Adriel Latorre Pérez

Optimizing Nanopore-based microbiome sequencing for characterizing biotechnologically-relevant ecosystems

Programa de Doctorado en Biomedicina y Biotecnología

Tesis Doctoral

Universitat de València

Programa de Doctorado en Biomedicina y Biotecnología

Optimizing Nanopore-based microbiome sequencing for characterizing biotechnologically-relevant ecosystems

Memoria presentada por Adriel Latorre Pérez, candidato al grado de Doctor por la Universitat de València

Codirectores: Dra. Cristina Vilanova Serrador y Dr. Manuel Porcar Miralles

Valencia, enero 2022

El Dr. MANUEL PORCAR MIRALLES, Investigador Indefinido Doctor de la Universitat de València (Instituto de Biología Integrativa de Sistemas I2SysBio, UV-CSIC), y la Dra. CRISTINA VILANOVA SERRADOR, Directora Científica de Darwin Bioprospecting Excellence S.L.:

AUTORIZAN la presentación de la memoria titulada "Optimizing Nanopore-based microbiome sequencing for characterizing biotechnologically-relevant ecosystems" y **CERTIFICAN** que los resultados que esta incluye fueron obtenidos bajo su codirección en el Instituto de Biología Integrativa de Sistemas (I2SysBio, UV-CSIC) y en Darwin Bioprospecting Excellence S.L., por ADRIEL LATORRE PÉREZ.

Y para que conste, firman el siguiente certificado.

Dr. Manuel Porcar Miralles (Director)	*Dra. Cristina Vilanova Serrador* (Directora)	*Dra. M.ª Pilar López García* (Tutora)

Paterna, a 10 de enero de 2022

Agradecimientos

Una vez escuché a un músico decir que una nota errónea tocada con insistencia y convicción se convierte en jazz. En mi opinión, esas mismas cualidades son las que se requieren para transformar una simple idea en una tesis y, al igual que un solista no puede brillar sin una buena banda detrás, un estudiante por sí solo no es nadie sin toda la gente que le acompaña y guía. Afortunadamente, he tenido el privilegio de coincidir con grandes personas que se han encargado de armonizar cada una de las notas de este largo solo que ha sido mi doctorado.

En primer lugar, quería agradecer a mis directores. Gracias, Manel: por tu pasión por la investigación, tu creatividad e innovación (tanto científica como eligiendo sobrenombres) y por haberme escuchado atentamente cada una de las veces que lo he necesitado, aunque te hablara de cosas ininteligibles. Gracias, Cristina: por ser una científica brillante, una mentora increíble y, en definitiva, un ejemplo a seguir en todos los aspectos. Gracias a los dos por apoyarme siempre y por sacar lo mejor de mí en todo momento. Es hora de salir del marsupio, aunque espero estar cerca de vosotros por muchos años más.

Gracias también a los dos por poner tanto empeño y cariño en impulsar DARWIN. Más que compañeros de trabajo, esta empresa me ha dado verdaderas amistades: Kristie, mi editora, guía espiritual y referente doctoral. Gracias por escucharme y ayudarme en tantas cosas. A Javi, mi Reviewer 2. Un gran maestro en lo científico y en lo personal. A Helena, mi compañera de bioprospección y de carretera, y a Marta, mi instructora durante mis pinitos en el wet lab. A Dani, Carmen y Alejandro, los número uno del pádel en Burjassot. Gracias en general a todos los que formáis o habéis formado parte de este equipo: Manu, Karla, Asier, María José, Laura, Morgane. ¡A por otros muchos años de éxitos (y celebraciones)!

A toda la gente del grupo de Manel en el I2SysBio. En especial a Esther y Àngela. Mis compis de doctorado: gracias por ser tan "boniques". Gracias también a Alba, porque trabajar contigo siempre es muy fácil. A Juli, una referencia intelectual, académica y gastronómica. Last but not least, thanks to Dr. Abendroth (a.k.a. the futbolín master) for teaching me all the secrets about biogas and for participating so actively in my PhD.

A todos los amigos que han estado presentes durante esta etapa de mi vida. Particularmente, gracias a PERANOIA por haberme regalado recuerdos y anécdotas impagables. A Pascual, porque aparte de ser un gran amigo, has contribuido (y mucho) en este trabajo. A mis amigos de Pinoso y de Valencia. Sois demasiados como para nombraros a todos, pero gracias por apoyarme en todo siempre y por hacerme sentir que tengo dos hogares increíbles.

En especial, gracias a mis padres, que sin tener las oportunidades académicas que he tenido yo, me han enseñado mejor que ninguna escuela el valor del conocimiento y la cultura. A mi hermano, el artista plástico de esta tesis. Gracias por ilustrarme con tu ejemplo, por aguantar todas mis "turras" científicas y por estar siempre tan cerca, aun estando tan lejos. No sería lo que soy si no fuera por haber nacido en una familia como la nuestra. Estoy muy orgulloso de vosotros.

A Irene. Gracias por acompañarme durante todo este camino y hacerme disfrutar de cada paso. No cambies nunca.

A todos vosotros y a los que no he mencionado aquí, pero formáis parte de mi vida de una u otra forma: ¡gracias!

Contents

General Introduction

1. A window to the microscopic life

Technology has always played a central role in human biological and cultural evolution [1]. Although the term 'biotechnology' was not introduced until 1919 by Karl Ereky [2], human beings had been already using, selecting, and taking advantage of other organisms for millennia. Indeed, the domestication of wild animals and crops could be considered examples of *avant la lettre* biotechnology. With the advent of the necessity for food storage, humans finally met microbial biotechnology in the form of fermentation [3]. This was not the first time that our species crossed paths with microorganisms, considering that it is likely that some pathogens such as *Helicobacter pylori* have accompanied humans since the speciation of *Homo sapiens* [4, 5]. Nevertheless, food fermentation is possibly the first precedent of the intentional use of microbial transformations to obtain value-added products. Despite the fact that knowledge about microbes has undoubtedly advanced since our ancestors intuitively discovered the enormous potential of these organisms, the *leitmotiv* of microbial biotechnology remains the same: to explore the available biological resources in order to find their practical applications. Under this scheme, microbial bioprospecting is the discipline that covers the first part of the process, that is, the search of biotechnologically-relevant microorganisms and their products from different environments. Fortunately, technical advances in multiple areas have led to the development of sophisticated tools that allow us to better understand how microscopic life thrives, thus facilitating the process of identifying and finding practical applications to the resources that microbes have to offer.

As microorganisms are invisible to the human eye, the discovery of the microbial world can be traced back to the invention of the microscope, and more accurately to the observations of Antony van Leeuwenhoek in 1674 [6]. In the following two centuries, knowledge about microbes advanced in a slow, but steady fashion until the second half of the 19th century, when discoveries of scientists such as Pasteur, Lister, Metchnikoff, Escherich and several others led to the emergence of microbiology as an essential scientific field [7]. In this environment, a new landmark in microbiology techniques was reached: Robert Koch demonstrated for the first time that bacteria could be isolated in pure cultures using artificial solid media [8]. The combination of microscopy and isolation techniques eventually opened the window to the microscopic life: humans were now able to see and handle -some- microorganisms. Indeed, these techniques -and their subsequent optimizations- were the basis for the great advances that took place during the 20th century (e.g., antibiotic discovery, pathogen description, industrial fermentation…). However, microbiology was soon challenged again by a new limitation: most of the microbes proved virtually uncultivable [9, 10]. With that perspective, it became evident that new methods were necessary to access and characterize this unexplored microbial diversity, the study of which was about to be revolutionized by nucleic acid sequencing.

2. DNA sequencing: a new game changer

Molecular biology techniques, such as DNA/RNA hybridization, recombination or polymerase chain reaction (PCR), in combination with traditional methods (i.e., culture and microscopy) can be considered the cornerstones of modern microbiology [11,12]. The development of DNA sequencing, indeed, constituted a clear turning point for expanding the boundaries of the knowledge about microbial diversity. The continuous revision and growth of the tree of life [13,14] or the isolation and characterization of previously uncultured bacteria based on genomic data [15] are only a few examples of the great achievements accomplished thanks to sequencing technologies. The following subsections provide a brief description of the history of DNA sequencing, evaluating its impact on microbiology and comparing the differences between the three generations of sequencing technologies **(Table GI.1)**.

2.1. Long story short: the way to DNA sequencing

Nowadays, DNA sequencing is the most commonly used technique in genomics. However, early efforts to ‘read’ nucleic acids focused on RNA. This can be explained by three main reasons [16]:

1. Ribosomal RNA (rRNA), transference RNA (tRNA) or viral RNA (e.g., bacteriophages) were relatively easy to purify and they could be produced in microbial cultures.
2. As opposed to DNA, the RNA molecules mentioned above were single-stranded.
3. Several ribonucleases were known and routinely used at that time.

The first RNA sequences were obtained from microorganisms. Specifically, the techniques introduced by Holley et al. (1964) [17] and Sanger et al. (1965) [18] led to the description of the alanine tRNA from *Saccharomyces cerevisiae* by the former [19], and the low molecular weight rRNA of *Escherichia coli* by the latter [20]. These methods were soon adapted and further developed to study DNA sequences [16], until the emergence of polyacrylamide gel electrophoresis (PAGE) fostered the design of more sophisticated protocols: the plus and minus system [21] and the chemical cleavage technique [22].

The importance of these procedures is often underestimated, as they were overshadowed by the DNA sequencing technique based on chain-terminating inhibitors proposed by Sanger et al. (1977) [23], and described in the next subsection. Nevertheless, it must be noted that the demonstration of the use of 16S rRNA for establishing phylogenetic reconstructions by Carl Woese and collaborators [24, 25], which is one of the biggest breakthroughs in microbial ecology, taxonomy and evolution, was achieved before this technique appeared [26].

2.2. The Sanger sequencing revolution

Frederick Sanger was one of the most prolific scientist in the early days of DNA sequencing. Although he conceived several protocols to read DNA and RNA molecules, nowadays the term ‘Sanger sequencing’ refers to a very specific method: the chain termination -or dydeoxi- sequencing [23]. This technology is based on the use of dideoxynucleotides (ddNTPs), DNA polymerase, DNA primers and deoxynucleotides (dNTPs) [27]. DNA extension occurs normally using dNTPs until a ddNTP gets randomly incorporated into the chain, which causes the termination of the polymerization. At the end, a pool of oligonucleotides of different lengths is obtained and can be resolved by PAGE. Initially, four different reactions -and electrophoresis lanes- were needed, one per each ddNTP used. This changed with the introduction of fluorescently labeled ddNTPS which, in combination with capillary electrophoresis (CE), stimulated the development of automated sequencing devices (**Figure GI.1**) [16].

These first-generation sequencers were employed for extending the catalogue of available 16S rRNA gene sequences [28], including those extracted from uncultured organisms [29], thus broadening the molecular tree of life [30]. Other research groups used this technology to obtain complete genomic sequences and, as always, it all started with microorganisms. *Escherichia coli* K-12 was the first organism to be proposed. Nonetheless, by the time its genome was first published in 1997 [31], there were already six other complete genomic sequences available. Among them, the first bacterium (*Haemophilus influenzae*) [32], archaeon (*Methanococcus jannaschii*) [33] and fungus (*Saccharomyces cerevisiae*) [34] to be sequenced. Not long afterwards, the results of the most ambitious project to the date, the Human Genome Project (HGP), were finally presented [35,36].

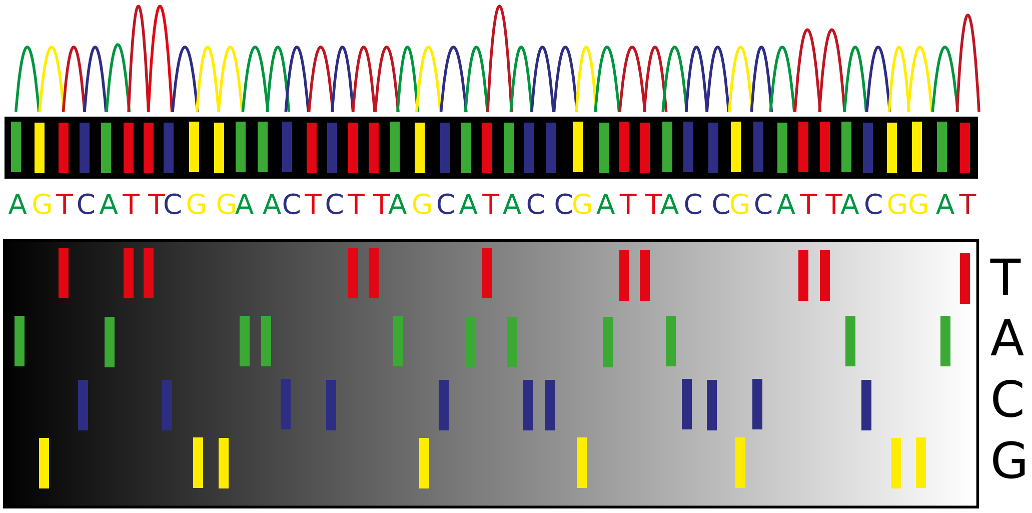


Figure GI.1 Overview of Sanger sequencing [23]. DNA elongation occurs normally, until a ddNTP (ddATP, ddTTP, ddCTP or ddGTP) is randomly incorporated to the chain, which stops the reaction. As a result, polynucleotides of different length are generated and can be separated using polyacrylamide gel electrophoresis (PAGE). Originally, four different reactions and PAGE lanes were used, one per each ddNTP (bottom panel). Each band indicates that elongation finished with this ddNTP. By 'reading' the bands, the sequence of nucleotides can be determined. Nowadays, ddNTPS are labeled with fluorescent dyes that emit light at different wavelengths. Hence, sequencing is performed in a single reaction and DNA fragments are read by coupling a laser to capillary electrophoresis (top panel).

It is worth highlighting that Sanger sequencing is still used routinely today, for instance to identify microbial isolates based on their 16S/18S rRNA gene sequences [37]. From the technical point of view, chain termination sequencing is characterized by its high accuracy (>99.9%) and moderate read length (~1,000 bp), at the cost of a relatively low throughput and a high price per base **(Table GI.1)** [38, 39]. In fact, despite the success of the first genome sequencing programs, it became obvious that the amount of infrastructure, time, money and personnel needed to retrieve the complete DNA sequence from a single organism was no longer sustainable [40]. A new revolution was needed.

2.3. The "-omics" era: Next-Generation Sequencing

Next-Generation Sequencing (NGS[1]) encompasses different technologies that were mainly developed at the turn of this century. These platforms are characterized by the mass parallelization of sequencing reactions, thus increasing the throughput of the sequencing run. For that reason, these technologies are also known as 'massively parallel sequencing', 'high-throughput sequencing', or 'second-generation sequencing'. NGS simplified the sequencing workflow and drastically reduced its costs. As a matter of fact, the HGP spent billions of dollars and several years to sequence a single human genome, while this task is currently accomplished in few days for only ~1.000 $ [41, 42], or even less [43]. These improvements enabled the expansion of different omics (i.e., genomics, epigenomics or transcriptomics), and they fostered the growth of other omic disciplines that are not based on nucleic acids (i.e., proteomics or metabolomics) [44].

The first NGS device was commercialized by 454 Life Sciences (later purchased by Roche). The early success of this technology was followed by other companies, which developed new sequencing strategies, such as sequencing

[1] Note that NGS will be strictly used as a synonym of second-generation sequencing throughout this thesis. Pacific BioSciences and Oxford Nanopore Technologies are therefore considered third-generation sequencing platforms.

Table GI.1. General overview of the most common sequencing technologies [45, 53–55, 57, 64] .

	Sanger	**Illumina**	**Pacbio[1] CLR**	**Pacbio[1] CCS**	**ONT[2]**
Generation	First	Second (or 'Next')	Third	Third	Third
Amplification	Often necessary	Necessary (i.e., bridge amplification)	Not necessary (single molecule)	Not necessary (single molecule)	Not necessary (single molecule)
Read length category	Moderate	Short	Long	Long	Long & Ultra-long
Direct epigenetic analysis	No	No	Yes	Yes	Yes
Direct RNA sequencing	No	No	No	No	Yes
Typical read length	400-900 bp	75-300 bp (x2)	25-50 Kbp	10-25 Kbp	10-100 Kbp[3]
Max. read length	1,000 bp	300 bp (x2)	>100 Kbp	>25 Kbp	No theoretical limit (record: 4.2 Mbp)
Read accuracy	>99.9%	~99.5-99.9%[4]	87-92%	>99% (up to ~99.9%)	87-98% (up to 99.3%[5] and 99,8%[6])[7]
Estimated cost[8]	444 €/Mbp[9]	9-56 €/Gbp[10]	12-173 €/ Gbp[10,11]	38-74 [3]/Gbp[10]	19-444 €/Gbp[10]
Max. throughput per run	96 Kbp[12]	1.2 Gbp[13] / 6 Tbp[14]	20-160 Gbp	35 Gbp	2.8 Gbp[15] / 14 Tbp[16]
Instrument Cost[17]	98,000 €[18]	19,000 / 900,000 €	350,000 / 500,000 €		900[19] / 268,000 €[16]
Portable sequencing	No	No	No		Yes[19]

1. PacBio: Pacific Biosciences; **2.** ONT: Oxford Nanopore Technologies; **3.** Depends on the sequencing kit used. Using the Ultra-Long DNA Sequencing Kit (cat. number SQK-ULK001) can increase the average read length to 50-100 Kbp; **4.** Depends on the instrument. Some of them reach an accuracy >99.9%. See Stoler & Nekrutenko **[47]**; **5.** Using the Q20+ early-access chemistry; **6.** Using the Q20+ early-access chemistry + 'Duplex' method; **7.** Data provided by ONT. See **[98]**; **8.** These estimations can greatly vary depending on the throughput of the run; **9.** Data from Frank et al. **[64]**; **10.** Data from Logsdon et al. **[55]**; 11. PacBio RS II is not considered; **12.** Considering a 96-capillary array and an average length of 1 Kbp; **13.** iSeq 100; **14.** NovaSeq 6000; **15.** Flongle; **16.** PromethION 48; **17.** These prices can greatly vary as they are often subjected to offers; **18.** For the Applied Biosystems ABI3730xl DNA Analyzer, price according to some vendors; (7 December 2021). **19.** MinION.

by oligonucleotide ligation and detection (SOLiD), Ion Torrent, or the combinatorial probe-anchor ligation (cPAL) developed by the Beijing Genomics Institute (BGI). How these technologies work and their impact on the sequencing market has been extensively discussed in the literature [16, 45, 46]. Nevertheless, there is a single company that has monopolized NGS over the last years, especially after Roche announced in 2013 the close down of 454. This company is Illumina.

Illumina sequencing is similar to Sanger's dydeoxi method, since it is also based on the use of modified dNTPs that prevent the elongation of the chain. In this case, however, the reaction is reversible, as the blocking group attached to the ribose 3'-OH can be removed. The four fluorescent and 3'-blocked dNTPs are provided at the same time and one of them will be added to the elongating strand. After the removal of the rest of dNTPs, fluorescence is measured and the incorporated dNTP is finally detected. Then, the blocking group is cleaved away and the process is iteratively repeated until the full oligonucleotide has been read [45]. These reactions occur simultaneously in millions of clusters, thus increasing the throughput of the device. Another characteristic of Illumina platforms is that sequencing is not produced from a single molecule. Instead, different pools of isothermally amplified oligonucleotides are used for detecting the incorporation of labeled dNTPs. Each pool comes from a single DNA fragment that is amplified within a single cluster in a process called bridge amplification [16, 45] (**Figure GI.2**).

Illumina offers a broad suite of instruments which output ranges from the 1.2 Gbp (4 million reads) delivered by the iSeq 100 device to the 6 Tbp (20 billion reads) that the NovaSeq 6000 machine can reach. Between these two models, and sorted by ascending output -and price-, Illumina's suite includes the MiniSeq, MiSeq, NextSeq and HiSeq series[2]. The reads produced by all these platforms are relatively short (25 – 300 bp; most common lengths: 150, 250 and 300 bp) and have an overall accuracy rate of 99.5-99.9% [45, 47] (**Table GI.1**). The most common error for Illumina sequencing is substitution [48], and its error profile is often considered as random. Nevertheless, systematic errors have been detected after homopolymeric regions [47], and under-representations of AT-rich and GC-rich regions have been also reported [49, 50]. Yet, Illumina's balance between throughput, cost and error profile is likely the best among sequencing technologies.

2.4. The end of a cycle? Third-Generation Sequencing

Illumina platforms have been the most widely used DNA sequencers for most applications during the last 8-10 years. Nevertheless, a new generation of sequencing technologies is becoming more and more popular. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are the companies behind these third-generation sequencers, which are mainly characterized by producing long reads (ranging from Kbp to Mbp) from single DNA molecules (i.e., without previous amplification) in real time [16, 45].

As the name of the company implies, ONT sequencing depends on the use of nanopores, which are engineered proteins that allow DNA and RNA molecules to pass through an electrically resistant membrane while changes in the electric current are measured. Thus, ONT sequencing is commonly called Nanopore sequencing. As this technology plays a central role in this thesis, a detailed explanation of Nanopore sequencing is given in **Section 4**.

On the other hand, PacBio single-molecule, real-time (SMRT) sequencing relies on a modified DNA polymerase that is attached to a circularized DNA molecule. This complex is then placed into microwells called Zero-Mode Waveguides (ZMWs) [51], and the DNA polymerase is immobilized at the bottom of the ZMW. Specifically, a single DNA molecule is placed in each ZMW. As DNA polymerase incorporates labeled nucleotides (one color per each dNTP), light is emitted and measured in real time. During the incorporation,

[2] Information obtained from the Illumina website: https://www.illumina.com/ (accessed 15 October 2021)

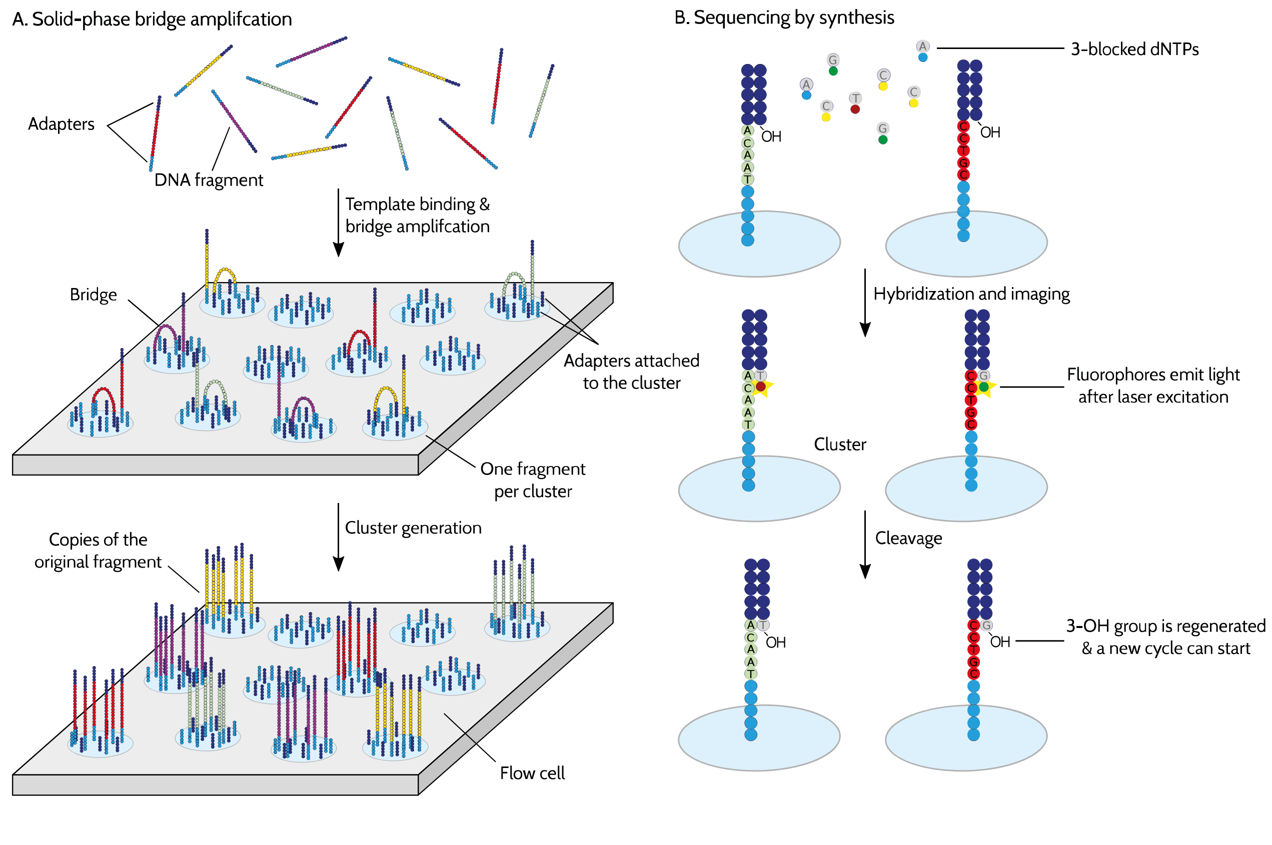


Figure GI.2. Overview of Illumina sequencing. **(A)** Solid-phase bridge amplification. Adapters (blue and dark blue) are ligated to the fragmented DNA. These adapters are complementary to the oligonucleotides attached to the flow cell. Therefore, free templates hybridize to the immobilized adapters. Then, bridge amplification begins. In this process, strands fold over and their adapters hybridize to another slide-bound oligonucleotide in the same cluster. Polymerases synthesize the complementary strand, creating a double-stranded bridge. After denaturation, two single-stranded copies of the original molecule are obtained. Bridge amplification is repeated over and over, and millions of clonal clusters are formed. **(B)** Sequencing by synthesis. Primers and fluorophore-labelled, terminally blocked dNTPs (3'-blocked dNTPs) are added. In each cycle, one 3'-blocked dNTPs is incorporated into the growing chain, based on the sequence of the template. A laser is used to excite the fluorophores and read which dNTP was incorporated in each cluster. For a given cluster, all the clones are read simultaneously. Then, dyes are cleaved and the 3'-OH is regenerated, thus allowing a new cycle to begin. The whole process is repeated until the entire DNA fragment is completely read. Adapted from Logsdon et al. **[55]** and Goodwin et al. **[45]**.

the fluorophore attached to the dNTP is cleaved by the DNA polymerase, allowing the fluorescent signal to vanish before the next dNTP is read [45, 52].

SMRT sequencing provides two alternative approaches: circular consensus sequencing (CCS) and continuous long read (CLR). In CCS mode, inserts of 10-25 Kbp can be read several times in the same ZMW, and a consensus sequence with up to 99.9% single-molecule read accuracy can be obtained [53,54]. Longer reads are generated with the CLR approach, where a DNA molecule is read a single time by the DNA polymerase [55]. Half of the CLR reads are above 50 Kbp in length[3],

[3] Information obtained from the PacBio website: https://www.pacb.com/ (accessed 16 October 2021)

at the cost of a higher error rate (ranging from 8% to 13%) [45, 55]. Regarding the throughput, CCS has an average output of 15-30 Gbp [53], in contrast to the 50–100 Gbp produced with the CLR mode [55] **(Figure GI.3)**.

