128/AEM.02627-17.
[40] Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, et al. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet 2012. https://doi.org/10.1038/nrg3129.
[41] Tedersoo L, Tooming-Klunderud A, Anslan S. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytol 2018. https://doi.org/10.1111/nph.14776.
[42] Bhyan SB, Wee Y, Zhao M, Liu Y, Lu J, Li X. The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. Brief Funct Genomics 2018;18:1–12. https://doi.org/10.1093/bfgp/ely037.
[43] Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR, et al. Evaluation of Oxford Nanopore's MinION sequencing device for microbial whole genome sequencing applications. Sci Rep 2018. https://doi.org/10.1038/s41598-018-29334-5.
[44] McNaughton AL, Roberts HE, Bonsall D, de Cesare M, Mokaya J, Lumley SF, et al. Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). Sci Rep 2019. https://doi.org/10.1038/s41598-019-43524-9.
[45] Prazsák I, Moldován N, Balázs Z, Tombácz D, Megyeri K, Szucs A, et al. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. BMC Genom 2018. https://doi.org/10.1186/s12864-018-5267-8.
[46] Jenjaroenpun P, Wongsurawat T, Pereira R, Patumcharoenpol P, Ussery DW, Nielsen J, et al. Complete genomic and transcriptional landscape analysis using third-generation sequencing: A case study of Saccharomyces cerevisiae CEN.PK113-7D. Nucleic Acids Res 2018. https://doi.org/10.1093/nar/gky014.
[47] Seki M, Katsumata E, Suzuki A, Sereewattanawoot S, Sakamoto Y, Mizushima-Sugano J, et al. Evaluation and application of RNA-Seq by MinION. DNA Res 2019. https://doi.org/10.1093/dnares/dsy038.

[48] Cusco A, Vines J, D'Andreano S, Riva F, Casellas J, Sanchez A, et al. Using MinION to characterize dog skin microbiota through full-length 16S rRNA gene sequencing approach. BioRxiv 2017.

[49] Hardegen J, Latorre-Perez A, Vilanova C, Gunther T, Porcar M, Luschnig O, et al. Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass. Bioresour Technol 2018;265:275–81. https://doi.org/10.1016/j.biortech.2018.06.005.

[50] Ma X, Stachler E, Bibby K. Evaluation of Oxford Nanopore MinION Sequencing for 16S rRNA Microbiome Characterization. BioRxiv 2017:99960. doi:10.1101/099960.

[51] Edwards A, Debbonaire AR, Nicholls SM, Rassner SME, Sattler B, Cook JM, et al. In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. BioRxiv 2019:73965. doi:10.1101/073965.

[52] Kai S, Matsuo Y, Nakagawa S, Kryukov K, Matsukawa S, Tanaka H, et al. Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION™ nanopore sequencer. FEBS Open Bio 2019;9:548–57. https://doi.org/10.1002/2211-5463.12590.

[53] Turner AD, Fenwick D, Powell A, Dhanji-Rapkova M, Ford C, Hatfield RG, et al. New invasive nemertean species (Cephalothrix Simula) in England with high levels of tetrodotoxin and a microbiome linked to toxin metabolism. Mar Drugs 2018;16:452. https://doi.org/10.3390/md16110452.

[54] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191.

[55] Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016;4:. https://doi.org/10.7717/peerj.2584e2584.

[56] Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 2014;42:D633–42. https://doi.org/10.1093/nar/gkt1244.

[57] Mitsuhashi S, Kryukov K, Nakagawa S, Takeuchi JS, Shiraishi Y, Asano K, et al. A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. Sci Rep 2017;7:5657. https://doi.org/10.1038/s41598-017-05772-5.

[58] Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 2016;26:1721–9. https://doi.org/10.1101/gr.210641.116.

[59] Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med 2015;7:99. https://doi.org/10.1186/s13073-015-0220-9.

[60] Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform 2018;19:1256–72. https://doi.org/10.1093/bib/bbx062.

[61] Magi A, Giusti B, Tattini L. Characterization of MinION nanopore data for resequencing analyses. Brief Bioinform 2017. https://doi.org/10.1093/bib/bbw077.

[62] Deshpande Reed, Sullivan Kerkhof, Beigel Wade. Offline next generation metagenomics sequence analysis using MinION detection Software (MINDS). Genes (Basel) 2019. https://doi.org/10.3390/genes10080578.