Overall, third-generation sequencing (TGS) technologies have several advantages over Illumina platforms **(Table GI.1)**:

- **Read length**. Paired-end reads longer than 300 bp (2x300 bp) are difficult to obtain by Illumina sequencing mainly due to a phenomenon called 'phasing'. As detailed above, Illumina sequencing is based on the clonal amplification of a single sequence within a cluster. Phasing describes the process in which some sequences are out of phase with the rest of the cluster (e.g., a group of sequences incorporates a dATP, and the rest of the cluster incorporates a dCTP). This can occur in two different ways: (1) two or more dNTPs are incorporated in the same

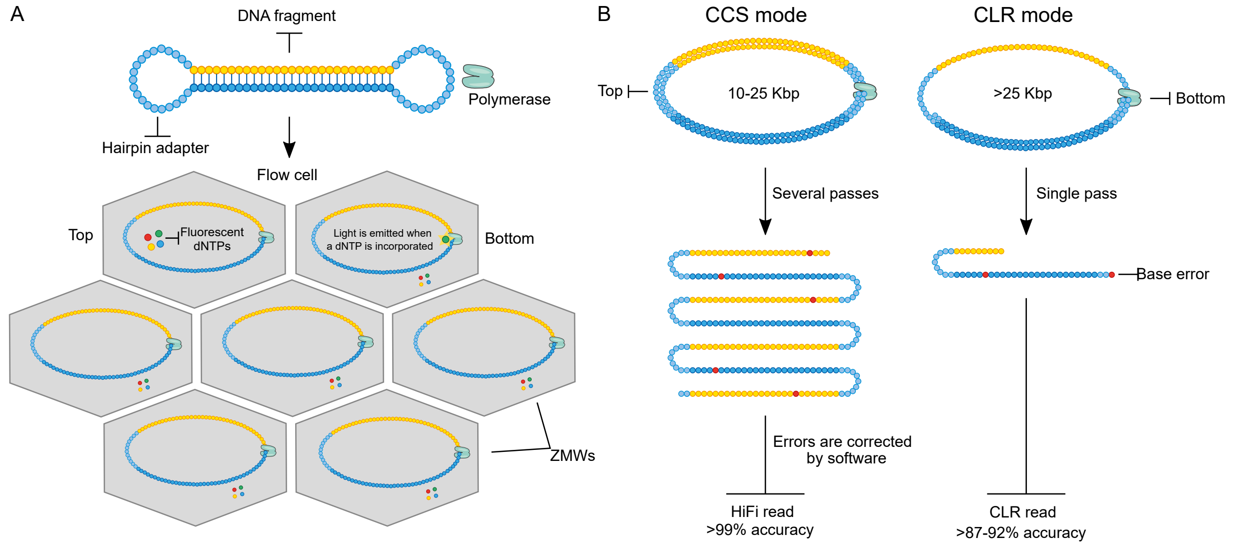


Figure GI.3. Overview of PacBio sequencing. **(A)** Hairpin adapters (light blue) are ligated to the forward (yellow) and reverse (blue) strands forming a circular molecule (SMRTbell). Polymerases are added to the library, which is loaded onto the PacBio instrument for sequencing. Each individual molecule of DNA is immobilized in a single microwell called Zero-Mode Waveguide (ZMW), where the sequencing takes place. Fluorescently labeled dNTPs are added to the ZMWs and the polymerase starts synthesizing a new strand. When a labeled dNTP is incorporated, the elongation is momentarily paused and a light pulse excites the fluorophore. The polymerase cleaves the fluorophore and the elongation continues. Hence, the DNA sequence is read in real time by a camera that records the color of emitted light. **(B)** Two different sequencing modes are available. In Continuous Long Read Sequencing (CLR), a long fragment of DNA (length: 25-50 Kbp; up to 100 Kbp) is read one or a few times, resulting in an accuracy of 87-92%. In Circular Consensus Sequencing (CCS), a shorter fragment (length: 10-25 kbp) is read several times. Then, a consensus sequence (HiFi read) is assembled from the reads, thus reducing the sequencing errors and increasing the accuracy up to 99.9%. Adapted from Logsdon et al. **[55]** and Goodwin et al. **[45]**.

cycle; (2) terminators bounded to dNTPs are not correctly removed, causing a lag in the synthesis of this individual clone [56]. Other technical limitations of Illumina sequencing include color or laser cross-talk, cross-talk between adjacent clusters and dimming (see [56] for a full explanation of these phenomena). Improvements introduced by TGS platforms, especially the fact that they rely on single-molecule sequencing, have led to an increase of read length.

- **No amplification is needed.** PCR is known to artificially introduce mutations and to have an amplification bias toward non-extreme GC content [57].
- **Reduced sequencing time**. In contrast to Illumina sequencing, where sequencing is paused after each base incorporation, in SMRT and ONT sequencing DNA is read in real time. Therefore, shorter sequencing runtimes are achieved in PacBio CLR mode [45], while in Nanopore sequencing reads are directly available for the analysis as they pass through the pore.
- **Direct epigenetic analysis**. Detecting base modifications with Illumina is possible only after inducing a chemical change in the DNA during the library preparation (i.e., bisulfite sequencing). On the other hand, both ONT and SMRT technologies can natively detect epigenetic modifications during the sequencing process without the need of previous steps [55, 57].

Despite these improvements, second-generation Illumina sequencing is still the most inexpensive way to obtain high-accuracy reads [55]. This can be explained by the general high throughput of Illumina instruments, and by the extensive implementation of this technology in large sequencing facilities. This allows to join samples from different costumers in a single run, thus optimizing the output and reducing the price per sample. Moreover, several strategies have been developed to generate synthetic long reads from short reads. Illumina synthetic long-read sequencing and 10X Genomics are examples of such approaches, and both rely on the same principle: individualizing long DNA molecules, obtaining barcoded short sequences from them, and reconstructing the original long fragments after sequencing by computational methods. These strategies have the same advantages and drawbacks associated with Illumina sequencing, but usually require additional equipment, higher throughput and longer run times [45, 57].

As described throughout this section, the current high-throughput sequencing market is diverse, with each sequencing generation having its own particularities which make them more or less suitable for different applications. In general, Illumina platforms are the gold standard of clinical and research sequencing [55], and the NGS era is far from over. TGS, however, is gaining ground in specific areas (e.g., genome/metagenome assembly, full-length RNA sequencing [58, 59], or epigenomics [60]) that are already having a considerable impact on microbiology [61–63].

3. Microbiome sequencing strategies

The term ‘microbiome’ was first defined by Whipps and colleagues in 1988, while they were studying the ecology of rhizosphere-associated microorganisms [65]. From that moment, many other authors have provided their own interpretations of the term. Recently, a panel of international experts have unified all these visions, proposing a new and standardized definition: “the microbiome is defined as a characteristic microbial community occupying a reasonable well-defined habitat which has distinct physio-chemical properties”, i.e., a biotope. “The microbiome not only refers to the microorganisms involved”, i.e., the microbiota, “but also encompass their theatre of activity”, i.e., genomes, metabolites, proteins, microbial structures, mobile genetic elements, etc., “which results in the formation of specific ecological niches. The microbiome, which forms a dynamic and interactive micro-ecosystem prone to change in time and scale, is integrated in macro-ecosystems including eukaryotic hosts, and here crucial for their functioning and health" [66]. Microbiomes can be studied from multiple perspectives and techniques, although sequencing-based methods are often preferred. In the following subsections, the most common strategies used for characterizing microbial communities are introduced. Moreover, a summary of the strengths and limitations of each sequencing-based technique is provided in **Table GI.2**.

3.1. Metataxonomics

Metataxonomic approaches rely on the amplification and sequencing of marker genes or loci, such as the 16S rRNA gene for prokaryotes, the 18S rRNA gene or the internal transcribed spacer fore fungi, or the cytochrome c oxidase subunit I (COI) for animals and other eukaryotes [66–69]. The marker region of choice must be conserved in the taxonomic group of interest, but it must include sufficient variations at the sequence to allow proper resolution. This is not a trivial decision, and even the most widely used marker gene (i.e., 16S rRNA gene) has some limitations. In this case, short-read sequencing fails to cover the full gene (~1.5 Kbp), so only a fraction of the sequence is analyzed (i.e., hypervariable regions). Despite the accuracy of NGS platforms, studying these regions allows for a robust identification of the microbiome at the genus level, but not at the species level [70]. TGS platforms are able to sequence the full-length 16S rRNA gene, but the error associated with these technologies still hampers the taxonomic resolution beyond the genus level [70, 71]. Nevertheless, improvements in TGS (i.e., PacBio CCS mode or sequencing the entire 16S-ITS-23S region of the *rrn* operon) have the potential to overcome this issue [72, 73].

3.2. Metagenomics

In metagenomics, total DNA is extracted from a sample and then sequenced, without the need of previously amplifying any marker region [74]. This process is also called shotgun metagenomic sequencing and it allows the direct detection of any taxonomic group. After gene prediction and annotation, the functional potential of the microbiome can be evaluated [75]. Taxonomic and functional profiles are obtained from metagenomic reads by directly comparing them against a database or by assembling and annotating the metagenomes. In the latter case, individual genomes can be recovered from the metagenomes in a process known as binning [76]. These metagenome-assembled genomes (MAGs) can be used to study the role of each microorganism within the microbial community. Overall, metagenomics is more informative than metataxonomics, but there are some, important exceptions to this rule. For instance, when studying the microbiome associated with a host (e.g., humans, plants...), metagenomic sequencing not only captures the microbial community, but also the host's genetic material, which can hamper the signal from microorganisms [77].

3.3. Metatranscriptomics

This strategy aims at studying the expression of RNA in a given sample, and it allows to retrieve a more accurate functional profile, as only the genes that are being transcribed in the microbial community are actually detected. This strategy can be used to analyze the dynamics of the microbiome and to detect metabolic functions and pathways that are activated or deactivated under certain stimuli. Metatranscriptomics is based on shotgun sequencing, so most of the advantages and limitations of metagenomic sequencing can be extended to this strategy. However, given that rRNA represents almost 85% of total RNA, rRNA depletion is necessary to capture the information about messenger RNA (mRNA) [78].

3.4. Other approaches

A microbial cell is a rather complex system whose components are regulated at different levels. The

Table GI.2. Approaches to microbiome sequencing [70, 78, 79].

Approach	Advantages	Limitations
Metataxonomics	- Cost-effective - Straightforward analysis - More complete databases - Targeted to a specific taxonomic group (i.e., avoids sequencing DNA derived from the host)	- Only the targeted taxonomic group is detected - No universal marker region for viruses - Limited taxonomic resolution (e.g., some bacterial species cannot be distinguished based on full-length 16S rRNA gene sequences) - PCR introduces an amplification bias that modifies the measured abundances of taxa with respect to the real ones - Intragenomic 16S gene copy variants are present in a significant proportion of bacterial taxa - No information about genomes or functions
Metagenomics	- Detects bacteria, archaea, viruses and eukaryotes in one experiment - Retrieves information about functions and other genetic elements - Allows the recovery of genomes directly from the metagenomes - Achieves higher taxonomic resolution (typically species- or strain-level)	- More sequencing depth is needed (i.e., more expensive) - Non-microbial DNA (i.e., host DNA) can hamper the detection of microorganisms - Databases are incomplete (i.e., some taxonomic groups are underrepresented) - It is difficult to calculate the abundance of each taxon, especially when comparing unrelated taxonomic groups (i.e., eukaryotes vs. prokaryotes) - Only the most abundant genomes can usually be assembled and recovered - Hampered by low DNA concentrations, as no PCR is performed
Meta-transcriptomics	- Retrieves information about active functions - Dynamic information - Detects multiple taxonomic groups in one experiment	- mRNA is unstable - Hampered by low RNA concentrations - rRNA depletion is necessary - Purification and amplification can add 'noise' to the results - These library preparation steps increase sequencing time and cost

complexity of a microbiome is even deeper, as various microorganisms are taking part in different transformation processes at the same time. Although widely accepted, sequencing-based approaches are not ideal for studying microbial communities, since the molecules of interest (i.e., peptides, proteins, metabolites...) are not directly detected, but inferred. In order to overcome this limitation, novel strategies have been developed: **metaproteomics** is focused on characterizing the whole protein content of a given sample, while **metametabolomics** attempts to analyze all the metabolites that are being produced by the microbiota. Both approaches depend on mass spectrometry and chromatography techniques, which are usually more expensive and experimentally challenging than sequencing [79].

4. Nanopore sequencing

The idea of Nanopore sequencing was first conceived in 1989 by Professor David Deamer, and its basis was conceptually very simple: if a protein channel (the 'nanopore') with the ability of translocating DNA molecules could be embedded into a lipid bilayer, then individual nucleotides would cause a blockade of ionic current when passing through the pore. This blockade should be specific for each nucleotide, as dNTPs differ in their chemical structure. [80]. Several years later, Kasianowicz et al. [81] proved that DNA and RNA molecules could actually be driven through a membrane channel, generating a decrease of ionic current the duration of which was proportional to polymer length. Further improvements introduced by the team led by Hagan Bayley, the co-founder of ONT, allowed the discrimination of DNA strands based on their sequence [82]. After the foundation of the company, which was first called Oxford Nanolabs, advances in Nanopore sequencing were mainly kept in secret, until the MinION sequencing device was finally launched for early access in 2014.

4.1. The anatomy of Nanopore sequencing

Nowadays, Nanopore sequencing relies on the same mechanism proposed by Professor Deamer back in 1989: a single-strand of DNA or RNA is passed through a protein nanopore, resulting in specific changes in the electric current depending on the nucleotide composition. Consequently, disruptions of the current can be measured and translated into DNA/RNA sequence data (**Figure GI.4**). ONT sequencing is based on several core components that have been actively optimized until reaching the current state of the art. These components are:

- **Nanopores:** protein-based or solid-state (synthetic) channels embedded in an electro-resistant membrane. Biological nanopores are currently preferred as they have a uniform pore size and conformation and they can be modified through protein engineering [83]. ONT products implement two different nanopore models: R9 and R10 (**Figure GI.4C**). R9 is the name given to the modified version of the *Escherichia coli* curli transport channel CsgG [84], while R10 refers to a new generation of dual-constriction nanopores with longer barrels that improve the resolution of homopolymeric regions [85]. Although the details of R10 design are not fully disclosed, it is likely that this nanopore is constructed by combining the CsgG channel with the accessory protein CsgF [86], as previously described [83, 87]. Regarding the membrane, lipid bilayers were rapidly replaced by electrically-resistant polymer membranes, which are more robust and stable [80].
- **Flow cells:** the hardware where the nanopores are placed. Flow cells are formed by an array of microscaffolds that include the membrane and the embedded nanopore. Each microscaffold has its own electrode connected to a sensor chip that is controlled by a bespoke Application-Specific Integrated Circuit (ASIC) [88].
- **Motor proteins**: molecules that control the speed at which polynucleotides pass through the nanopore. Motor proteins enable a better

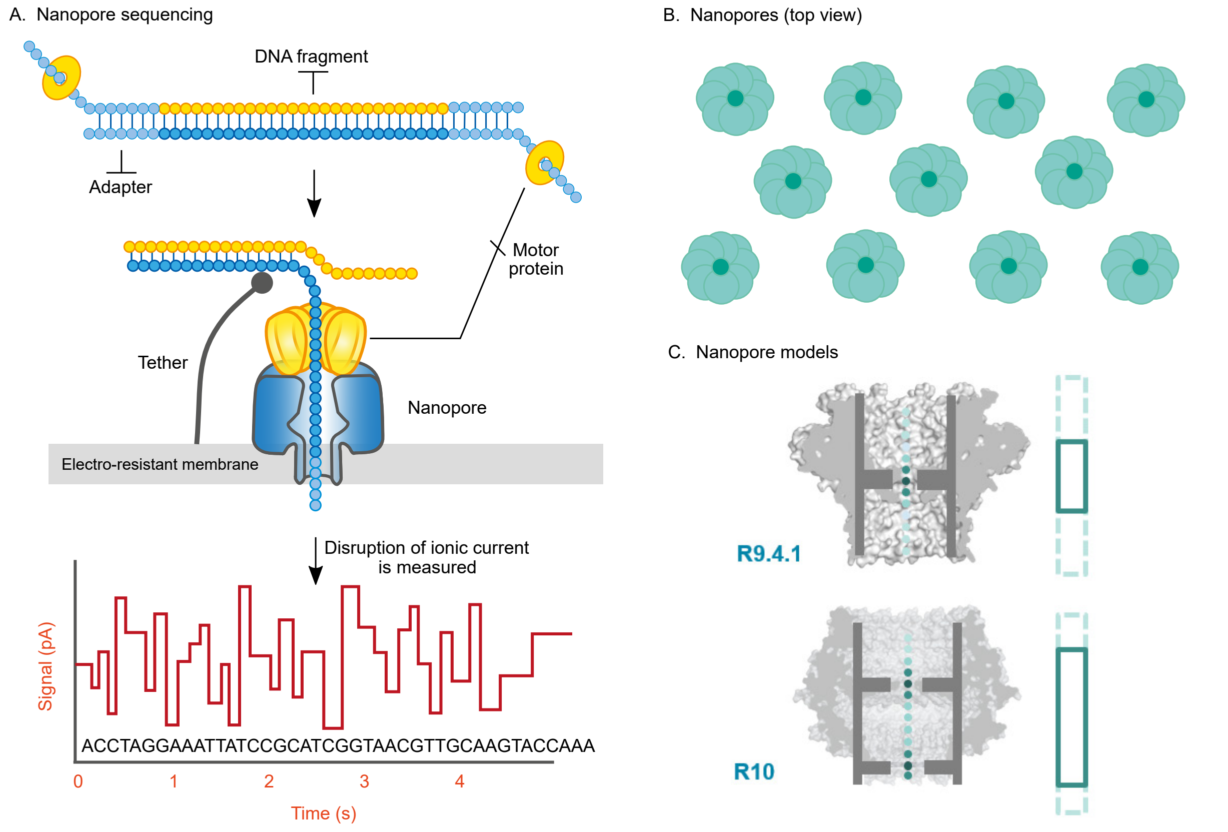


Figure GI.4. Overview of Nanopore sequencing. **(A)** Motor proteins and adapters are ligated to DNA fragments. Tether proteins help the molecules to find the nanopores, which are embedded in a synthetic membrane. Motor proteins control the translocation of DNA molecules. As DNA moves through the pore, it produces specific changes in the ionic current. Electric signals are recorded and converted into nucleotide sequences by basecallers. **(B)** Top view of the nanopores in the flow cell. **(C)** Two different nanopore models are available (R9.4.1 and R10). R10 pores have a longer barrel and a dual reader, thus improving the resolution of homopolymeric regions. Adapted from Logsdon et al. **[55]** (Panels A & B) and Oxford Nanopore Technologies web page **[85]** (Panel C).

resolution of individual nucleotides, as accuracy depends on translocation speed. Without these proteins, DNA/RNA molecules would cross the membrane so fast that sequence determination would not be possible [83]. A slower translocation speed increases the reading accuracy, but at the cost of decreased sequencing yield. For that reason, motor proteins are designed to meet a compromise between accuracy and throughput.

- **Basecaller:** a software that converts the electrical signals into nucleotide sequences. This is accomplished via recurrent neural networks (which have replaced algorithms based on hidden markov models) and machine learning. Different neural network structures and algorithms are constantly tested, and the models showing improved characteristics are then implemented into Guppy, the ONT production basecaller [89]. The basecalling algorithms 'learn' how to transform electrical signals into reads by using training datasets of known sequences. Default models are trained on a mixture of plant, animal, bacterial and viral genomes. Nevertheless, species-specific and modification-aware basecallers can be obtained by using custom datasets [89, 90].

- **Computer components**: central processing unit (CPU), graphics processing unit (GPU), RAM and disk storage. These components are necessary for data acquisition and storage. Basecalling models can be run on CPUs and GPUs, although the use of the latter accelerates the process [90].

4.2. Nanopore-based sequencers

ONT hase developed three different sequencing devices: MinION (two models: Mk1B and Mk1C), GridION, and PromethION (two models: P24 and P48) – sorted by increasing output and device's capital cost. All the sequencers but MinION Mk1B include the computer components in their hardware. In the case of MinION Mk1B, the hardware is managed by an external computer that needs to meet the IT requirements [91]. ONT sequencing devices differ in the number and type of flow cells that they can control at the same time and, in consequence, they also differ in the final throughput. MinION is a palm-sized, portable device which is compatible with a single flow cell. GridION relies on the same flow cells as MinION, but it can run up to five cells at the same time. PromethION can operate up to 24 (P24 model) and 48 (P48 model) flow cells at once. PromethION flow cells are platform-specific: they provide a higher number of available nanopores, thus increasing the sequencing yield per flow cell. Finally, Flongle is not a sequencer itself, but

Table GI.3. General overview of Oxford Nanopore Technologies sequencing platforms in 2021[1].

	Flongle[2]	MinION Mk1b/1c	GridION Mk1	PromethION 24/48
Flow cells per device	1	1	5	24/48
Channels per flow cell[3]	126	512	512	2,675
Guaranteed nanopores per flow cell	50	800	800	5,000
Theoretical Maximum Output per flow cell	2.8 Gbp	50 Gbp	50 Gbp	290 Gbp (record: 245 Gbp)
Storage	-	- / 1 TB SSD	4 TB SSD	32 / 64 TB SSD
RAM	-	- / 8 GB	64 GB	384 GB
Price per flow cell[4]	61.3 €[5]	810 – 430 €	810 – 430 €	1,800 – 562.5 €
Starter Pack Cost	1,320 €[6]	900 / 4,410 €	44,960 €	175,910 / 267,710 €

1. Information was obtained from the ONT website: https://nanoporetech.com/products (accessed 6 December 2021); **2.** Flongle is not a sequencing platform, but an adapter for MinION and GridION; **3.** Channels contain several pores; **4.** Flow cell prices depends on the number of flow cells ordered; **5.** Minimum order: 12 flow cells; **6.** It includes the adapter and 12 flow cells.

an adapter for the Flongle Flow Cells that can be used on MinION and GridION. This product is meant to sequence individual samples in inexpensive, single-use flow cells. The particularities of each ONT device have been summarized in **Table GI.3.**

4.3. Advantages and drawbacks

While the advantages and drawbacks of TGS platforms compared to NGS technologies were discussed in **subsection 2.4**, in this subsection the particularities of ONT sequencing will be analyzed.

- **Read length**. Nanopore sequencing holds the record for the longest DNA fragment sequenced to date: ~2.3 Mbp [92] (~4.2 Mbp if considering ONT internal tests [93]). In fact, there is no theoretical length limit in the molecules that can pass through the nanopore. The main factor that hampers the attainment of longer reads is the DNA extraction and library preparation protocols [55]. The average read length obtained by Nanopore sequencing ranges from 10 to 60 Kbp [55, 89], which is similar to the results from PacBio CLR mode.
- **Portability.** As Nanopore sequencing does not depend on optics, but rather on electronics, ONT devices can be significantly miniaturized. In contrast to the other sequencing platforms, the most widely used Nanopore-based sequencer, MinION, is easily portable and has thus been used for numerous in-field applications [94]. Moreover, ONT is developing SmidgION, a sequencing device that can be controlled from a smartphone [95].
- **Real real-time analysis.** In TGS platforms, sequencing is not paused after the detection of the bases (as in NGS), but the molecules are read in real time [55]. In the particular case of ONT platforms, sequences become available as they pass through the pore. Therefore, bioinformatic pipelines can be run in parallel to the sequencing run, and final results can be obtained in a very short time -a few minutes-, which is key for some applications (i.e., identification of pathogens, genes, mutations…) [96, 97].
- **Accuracy**. This is the main drawback of Nanopore sequencing. At this moment, the overall accuracy of ONT platforms depends on multiple factors (e.g., basecalling algorithm and model, type of sequencing, type of sample, etc.). For that reason, Nanopore sequencing error is estimated to range from 13% to 2% [55], which is far from the accuracy provided by Illumina or PacBio CCS. Nevertheless, errors are decreasing with improved basecalling models, even when using data generated years ago, and ONT now claims a raw read modal accuracy of 98.3% (>99.3% and >99.8% if using the Q20+ early-access chemistry and the 'Duplex' method, respectively) [98, 99]. The main problem of Nanopore sequencing is that errors are not uniformly distributed: the error rate systematically increases in homopolymers. This bias has also been reported in data produced by PacBio CLR mode [55]. Despite the fact that R10 pores increase the accuracy of homopolymers, the error rate is still higher than for other sequencing technologies (i.e., Illumina and PacBio CCS) [55, 89].
- **Cost.** Regarding this aspect, there are two main points to be discussed: (1) capital cost of the platforms; (2) cost per Gbp of data. ONT offers, by far, the most inexpensive and accessible sequencing platform on the market: the MinION (**Tables GI.1 & GI.3**). Nonetheless, the costs of sequencing derived from using this device are high compared to Illumina [55]. More cost-effective sequencing is obtained with PromethION, but the capital cost of acquiring this platform is substantially higher (**Table GI.3**). Although the price of PromethION is lower than Illumina's NovaSeq 6000, only large sequencing centers can afford this investment.

To summarize, Nanopore-based platforms have 'democratized' the access to sequencing, as almost any user can acquire and use ONT products (i.e., MinION or Flongle) for their own applications. However, low-cost sequencing will still depend -at least in the near future- on centralized sequencing facilities, that mainly rely on Illumina platforms at this moment.

- **Native RNA-seq and epigenomics.** Nanopore sequencing is the only technology that allows reading RNA fragments directly, avoiding the previous conversion of RNA to DNA [100]. Similar to PacBio, DNA -and RNA- base modifications can be also determined by ONT platforms. As modified bases produce a specific change in the electric current, basecallers can be trained to identify a particular modification, thus increasing the range of epigenetic markers that can be studied [101, 102].

4.4. The starting point for Nanopore-based microbiome sequencing

Soon after being released, several reports demonstrated the potential applications of MinION to microbiology and microbiome characterization [103, 104]. These studies were mainly proofs of concept and authors made use of available bioinformatic tools, although they were not specifically designed for Nanopore data. For example, Loman et al. [105] used the Celera assembler, which was published in 2000 [106], for reconstructing the genome of *Escherichia coli* K-12 from corrected Nanopore reads. Other researchers tuned the parameters of different alignment tools such as BLAST, LAST or Centrifuge in order to use them for microbial composition analyses and pathogen detection [75, 103, 107, 108]. Overall, the resources to analyze Nanopore data were very limited at the time this thesis was conceived: the first long-read assembler that supported ONT sequences (Canu) had just been released in 2017 [109], while the most popular alignment tool for long and error-prone reads (minimap2) had not even been published yet [110]. This was in clear contrast with the bioinformatic landscape for Illumina sequencing analysis, which was characterized by the availability of hundreds of technology-specific tools and pipelines. Not only that, but systematic benchmarks assessing the performance of different tools for specific microbiome applications had been already published. These works included evaluations for metagenome assembly and binning [111, 112], taxonomic and functional classification [113, 114] or taxonomic assignment of 16S rRNA gene sequences [115–117], among other approaches. All these results had been used to create best practices guidelines for analyzing microbiomes [118, 119]. Although novel tools that improve some aspects of the current bioinformatic protocols are continuously being released (see [120] for an example), the basis for short-read data analysis was -and is still- solid and unlikely to be drastically changed. On the contrary, the first successful applications of ONT platforms were certainly going to encourage the development of new software specifically designed to handle the particularities of this data. Therefore, the efforts needed to standardize Nanopore sequencing and data analysis while demonstrating the convenience of this technology for studying microbiomes could be anticipated. For that reason, the main motivation of the present work was to track new algorithms, implement them into pipelines, test their performance on metataxonomic (**Chapter I**) or metagenomic data (**Chapter II**), and use these novel analytical tools to characterize biotechnologically-relevant samples.

5. Nanopore sequencing from a business perspective

Considering that this thesis was carried out in the framework of an industrial doctorate programme, the benefits of optimizing Nanopore sequencing deserve to be evaluated, not only from the scientific point of view, but also from the business perspective. As discussed in previous sections, ONT platforms are in constant evolution, so developing and adapting state-of-the-art protocols requires significant investment in terms of money and human resources. On the other hand, a company able to lead the implementation and optimization of Nanopore sequencing can place itself in a privileged position within the market. This position can be used to obtain an economic return by two different mechanisms: (1) offering Nanopore sequencing as a service; (2) designing and selling applications that take advantage of the particularities of ONT platforms (i.e., portability, real-time analysis, etc.). The first option depends on the acquisition of equipment of high capital cost (i.e., PromethION) to become competitive, which is not necessarily compatible with the financial strategy of a start-up company. Conversely, the second option relies on creating and consolidating a know-how that can be valuable for other commercial partners interested on the technology but with a lack of technical knowledge. This alternative is more attractive for a small company, as the scope of Nanopore sequencing is almost unlimited. The following subsections describe a range of applications in which it is key to have portable, low-cost, fast, and robust technologies allowing an *in situ* analysis of samples[4].

5.1. Real-time analysis of clinical samples

Pathogen identification in hospitals is still mainly dependent on microbial cultures, which have several limitations regarding specificity, bias, sensitivity, and time to diagnosis. For instance, in the case of sepsis, patients are usually treated with broad-spectrum antibiotics until the first results of culture-based analysis (including determination of antibiotic susceptibility) are obtained 36–48 h later. In this context, MinION sequencing paves the way towards a diagnostic alternative in a clinically critical time frame, which could reduce the morbidity and mortality associated with major microbial infections.

The first reports on Nanopore sequencing in clinical diagnosis aimed at detecting pathogens during outbreaks. Flagship examples of such applications are the fast (<24 h) detection of Ebola virus during the 2015 outbreak in West Africa [121], or the fast (<6 h) phylogenomic analysis of *Salmonella* strains during a hospital outbreak [122]. More recently, ONT platforms have been used for sequencing SARS-CoV-2 genomes [123] and for developing novel diagnostic tests for COVID-19 [124]. Other significant efforts have focused on the fast identification of single clinical isolates [125], including the analysis of antibiotic resistance genes (ARGs) in less than 6 h [126, 127]. However, a range of applications in the clinical field requires the use of microbiome sequencing to unveil the identity of viral or microbial communities rather than single isolates. Greninger et al. [128] reported the detection of several viral pathogens in human blood in <6 h since obtaining the samples. A similar approach was reported for the rapid identification of mosquito-borne arbovirus [129], and other viruses causing co-infections, including dengue, from human serum samples [130].

An extensive number of reports have focused on the analysis of infections caused by bacterial communities. PCR-based approaches targeting the 16S rRNA gene proved the most rapid method to identify pathogenic agents from human samples, as they avoid sequencing DNA derived from the host, whose concentration may be overwhelming in some sample types. Using this approach, pathogen detection can be achieved in only 2 h in patients with pleural effusion [108] or with acute respiratory distress syndrome [131]. The use of human cell-free samples allows the application of metagenomic protocols for the analysis of the communities, yielding not only taxonomic information but also the identification of putative ARGs, which are of outstanding relevance for the selection of effective treatments. In 2017, Pendleton et al. [132] analyzed lavage fluids from patients with pneumonia and managed to identify the bacterial pathogens in the lungs in <9 h. Similar approaches performed on urine samples [133] and resected valves from patients with endocarditis [134] yielded a diagnosis in 4 h. For the analysis of bacterial sepsis, recent reports describe the

[4] Some of the content of the following subsections has been published in Biology Methods and Protocols (see Publication VII in Appendix C).

application of MinION metagenomic sequencing on cell-free samples (<6 h from samples to results) [135] and on faecal samples from preterm infants (obtaining results in <5 h) [136]. The depletion of human DNA prior to metagenomic sequencing proved also a useful alternative to reduce total analysis time [137].

5.2. Supporting microbiome-driven industrial processes

Microbiome sequencing has been widely applied to shed light on the microbial transformations occurring on different industrial processes (e.g., preparation of fermented foods or compost production). Portable sequencers are not only a valuable tool for characterizing industrial microbiomes, but also for detecting microorganisms in real time. For instance, water quality and wastewater management are an area of great interest for microbial monitoring. It has been proposed that sewage can be used for tracking infectious agents excreted in urine or faeces, such as SARS-CoV-2 [138]. Moreover, Hu et al. [139] reported correlations between *E. coli* culturing counts and the proportion of nanopore reads mapping a comprehensive human gut microbiota gene dataset, highlighting the potential use of this molecular technique as an indicator of faecal contamination. Nanopore sequencing can be also employed for evaluating ARGs and antimicrobial-resistant pathogens present in wastewater treatment plants [140].

Agro-food industry would also benefit from real-time sequencing. Indeed, Hu et al. [141] were able to identify the fungal species causing diseases on wheat plants, which were previously infected with known microbes. Viral infectious diseases can be also monitored by using this technology, allowing a rapid and improved response to outbreaks [142]. Other successful applications of ONT in the food industry include the characterization of the microbiome of a salmon ectoparasite (*Caligus rogercresseyi*), revealing its potential role as a reservoir for fish pathogens [143]; and the determination of the fish species present in complex mixtures, which would help to prevent—and rapidly detect—food fraud [144].

5.2.1. The particular case of anaerobic digestion

Anaerobic digestion (AD) is the process of converting several complex substrates, typically waste (e.g., sewage sludge, food waste, manure...) into biogas (methane and carbon dioxide), which is an industrially relevant biofuel. The transformation of biomass into biogas is driven by microorganisms, mainly bacteria and archaea, and it can be divided into four phases: hydrolysis, acidogenesis, acetogenesis, and methanogenesis (**Figure GI.5**). All the phases typically occur in the same anaerobic digester, although two-stage systems are common too. In this case, one reactor is used for hydrolysis and acidogenesis (acidification stage), while acetogenesis and methanogenesis take place in a different digester (methane synthesis stage) [145]. Separated acidification causes the accumulation of volatile fatty acids (VFAs), which are short-chain (C2–C6) organic acids (e.g., propionic acid, butyric acid, valeric acid, etc.) [146], since methanogenesis is intentionally inhibited at this stage. Therefore, the output of acidification is a high-strength liquor, rich in VFAs, that can be used for methane synthesis (in a different reactor) or other industrial activities such as bioplastic production or bioelectricity generation [146].

AD is highly variable, as it depends on multiple factors such as type of substrate, reactor configuration (e.g., leach bed reactor, anaerobic filter, upflow anaerobic sludge blanket, expanded granular sludge blanket, etc.), and other operating parameters (e.g., pH or temperature). All these factors also affect the microorganisms that drive biogas production and, in consequence, there is not a universal AD microbiome. In the present thesis, both bacterial and archaeal communities associated

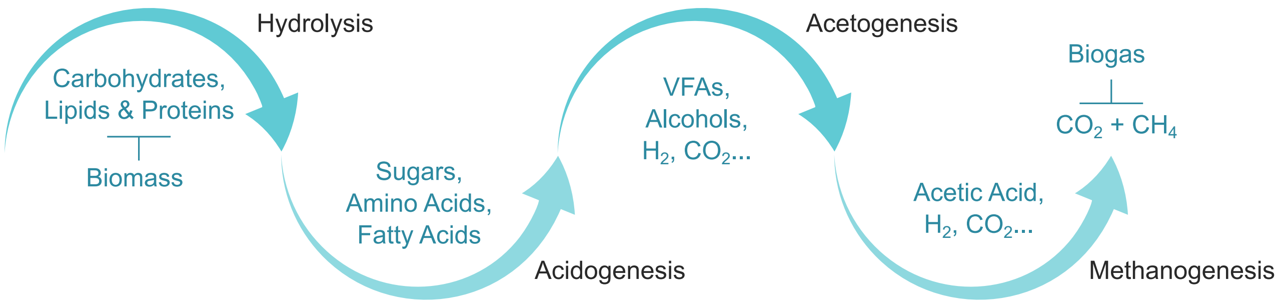

Hydrolysis
Carbohydrates, Lipids & Proteins
Biomass
Sugars, Amino Acids, Fatty Acids
Acidogenesis
VFAs, Alcohols, H_2, CO_2...
Acetogenesis
Acetic Acid, H_2, CO_2...
Methanogenesis
Biogas
$CO_2 + CH_4$

Figure GI.5. The four phases of anaerobic digestion. First, complex polymers are converted into monomers (hydrolysis), which are subsequently broken down into volatile fatty acids (VFAs), alcohols, hydrogen (H_2) and carbon dioxide (CO_2) (acidogenesis). Then, acetic acid, CO_2 and H_2 are produced (acetogenesis). Finally, methane (CH_4) is synthesized from these products (methanogenesis). Adapted from Prajapati et al. **[147]**.

with different AD processes have been analyzed using Nanopore sequencing (**Chapter IA**)[5].

5.3. Portable sequencing in natural environments

Biodiversity assessment studies are usually carried out in remote locations with limited access to DNA sequencing services, forcing scientists to design intensive sampling expeditions and returning to their home institutions to perform the sequencing and the data analysis. ONT sequencers have emerged as an alternative to these traditional approaches, allowing the creation of mobile, in-field laboratories. Pomerantz et al. [148] and Menegon et al. [149] designed portable laboratories that included thermocyclers and centrifuges powered by external batteries, and a MinION device connected to a laptop to perform *in situ* DNA sequencing. Both works were not focused on metagenomic applications, but on evaluating the taxonomic identity of different animal specimens (reptiles and amphibians) via targeted sequencing of the 16S rRNA gene or other mitochondrial genes. However, the applied methodologies and lab configurations could be easily adapted to perform metataxonomic approaches relying on the amplification and massive sequencing of marker genes.

The feasibility of MinION-based metagenomic sequencing protocols has specially been tested in very cold environments. Edwards et al. [150] reported for the first time the use of mobile laboratories for the *in situ* characterization of the microbiota of a High Arctic glacier, whereas Goordial et al. [151] were also able to perform MinION sequencing in the McGill Arctic Research Station. Johnson et al. [152] used portable field techniques to isolate DNA from desiccated microbial mats collected in the Antarctic Dry Valleys, to construct metagenomic libraries, and to sequence the samples outdoors (Taylor Valley; Temperature = −1°C) and in the McMurdo Station (Room Temperature). Finally, Gowers et al. [153] designed and transported a miniaturized lab across Europe's largest ice cap (Vatnajökull, Iceland) by ski and sledge. They adapted DNA extraction and sequencing protocols to be performed in a tent during the expedition, using solar energy and external batteries to power the hardware.

In addition to cold environments, ONT sequencers have

[5] AD requires a complex infrastructure and a deep technical knowledge about the process. For that reason, these studies where performed in collaboration with the Robert Boyle Institut e.V. (Jena, Germany) and the Technische Universität Dresden (Dresden, Germany), among other research centers. Researchers from these institutions set the AD experiments and analyzed the chemical data, while we carried out all the work related to Nanopore sequencing.

also been applied to sequence a biofilm sample at a depth of 100 m within a Welsh coal mine [154]. Even more interestingly, MinION has allowed DNA sequencing off the Earth. A first study from Castro-Wallace et al. [155] compared the performance of Nanopore sequencing in the International Space Station (ISS) with experiments carried out on Ground Control, obtaining similar results. Recently, Burton et al. [156] have reported that the preparation and sequencing of 16S rRNA gene libraries are also achievable at the ISS. Remarkably, Carr et al. [157] determined that ONT sequencers performed consistently in reduced gravity environments, which would allow the use of Nanopore sequencing in space expeditions to Mars or icy moons.

5.3.1. The particular case of bioprospecting

As previously defined, microbial bioprospecting is the search of biotechnologically-relevant microorganisms and their products from different environments, which can be both natural (e.g., deserts or oceans) or artificial (e.g., solar panels or wasted chewing gums). Samples collected during bioprospecting expeditions are generally screened upon arrival at the laboratory, and results are obtained after several weeks or months. Nonetheless, portable sequencers (i.e., MinION) can be used to characterize microbial communities directly in the field, thus providing scientists with valuable information about the samples and their biotechnological potential. In this thesis, the suitability of applying *in situ* Nanopore sequencing to inform sampling during a bioprospecting expedition has been evaluated (**Chapter IB**).

6. Motivation

Throughout this **General Introduction**, the potential of Nanopore sequencing has been proved both from a scientific and from a business points of view. Nevertheless, moving from the potential towards the actual adoption of Nanopore sequencing require a tremendous effort in terms of validation and optimization. At the time this thesis was conceived, the bioinformatic tools and experimental protocols available for applying Nanopore sequencing to characterize microbiomes were very limited. For that reason, the main motivation of the present work was to identify, optimize, and validate new methodologies for Nanopore-based metagenomics and metataxonomics, while designing and implementing novel applications of this technology to address problems of industrial or biotechnological relevance.

Objectives

This thesis aims at optimizing Nanopore sequencing to study microbial communities of industrial or biotechnological interest. It has to be noted that this technology is rapidly evolving, and its use for characterizing microbiomes can be approached from several perspectives (i.e., metataxonomics and metagenomics) and applied to multiple fields. In the context of this broad framework, the present thesis has been designed to address the following objectives:

- Developing experimental protocols and bioinformatic pipelines for the metataxonomic analysis of both archaeal and bacterial communities using Nanopore sequencing (**Chapter I**), and:
    - Applying these novel protocols to characterize microbiomes of industrial relevance, focusing on the microbial communities associated with the production of biogas as a case study (**Chapter IA**).
    - Testing the potential of *in situ* metataxonomic sequencing to improve the sampling strategy during a bioprospecting expedition (**Chapter IB**).
- Evaluating the performance of different assembly methods for Nanopore-based metagenomic sequencing and defining the advantages and limitations of this technology compared to the state of the art of metagenome assembly (**Chapter II**).

General Materials and Methods

1. Metataxonomic sequencing and analysis

Chapter I describes metataxonomic applications for studying biotechnologically-relevant ecosystems using Nanopore sequencing. For that purpose, several experimental and bioinformatic protocols were adapted, developed and/or implemented. The main steps of these protocols are explained in the subsections below and summarized in **Figure GMM.1**. However, all the materials and methods that are specific to a particular study will be described in the corresponding chapter.

1.1. DNA extraction and quantification

Two commercial extraction kits were used for DNA isolation:

- **DNeasy Power Soil Kit** (QIAGEN, Germany, Cat. No.: 12888). In all cases, approximately 0.25 g of each sample was used to perform DNA extraction according to the manufacturer's instructions, with an additional incubation step at 65 ºC after the addition of the C1 solution in order to improve cell lysis.
- **FastDNA Spin Kit for Soil** (MP Biomedicals GmbH, Germany, Cat. No.: 116560200-CF). For liquid samples, DNA isolation was performed as described by Bergmann et al. [158], using 500 µL of sample and a FastPrep 24 instrument (MP Biomedicals GmbH, Germany, Cat. No.: 116004500). DNA was eluted in 100 µL of warm DES solution (55 ºC).

In order to reduce the concentration of inhibiting substances, solid samples collected from anaerobic digesters were sedimented by centrifugation (10 min at 20,000 g) and washed several times with sterile Phosphate Buffered Saline (PBS) until a clear supernatant was observed.

DNA was quantified using either the Nanodrop 1000 Spectrophotometer (Thermo Scientific, DE, United States) or the Qubit 1X dsDNA High-Sensitivity Assay kit (Qubit 2.0 Fluorometer, Thermo Fisher, Waltham, United States, Cat. No.: Q33230), although the latter was preferred.

1.2. 16S rRNA gene amplification

All the practical applications reported in this thesis were based on analyzing the prokaryotic fraction of the microbial communities, hence the 16S rRNA gene was the marker region of choice for metataxonomic analyses. As Nanopore sequencing produces long reads, the full-length 16S rRNA gene can be read with this technology. Therefore, primers able to amplify the entire gene (V1-V9 regions; ~1.5 Kbp) were selected according to the literature [159]:

- **Archaea-specific primers**
  - Forward primer: Arch8F (5′- TCC GGT TGA TCC TGC C -3′).
  - Reverse primer: Arch1492R (5′- GGC TAC CTT GTT ACG ACT T -3′).
  - PCR mix: 1× Taq Polymerase Buffer (VWR, WR International bvba/sprl, Belgium), 200 µM dNTPs, 200 nM primers, 1 U of Taq DNA polymerase (VWR, WR International bvba/sprl, Belgium), and 10 ng of DNA template in a final volume of 50 µL.
  - PCR conditions: initial denaturation (94 ºC; 1 min); amplification (35 cycles) comprising denaturation (95 ºC; 1 min), annealing (49 ºC; 1 min) and extension (72 ºC; 2 min); final extension (72 ºC; 10 min).
- **Bacteria-specific primers**
  - Forward primer: S-D-Bact-0008-a-S-16 (5′-AGR GTT YGA TYM TGG CTC AG -3′).
  - Reverse primer: S-D-Bact-1492-a-A-16 (5′-TAC CTT GTT AYG ACT T -3′).
  - PCR mix: the PCR mix described for the archaea-specific primers was used for the bacteria-specific reaction (**Chapter IA; Study II)**. This mix was simplified for the in-field application described **in Chapter**

IB. In this case, the PCR reaction mix for each sample consisted of 22 µL of H_2O, 25 µL of NZYTaq II 2X Green Master Mix (NZYTech, Portugal, Cat. No.: MB358), 1 µL of both forward and reverse primers and 1 µL of template DNA.
- <u>PCR conditions</u>: same as for Archaea-specific primers.

Both archaea- and bacteria-specific primers were tailed with the ONT Universal Tags: 5′- TTT CTG TTG GTG CTG ATA TTG C -3′ for forward primer, and 5′ -ACT TGC CTG TCG CTC TAT CTT C -3′ for reverse primer. To assess possible reagent contamination, each PCR reaction included a negative control using Milli-Q water instead of template DNA. It must be noted that archaea-specific oligos can result in the amplification of some bacterial groups, and *vice versa*, as both primers are based on the 16S rRNA gene.

Amplicons were purified with either the Agencourt AMPure XP beads kit (Beckman Coulter, CA, United States, Cat. No.: A63880) (**Chapter IA**) or the NucleoMag kit for PCR clean up with magnetic beads (Macherey-Nagel, Germany, Cat. No.: 744100.4) (**Chapter IB**). Magnetic beads were used at 0.5X concentration, and manufacturer's instructions were followed. DNA was recovered and quantified using the the Qubit 1X dsDNA High-Sensitivity Assay kit.

1.3. Barcoding

Barcodes were added by employing the PCR Barcoding Expansion 1-12 (ONT, United Kingdom, Cat. No.: EXP-PBC001) for less than twelve samples, or the PCR Barcoding Expansion 1-96 (ONT, United Kingdom, Cat. No.: EXP-PBC096) for thirteen samples or more. The amplification protocol was as follows:

- **Chapter IA**
  - <u>PCR mix</u>: 0.5 nM of the purified PCR product, 1X Taq Polymerase Buffer, 200 µM of dNTPs, 1 U of Taq DNA polymerase, and 1 µL of the specific barcode (final volume of 50 µL).
  - <u>PCR conditions</u>: initial denaturation (98 ºC; 30 s); amplification (15 cycles) comprising denaturation (98 ºC; 15 s), annealing (62 ºC; 15 s) and extension (72 ºC; 90 s); final extension (72 ºC; 7 min).
- **Chapter IB**
  - <u>PCR mix</u>: 22 µL of H_2O, 25 µL of NZYTaq II 2X Green Master Mix, 1 µL of the specific barcode and 2 µL of the purified PCR product.
  - <u>PCR conditions</u>: initial denaturation (95 ºC; 3 min); amplification (15 cycles) comprising denaturation (95 ºC; 15 s), annealing (62 ºC; 15 s) and extension (72 ºC; 90 s); final extension (72 ºC; 5 min).

Amplicons were purified with the Agencourt AMPure XP beads kit (**Chapter IA**) or the NucleoMag kit (**Chapter IB**) and quantified with the Qubit 1X dsDNA High-Sensitivity Assay kit, as described above. Finally, an equimolar pool of amplicons was prepared for the subsequent library construction.

1.4. Library preparation

The sequencing libraries were prepared using two different versions of the Ligation Sequencing Kit (ONT, United Kingdom): SQK-LSK108 (**Chapter IA**), and SQK-LSK109 (**Chapter IB**). Briefly, the NEBNext FFPE DNA Repair Mix (New England Biolabs, Ipswich, United States, Cat. No.: M6630) was used for DNA repair and end-prep. Then, a purification with the Agencourt AMPure XP beads kit (**Chapter IA**) or the NucleoMag kit (**Chapter IB**) was carried out. Finally, adapter ligation and clean-up were performed by following the ONT SQK-LSK108 or SQK-LSK109 protocol.

1.5. Nanopore sequencing and basecalling

All the experiments were performed using R9.4.1 MinION flow cells (ONT, United Kingdom, Cat. No.:

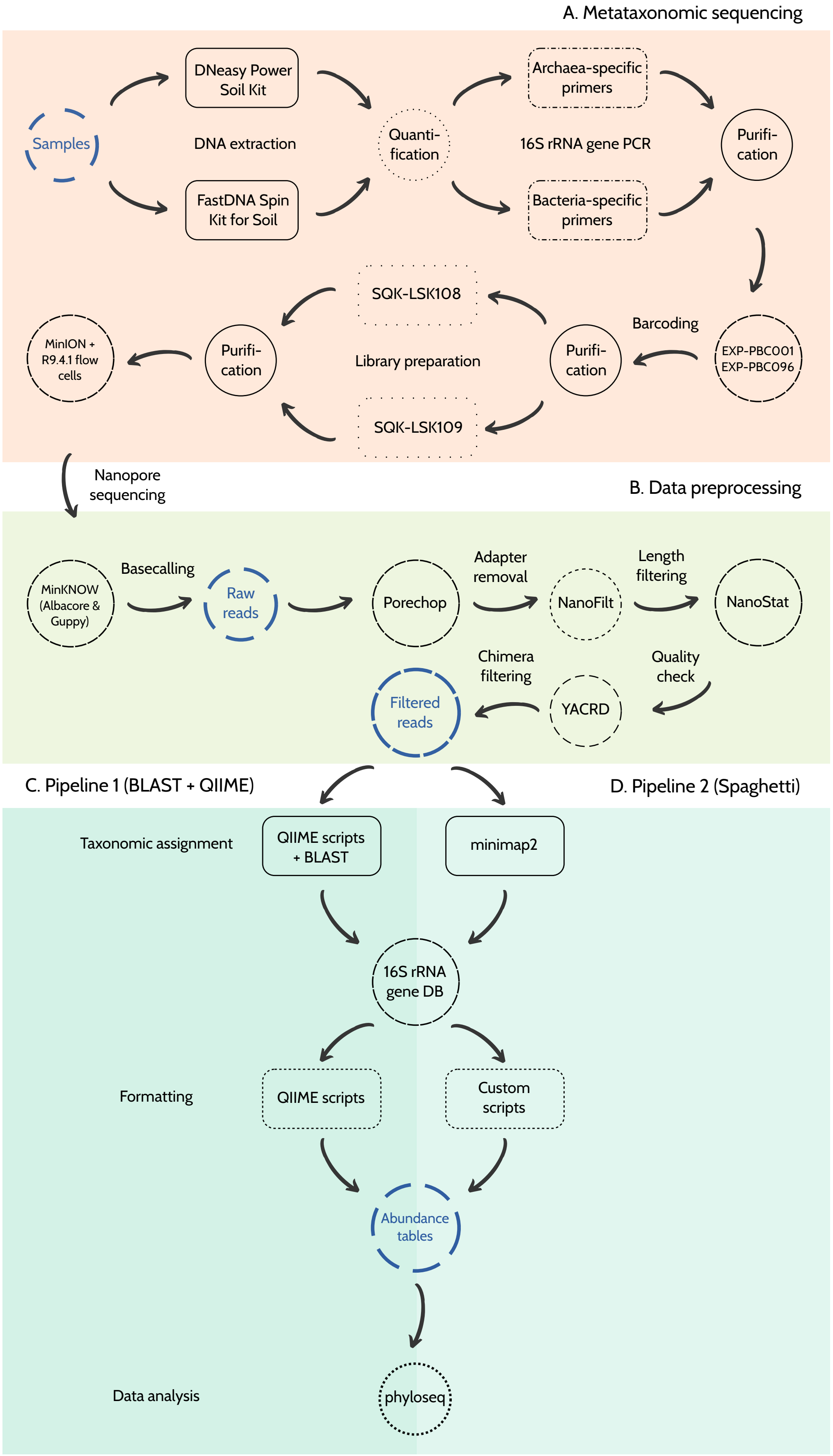


Figure GMM.1. Summary of the protocols used for metataxonomic sequencing and data analysis.

FLO-MIN106D). Reads were basecalled using the most updated version of the MinKNOW software (ONT, United Kingdom) that was available at the time when sequencing was performed: version 1.13 (**Chapter IA**) and version 20.06.5 (**Chapter IB**). The former version included Albacore as basecalling tool [160], while the latter relied on Guppy [161]. Sequences with Phred quality score (or Q score) under 7 (default threshold implemented in MinKNOW) were discarded for further analysis. Demultiplexing was carried out with either Porechop (https://github.com/rrwick/Porechop) or MinKNOW.

1.6. Bioinformatic analysis

16S rRNA gene sequences were analyzed using two different bioinformatic pipelines (**Figure GMM.1**). The first approach (used in **Chapter IA**) was mainly based on BLAST as implemented in QIIME (v. 1.9.1; http://qiime.org/) [162]. The second pipeline (used in **Chapter IB**) was called 'Spaghetti' and it mainly relied on minimap2 [110]. This approach was more suitable for *in situ* applications, as detailed in the **General Discussion**. Nevertheless, preprocessing steps were similar in the two pipelines.