[63] Escobar-Zepeda A, Godoy-Lozano EE, Raggi L, Segovia L, Merino E, Gutiérrez-Rios RM, et al. Analysis of sequencing strategies and tools for taxonomic annotation: defining standards for progressive metagenomics. Sci Rep 2018. https://doi.org/10.1038/s41598-018-30515-5.

[64] Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare?. BMC Genomics 2017;18:114. https://doi.org/10.1186/s12864-017-3501-4.

[65] 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 2013;41:D590–6.

[66] McDonald D, Price MN, Goodrich J, Nawrocki EP, Desantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 2012. https://doi.org/10.1038/ismej.2011.139.

[67] Federhen S. The NCBI taxonomy database. Nucleic Acids Res 2012. https://doi.org/10.1093/nar/gkr1178.

[68] Park S-C, Won S. Evaluation of 16S rRNA Databases for Taxonomic Assignments Using Mock Community. Genomics Inform 2018;16:e24–e24. doi:10.5808/GI.2018.16.4.e24.

[69] Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 2017. https://doi.org/10.1099/ijsem.0.001755.

[70] Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, et al. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res 2019. https://doi.org/10.1093/nar/gkz569.

[71] Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, et al. Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes 06 Biological Sciences 0604 Genetics 06 Biological Sciences 0605 Microbiology. Microbiome 2018. https://doi.org/10.1186/s40168-018-0569-2.

[72] Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infect 2018. https://doi.org/10.1016/j.cmi.2017.10.013.

[73] Rosen MJ, Callahan BJ, Fisher DS, Holmes SP. Denoising PCR-amplified metagenome data. BMC Bioinf 2012. https://doi.org/10.1186/1471-2105-13-283.

[74] Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. Nat Rev Genet 2018. https://doi.org/10.1038/s41576-018-0003-4.

[75] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010. https://doi.org/10.1093/bioinformatics/btq461.

[76] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9. https://doi.org/10.1093/bioinformatics/btl158.

[77] Boza V, Brejova B, Vinar T. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLoS ONE 2017;12:. https://doi.org/10.1371/journal.pone.0178751e0178751.

[78] Hall MB, Cao MD, Duarte T, Teng H, Coin LJM, Wang S. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. Gigascience 2018;7. doi:10.1093/gigascience/giy037.

[79] Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol 2019. https://doi.org/10.1186/s13059-019-1727-y.

[80] Marchet C, Lecompte L, Da Silva C, Cruaud C, Aury J-M, Nicolas J, et al. De novo clustering of long reads by gene from transcriptomics data. Nucleic Acids Res 2018. https://doi.org/10.1093/nar/gky834.

[81] Calus ST, Ijaz UZ, Pinto AJ. NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. Gigascience 2018;7. doi:10.1093/gigascience/giy140.

[82] De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 2018;34:2666–9.

[83] Santoyo-Lopez J, Risse J, Gharbi K, Thomson M, Blaxter M, Watson M, et al. poRe: an R package for the visualization and analysis of nanopore sequencing data. Bioinformatics 2014;31:114–5. https://doi.org/10.1093/bioinformatics/btu590.

[84] Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics 2014;30:3399–401. https://doi.org/10.1093/bioinformatics/btu555.

[85] Breitwieser FP, Salzberg SL. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. BioRxiv 2016.

[86] Bik HM. Phinch: An interactive, exploratory data visualization framework for –Omic datasets. BioRxiv 2014.

[87] Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinf 2011;12:385. https://doi.org/10.1186/1471-2105-12-385.

[88] Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. Community edition - interactive exploration and analysis of large-scale microbiome sequencing. Data. PLOS Comput Biol 2016;12:e1004957.

[89] Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res 2017;45:W180–8. https://doi.org/10.1093/nar/gkx295.

[90] Hsieh TC, Ma KH, Chao A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods Ecol Evol 2016;7:1451–6. https://doi.org/10.1111/2041-210X.12613.

[91] Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB. Package vegan. R Packag Ver 2013.

[92] Krehenwinkel H, Pomerantz A, Henderson JB, Kennedy SR, Lim JY, Swamy V, et al. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. GigaScience 2019. https://doi.org/10.1093/gigascience/giz006.