Pipeline 1: BLAST + QIIME

This pipeline was designed to assign the taxonomy of the sequences through BLAST searches against a 16S rRNA gene database. For that purpose, the QIIME interface was chosen, since it provided a suite of scripts for analyzing metataxonomic data and several ready-to-use databases (http://qiime.org/home_static/dataFiles.html). However, this tool was developed for NGS data, and hence some parts of the default pipeline had to be adapted, as indicated below. The bioinformatic protocol comprised the following steps:

1. Porechop (v. ≤ 0.2.4) (https://github.com/rrwick/Porechop) was run with default parameters for removing sequencing adapters from reads.
2. Reads shorter than 700 bp or longer than 1,700 bp were filtered with Nanofilt [163] (v. 2.7.1).
3. Quality check was performed with NanoStat (v. 1.4.0) using default parameters [163].
4. Chimeras were detected and removed by using yacrd (v. 0.6.2) [164] with "-c" and "-n" parameters set to 4 and 0.4, respectively, as suggested by the authors for Nanopore data.
5. FASTQ reads were converted to FASTA using the following command: sed -n '1~4s/^@/>/p;2~4p' in.fastq > out.fasta
6. add_qiime_labels.py (http://qiime.org/scripts/add_qiime_labels.html) was applied to make the FASTA files compatible with QIIME.
7. QIIME worked with operational taxonomic units (OTUs). An OTU is a group of sequences that shares a certain similarity (typically ~97% when working at the species level). Since the accuracy of Nanopore sequences was lower than this threshold, OTU-based approaches were not directly applicable. Instead, taxonomic assignment was performed at the read level. In order to adapt this part of the QIIME pipeline, the in-house fakePickOTUs.py script (https://github.com/adlape95/ONT-16S-BLAST-and-QIIME/blob/main/fakePickOTUs.py) was created and used. After running this script, QIIME treats each individual sequence as an OTU.
8. pick_rep_set.py (http://qiime.org/scripts/pick_rep_set.html) was applied to pick a representative sequence for each OTU. In this case, each OTU was already constituted by a single sequence (see step 7), so this step was added only to meet the QIIME file format.
9. Taxonomy was assigned to each sequence with parallel_assign_taxonomy_blast.py (http://qiime.org/scripts/parallel_assign_taxonomy_blast.html), using eight threads ("-O" option). GreenGenes (v. 13.8) [165] and SILVA (v. 132) [166] databases were used for **Study I** and **Study II**, respectively.

10. Abundance tables (matrix containing all the microorganisms detected and their absolute abundances) were created with make_otu_table.py and summarize_taxa.py. These tables were imported to R for statistical analysis (see **subsection 1.7**).

This pipeline is available on GitHub (https://github.com/adlape95/ONT-16S-BLAST-and-QIIME), along with an extended description of each step.

Pipeline 2: Spaghetti

Spaghetti is a custom pipeline, developed in the framework of this thesis, for the automated bioinformatic analysis of Nanopore sequencing data and semi-automatic exploratory analysis and data visualization. Spaghetti bioinformatic pipeline is inspired by previous works [73, 167, 168], and it consists of the following steps:

1. Porechop (v. 0.2.4) is applied as described in **Pipeline 1**.
2. Nanofilt (v. 2.7.1) [163] is used to filter reads shorter than 1,200 bp or longer than 1,800 bp.
3. Quality check is carried out as described in **Pipeline 1**.
4. Chimeras are detected as noted in **Pipeline 1**.
5. Filtered reads are mapped against the SILVA database (v. 138) [166], as formatted and provided by Qiime2 (https://docs.qiime2.org/2020.8/data-resources/), by using minimap2 (v. 2.17-r9419) [110] with "-x map-ont" and "--secondary=no" options. In order to reduce minimap2's memory usage, -K option was set to 10M, as previously suggested [169].
6. Alignments are subsequently filtered with in-house python scripts (included in the pipeline), and taxonomy and abundance tables are obtained.

A detailed explanation of the pipeline and the specific commands that were implemented can be found on Spaghetti's GitHub repository (https://github.com/adlape95/Spaghetti).

1.7. Statistical analysis and data visualization

All the statistical analyses were mainly performed using the phyloseq R package (v. ≤ 1.30.0) [170]. For alpha diversity tests, samples were rarefied to the lowest library size of each experiment to mitigate uneven sequencing depth. For beta diversity, Principal Coordinates Analysis (PCoA) plots were created using the Bray-Curtis dissimilarity metric and relative abundances. Heatmaps showing the relative abundance of microorganisms at different taxonomic levels were produced with ampvis2 (v. 2.6.5) [171]. Other custom figures (i.e., barplots, line plots, boxplots...) were created using ggplot2 (v. ≤ 3.3.1). Plotly (v. 4.9.2.1) was used for producing interactive plots. Venn diagrams were obtained with an online tool (http://bioinformatics.psb.ugent.be/webtools/Venn/). Differential abundance analyses were carried out using the DESeq2 package (v. ≤ 1.26.0) [172]. Briefly, the 'phyloseq_to_deseq2' function (http://joey711.github.io/phyloseq-extensions/DESeq2.html) was applied to convert the phyloseq object into a DESeq2 object. Then, the DESeq2 main function was used with the 'parametric' option for fitting the dispersion and the 'Wald test' option for calculating the significance of the resulting coefficients. When applicable, the Benjamini-Hochberg method was used for adjusting the p-values, and only features with an adjusted p-value lower than 0.05 were generally considered significant.

It is worth highlighting that some of the analyses described above were included into the Spaghetti pipeline (see https://github.com/adlape95/Spaghetti/blob/main/module2/spaghetti.md for a full explanation).

2. Metagenomic analysis

Chapter II reports one of the first systematic evaluations of metagenomic assembly using Nanopore sequencing. A general overview of the procedures used in this work is provided in the subsections below, although some specific details about the methods will be further described within the chapter.

2.1. *De novo* assembly

As proposed by Lindgreen et al. [114], the tools selected for the present benchmarking were required to meet the following criteria:

- The tool should be freely available.
- The tool should include a suitable user guide, both for installation and usage.
- The tool should have been extensively used or show potential to become widely used.

A total of three widely used metagenomic short-read assemblers and ten long-read tools (or different versions of the same tool) were taken into consideration. Nevertheless, it was not possible to install and/or run all the software due to different reasons (**Table GMM.1**).

The commands used for running each assembler are provided in **Table GMM.2**. It is worth highlighting that all the tools were run with default parameters when no metagenomic configuration was explicitly recommended in the user guide.

2.2. Evaluation of the assemblies

Completeness and contiguity of *de novo* assemblies were first evaluated via QUAST (v. 5.0.2) [182]. MetaQUAST (v. 5.0.2) [112] was used for obtaining assembly statistics based on the alignment of the generated contigs against the reference genomes. Only contigs longer than 500 bp and with >X10 coverage were selected for calculating the general statistics. MetaQUAST failed to run with some draft metagenomes and, for that reason, minimap2 (v. 2.15) [110] was used instead to align the assemblies to the reference metagenome. Then, the percentage of metagenome covered by the draft assemblies was calculated using the pileup.sh script from BBTools suite (http://sourceforge.net/projects/bbmap/).

The resulting assemblies were further evaluated to determine their error profile. Due to the lack of a standard methodology, the presence of SNPs and indels was analyzed using two different strategies. The first one consisted of the alignment of the contigs against the reference metagenome via minimap2. BAM files were then analyzed using bcftools (https://samtools.github.io/bcftools/; v. 1.9) and the in-house script indels_and_snps.py (https://doi.org/10.5281/zenodo.3935763) was applied to quantify the variants. The second strategy was based on the use of MuMmer4 (https://sourceforge.net/projects/mummer/files/; v. 3.23). This tool was applied to align the draft assemblies to the reference metagenome. Then, the script 'count_SNPS_indels.pl' from Goldstein et al. [183] was used to calculate the final number of SNPs and indels. In both strategies, the number of variants was normalized to the total assembly size of each metagenome.

Biosynthetic gene clusters (BGCs) are usually composed of repetitive genetic structures that are hard to assemble with short reads, with long-read technologies being, therefore, more suitable to overcome this issue. However, BGCs are also very sensitive to frameshift errors, which have been reported to occur frequently in nanopore data [183]. For that reason, AntiSMASH web service (v. 5.0) [184] was used to compare the performance on BGC prediction between the different assembly tools.

2.3. Assembly polishing

Draft assemblies were further polished with Racon [185] and Medaka (https://github.com/nanoporetech/medaka), using the commands specified in **Table GMM.2**. As the Medaka model for the specific version of Guppy (v2.2.2 GPU basecaller) that was originally used for basecalling the data was not available, the Medaka default model (r941_min_high_g351) was applied instead. ONT or Illumina reads were used for iteratively running 4 rounds of Racon. Polishing was carried out using the same ONT input reads as those used for assembling each dataset, whereas Illumina reads (MiSeq plataform) were retrieved from the shotgun metagenomic sequencing data available for the Even mock community (ENA Run Accession: ERR2984773)

[186]. Only the draft assemblies corrected with ONT reads were further polished with Medaka, again using the original ONT sequences as input. Indels and SNPs were evaluated after each polishing step using the MumMer-based strategy, as detailed above[6].

[6] Some of the content included in this section has been published in the articles reported in Chapter I and Chapter II

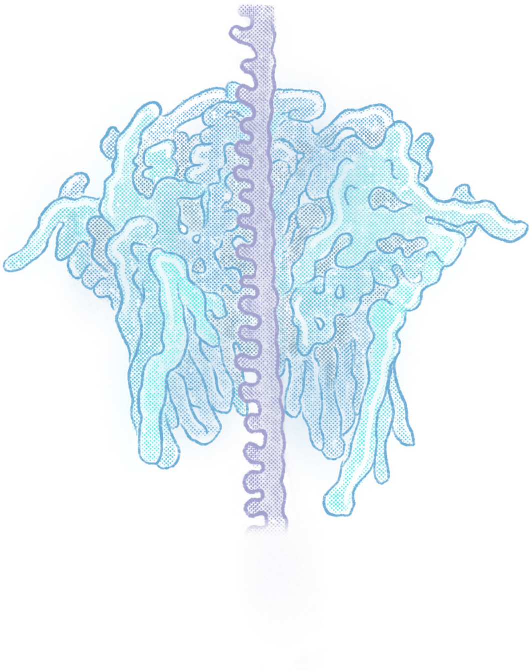
Table GMM.1. List of the assemblers selected for the benchmark

Assembler	Version	Type	Problems encountered	Intended scope
MetaSPAdes	v3.13.0	Short-read	RAM memory error	Metagenomic assembly
Megahit	v1.1.4-2	Short-read	No problems reported	Metagenomic assembly
Minia	v.2.0.7	Short-read	No problems reported	Metagenomic assembly
Canu	v1.8	Long-read	No problems reported	General usage
HINGE	---	Long-read	Config files could not be properly modified	General usage
Miniasm	v0.3(r179)	Long-read	Failed to run with the 6 Gbp Log datasets	General usage
MetaFlye	v2.4	Long-read	No problems reported	Metagenomic assembly
MetaFlye	v2.7	Long-read	No problems reported	Metagenomic assembly
Pomoxis	v0.3.2	Long-read	Failed with the 6 Gbp Log datasets and the 14 Gbp Even GridION dataset	General usage
Raven	v0.0.8	Long-read	No problems reported	General usage
Redbean	Wtdbg v2.5	Long-read	No problems reported	General usage
Shasta	v.0.4.0	Long-read	No problems reported	General usage
Unicycler	v0.4.8-beta	Long-read	Failed with the 3 Gbp Log datasets	Isolated bacteria

Table GMM.2. Detailed list of the commands used.

Assembly	
Software	**Commands**
Megahit	**1.** megahit -t *number_threads* -r *dataset_trimmed.fastq* --out-prefix *prefix_output_files* -o *output_folder*
Minia	**1.** minia *dataset_trimmed.fastq kmer_size min_abundance estimated_genome_size prefix_output_files*
Canu	**1.** canu -d *output_folder* -p *prefix_output_files* genomeSize=*genome_size(number[g|m|k])* -nanopore-raw *dataset_trimmed.fastq*
MetaFlye (v2.4 and v2.7)	**1.** flye --nano-raw *dataset_trimmed.fastq* --out-dir *output_folder* --genome-size *genome_size(number[g|m|k])* --threads *number_threads* --meta --plasmids
Miniasm	**1.** minimap2 -x ava-ont -t *number_threads dataset_trimmed.fastq* dataset_trimmed.fastq | gzip -1 > *[prefix_input_files].paf.gz* **2.** miniasm -f *dataset_trimmed.fastq [prefix_input_files].paf.gz* > *[prefix_input_files].gfa* **3.** awk '/^S/{print ">"$2"\n"$3}' *[prefix_input_files].gfa* | fold > *[prefix_output_files].fa*
Pomoxis	**1.** source [Absolute_path_pomoxis_software_folder]/venv/bin/activate **2.** mini_assemble -i *dataset_trimmed.fastq* -o *output_folder* -p *prefix_output_files* -t *number_threads*
Raven	**1.** raven -t *number_threads* dataset_trimmed.fastq > [prefix_output_files].fa
Redbean	**1.** wtdbg2 -x ont -g *genome_size(number[g|m|k])* -t *number_threads* -i *dataset_trimmed.fastq* -fo *prefix_output_1_files* **2.** wtpoa-cns -t *number_threads* -i *[prefix_output_1_files].ctg.lay.gz* -fo *[prefix_output_2_files].ctg.fa*
Shasta	**1.** shasta-Linux-0.4.0 --threads *number_threads* --memoryBacking *arg([disk|4K|2M)* --memoryMode *arg([anonymous|filesystem])* --input *dataset_trimmed.fastq* --assemblyDirectory *output_folder*
Unicycler	**1.** unicycler -l *dataset_trimmed.fastq* -t *number_threads* -o *output_folder*
Polishing	
Software	**Commands**
Racon	**1.** minimap2 -x map-ont -d *assembly_unpolished.mmi assembly_unpolihed.fasta* **2.** minimap2 -ax map-ont -t *number_threads assembly_unpolished.mmi reads_(ONT/Illumina).fastq* > *overlaps.sam* **3.** racon -t *number_threads reads_(ONT/Illumina).fastq overlaps.sam assembly_unpolished.fasta* > *assembly_polished.fasta*
Medaka	**1.** source activate medaka **2.** medaka_consensus -i *reads_ONT.fastq* -d *assembly_unpolished.fasta* -o *output_folder* -t *number_threads*

Chapter I. Novel applications of Nanopore-based metataxonomic sequencing



Chapter IA. Unraveling industrial microbiomes: the case of anaerobic digestion

Chapter IB. *In situ* microbiome sequencing to inform targeted bioprospecting

Chapter IA. Unraveling industrial microbiomes: the case of anaerobic digestion

Abstract:

Anaerobic digestion (AD) of organic matter is a robust technology for biogas synthesis from different types of waste. AD is essentially considered a black-box process, as the microbial transformations leading to the biogas production are not yet fully understood. In this chapter, Nanopore-based metataxonomic sequencing methods were designed, implemented and applied to study the impact of different parameters on the archaeal and bacterial communities associated with AD. In a **first study,** liquid and solid fractions of grass biomass were used as co-substrates for anaerobic co-digestion of sewage sludge, and the methanogenic microbiome was monitored over time. Liquid-fed batches developed a more stable microbiome, enriched in *Methanosarcina*, and resulted in higher methanogenic yield. In contrast, solid-fed batches were highly unstable at higher substrate concentrations, and *Methanosaeta* –typically associated with sewage sludge– remained as the most abundant methanogenic archaea. The **second study** aimed at evaluating the use of nitrogen removal methods during anaerobic acidification. To eliminate nitrogen, NH_3-stripping and MAP (magnesium ammonium phosphate hexahydrate) precipitation were compared. Despite the treatments, bacterial communities were very similar in all the experiments, which proved the high robustness and resilience of AD microbiomes. Nevertheless, MAP had a stronger effect on the bacterial communities according to Nanopore sequencing. Moreover, the abundance of *Acholeplasma* and *Erysipelotrichaceae* UCG-004, which are predominant in AD processes with high ammonia concentrations, tended to increase during acidification. This suggests that bacterial communities were able to progressively adapt to high ammonia levels. **Altogether**, these works demonstrate for the first time that Nanopore sequencing can be effectively used to study AD microbiomes, thus paving the way towards a more ambitious goal: using this technology to monitor any microbial-based industrial process *in situ*.

Background

Anaerobic digestion (AD) is a widely used technology that allows microbial conversion of biomass into biogas, which is a mixture of methane (CH_4), carbon dioxide (CO_2) and traces of other gases (e.g., hydrogen sulphide). Anaerobic digesters can be fed with a wide range of substrates, such as grass biomass [187], sewage sludge from water treatment [188], microalgal biomass [189], and food waste [190], among others. Consequently, biogas is considered an environmentally sustainable source of energy and a key element in the frame of circular bioeconomy [191]. The AD industry has been in continuous expansion over the last few years. For instance, the total number of biogas plants increased from 6,227 to 17,662 between 2009 and 2016 [192]. Moreover, it is estimated that about 30-40% of EU gas needs can be fulfilled with biogas by 2050 [193].

Basically, AD consists of four phases: hydrolysis (biomass fragmentation), acidogenesis (formation of organic acids, alcohols, CO_2, and hydrogen), acetogenesis (formation of acetic acid), and methanogenesis (last phase of the process, in which acetic acid, hydrogen, and carbon dioxide are the main substrates for the formation of methane) (**Figure GI.5**). A diverse number of bacteria are known to be involved in the hydrolysis and further acidogenesis of complex polymers, whereas the oxidation of intermediate metabolites to acetate (acetogenesis) is performed by either hydrogen- or formate-producing acetogens [194]. Lastly, methane synthesis is mainly derived from acetate and H_2/CO_2 by acetoclastic and hydrogenoclastic methanogenic archaea [195].

Different techniques based on sequencing technologies (e.g., 16S rRNA gene sequencing or shotgun metagenomic sequencing) have been applied to study the structure and composition of microbial communities in different types of anaerobic digesters [187, 188, 196, 197]. These works have shown that each particular community is influenced by parameters such as the type of feedstock [196], temperature [198], retention time [199], salt content [200], pH [201], or the loading rate [202]. In other words, microbiomes involved in the production of biogas are heterogeneous and their dynamics are influenced by multiple factors. This explains why AD is still considered a black-box process [203].

Nevertheless, metataxonomic and metagenomic studies have allowed the identification of different microorganisms that can be associated with specific operating parameters. For example, mesophilic reactors tend to be richer in *Methanosaeta*, *Methanoculleus*, and *Methanosarcina*, while thermophilic digesters exhibit an increased abundance of *Methanothermobacter* or *Methanobacterium* [188, 204]. On this basis, DNA sequencing could be used as a monitoring tool for industrial processes (i.e., fermentations), albeit the economic investment needed to acquire a sequencer, the technical complexity of the sequencing process and the further bioinformatic analysis hamper its adoption.

As demonstrated throughout the **General Introduction**, Nanopore sequencing has the potential to overcome these limitations. For that reason, we aimed at applying the in-house protocols described in **General Material and Methods** to study real-life AD processes, as a proxy for assessing the suitability of ONT platforms to characterize and monitor industrial microbiomes. Specifically, Nanopore-based metataxonomic sequencing was used in four different works. The full description of all these studies requires addressing several complex concepts about AD, which is not the main goal of this thesis. Therefore, and for the sake of conciseness and clarity, only two of these works have been included in this chapter:

- **Study I: Analysis of archaeal communities associated with the digestion of sewage sludge**. Under mesophilic conditions, digestion processes with high amounts of chemical oxygen demand (COD) tend to have high amounts of *Methanosarcina* and *Methanoculleus*. On the other hand, sewage sludge, which has typically lower amounts of COD compared to typical industrial co-digesters, tends to have higher amounts of archaea corresponding to the genus *Methanosaeta* [188, 204]. A common strategy to increase the efficiency of wastewater treatment plants is to co-digest sewage sludge with other substrates such as food waste or grass biomass, which have higher COD. Nevertheless, the transition from mono-digestion to co-digestion of sewage sludge has not been sufficiently characterized at the microbiome level. Hence, this work aimed to investigate the impact of slowly increasing concentrations of COD on the underlying archaeal community of sewage sludge digesters by means of Nanopore sequencing. The effect of different feeding strategies (feeding with liquids or solids) was also analyzed[7].
- **Study II: Effect of ammonia removal methods on bacterial communities**. AD of poultry manure -probably the largest residue in the farming industry- is problematic due to its high content of nitrogen. During the hydrolysis of complex compounds, the initial step of the anaerobic digestion process, organically-bound nitrogen is released in the form of ammonia, which causes the inhibition of anaerobic microorganisms involved in

[7] See Publication I in Appendix C for more context about this study

biomethanation. To reduce the toxic effects of poultry manure, different experimental approaches have been previously investigated: (1) to apply physico-chemical methods for ammonia removal, such as ammonia stripping [205, 206] or precipitation of ammonia and phosphate together with magnesium salts (MAP precipitation) [207]; (2) to perform a biological pre-treatment in a separated acidification stage[8] [208, 209], which allows a rapid formation and accumulation of volatile fatty acids in a high-strength liquor thanks to the intended inhibition of methanogenic archaea[9] ; (3) to use leach bed reactors (LBR) for separating the acidic high-strength liquor from solids prior to methanogenesis, which facilitates pumping, and allows the application of a high-performance methane reactor, such as an anaerobic filter (AF), upflow anaerobic sludge blanket (UASB) or expanded granular sludge blanket (EGSB) for efficient methanation in a second stage [210]. This work aimed at demonstrating the possibility of combining leach-bed acidification of poultry manure with ammonia stripping or ammonia precipitation while monitoring the influence of these ammonia removal methods on the underlying bacterial communities by using Nanopore sequencing[10].

The other two studies were also focused on bacterial communities and can be found in **Appendix C (Publication III and Publication IV)**. In the first one, Nanopore-based 16S rRNA gene analysis was used in combination with short-read metagenomic sequencing to study the taxonomic and functional structure of a bacterial biofilm that had grown on the inner wall of a laboratory-scale transparent anaerobic digester illuminated with natural sunlight. Finally, the last work describes the use of the Lotka–Volterra model coupled to Nanopore sequencing results to analyze the bacterial interactions occurring during AD.

Study I: Analysis of archaeal communities associated with the digestion of sewage sludge

Materials and methods

Substrates

Sewage sludge was collected from a mesophilic anaerobic digester of a municipal wastewater treatment plant in Jena (Germany) and stored for one week. The sludge contained 3.52% of total solids (TS) and 56.30% of volatile solids (VS, percentage of TS). Fresh grass (*Gramineae*) biomass was chosen as co-substrate. It was collected from the front garden of the Robert Boyle Institute (Jena, Germany) and stored at 0 °C. To separate liquids from solids, a conventional juicer (Angel Juicer 8500s, Angel Co. LTD., Korea) was used for pretreatment of grass biomass. The solid fraction contained a COD of 366 mg O_2 per g of substrate (according to the German guideline DIN 3814-S9), and the liquid fraction had a COD of 82 mg O_2 per g of substrate (according to DIN 38409-H41). The produced liquid fraction contained 11.63% of TS with 76.63% of VS. The remaining solid fraction contained 55.80% of TS and 90.16% of VS.

Experimental design

Five batch reactors (A–E) were set up according to the German guideline VDI 4630. Feeding occurred semi-continuously with gradually increasing loading rates as shown in **Figure IA.1**. Incubation occurred at 37 °C without stirring and incubation bottles were agitated manually before biogas measurements or sampling. Reactor A was used as a negative control (without co-substrate input); reactors B and C were fed with liquid co-substrate (liquids separated from grass biomass);

[8] Acidification is the first step of a two-phase AD system. The objective of this stage is to convert biomass into volatile fatty acids (VFAs), which can be used for producing biogas or other relevant products.

[9] Methanogesis is unwanted during acidification, as the objective of this phase is to obtain VFAs, not biogas.

[10] See Publication II in Appendix C for more context about this study.

and reactors D and E were fed with solid co-substrate input (remaining solids after liquid separation). Each reactor was filled with 300 mL of sewage sludge. The reactors were opened every 3–4 days –twice a week– to take samples and to add substrate. Afterwards, the reactors were closed and flushed with nitrogen to ensure an anaerobic atmosphere. Gas was collected in a liquid displacement device (eudiometer) and the volume of biogas, as well as the ratio of CO_2 and CH_4 were measured daily. Produced gas was analyzed using the "COMBIMASS 99 GA-m" gas measurement device (Binder, Germany) to determine the ratio of CO_2 and CH_4. The amount of total volatile fatty acids (TVFAs) and the solubilization of COD were monitored using conventional photometer-based assays (Nanocolor CSB15000 and Nanocolor organische Säuren 3000, Macherey-Nagel, Germany).

At the beginning (**Figure IA.1**, phase I), the loading rate was adjusted in such a way that liquid- and solid-fed reactors produced similar amounts of methane. After running the reactors with a constant input (phase II), solid-fed batches (reactors D and E) reached an extremely high viscosity, and 100 mL of water was added to each

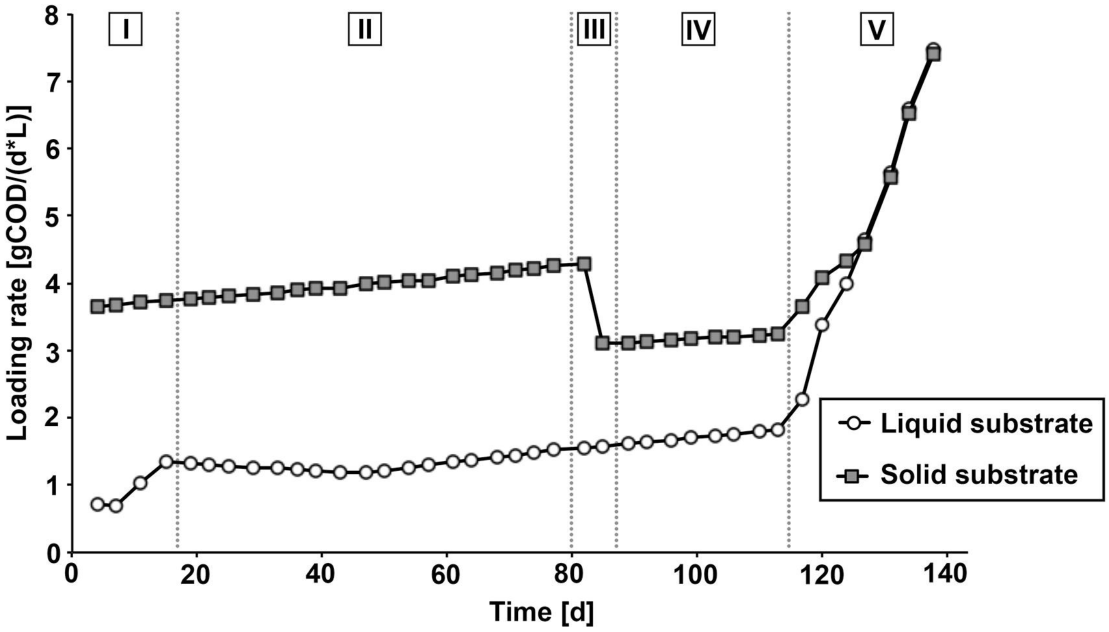


Figure IA.1. Substrate load (chemical oxygen demand -COD- per day -d- and litre -L-) in time. The addition of liquid (batch reactions B and C) and solid (batch reactions D and E) grass co-substrate was performed in five different phases, as indicated in roman numbers. Phase I: COD input concentration was adjusted to a value in which similar amounts of methane were produced in batches fed with liquid or solid substrates. Phase II: A stable period of feeding occurred. Phase III: solid-fed batches (reactors D and E) reached an extremely high viscosity, and small amounts of water were added to enable stirring. Phase IV: Another phase of stable conditions followed. Phase V: in order to drastically increase the organic loading rate, the substrate was changed to molasses in all the reactors.

reactor to enable stirring (phase III). After another phase of stable conditions (phase IV) and from day 113 onwards, the input of both liquid and solid substrates was reduced by 25% each cycle, reaching a reduction of 100% at day 124. At the same time, the grass biomass was replaced with molasses to induce a shock loading (i.e., a rapid change in substrate concentration that can lead to the inhibition of AD). The molasses input was increased by 0.5 g per cycle (phase V).

Fluorescent microscopy

Methanogenic archaea were quantified with an epifluorescent microscope (Axio Lab.A1, Carls Zeiss, Germany), adjusting the optical filters and excitation wavelengths to the quantification of cofactor F420, which is associated with methanogenic archaea. Excitation occurred with wavelengths ranging from 400 to 440 nm and light emitted with wavelengths between 500 nm and 550 nm was collected. Samples were diluted 1:2 with a mounting solution (10 µL each) (RotiR-Mount FluorCare, Carl-Roth, Germany) and 3 µL of the suspension were applied between the cover slip and the slide. Pictures were taken with 400× magnification and 126 ms exposure time. For each time point, 48 pictures ware taken and evaluated using the ImageJ software (*https://imagej.nih.gov/ij/download.html*). *Methanosarcina*-like complexes were identified from the images and used as a semi-quantitative estimator of the abundance of *Methanosarcina* in the samples. On the other hand, F420-signals emitted from rod-shaped archaea were used to evaluate the abundance of *Methanosaeta*.

Metataxonomic sequencing and analysis

DNA extraction, 16S rRNA gene amplification, Nanopore sequencing and bioinformatic analysis were performed as described in **General Material and Methods**. Briefly, samples were washed with PBS and centrifuged (10 min at 20,000 g) several times, until the supernatant was clear. Then, DNA was extracted with the DNeasy Power Soil Kit. The full-length 16S rRNA gene of archaea was amplified by PCR with the Arch8F and Arch1492R primers. Libraries were prepared using the SQK-LSK108. All the purification steps included in the protocol were carried out with the Agencourt AMPure XP beads kit. Sequencing was performed during 12 h using the MinION and a R9.4.1 flowcell. Reads were basecalled with MinKNOW (v. 1.13; basecaller: Albacore) and sequences with an average Q < 7 were discarded. Porechop (https://github.com/rrwick/Porechop) was applied to demultiplex the sequences according to the barcodes. **Pipeline 1** (https://github.com/adlape95/ONT-16S-BLAST-and-QIIME; detailed in **General Material and Methods**) was followed for taxonomic assignment using the GreenGenes database (v. 13.8) as reference [165]. Principal Coordinates Analysis (PCoA) plots were created using the Bray-Curtis dissimilarity metric and relative abundances. Barplots and line plots were generated with ggplot2 (v. 2.2.1).

Results and discussion

Reactor performance: liquid vs. solid feeding

Two different strategies for the repowering of sewage sludge involving co-digestion were compared: (1) using a liquid co-substrate with very low percentage of total solids (TS) and, therefore, with low amounts of lignocellulose (batch reactions B and C); and (2) using a solid co-substrate with a very high percentage of TS and, therefore, with high amounts of lignocellulose (batch reactions D and E). Both co-substrates were obtained from fresh grass (*Graminidae*) biomass. Additionally, a control digester was kept without co-substrate input (batch reaction A). The loading rate of both experimental approaches was adjusted in such a way that both systems produced similar volumes of methane per working volume, as described in “Materials and methods” (**Figures IA.1**).

In both co-digestion strategies, the amount of biogas ranged between approximately 100 and 200 mL of methane per day during phase I and II (day 1–82) (**Figure IA.2**). By the end of phase II, the liquid fed batches reached a solubilized COD of 5.9 ± 0.2 g COD/L, and a TVFA (intermediates in the methane formation pathway) concentration of 2.81 ± 0 g TVFA/L. Although the solubilized COD and TVFA of the solid fed system were higher at that time (12.02 ± 2.98 g COD/L, and 3.98 ± 0.02 g TVFA/L), the amount of methane produced was slightly higher in the liquid fed system (**Figure IA.2**). Moreover, methane production within the liquid fed system proved more stable in time.

By the end of phase II, the digestion sludge in the solid fed system reached such high viscosity that no stirring was possible. In order to ensure a better substrate distribution and to facilitate to movement of bubbles, a small amount of water was added to the solid fed batch systems D and E. Due to the dilution, the loading rate of the solid fed batches was slightly reduced during phase IV (**Figure IA.1**).

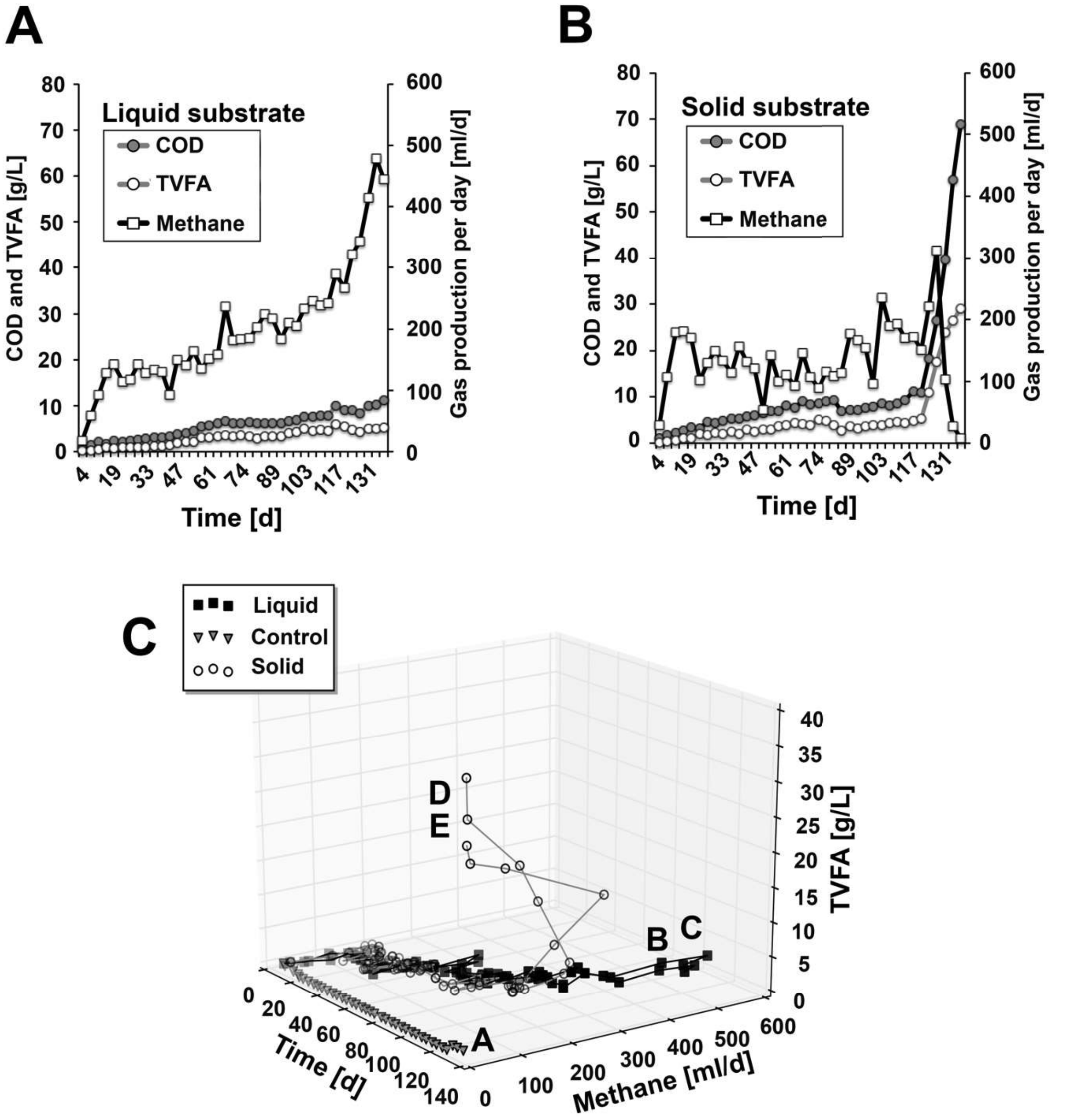


Figure IA.2. Chemical parameters measured for the different experimental set-ups. Data on chemical oxygen demand (COD), total volatile fatty acids (TVFA) and methane production are represented for one of the liquid fed reactors (A) and one of the solid fed reactors (B). A three-dimensional representation of the evolution of all the reactors is also shown (C).

Since the high viscosity of the solid fed batch prevented any further increase of the loading rate, the substrate was changed stepwise to molasses for both digestion experiments (liquid and solid fed batches), starting at day 117 (phase V). In parallel, the loading rate of the other substrates (liquid and solid grass biomass) was lowered stepwise until both experimental set-ups were fed exclusively with molasses. The solubilized COD increased drastically in both digestion approaches, indicating that reactors reached the intended substrate overload. However, from day 132 onwards (**Figure IA.1**, phase V), the produced amount of methane became drastically reduced in the solid fed batches (**Figure IA.2**). In contrast, the liquid fed batch systems displayed higher stability, with continuously increasing levels of methane production. Moreover, a sudden acid shock was detected in the concentration of TVFAs in the solid fed batches, reaching more than 20 g TVFA/L at day 132 (Fig. 2B). The liquid fed batches remained with a lower concentration of TVFAs, reaching 5.52 ± 0.54 g TVFA/L at day 132, indicating a much more efficient conversion of TVFAs into methane.

Changes in the composition of the methanogenic microbiome

Changes in the relative abundance of the main genera involved in methane production were monitored using fluorescent microscopy and confirmed by full-length 16S rRNA gene high-throughput sequencing, using specific primers targeting archaea. As shown in **Figure IA.3**, microscopic analysis revealed that the number of rod-shaped methanogens in solid-fed batches remained high throughout the experiment, indicating a high number of *Methanosaeta*, a genus typically observed in sewage sludge [188, 204]. Interestingly, liquid-fed batches showed decreasing numbers of rod-shaped methanogens and increasing numbers of *Methanosarcina*-like complexes. Microscopic analysis made it difficult to quantify *Methanosarcina* species, as they tend to form large complexes, which prevent the identification of single cells. *Methanosarcina* is a genus found in co-digesters with higher loading rate than sewage sludge [188].

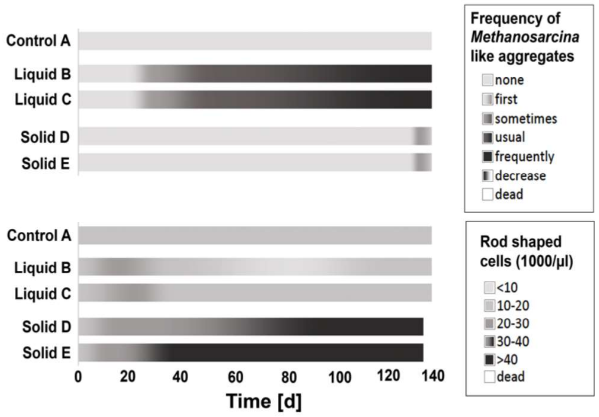


Figure IA.3. Microscopic analysis of methanogenic communities. Methanogens were screened by quantifying co-factor F420. *Methanosarcina*-like cell aggregates and rod shaped F420-signals were analysed semi-quantitatively. High amounts of rod-shaped F420-signals were used as indicators for high concentrations of *Methanosaeta* spp., which is typical for sewage sludge with low COD content.

In parallel, the archaeal communities present in the different reactor configurations were analyzed by means of full-length archaeal 16S rRNA gene high throughput analysis using the MinION device. In accordance with the chemical data, the principal coordinates analysis performed with the archaeal profile of each sample showed a different microbial evolution for the solid and liquid-fed reactors (**Figure IA.4**). **Figure IA.4A** shows the taxonomic composition in each experimental condition. At day 29, *Methanosaeta* was the most abundant genus in all cases, accounting for 23–75% of sequences. *Methanosaeta* remained the majority genus throughout the experiment in solid-fed batches, as well as in the control digester. However, the abundance of *Methanosarcina*. increased in liquid-fed batches, with it becoming the most abundant genus at days 103 (batch C) and 139 (batch B and C). **Figures IA.4A**

and IA.4C show the decline of *Methanosaeta* spp., and the enrichment of *Methanosarcina* spp. Moreover, the increase in *Methanosarcina* was accompanied by a transient increase of *Methanomethylovorans* at day 29 and 103 (**Figure IA.4A**). Consistently with these findings, an earlier work reported a high abundance of *Methanomethylovorans* in sewage digesters which received co-ferments from biodiesel production [188].

It has to be noted that *Methanosarcina* spp. are methanogens which can be found in co-digesters with high loading rates, especially in the leachate of leach-bed systems [188, 211, 212]. Therefore, the observed results indicate that the methanogenic microbiome of liquid-fed systems can be successfully shaped and adapted to higher loading rates. This is in contrast with solid-fed systems, which remained enriched in *Methanosaeta*, a typical genus from sewage sludge digesters, which are usually operated at lower loading rates [188].

The present results show that the successful adaption of archaeal communities of sewage sludge to higher loading rates is a time-consuming process and it depends on the

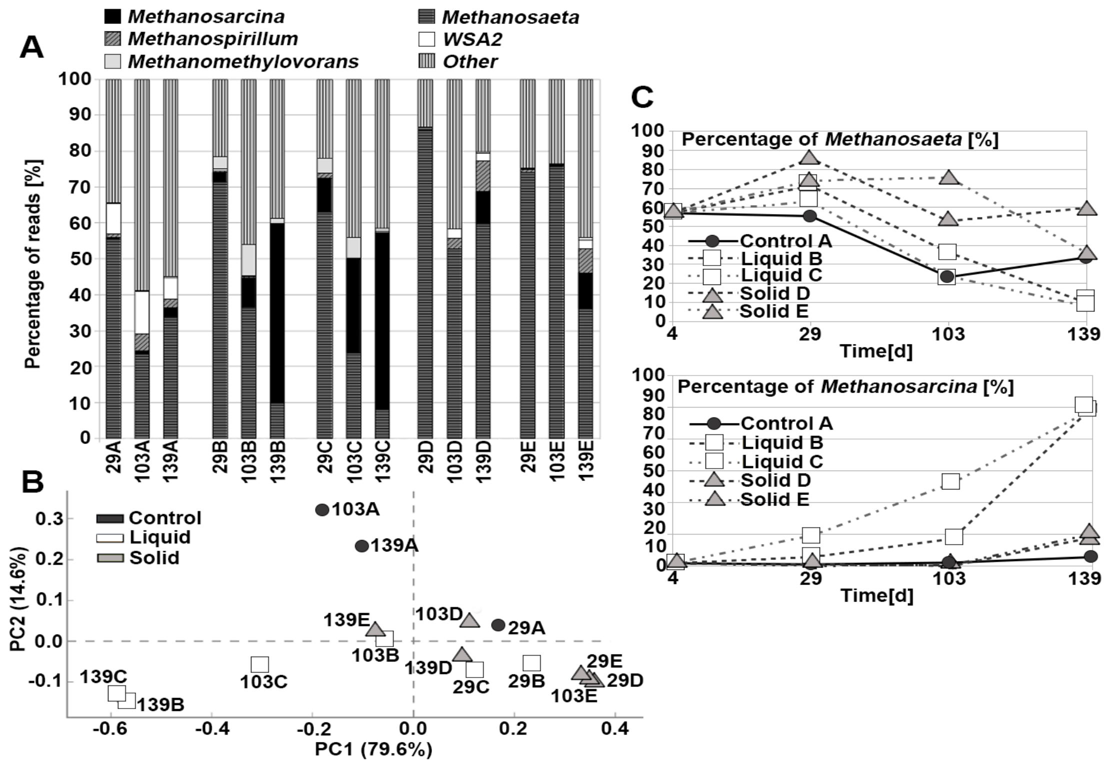


Figure IA.4. Taxonomic analysis of the archaeal communities based on full-length 16S rRNA gene Nanopore sequencing. (A) Relative frequency of the most abundant taxa in each sample. Numbers in the X axis indicate the day in which samples were taken. Letters indicate the type of feeding (A: control; B and C: liquid-fed batches; D and E: solid-fed batches). (B) Principal Coordinate Analysis (PCoA) based on the full taxonomic profile of the samples. (C) Evolution of the frequency of *Methanosaeta* (up) and *Methanosarcina* (down) in the different reactors.

feeding strategy. To ensure stable process conditions, it is recommended to intensively screen the taxonomic profile of industrial sewage digesters, when increasing the loading rate due to the application of co-substrates. As lignocellulose-enriched substrates are problematic during the digestion process, it is preferred to use liquid substrates and to remove the lignocellulolytic fraction. Alternatively, it is also possible to liquefy the lignocellulolytic fraction prior to the digestion process, as discussed recently by Rajesh Banu et al. [213].

Overall, liquid-fed reactors were re-shaped, changing from a *Methanosaeta*-dominated community, to a *Methanosarcina*-enriched community. For the first time in this field, 16S rRNA gene high-throughput sequencing was performed with the Nanopore-based MinION device, allowing the identification of changes in the taxonomic profiles during the AD process. This work reports that the addition of liquid co-substrates resulted in a more effective methanogenic microbiome, and allowed higher biogas production. Altogether, these results confirm the high potential for increasing the efficiency of sewage sludge digesters from wastewater treatment plants by using relatively simple strategies such as co-digestion with liquid substrates.

Study II: Effect of ammonia removal methods on bacterial communities[11]

Materials and methods

Substrates

The solid part of a digestate was obtained from a local biogas plant in Woltow (Germany) and used as seed sludge (i.e., microbial starter). This biogas plant was linked to an egg-producing poultry farm. Chicken manure from laying hens was used as the main substrate, which was collected from an egg-producing poultry farm (NEN Marth GmbH, Woltow 19, Selpin-Woltow, Germany). Additionally, wheat straw was used as the packing material for the leach bed. The applied raw materials were chemically characterized (**Table IA.1**).

Experimental design

Leach-bed acidification was performed in custom-made, stainless-steel reactors with a total volume of 100 L (number 1 in **Figure IA.5a**). The leach-bed, a mixture of 5 kg seed sludge and 5 kg chicken manure, was incubated at a mesophilic temperature (37 ºC) for 7 days. A seven-day period was chosen as treatment time in the leach-bed reactor since this duration had been

[11] See Background and Publication II (Appendix C) for more context about the study.

Table IA.1. Characterization of the seed sludge and substrates used in this study.

	Seed sludge	Substrate	Packing material
TS (% of FM)	18.03 + 1.04	35.71 + 1.14	88.81 + 2.40
VS (% of TS)	70.25 + 2.28	51.30 + 2.45	94.71 + 2.73
Total nitrogen (Kjeldahl, g kg^{-1} of FM)	3.10 + 0.29	8.31 + 0.82	5.89 + 1.20
COD (g kg^{-1} of FM)	183.09 + 15.86	257.18 + 10.26	1074.92 + 17.87
pH	8.81 + 0.10	7.77 + 0.24	6.91 + 0.15
Conductivity (mS cm^{-1})	27.41 + 11.79	54.40 + 5.19	8.08 + 3.59

TS: total solids; FM: fresh matter; VS: volatile solids; COD: chemical oxygen demand.

previously found suitable to yield high volatile fatty acid concentrations and to prevent conversion of acids into methane (data not shown). As packing material for the leach-bed, 0.5 kg wheat straw was used and retained in a strainer. A custom-made storage tank with a total volume of 60 L (number 2 in **Figure IA.5a**) was filled with 20 L of fresh tap water at the beginning of the experiment. During the incubation period of 7 days, the liquor was circulated between the storage tank and the LBR with a flow rate of 193.8 mL s^{-1} resulting in a high-strength liquor (leachate). Circulation was achieved by using a pump (Wilden Model P1 Plastic/P1/PPPPP/WFS/WF/KWF"; Wilden Pumps Grand Terrace, CA, U.S.A.). Liquor was pumped from the storage tank and percolated over the leach-bed every 3 h. After passing the leach-bed, the liquor was returned to the storage tank. Percolation was adjusted to achieve a complete cycle of circulation of the liquor each 24 h. The storage tank was maintained at 37 ºC.

The volume of produced gas was measured in a multi-chamber rotor gas meter (TG 05/5 model, Dr.-Ing. Ritter Apparatebau GmbH & Co. KG, Bochum, Germany) and collected in a gas bag (Tesseraux

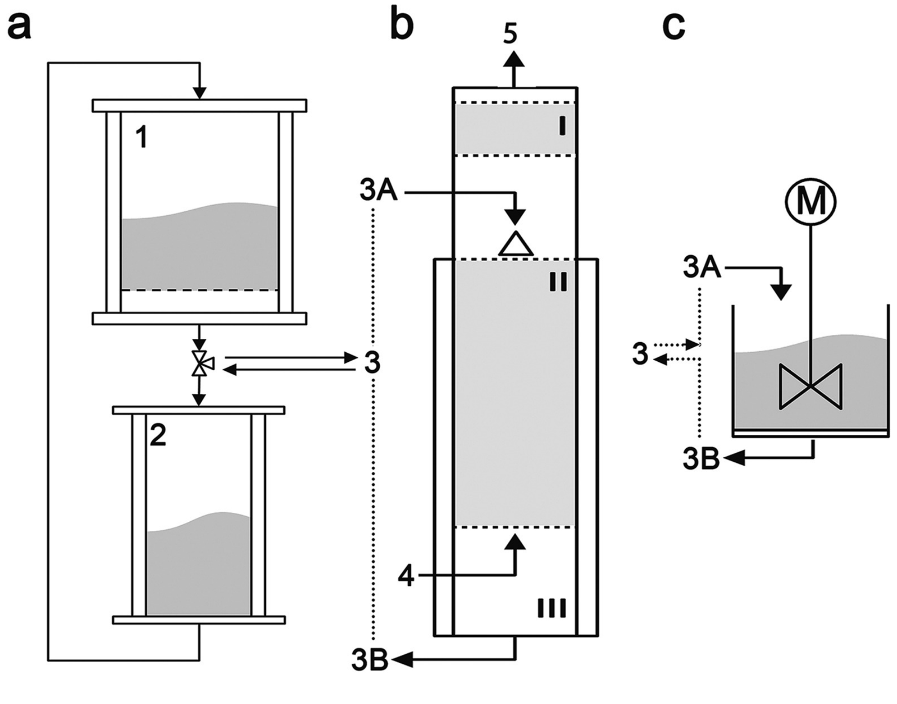


Figure IA.5. Experimental set-up: Acidification of chicken manure in a leach-bed configuration (a). Leachate was collected in a storage tank (2), and continuously recirculated into the leach-bed reactor (1). To apply NH3-stripping and MAP precipitation, liquor was removed from leach-bed system and then returned back after nitrogen removal (3); NH3-stripping occurred in a stripping column (b). For this, leachate was pumped into segment I of the column (3A), which was equipped with a droplet separator. From there, the leachate was transferred through a packed bed in segment II. In the opposite direction to the leachate flow, atmospheric air was pumped into the column in section III through the packed bed (4), which was used as trickling filter. The treated leachate was collected in the lower part of the column, from where it was removed (3B) and, finally, refilled into the leach-bed acidification (A). NH3-enriched gas escaped through the upper part of the column (5). MAP precipitation occurred in an open tank coupled to a laboratory stirrer (c).

Spezialverpackungen GmbH, Bürstadt, Germany), where the gas was kept until analysis. The measured biogas volume was normalized to standard conditions (temperature 273.15 K, pressure 1013.25 mbar).

Leach-bed acidification was performed without nitrogen removal (control), and with nitrogen removal by either NH_3-stripping or MAP precipitation. Acidification experiments were repeated twice for each configuration (Duplicate A and B).

NH_3-stripping

Stripping of ammonia involves blowing air or biogas through the fermentation liquid, resulting in the transfer of free ammonia into the gas phase. The removal of ammonia is facilitated at high pH and temperature, since ammonic nitrogen is present in the shape of free ammonia under these process conditions [214]. The ammonia can subsequently be captured and recovered from the gas phase by absorption (e.g., with sulphuric acid) to yield ammonium sulphate.

A custom-made stripping column was used for NH_3-stripping. The column had an inner diameter of 0.168 m and a total length of 2.05 m. It was divided into three segments: segment I = leachate feed, segment II = packed bed, and segment III = gas carrier feed (**Figure IA.5b**). The packed bed of the column contained a loose bulk of plastic rings (Pall-Ring 15; Raschig GmbH, Ludwigshafen, Germany) showing a diameter of 0.015 m, a specific surface of 350 m^2 m^{-3}, and a porosity of 0.88 m^3 m^{-3}. The bulk height was 0.81 m. The column was heated with a water bath (Lauda Alpha RA8; Dr. R. Wobser GmbH & Co. KG, Lauda-Königshofen) up to 90 ºC.

The column was loaded with approximately 0.02 m^3 h^{-1} of leachate from the storage tank of the leach-bed acidification (number 3 in **Figure IA.5a**) using a pump (Watson-Marlow 323du Drive 400 RPM EU; Spirax-Sarco Engineering group, Falmouth Cornwall, England). To facilitate ammonia release, the pH of the leachate was adjusted to 11.6 using approximately 50 g L^{-1} of sodium hydroxide (32%). Released ammonium was removed from the column by a constant air stream, which was used as gas carrier. The gas carrier was pumped at a rate of 20 m^3 h^{-1} through the stripping column, using an air ventilator (RL65-21/14; ebm-pabst, St. Georgen, Hungary). After passing through the stripping column, leachate was taken from the bottom of the column and returned back into the storage tank, the pH was re-adjusted to its former value using approximately 12.3 g L^{-1} of sulphuric acid (75%). NH_3-stripping was performed only once during the acidification of chicken manure (the treatment was applied after three days of the 7-day leach-bed incubation time). In the applied treatment procedure, the leachate was passed two times through the stripping column to achieve an appropriate removal of ammonia.

MAP precipitation

MAP (magnesium ammonium phosphate hexahydrate) precipitation, also known as struvite precipitation, is based on the ability of ammonia to bind with magnesium and phosphate [215]. Under weakly alkaline conditions, and if phosphate and magnesium salts are added at required amounts, magnesium ammonium phosphate crystallizes, and it can be withdrawn from the liquid phase for further utilization as fertilizer.

For the MAP precipitation (precipitation of $MgNH_4PO_4{\cdot}6H_2O$), an open tank coupled to a laboratory stirrer for continuous mixing was used (**Figure IA.5c**). The working volume of the tank was 10 L. The stirrer was a paddle agitator, which was powered by a Eurostar 40 digital stirrer drive (IKA Werke GmbH & Co. KG, Staufen, Germany). The reaction was conducted at a constant temperature of 22 ºC. For each MAP precipitation step, 30.5 g L^{-1} of potassium hydrogenphosphate and 89 g L^{-1} of magnesium chloride hexahydrate (both previously dissolved in deionized water) were applied to the leachate taken from the storage tank of the leach-bed acidification (number 3 in **Figure IA.5a**). After 30 min of incubation with agitation (100 rpm), a sedimentation step of 60 min without agitation was carried out in order to achieve a separation between the precipitated MAP crystals and the leachate. After sedimentation, the supernatant was released from the open tank by a lateral outlet and returned back into the storage tank. The MAP sludge was discarded. The dry weight of the MAP sludge was approximately 4.7% (in weight) of the treated leachate.

Two different MAP precipitation experiments were performed: in one of them, MAP precipitation was applied only once, after three days of the 7-day leach-bed incubation time (further referred as 1x MAP-P); while in the other, MAP precipitations were performed after days 1, 2 and 4 (further referred as 3x MAP-P).

Chemical analyses

pH and conductivity were analyzed using the WTW pH/Cond 340i device (WTW, Weilheim, Germany), equipped with the WTW pH Electrode SenTix41 and the WTW TetraCon 325 electrodes. The TS, VS, NH_4-N, total nitrogen and trace elements were quantified according to the VDLUFA guideline (1983/2006). The COD was analyzed according to the guideline DIN 38409-41:1980-12. Volatile fatty acids (VFAs), including acetic, propionic, butyric, iso-butyric, valeric, iso-valeric, and caproic acid, were analyzed as described by Herrmann et al. [216]. The composition of the produced biogas was measured on a daily basis using the portable gas analyzer GA 2000 (Geotechnical Instruments (UK) Ltd., Warwickshire, England).

DNA extraction and qPCR

DNA isolation was performed with the FastDNA Spin Kit for Soil, as described in **General Material and Methods**. qPCR was performed as described by Bergmann et al. [158], using the following primer and TaqMan sets: BACfw 5'-ACT CCT ACG GGA GGC AG-3', BACrev 5'-GAC TAC CAG GGT ATC TAA TCC-3', and BAC TaqMan 5'-TGC CAG CAG CCG CGG TAA TAC-3' for quantifying bacteria; and ARCfw 5'-ATT AGA TAC CCS BGT AGT CC-3', ARCrev 5'-GCC ATG CAC CWC CTC T-3', and ARC TaqMan 5'-AGG AAT TGG CGG GGG AGC AC-3' for quantifying archaea. The The qPCR was conducted utilizing the 2x qPCR Probe Mix (Bioenzym Scientific GmbH, Hess. Oldendorf, Germany), TaqMan probes, primer-fw/rev (biomers.net GmbH, Ulm, Germany) and a CFX96 Real-Time System (Bio-Rad Laboratories GmbH, München, Germany). The PCR mixture consisted of 10 µL 2x pPCR Probe Mix,1.8 µL primer-fw,1.8 µL primer-rev, 0.4 µL TaqMan probe, 1000 pg of sampled DNA (2 µL of the isolated DNA with a concentration of 500 pg μL^{-1}) and 4 µL PCR-grade water. The PCR program for bacteria started with a 7 min step 95ºC, followed by 45 cycles with 15 s at 95 ºC for denaturation, 30 s at 57 ºC for primer annealing, and 60 s at 60 ºC for elongation. The thermocycler program for archaea started with 7 min at 95 ºC too and was followed by 40 cycles with 15 s at 95 ºC for denaturation, and 60 s at 60 ºC for primer annealing and DNA elongation. Evaluation of the PCR results was supported by the program CFX Manager 3.1 (Bio-Rad Laboratories GmbH, München, Germany).

Metataxonomic sequencing and analysis

DNA extraction, 16S rRNA gene amplification, Nanopore sequencing and bioinformatic analysis were performed as described in **General Material and Methods**. Briefly, the full-length 16S rRNA gene of bacteria was amplified by PCR with the S-D-Bact-0008-a-S-16 and S-D-Bact-1492-a-A-16 primers. Libraries were prepared using the SQK-LSK108. All the purification steps included in the protocol were carried out with the Agencourt AMPure XP beads kit. Sequencing was performed using the MinION and a R9.4.1 flowcell. Reads were basecalled with MinKNOW (v. 1.13; basecaller: Albacore) and sequences with an average Q < 7 were discarded. Porechop (https://github.

com/rrwick/Porechop) was applied to demultiplex the sequences according to barcodes. **Pipeline 1** (https://github.com/adlape95/ONT-16S-BLAST-and-QIIME; detailed in **General Material and Methods**) was followed for taxonomic assignment using the SILVA database (v. 132) as reference [166]. As recommended by the authors of the phyloseq package [170], taxa which were not detected more than 3 times in at least 20% of the samples were removed, and abundances were standardized to the median sequencing depth in order to correct different library sizes. Principal Coordinates Analysis (PCoA) was carried out based on Bray-Curtis dissimilarities and relative abundances. For alpha diversity analyses, samples were rarefied to the lowest library size to mitigate uneven sequencing depth, and plots were created with the plot_ordination function included in phyloseq. Additionally, a comparative analysis of the microbial profiles observed at day 7 was carried out using DESeq2 package [172]. Barplots were generated with ggplot2 (v. 2.2.1).

Results and discussion

Nitrogen removal during acidification

In this work, a leach-bed configuration was used for acidification, and NH_3-stripping and MAP-precipitation[12] were compared to a control reaction, which was acidified without any nitrogen removal. Nitrogen was removed generally at day 3, in order to ensure that a significant fraction of nitrogen was converted into ammonia. Additionally, a fourth experiment was performed, where the MAP-precipitation was repeated three times (3x MAP-P).

After 7 days of acidification, the treated liquors always showed lower nitrogen concentrations than the untreated control. The highest nitrogen removal was achieved with the triple MAP-precipitation, where the total nitrogen (Kjeldahl) concentration was around 29% lower than the control. The concentration of NH_4-N was reduced by 38% (**Figure IA.6a**). The amount of biogas produced strongly varied between duplicates, and this observation can be attributed to the natural

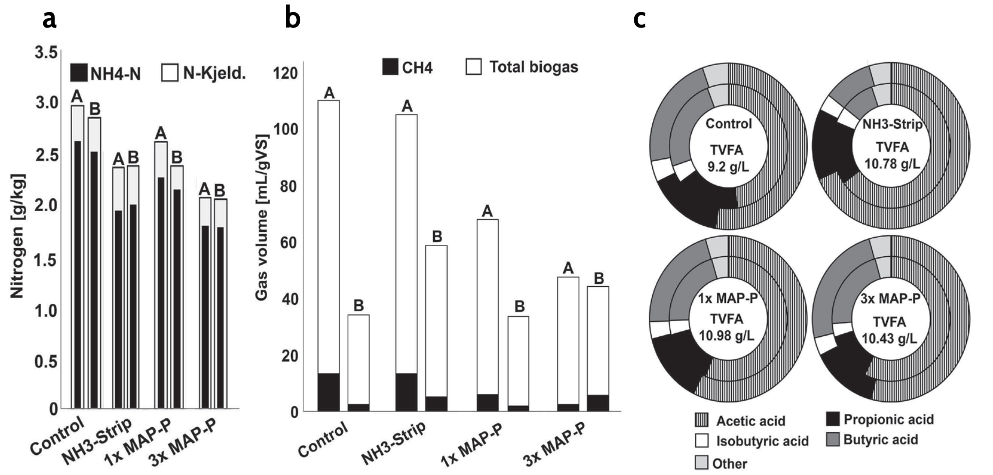


Figure IA.6. Characterization of the high-strength liquor produced with the different treatments. (a) Total nitrogen content (N-Kjeld.) and nitrogen content corresponding to ammonia. (b) Biogas and methane production. (c) Total amount of volatile fatty acids (TVFAs) and percentage of each different organic acid analysed.

[12] NH3-stripping and MAP-precipitation are two different methods for removing nitrogen from substrates.

fluctuations of the used raw materials (see **Table IA.1**). Particularly low biogas production was detected in both replicates when three MAP precipitations were performed (**Figure IA.6b**). Taken together, our results indicate that MAP-precipitation could be an effective treatment to prevent unwanted methane production during acidogenesis.

The content of TVFAs increased between 13 and 19% when N-removal methods were applied. High ammonia concentrations can inhibit hydrolysis and acidification, and thus the formation of VFAs. Methanogenesis is more susceptible to high ammonia concentrations, but hydrolysis and acidification have also been found to be negatively affected by high ammonia concentrations [217, 218]. In the present study, ammonia removal during the acidification process correlated with enhanced acidification at high nitrogen input concentrations. TVFA concentration were similar in all three approaches for nitrogen removal (averaged 10.73 ±0.28 gTVFAs L^{-1}), but the ratio between different kinds of VFAs differed (**Figure IA.6c**). While in the untreated process the amount of acetic acid ascended to 59% of TVFAs, the liquid treated with MAP precipitation and NH_3-stripping showed values of approximately 65% and 74%, respectively. This enhancement in the formation of acetic acid is highly convenient for a subsequent use of the process liquor in methanation[13].

Quantification of archaea and bacteria

The absolute abundance of bacteria and archaea was analyzed on the last day of each leach-bed experiment by means of qPCR (**Figure IA.7**). The number of bacteria and archaea found in the control was in concordance with other studies [158, 216]. However, all experiments, except the control, showed a strong decrease in the number of archaea. This indicates that the harsh N-removal method might contribute to prevent contamination with methanogenic microorganisms, which is a positive outcome during acidification. In the case of x3 MAP-precipitation, qPCR results show that the bacterial community was strongly affected too. Surprisingly, this reduction in the bacterial cell number did not result in an apparent reduction of TVFAs production. However, it has to be noted that TVFAs content reached values of 8.9 and 8.6 g L^{-1} at day 3 (data not shown), indicating that most of the TVFAs were produced before the second MAP-precipitation was conducted.

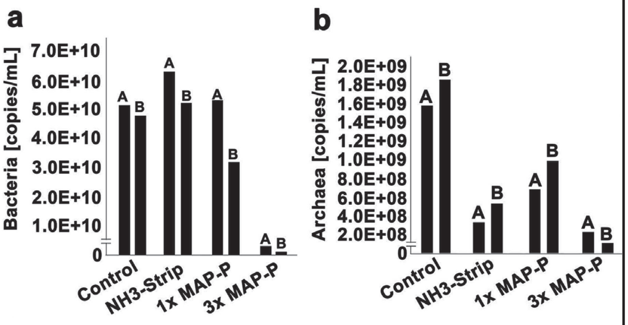


Figure IA.7. Quantification of bacteria (a) and archaea (b) in the high-strength liquor produced with the different treatments. The copy numbers of 16S rRNA genes were analysed based on qPCR using domain-specific primers.

Characterization of taxonomic profiles

Besides qPCR, high-throughput sequencing of the 16S rRNA gene was performed with a MinION sequencer. Samples were collected from each acidification experiment (control, NH_3-stripping, 1x MAP-P and 3x MAP-P) at days 1, 3 and 7. Additional samples were taken at day 5 for the 3x MAP-P. As each experiment was performed in duplicate, a total of 26 samples were analyzed. Overall, 288,222 reads were generated and taxonomically assigned, accounting for an average of 11,085 reads per sample (min: 1,414; max: 30,596).

Taxonomic profiles were similar in all the experiments (**Figure IA.8**). Similarly, other studies have found that AD microbiomes are highly robust and show resilience against alterations in different operating parameters[14]

[13] See Publication II in Appendix C for an extended discussion on the efficiency of Nitrogen removal methods

[14] See Publication IV in Appendix C

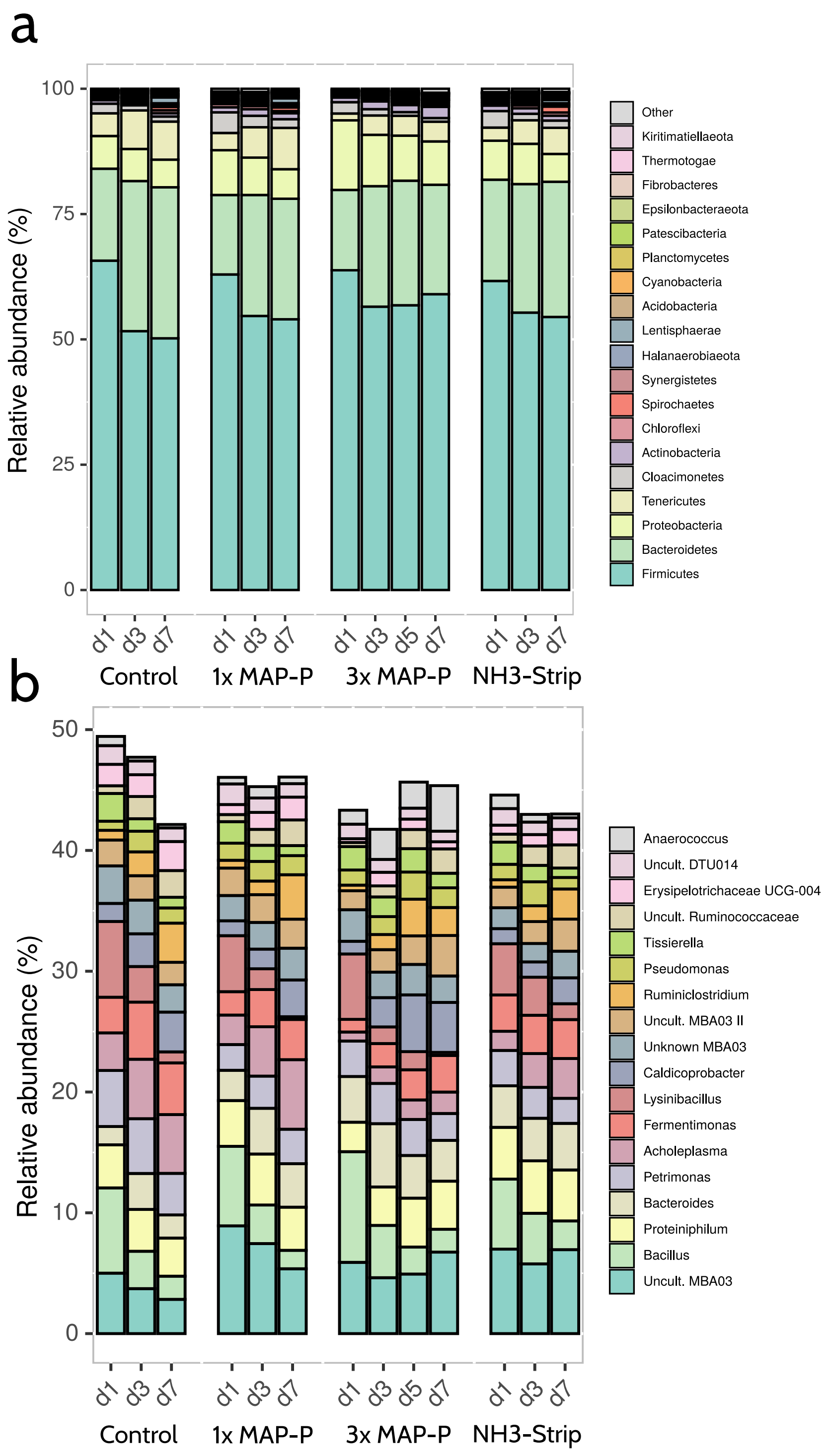


Figure IA.8. Taxonomic analysis of bacterial communities by Nanopore sequencing at the phylum (a) and genus (b) level. Each bar displays the average relative abundance calculated from duplicates. Only the most abundant taxa are shown. 'd': day.

[219]. At the phylum level, samples were dominated by *Firmicutes* (~57.4% of average relative abundance), *Bacteroidetes* (23.2%), *Proteobacteria* (~8%), *Tenericutes* (~4.9%), and *Cloacimonetes* (1.78%) (**Figure IA.8a)**, which was concordant with previous investigations [188, 204, 220–222]. *Firmicutes, Bacteroidetes* and *Proteobacteria* are also predominant in the chicken gut microbiota [223, 224]. Moreover, according to Abendroth (2017) [219], a high abundance of *Firmicutes* and *Bacteroidetes* is expected in reactors showing high concentrations of TVFAs, while *Proteobacteria* is usually found in processes subjected to harsh conditions, which matches the results obtained in our experiments.

At the genus level, taxonomic profiles proved to be diverse, with no specific genus dominating the microbial community. Instead, all the samples were rich in various taxa including an uncultured MBA03 bacterium (~5.8 of average relative abundance), *Bacillus* (~4.1%), *Proteiniphilum* (~3.7%), *Bacteroides* (~3.3%), *Petrimonas* (~3%), *Acholeplasma* (~3%), *Fermentimonas* (~2.9%) or *Lysinibacillus* (~2.6%), among others (**Figure IA.8b)**. These genera are commonly found in anaerobic digesters or in the chicken gut microbiota. For instance, it has been previously demonstrated that members of the MBA03 taxonomic group are associated with AD processes using farm waste (i.e., animal manure) as substrate [225] or with high amounts of ammonia [226]. Interestingly, other studies have proved that the genera *Fermentimonas*, *Proteiniphilum* and *Petrimonas* -common in biogas reactors [227]- were associated with nitrogen rich substrates or slightly elevated pH values and TVFAs [226, 228]. *Bacteroides,* a well-known fermentative and acid-producing bacterium, was found to be abundant in both chicken gut microbiomes [229] and digesters fed with chicken manure, which were also rich in *Ruminiclostridium* and *Caldicoprobacter* [230], as also shown in this work. Finally, *Acholeplasma* and *Erysipelotrichia* proved predominant in microbiomes showing a robust performance under high ammonia concentrations [231]. In concordance with these results, the relative abundance of *Acholeplasma* and *Erysipelotrichaceae* UCG-004 increased during acidification, with the exception of *Erysipelotrichaceae* UCG-004 after 3x MAP-P (**Figure IA.9**), suggesting that microbial communities progressively adapt to high ammonia levels.

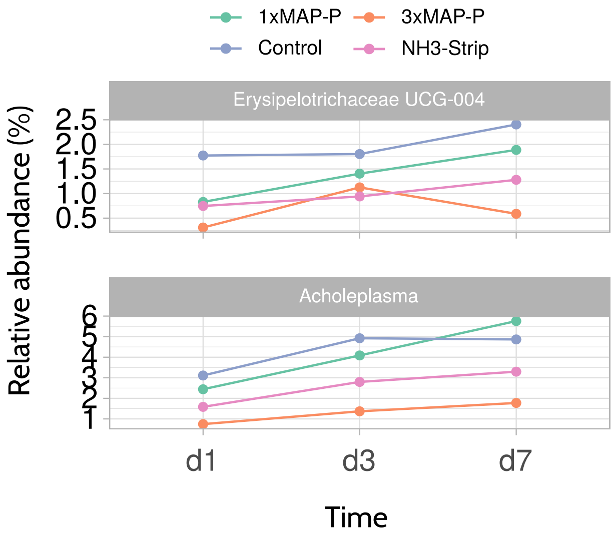


Figure IA.9. Evolution of the relative abundance of *Erysipelotrichaceae* UCG-004 (top) and *Acholeplasma* (bottom) with time. Relative abundance was calculated from Nanopore sequencing results. 'd': day.

Beta diversity analysis revealed that all the microbiomes showed a similar evolution in time (**Figure IA.10**; Axis 1). This probably reflected the adaptation of the microbial communities to the conditions, which were similar in all cases despite the different ammonia removal methods applied (same type of substrate, same reactor configuration, similar levels of TVFAs, etc.). In agreement with qPCR results, the triple MAP-precipitation resulted in a more noticeable modification of the microbial community (**Figure IA.10**; Axis 2). To further investigate this phenomenon, a statistical comparison was performed using the microbial profiles corresponding to the last day of each experiment. Microbial communities were only slightly affected by NH_3-stripping, with a significant reduction in the abundance of only one genus in comparison to the control experiments (FDR-adjusted p-value < 0.05; DESeq2 test). The communities proved much more affected by MAP-precipitation, which resulted in a significant change in nine (1x MAP-P) and twenty (3x MAP-P) different genera (FDR-adjusted p-value < 0.05; DESeq2 test) (**Supplementary Figure IA.S1**). Finally, alpha diversity analysis showed that reactors subjected to ammonia removal methods displayed higher richness and Shannon indices (**Supplementary Figure IA.S2**). Nevertheless, this increase in the alpha diversity metrics could not be directly attributed to the ammonia removal, as control samples showed reduced alpha diversity at day 1, when no removal method had been applied yet. (**Supplementary Figure IA.S3**).

The present work evaluates the suitability of using NH_3-stripping or MAP-precipitation during an acidification treatment of chicken manure in a leach-bed configuration. In general, all the investigated procedures showed comparable efficiencies for ammonia removal. Denitrification techniques lead to an up to 19% increase of production of VFA, however, the highest production of acetic acid (favourable for methane production) was observed when applying NH_3-stripping. To the best of our knowledge, this is the first study in which Nanopore sequencing has been used to analyze the bacterial communities associated with AD. Overall, microbiomes were similar in all the experiments and concordant with previous studies. Nevertheless, 3x MAP-P proved to have a higher influence on the bacterial composition according to both qPCR and microbiome sequencing. The effect of these changes in the production of biogas (in a second reactor) from the high-strength liquor generated after ammonia removal remains unknown and deserves further study.

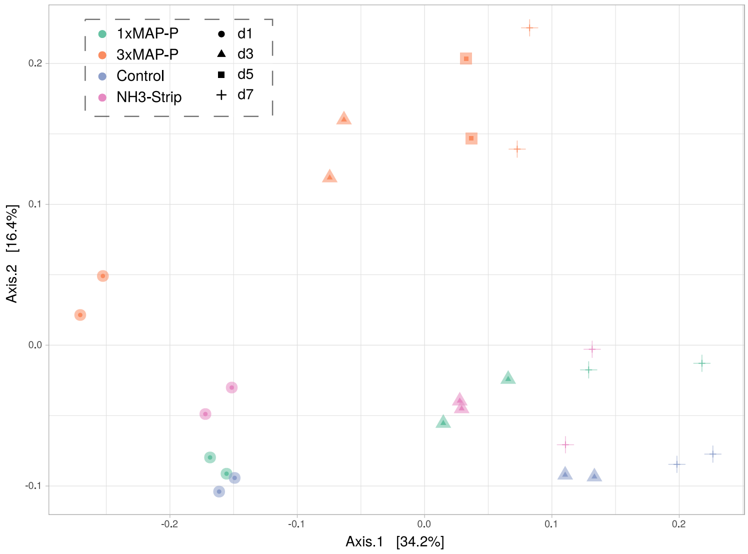


Figure IA.10. Principal Coordinates Analysis (PCoA) of bacterial taxonomic profiles (genus level) based on the Bray-Curtis dissimilarity metric. Relative abundances were used as input data. ‘d’: day.

The content of this chapter has been adapted from the following publications (available in **Appendix C**):

Publication I: Hardegen J, **Latorre-Pérez A**, Vilanova C, et al. Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass. *Bioresour Technol* 2018;**265**:275-81. doi:10.1016/j.biortech.2018.06.005
Publication II: Ramm P*, Abendroth C*, **Latorre-Pérez A**, et al. Ammonia removal during leach-bed acidification leads to optimized organic acid production from chicken manure. *Renew Energy* 2020;**146**:1021-30. doi:10.1016/j.renene.2019.07.021
Publication III: **Latorre-Pérez A***, Abendroth C*, Porcar M, et al. Shedding light on biogas: Phototrophic biofilms in anaerobic digesters hold potential for improved biogas production. *Syst Appl Microbiol* 2020;**43**. doi:10.1016/j.syapm.2019.126024
Publication IV: Schwan B*, Abendroth C*, **Latorre-Pérez A**, et al. Chemically Stressed Bacterial Communities in Anaerobic Digesters Exhibit Resilience and Ecological Flexibility. *Front Microbiol 2020*;**11**. doi:10.3389/fmicb.2020.00867

* Equal contributions
In all the publications listed above, the work related to Nanopore sequencing and microbiome data analysis was carried out by **Latorre-Pérez A**.

Chapter IB. *In situ* microbiome sequencing to inform targeted bioprospecting

Abstract:

Bioprospecting expeditions are often performed in remote locations, in order to access previously unexplored samples. Nevertheless, the actual potential of those samples is only assessed once scientists are back in the laboratory, where a time-consuming screening must take place. This work evaluates the suitability of using Nanopore sequencing during a journey to the Tabernas Desert (Spain) for forecasting the potential of specific samples in terms of bacterial diversity and prevalence of radiation- and desiccation-resistant taxa, which were the target of the bioprospecting activities. Samples collected during the first day were analyzed through 16S rRNA gene sequencing using a mobile laboratory. Results enabled the identification of locations showing the greatest and the least potential, and a second, informed sampling was performed focusing on those sites. After finishing the expedition, a culture collection of 166 strains belonging to 50 different genera was established. Overall, Nanopore and culturing data correlated well, since samples holding a greater potential at the microbiome level also yielded a more interesting set of microbial isolates, whereas samples showing less biodiversity resulted in a reduced (and redundant) set of culturable bacteria. Thus, we anticipate that portable sequencers hold potential as key, easy-to-use tools for *in situ*-informed bioprospecting strategies.

Background

Scaling laws have predicted that the Earth is home to 1 trillion (10^{12}) microbial species [232]. A large fraction of this biodiversity still remains to be explored and very likely harbors novel molecules, enzymes and/or biological activities with potential applications in industrial processes, drug development, cosmetics or environment-related issues (i.e., bioremediation). The search for these novel products from biological sources and, in particular, from microorganisms, is known as microbial bioprospecting. Extreme environments, such as the deep sea or hyper-arid deserts, are of special interest for bioprospecting studies, as they tend to be sources of undiscovered biodiversity [233].

The characteristics (i.e., nutrient and oxygen availability, humidity, irradiation, pH, etc.) of a given environment shape the composition of its microbiota, often leading to the existence of temporal and spatial variations in the microbial community composition [234, 235]. Spatial changes have also been observed at microscale: for example, in gradients of soil depths as recently demonstrated with the SoilBox system [236].

In this context, sequencing technologies can be used to elucidate whole microbial profiles from samples, thus revealing changes in microbiome composition which are usually not detected with culture-based approaches. Illumina sequencing platforms –such as the MiSeq System– are the current standard for microbiome sequencing. Nevertheless, this technology is time-consuming and usually requires shipping the samples to a centralized sequencing facility. Therefore, *in situ* TGS strategies emerge as a promising alternative to this traditional approach.

Among TGS technologies, the MinION system is especially relevant for *in situ* sequencing as it is the smallest sequencing device currently available, it is inexpensive in comparison to other TGS devices, and the generation of long reads can be assessed in real time. Thus, sequencing data can be directly analyzed through bioinformatic pipelines that can be run on servers, laptops, or even mobile phones [108, 237].

Nanopore sequencing has previously been used in range of real-time applications, such as pathogen detection and surveillance [121, 137, 238]; forensic identification

[239, 240] or industrial process monitoring [241, 242]. Among all the potential uses of MinION, *in situ* sequencing is especially interesting for those situations where no alternative analyses are feasible due to a lack of equipment (i.e., second-generation sequencing platforms, qPCR instruments...). This is the case for most bioprospecting expeditions, which are usually carried out far away from microbiology laboratories. Previous works have demonstrated that both sample preparation and microbiome sequencing can be achieved using a reduced, mobile laboratory. Indeed, Nanopore sequencing has been successfully applied in extremely remote locations such as the Antarctic Dry Valleys, Canadian High Arctic, the largest European ice cap or the International Space Station (**see subsection 5.3 of General Introduction**). Beyond the undoubtedly scientific interest of analyzing microbial samples up to hundreds of kilometers away from the nearest laboratory, microbial bioprospecting could further benefit from *in situ* sequencing, as it would allow for a more directed and evidence-based sampling procedure focused on those sampling locations that prove to be enriched with the microbial taxa and/or biological activities of interest.

To test this hypothesis, we planned a two-night expedition to the Tabernas Desert (Almería, Spain). This dryland has recently been reported to harbor a previously unexplored high bacterial biodiversity, significantly enriched in radiation- and desiccation-resistant microorganisms, which were the target of our study [37]. A minimum setup of both laboratory and bioinformatic tools was designed for analyzing biocrust and soil samples via 16S rRNA gene sequencing throughout the expedition. The obtained taxonomic profiles were used to identify sample types enriched in taxa that have been described to be radiation resistant, allowing us to collect additional samples before ending the journey. Overall, this work demonstrates the feasibility of using portable, Nanopore-based sequencing devices to study microbial communities without the need of returning to the lab, which could potentially inform decision-making during sampling.

Materials and methods

Sample collection

Sampling was carried out in November 2020 at the Tabernas Desert Natural Park (Almeria, Spain), under the permission of the competent authorities. Biocrust and bulk soil samples were collected in two different days. Biocrust samples were gathered using a laboratory spatula that was sterilized with ethanol 96% immediately before collecting each sample. Bulk soil (~5 cm deep) was directly introduced into sterile falcon tubes. On the first day, fourteen different samples were taken, and then analyzed through *in situ* microbiome sequencing. Based on the results, six additional samples were gathered during the second sampling day. These samples were, indeed, biological replicates of the least and most promising samples based on sequencing data. Metadata (geolocation, type of sample, appearance and pictures) was collected and associated to each sample (**Supplementary Figure IB.S1**).

Laboratory setup

Requirements for DNA extraction, PCR amplification, library preparation and sequencing were evaluated, and a minimum laboratory setup was designed accordingly (**Supplementary Table IB.S1**). The necessary equipment fitted in the trunk of a compact car, and it was transferred to an apartment in Viator (Almería, Spain), 15 km away from the Tabernas Desert, where the mobile laboratory was established and all the experimental and data analysis procedures were carried out. The apartment was equipped with electricity, internet connection, a fridge and a freezer.

Metataxonomic sequencing and analysis

DNA extraction, 16S rRNA gene amplification, Nanopore sequencing and bioinformatic analysis were performed as described in **General Material and**

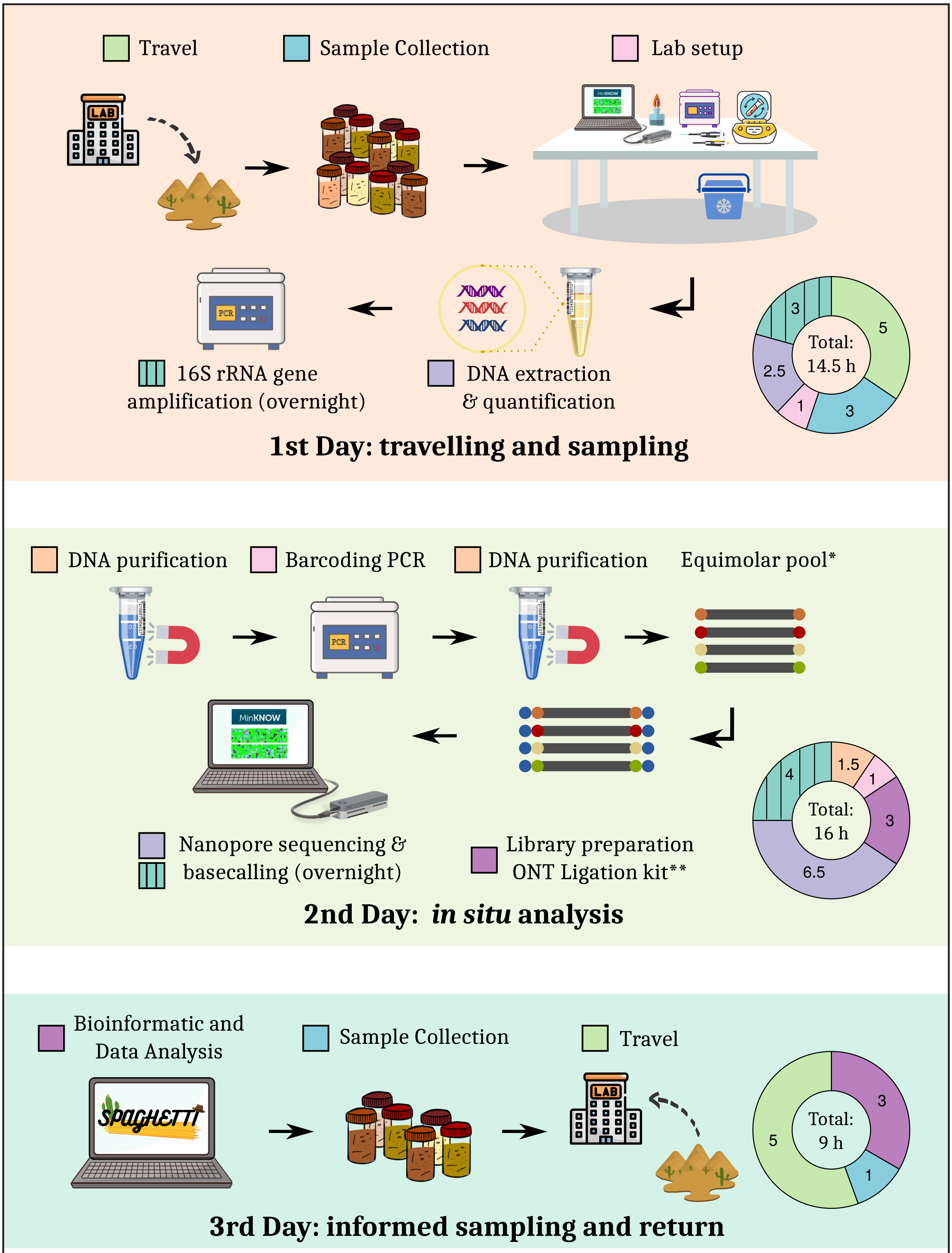

Travel
Sample Collection
Lab setup
MinKNOW
PCR
16S rRNA gene amplification (overnight)
DNA extraction & quantification
3
5
2.5
Total: 14.5 h
1
3
1st Day: travelling and sampling
DNA purification
Barcoding PCR
DNA purification
Equimolar pool*
PCR
MinKNOW
4
1.5
1
Total: 16 h
3
6.5
Nanopore sequencing & basecalling (overnight)
Library preparation ONT Ligation kit**
2nd Day: *in situ* analysis
Bioinformatic and Data Analysis
Sample Collection
Travel
SPAGHETTI
LAB
3
Total: 9 h
5
1
3rd Day: informed sampling and return

Figure IB.1. Detailed roadmap of the sampling expedition.

Methods. Briefly, 0.25 g of the samples were used to perform DNA extraction with the DNeasy Power Soil Kit. DNA was resuspended in 30 µL of sterile Milli-Q water and Qubit 1X dsDNA High-Sensitivity Assay kit was used for in-field DNA quantification. PCR amplification of the full-length bacterial 16S rRNA gene (V1-V9; ~1.45 kbp) was carried out by using the S-D-Bact-0008-a-S-16 and S-D-Bact-1492-a-A-16 primers and a simplified PCR mix (see **General Material and Methods**). All the purification steps included in the protocol were carried out with the NucleoMag kit for PCR clean up with magnetic beads. Barcodes were added by employing the PCR Barcoding Expansion Pack 1-96. Libraries were prepared using the SQK-LSK109. A R9.4.1 MinION flow cell was primed and loaded as indicated by the manufacturer. Sequencing was performed during ~6.5 h. Reads were basecalled with MinKNOW software (v. 20.06.5; core v. 4.0.5) using Guppy's (v. 4.0.9) fast basecalling model, and sequences with $Q < 7$ were discarded. 16S rRNA gene sequences were analyzed with Spaghetti (see **Pipeline 2** or https://github.com/adlape95/Spaghetti), a custom pipeline for automatic bioinformatic analysis of Nanopore sequencing data and semi-automatic exploratory analysis and data visualization.

All the analyses were run on a MSI GF63 Thin 9SC-047XES laptop (CPU: Intel Corei7-9750H, 6 core, 12 threads; RAM: 16GB; SSD: 512 Gb; Graphics Card: GeForce GTX 1650).

Isolation of bacterial strains

Upon arrival at the laboratory, the samples were homogenized by mixing 1 g of the sample with 1 mL of sterile Phosphate Buffered Saline (PBS) and serial dilutions up to 10^{-7} were performed. Then, 50 µL of the 10^{-2} to 10^{-7} dilutions were spread on Petri dishes containing either TSA medium (composition in g/L: 15.0 tryptone, 5.0 soya peptone, 5.0 sodium chloride, 15.0 agar) or SSE/HD 1:10 medium (composition detailed on the DSMZ media database, medium number 1,426). In the case of the SSE/HD 1:10 medium, duplicates of each dilution were cultured, with one of the replicates being incubated under uninterrupted artificial light and the other replicate being incubated, together with the TSA plates, under natural light. All plates were incubated in oxygenic conditions and at room temperature.

Individual colonies were selected based on their color and morphology from the TSA and SSE/HD 1:10 plates incubated under natural light after 6, 11, 18, 30 and 35 days of incubation (**Supplementary Table IB.S2,** available on https://doi.org/10.5281/zenodo.5771104). These colonies were re-streaked on fresh culture medium to isolate them in pure culture. Most of the isolates were obtained from TSA medium and from the more concentrated dilutions (10^{-2} and 10^{-4}). Regarding the samples cultured on SSE/HD 1:10 under uninterrupted artificial light, these were removed from the artificial light after 4 weeks of incubation as they did not display any microbial growth. A few days after removal, different bacterial colonies started to grow. These colonies were re-streaked on fresh culture medium and isolated in pure culture. All pure strains were cryo-preserved in glycerol (20% glycerol in an over-night culture of the strain) at −80 ºC for further uses.

Molecular identification of isolates

A loopful of each isolate, grown on solid medium, was resuspended in 100 µL of sterile Milli-Q water and subjected to a rapid DNA extraction that consisted of three cycles of boiling and freeze-thawing. Then, a PCR was performed to amplify the 16S rRNA gene using the following universal primers: 8F (5′-AGA GTT TGA TCC TGG CTC AG-3′) [243] and 1492R (5′-GGT TAC CTT GTT ACG ACT T-3′) [244]. The following conditions were used for PCR: initial denaturation (95 ºC; 5 min); amplification (24 cycles) comprising denaturation (94 ºC; 15 s), annealing (48 ºC; 15 s) and extension (72 ºC; 90 s); final extension (72 ºC; 5 min).

Amplicons were visualized by electrophoresis in a 1% agarose gel stained with GoldView DNA Safe Stain (UVAT Nerium Scientific, Spain) (100 V, 30 min). Amplicons were precipitated overnight at –20 ºC in a mixture of isopropanol 1:1 (vol:vol) and potassium acetate 1:10 (vol:vol) (3M, pH 5). The next day, DNA was pelleted by centrifugation for 10 min at 12,000 rpm, then washed with 70% ethanol and resuspended in 15 µL of sterile Milli-Q water. Amplicons were tagged using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, CA, United States) and sent for Sanger sequencing of the partial 16S rRNA gene at the SCSIE (Serveis Centrals de Suport a la Investigació Experimental) of the University of Valencia (Spain), using the same universal primers as previously mentioned (8F and 1492R).

All resulting sequences were edited with UGENE v.33 [245] to remove low quality base calls, and taxonomic identification was performed using the BLASTn tool and the 16S ribosomal RNA sequences (Bacteria and Archaea) database (NCBI). Finally, clones were dereplicated using the BLASTn tool to compare each partial 16S rRNA sequence to the rest of strains belonging to the collection of microorganisms established in this project. Any strain displaying >99.9% similarity to another strain already in the collection and isolated from the same sample was considered to be a replicate, and therefore duplicates were discarded from the collection. This was performed to avoid an overestimation of the culturable diversity, as bacterial clones of the same species are not relevant for the microbial collection.

The comparison between results from Nanopore sequencing and microbial culture collection was based on taxonomic information. Nanopore and Sanger 16S rRNA gene sequences were taxonomically classified independently, as described above. Then, the genus-level profiles were evaluated to find those taxa that had been identified by both approaches.

Data availability

The datasets generated in this study were submitted to the NCBI Bioproject database (acc. number: PRJNA749463): https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA749463.

Results

Sampling expedition roadmap

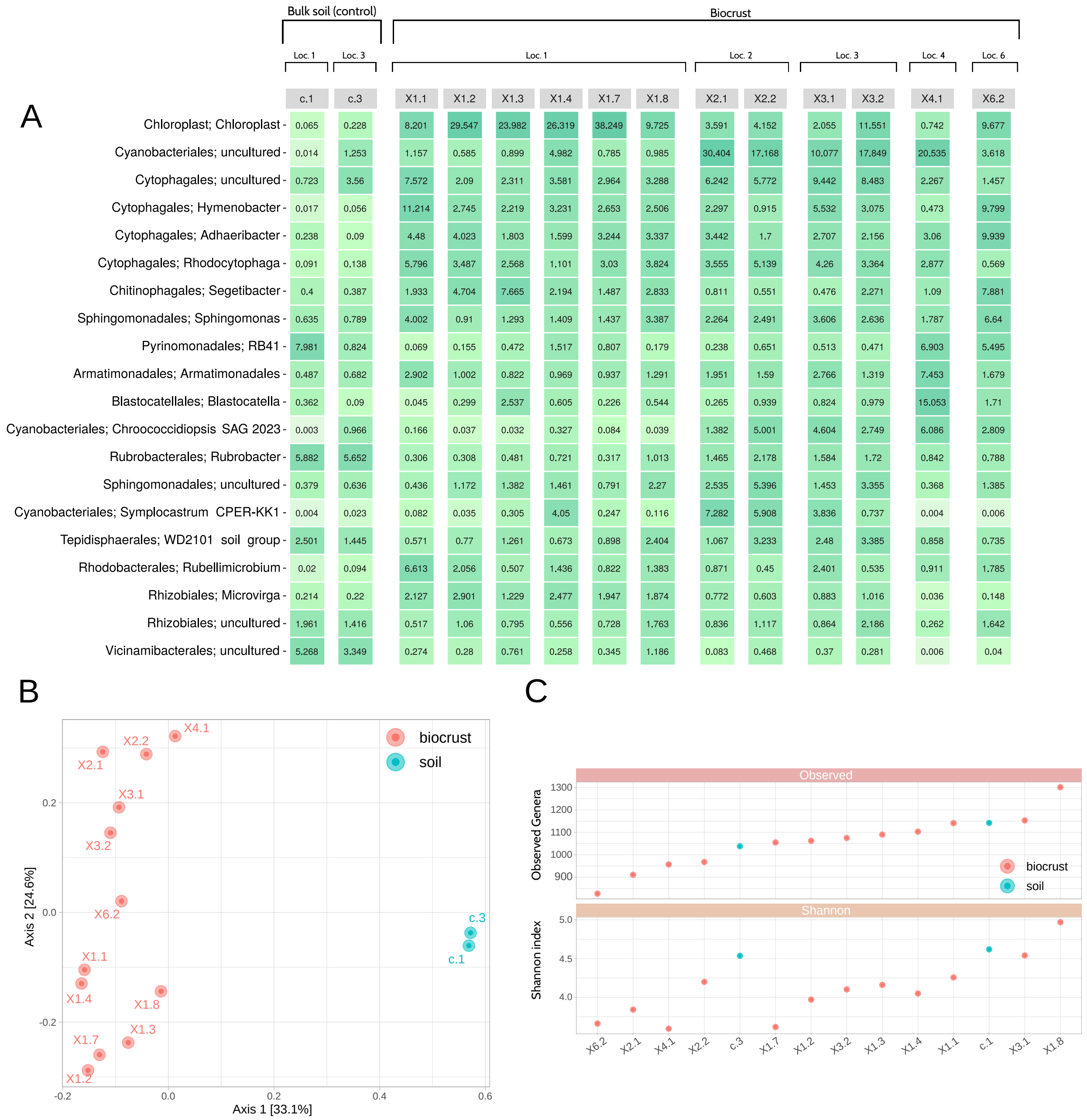
Based on previous sampling experiences in the Tabernas Desert and sequencing tests performed in the laboratory, a detailed roadmap for the expedition and the experimental procedures was designed (**Figure IB.1**). The total duration of the expedition was less than 60 h, including travelling (~25% of hands-on time) and two nights. The rest of hands-on time was spent on library preparation (~28%), sequencing and basecalling (~26%), sampling and setup (13%), and data analysis (8%). The first set of sequencing data was generated approximately 24 h after sample collection.

Microbiome sequencing and bioinformatic analysis

Twelve biocrust and two bulk soil samples ("control" samples) were collected and analyzed through full-length 16S rRNA gene sequencing using the MinION platform. A total of 1,657,804 raw reads were generated. After length and quality filtering, an average of 101,972 ± 20,949 sequences per sample were obtained (min: 50,051; max: 128,282; median Q = 10.3). Reads were subsequently analyzed by using Spaghetti (**see General Materials and Methods**), which was inspired by previous works [73, 167, 168]. Spaghetti relied on minimap2 [110] alignments against the SILVA v. 138 database [166], and taxonomic assignments were obtained in ~2 h. Other alignment tools were tested as alternatives to minimap2, but they were discarded for different reasons: BLAST took ~26 h to finish a ~1M reads analysis, while LAST exceeded the available laptop's RAM (16 Gb).

Taxonomic and diversity analysis

Spaghetti data analysis and visualization pipeline



	Bulk soil (control)		Biocrust											
	Loc. 1	Loc. 3	Loc. 1						Loc. 2		Loc. 3		Loc. 4	Loc. 6
	c.1	c.3	X1.1	X1.2	X1.3	X1.4	X1.7	X1.8	X2.1	X2.2	X3.1	X3.2	X4.1	X6.2
Chloroplast; Chloroplast	0.065	0.228	8.201	29.547	23.982	26.319	38.249	9.725	3.591	4.152	2.055	11.551	0.742	9.677
Cyanobacteriales; uncultured	0.014	1.253	1.157	0.585	0.899	4.982	0.785	0.985	30.404	17.168	10.077	17.849	20.535	3.618
Cytophagales; uncultured	0.723	3.56	7.572	2.09	2.311	3.581	2.964	3.288	6.242	5.772	9.442	8.483	2.267	1.457
Cytophagales; Hymenobacter	0.017	0.056	11.214	2.745	2.219	3.231	2.653	2.506	2.297	0.915	5.532	3.075	0.473	9.799
Cytophagales; Adhaeribacter	0.238	0.09	4.48	4.023	1.803	1.599	3.244	3.337	3.442	1.7	2.707	2.156	3.06	9.939
Cytophagales; Rhodocytophaga	0.091	0.138	5.796	3.487	2.568	1.101	3.03	3.824	3.555	5.139	4.26	3.364	2.877	0.569
Chitinophagales; Segetibacter	0.4	0.387	1.933	4.704	7.665	2.194	1.487	2.833	0.811	0.551	0.476	2.271	1.09	7.881
Sphingomonadales; Sphingomonas	0.635	0.789	4.002	0.91	1.293	1.409	1.437	3.387	2.264	2.491	3.606	2.636	1.787	6.64
Pyrinomonadales; RB41	7.981	0.824	0.069	0.155	0.472	1.517	0.807	0.179	0.238	0.651	0.513	0.471	6.903	5.495
Armatimonadales; Armatimonadales	0.487	0.682	2.902	1.002	0.822	0.969	0.937	1.291	1.951	1.59	2.766	1.319	7.453	1.679
Blastocatellales; Blastocatella	0.362	0.09	0.045	0.299	2.537	0.605	0.226	0.544	0.265	0.939	0.824	0.979	15.053	1.71
Cyanobacteriales; Chroococcidiopsis SAG 2023	0.003	0.966	0.166	0.037	0.032	0.327	0.084	0.039	1.382	5.001	4.604	2.749	6.086	2.809
Rubrobacterales; Rubrobacter	5.882	5.652	0.306	0.308	0.481	0.721	0.317	1.013	1.465	2.178	1.584	1.72	0.842	0.788
Sphingomonadales; uncultured	0.379	0.636	0.436	1.172	1.382	1.461	0.791	2.27	2.535	5.396	1.453	3.355	0.368	1.385
Cyanobacteriales; Symplocastrum CPER-KK1	0.004	0.023	0.082	0.035	0.305	4.05	0.247	0.116	7.282	5.908	3.836	0.737	0.004	0.006
Tepidisphaerales; WD2101 soil group	2.501	1.445	0.571	0.77	1.261	0.673	0.898	2.404	1.067	3.233	2.48	3.385	0.858	0.735
Rhodobacterales; Rubellimicrobium	0.02	0.094	6.613	2.056	0.507	1.436	0.822	1.383	0.871	0.45	2.401	0.535	0.911	1.785
Rhizobiales; Microvirga	0.214	0.22	2.127	2.901	1.229	2.477	1.947	1.874	0.772	0.603	0.883	1.016	0.036	0.148
Rhizobiales; uncultured	1.961	1.416	0.517	1.06	0.795	0.556	0.728	1.763	0.836	1.117	0.864	2.186	0.262	1.642
Vicinamibacterales; uncultured	5.268	3.349	0.274	0.28	0.761	0.258	0.345	1.186	0.083	0.468	0.37	0.281	0.006	0.04

Figure IB.2. Nanopore sequencing results. (A) Heatmap showing the top 20 genera detected in the samples and their relative abundances. (B) Principal Coordinates Analysis (PCoA) using the Bray-Curtis dissimilarity metric. (C) Alpha diversity analysis: Observed genera (richness) (top); Shannon index (bottom). Samples are ordered by richness. Loc. = Location.

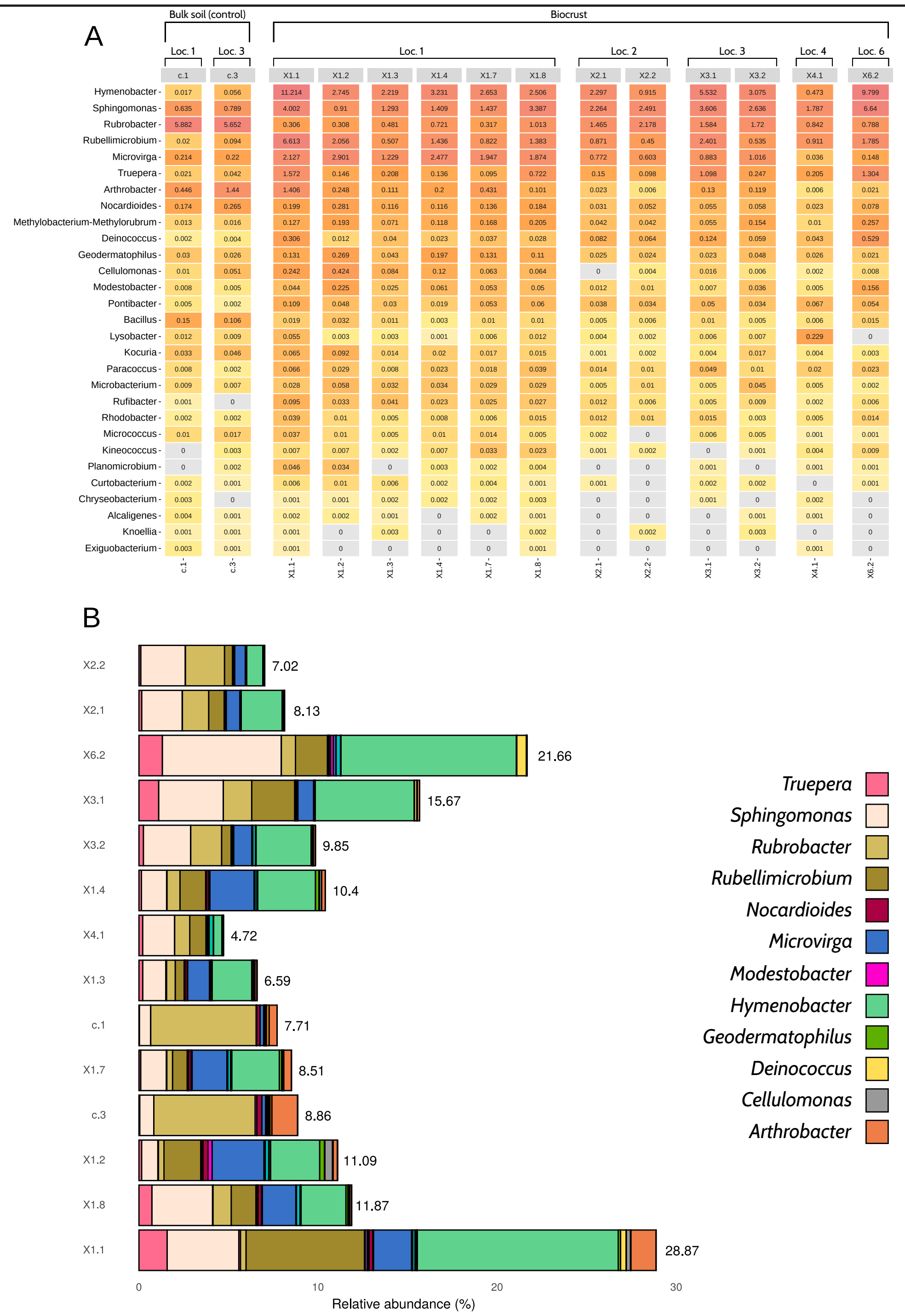


Figure IB.3. Profile of desiccation- and radiation-resistant bacteria according to Nanopore sequencing data. (A) Heatmap showing the 29 genera of interest and their relative abundances (%). (B) Barplot displaying the cumulative relative abundances of the selected taxa (n = 29). Only twelve genera have been colored in order to improve visualization, as the abundance of some taxa was so low that they could not be properly distinguished in a figure. An interactive version of this figure including the 29 genera of interest can be found in **Supplementary Figure IB.S4** (see https://doi.org/10.5281/zenodo.5771104). The relative abundance of desiccation- and radiation-resistant genera was calculated considering the whole microbial community, not only the taxa of interest. Loc. = Location.

generated several plots designed to provide a rapid overview of the taxonomy and the diversity of the samples (**Supplementary File IB.S1,** available on https://doi.org/10.5281/zenodo.5771104). At the phylum level, biocrust samples were dominated by *Cyanobacteria* (~34.5% of average relative abundance), *Bacteroidota* (~22.7%), *Proteobacteria* (~19.2%), *Acidobacteriota* (~6.0%) and *Actinobacteriota* (~4.7%), while soil samples were mainly characterized by *Actinobacteriota* (~24.8%), *Acidobacteriota* (~18.6%), *Proteobacteria* (~14.2%). *Planctomycetota* (~14.2%) and *Gemmatimonadota* (~7.3%) (**Supplementary Figure IB.S2; Supplementary Table IB.S3**).

As expected, a higher variability in the microbiome composition was detected at the genus level, with an uncultured *Cyanobacteriales* (~4.7% of average relative abundance), *Hymenobacter* (~3.9%), an uncultured *Chroococcidiopsaceae* (~3.8%), an uncultured *Spirosomaceae* (~3.7%) and *Adhaeribacter* (~3.5%) being the most dominant taxa for biocrust samples. Moreover, a considerable number of reads (~14.0%) were assigned to chloroplasts in these samples. On the other hand, soil samples were mainly characterized by *Rubrobacter* (~5.8%), *Vicinamibacteraceae* (~4.8%), an uncultured *Pirellulaceae* (~4.7%), *Pyrinomonadaceae* RB41 (~4.4%), an uncultured *Vicinamibacterales* (~4.1%), and a low presence of reads assigned to chloroplasts (~0.15%) (**Figure IB.2A; Supplementary Table IB.S4,** available on https://doi.org/10.5281/zenodo.5771104).

Beta diversity analyses showed that biocrust and soil samples were clearly distinguishable at the microbiome level. Moreover, samples tended to cluster based on their sampling location (X1, X2, X3, X4 or X6), instead of other characteristics (i.e., color and shape of the biocrust) (**Figure IB.2B**). Alpha diversity indices were used to identify the most and least rich and diverse samples, which were X1.8/X1.3/c.1 and X6.2/X2.1/X4.1, respectively (**Figure IB.2C**).

Radiation- and desiccation-resistant bacteria detection

Once the general taxonomic and diversity profiles were obtained, special attention was paid to 29 bacterial genera that had proven to be radiation- and/or desiccation-resistant according to the literature [246–252]. The objective of this analysis was to identify those samples which maximized the richness and abundance of those radiation- and/or desiccation- resistant taxa, since they should hold a greater potential for isolating and discovering microbial strains and substances of biotechnological interest.

Overall, the number of radiation- and desiccation-resistant genera detected in the samples by Nanopore sequencing was high, ranging from 23 (X2.1 & X2.2) to 29 (X1.1 & X1.8) (**Figure IB.3A**). Although some of the taxa were present in low abundance (<0.01%), the selected bacteria accounted for 11.5% of the relative abundance of the samples, in average (**Figure IB.3B**). Biocrust profiles were dominated by *Hymenobacter* (~3.9% of the total relative abundance), *Sphingomonas* (~2.7%), *Rubellimicrobium* (~1.7%), *Microvirga* (~1.3%) and *Rubrobacter* (~1%). The two bulk soil samples were mainly characterized by the presence of *Rubrobacter* (~5.8%), *Arhtrobacter* (~0.9%) and *Sphingomonas* (~0.7%) (**Figure IB.3; Supplementary Table IB.S4**).

After analyzing all the results provided by the pipeline, additional samples were collected. This time, bioprospecting activities focused on obtaining biological replicates of three selected samples: **(a)** biocrust X1.1, with the highest number of radiation- and desiccation-resistant genera (29); **(b)** biocrust X2.1, with the lowest number of radiation- and desiccation-resistant genera (23); and **(c)** bulk soil c.1, taken as a control for comparisons between biocrust and bulk soil samples.

GPS positions of the original samples were traced back and samples were identified based on the pictures that were taken on the first sampling day. Finally, two additional replicates were collected for each type of sample.

Microbial collection establishment and identification

Back in the laboratory after the expedition, all the collected samples (n = 20) were cultured under three different conditions: (1) Tryptic Soy Agar (TSA) medium, (2) SSE/HD medium (SSE/HD), and (3) SSE/HD medium + uninterrupted artificial light (SSE/HD + light). A total of 166 strains comprising 50 different genera were isolated and identified through Sanger sequencing of the partial 16S rRNA gene. The bacterial colonies displayed differences in morphology and appearance, with white, yellow, pink, red, orange and brown being the most predominant colors (**Supplementary Table IB.S2**). Initially, samples cultured on SSE/HD + light did not display any microbial growth after four weeks of

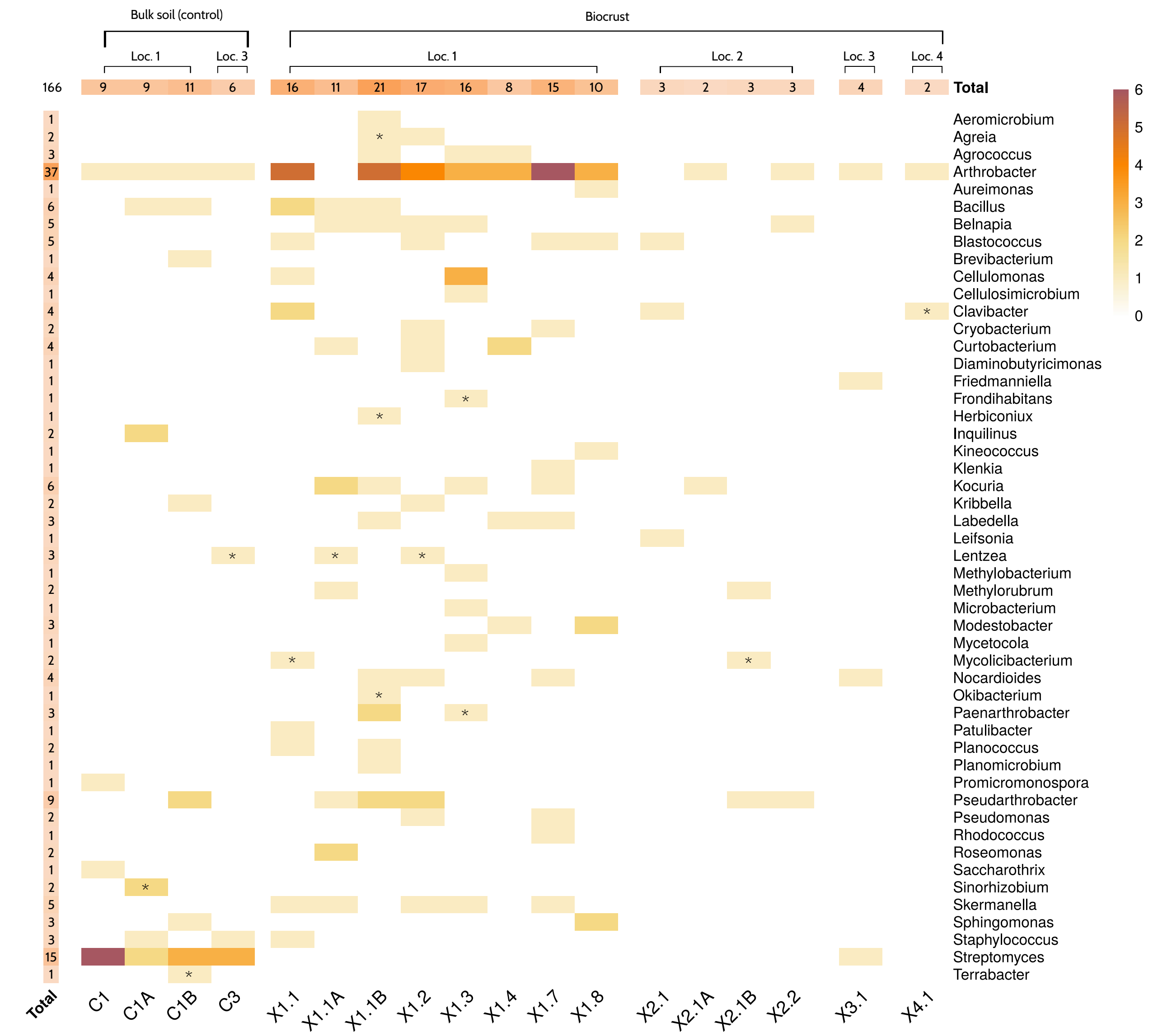


Figure IB.4. Culture collection description. Heatmap showing the number of strains isolated from each sample. Genus-level taxonomy of the strains was obtained by partial 16S rRNA gene sequencing of isolates. Letters “A” and “B” indicate the samples that were collected on the third day, after analyzing the original samples by Nanopore sequencing. Symbol “*” highlights those genera that were not originally detected in that sample by *in situ* sequencing. Only genera with a relative abundance higher than 0.001% were considered as detected. Samples 3.2 and 6.2 are not shown, since no bacterial strains were isolated from them. Loc. = Location.

incubation. For that reason, plates were removed from the artificial light, and a few days later, different bacterial colonies started to grow.

The genus *Arthrobacter* was the most represented in the microbial collection, with up to 37 isolates belonging to this taxonomic group (**Figure IB.4**). A total of 15 strains, which were mainly isolated from soil samples, were classified as *Streptomyces*. Other predominant genera in the collection were *Pseudoarthrobacter* (9 isolates), *Kocuria* (6), *Bacillus* (6), *Skermanella* (5), *Blastococcus* (5), and *Belnapia* (5). At the sample level, biocrusts collected from Location 1 (X1) presented the highest number of bacterial isolates. Specifically, samples X1.1B (21 isolates/15 unique genera), X1.2 (17/13), X1.3 (16/12), and X1.1 (16/10) showed the highest diversity of cultured bacteria. On the other hand, samples X3.2 (0/0), X6.2 (0/0), X4.1 (2/2), and X2.1A (2/2) presented the lowest diversity of isolates (**Figure IB.4**).

The taxonomic profiles obtained by Nanopore sequencing were compared to the results from the molecular identification of the isolated strains. Overall, Nanopore sequencing and culture-based data correlated well. In fact, only 14 out of the 166 isolated strains belonged to genera that were not detected in the original sample by *in situ* microbiome sequencing (**Figure IB.4; Supplementary Table IB.S2; Supplementary Table IB.S4**). Interestingly, three of the isolated genera (*Mycolicibacterium*, *Lentzea* and *Sinorhizobium*) were not detected in any sample of the dataset. After revising the database used for assigning the taxonomy of the reads, a mislabeling of those taxa at the genus level was detected. Specifically, *Mycolicibacterium* was labeled as *Mycobacterium*, *Lentzea* as *Lechevalieria* and *Sinorhizobium* as *Ensifer*. These three genera were indeed detected by Nanopore sequencing in all the samples where the strains were isolated from (**Supplementary Table IB.S4**). Finally, it is worth highlighting that

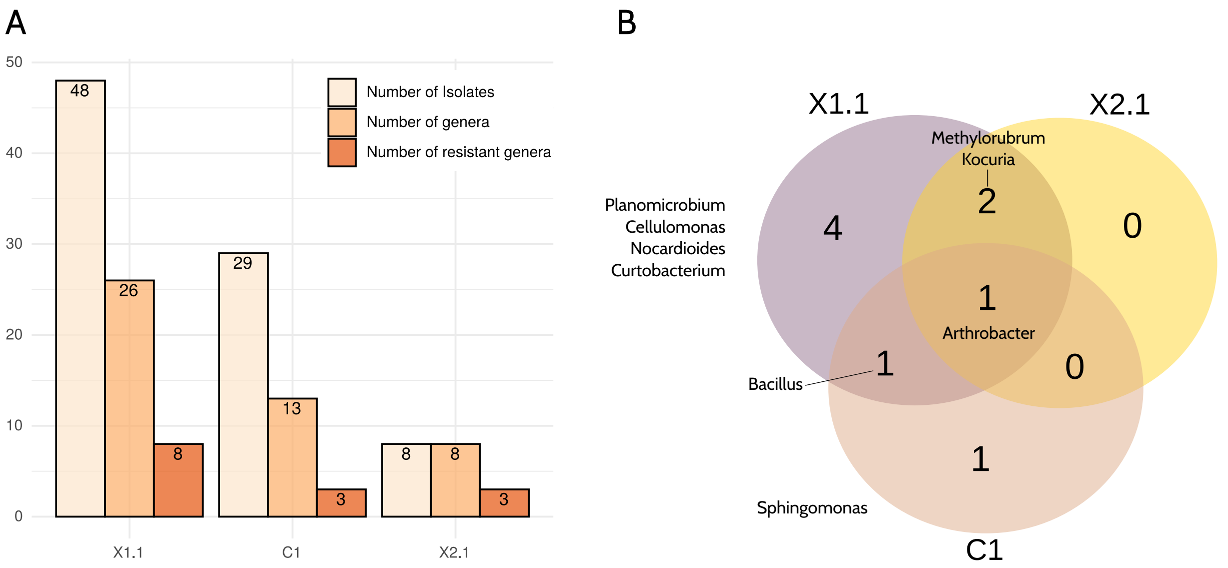


Figure IB.5. Summary of the strains isolated from the selected samples (X1.1, X2.1 and C1). (A) Number of different strains, genera and desiccation-/radiation-resistant genera isolated from each sample. (B) Venn diagram comparing the desiccation- and radiation-resistant genera isolated from any replicate of each selected sample. Note that all the identifications were obtained by partial 16S rRNA gene Sanger sequencing.

some of the most abundant radiation-resistant bacteria detected by *in situ* 16S rRNA sequencing (i.e., *Hymenobacter, Rubrobacter, Rubellimicrobium, Microvirga, Truepera*...) could not be cultured from any sample. Indeed, only 50 out of the 441 genera (11.3%) with an average relative abundance >0.01% according to Nanopore sequencing were represented in the microbial culture collection.

Among the selected samples, X1.1 yielded the highest number of total cultured strains (48), the highest number of different cultured genera -as deduced by partial 16S rRNA gene Sanger sequencing- (26), and the highest number of cultured genera classified as radiation- or desiccation-resistant according to literature (8) (**Figure IB.5**). In contrast, and as expected considering the results from *in situ* sequencing (**Figure IB.3B**), X2.1 samples displayed the lowest diversity of cultured bacteria and radiation- and desiccation-resistant genera. Moreover, almost all the genera isolated from X2.1 samples were also isolated from X1.1 (**Figure IB.5B; Supplementary Figure IB.S3; Supplementary Tables IB.S5 & IB.S6**), thus confirming the hypothesis that this sample was less valuable from the bioprospecting point of view. A different profile of bacteria was isolated from C1 bulk soil samples (**Supplementary Figure IB.S3**), with only one radiation-resistant genus -*Sphingomonas*- cultured exclusively from this type of sample (**Figure IB.5B**). Interestingly, the relative abundance of *Sphingomonas* was higher in all the biocrust samples than in bulk soil (**Figure IB.3A**), although this genus was isolated only from samples C1B and X1.8.

Focusing on the culture conditions, 124 strains were isolated from TSA (38 different genera), 24 from SSE/HD (17 different genera), and 18 from SSE/HD + light (13 different genera) (**Figure IB.6; Supplementary Table IB.S7**). Nevertheless, strains isolated from SSE/HD + light presented a significantly lower similarity to their closest type strain than strains isolated from TSA (FDR adjusted p-value < 0.05; Mann–Whitney U test), based on partial 16S rRNA gene sequencing. In fact, ~89% of the strains isolated from SSE/HD + light showed a similarity lower than 98.7% to their closest neighbor, a common threshold for defining new species [253], compared to ~46% and ~66% displayed by TSA- and SSE/HD-isolated strains, respectively (**Figure IB.6B**).

Discussion

Bioprospecting is often a unidirectional process, with scientists leaving their research institute for several days or weeks to collect samples that are only screened upon arrival at the laboratory. This is usually a blind task, since the screening results are obtained once the expedition is over. As sampling sites are generally remote and far from the researcher's laboratory, returning to the locations where bioprospecting occurred is not always viable, thus preventing further exploitation of the samples that showed a greater potential based on the screening. This work is a proof of concept of the use of portable Nanopore sequencing as a tool for guiding and informing bioprospecting activities during a sampling expedition, in our case to the only European desert, the Tabernas Desert (Almería, Spain).

ONT sequencing is a well-established technique for studying microbial communities [254], and portable sequencing (i.e., MinION) has indeed been applied to characterize microbiomes in some of the most remote places of the universe that are accessible to human beings (**see subsection 5.3. of General Introduction**). Although some authors have demonstrated the utility of *in situ* sequencing to assess the animal biodiversity in the rainforest [148, 149], the present work is, to the best of our knowledge, the first confirmation that this technology can be applied during a microbial bioprospecting expedition to improve the bioprospecting strategy itself.

Our results demonstrate that DNA analyses can be integrated into the sampling roadmap, while keeping the duration of the journey under 72 h (**Figure IB.1**).

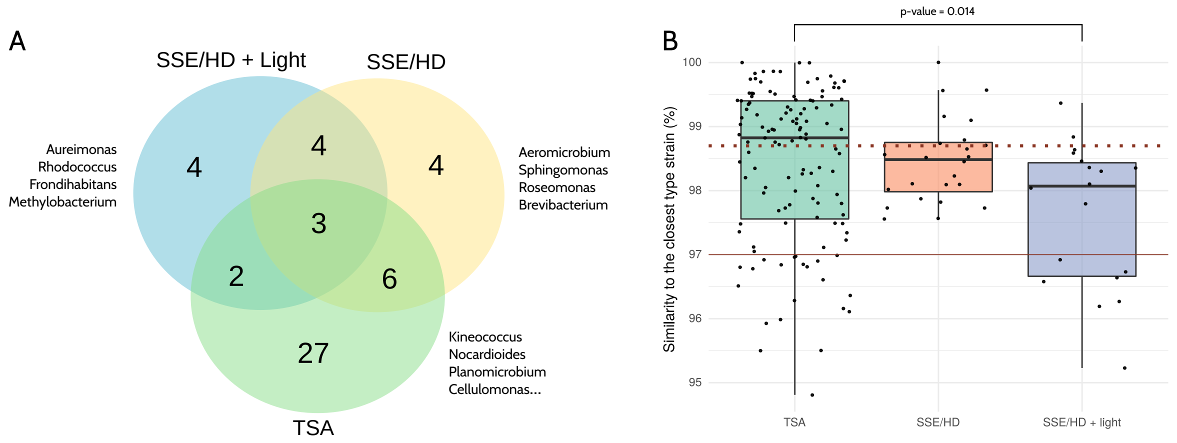


Figure IB.6. Comparison of the different culture conditions. (A) Venn diagram showing the bacterial genera isolated from each culture condition. The complete list of genera isolated from each condition can be found in Supplementary Table IB.S7. (B) Percentage of similarity shared by each strain and its closest phylogenetic neighbor according to partial 16S rRNA gene sequencing. The dotted and the solid red lines are drawn on 98.7% and 97% of similarity, respectively. The Mann–Whitney U test was applied for comparing between groups, and p-values were corrected using the Benjamini-Hochberg method. Only significant results are highlighted. Note that all the identifications were obtained by partial 16S rRNA gene Sanger sequencing.

The obtained sequencing yield was substantially higher than the output reported in other on-site studies, and it was comparable to the yield of runs performed in fully-equipped laboratories [168, 255]. It must be noted that instead of directly sequencing in the field, we decided to set up a mobile laboratory 15 km away from the sampling location in an apartment with internet and electricity access. This allowed us to apply the same protocols that we routinely use in the laboratory with little modifications, thus reducing the risk of failure during the expedition. Nevertheless, simplified protocols (i.e., Field Sequencing Kit; ONT, Oxford, UK, Cat. No.: SQK-LRK001) involving shorter preparation time and less equipment could be employed, even with the lack of electricity or internet, as has been previously demonstrated [153, 256]. Indeed, Spaghetti does not require an internet connection, so this pipeline could be also used for on-site analyses.

Different sample types (i.e., biocrust and bulk soil) were clearly distinguishable according to microbial profiles (**Figure IB.2**). As expected, *Cyanobacteria* was more abundant in biocrust samples, since these microorganisms are a crucial part of biological soil crusts, which often also harbor other organisms such as lichens, microalgae, microfungi or mosses [257,258]. This would explain the higher presence of sequences assigned to chloroplasts in this type of samples. Overall, phylum-level taxonomy was concordant with the microbial profiles expected for soil samples, with *Bacteroidota, Proteobacteria, Acidobacteriota, Actinobacteriota, Planctomycetota, Verrucomicrobiota* and *Gemmatimonadota* dominating the microbiomes [259–265]. At the genus level, differences and similarities between samples were resolved. In consequence, *in situ* Nanopore sequencing could be especially helpful for choosing those samples that maximize the microbial diversity -according to beta diversity or any other metric-, preventing the selection of samples with poor diversity or little variation for further screening, thus saving time and resources.

Taxonomic information could also be used for identifying those samples that contain the microorganisms of interest. As a proof of concept, we focused on genera that were previously described to be desiccation- and/or radiation-resistant, and which thus hold potential for biotechnological applications [37, 266]. The prevalence of these taxa in the samples collected from the Tabernas Desert was high (**Figure IB.3**). This was expected, since most of these bacteria are often found in or isolated from other arid soils and biocrusts [267–274]. Nevertheless, *in situ* sequencing in combination with our analysis pipeline led to the categorization and identification of samples that showed a greater diversity and abundance of the genera of interest. Thanks to such information, those samples -technically, biological replicates of the samples- could be further collected and thoroughly analyzed back in the laboratory.

It is well known that detecting a certain taxon by high-throughput sequencing does not necessarily mean that this taxon can be successfully isolated from the sample. In our case, culture-based and Nanopore sequencing data correlated well (**Figure IB.4**), although an important fraction of the genera detected with the sequencing approach was not represented in the microbial culture collection. This could be expected given that a significant number of prokaryotic taxa are virtually 'unculturable'. In any case, the sample that held the greatest potential at the microbiome level -according to Nanopore data- (X1.1) also resulted in the most interestingly complex set of culturable bacteria (**Figure IB.5**), despite using a relatively simple culturing approach. On the other hand, some of the most dominant bacteria according to sequencing data could not be isolated from any sample (i.e., *Hymenobacter* or *Rubrobacter*) very likely due to culturing biases. Although this limitation is inherent to bioprospecting strategies that rely on obtaining microbial cultures, knowing the presence of a certain taxonomic group in the sample would allow for the use of microorganism-specific culture conditions or enrichment methods, thus increasing the chances of success.

In general, the profile of bacteria isolated from the Tabernas Desert was similar to the one previously described [37]. *Arthrobacter* was the predominant genus, which is consistent with the observations of da Rocha et al. [275]. Other bacteria, such as *Belnapia*, *Kocuria* or *Skermanella* were also recurrent in biocrust samples. Nevertheless, up to 29 genera isolated in this study were not recovered by Molina-Menor et al. [37].

Interestingly, some of the isolated bacteria may represent new species according to partial 16S rRNA gene sequencing, showing the great, yet to be discovered, ecological and biotechnological potential hidden in the Tabernas Desert. Although full 16S rRNA gene sequences and genomes should be retrieved for circumscribing new taxa [253], bacteria isolated from SSE/HD + light displayed a lower similarity to any other previously described type strain (**Figure IB.6**). These results were indeed obtained by serendipity, as bacterial growth was only detected after removing the culture plates from artificial light (~4 weeks after plating), which was not the original idea.

Despite the promising results obtained in this proof of concept, we have identified some limitations of *in situ* Nanopore sequencing. The first one is the taxonomic resolution of 16S rRNA gene sequencing. Although long-read platforms have the ability to sequence the full-length 16S rRNA gene, the intrinsic error associated with ONT sequencing hampers species-level identification. This error also hinders the direct comparison between Nanopore-based microbiome sequencing and the 16S rRNA gene sequences obtained from the isolates by Sanger sequencing, as it would be difficult to discern if a particular fraction of Nanopore reads actually comes from a specific strain in the collection or from a phylogenetically related strain (or even species) that may or may not have been isolated. For that reason, we decided to perform the analyses at the genus level and to compare the taxonomic profiles instead of comparing the sequences. Nanopore-based, 16S rRNA gene sequencing has proved to be robust for microbiome characterization at this taxonomic level, showing a performance similar to Illumina sequencing [71, 73, 255, 276, 277]. However, as the final objective of bioprospecting is to actually isolate (culture) the bacterial strains, it must be noted that phenotype can greatly vary among members of the same genus or even species, so genus-resolved taxonomy could be insufficient in some cases. Recent studies have shown that species-level resolution is feasible thanks to advances in software [278, 279], while other works demonstrated that improved taxonomic resolution could be achieved by using longer amplicons (16S-ITS-23S) [73, 107]. Moreover, Nanopore sequencing errors are also decreasing due to improvements in basecallers and chemistries, which have allowed to reach up to 99.3% of modal accuracy on raw reads [99]. If accuracy continues to increase at this rate, it is reasonable to think that species-level identifications, and even strain-level resolution in some cases, may be achieved in the near future. Nevertheless, high-accuracy basecalling models are based on complex machine learning methods that require longer execution time, so improvements on the speed of these models are still required for being used in real-time applications [280].

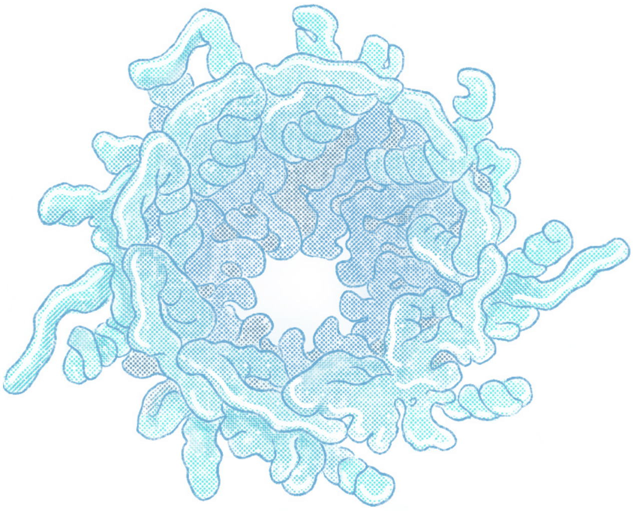
It must be highlighted that this study was focused on the detection and isolation of potential radiation- and desiccation-resistant bacteria according to their taxonomic affiliation and according to the previous bibliography describing this type of features in particular genera. Our approach is thus a proof of concept that a wide taxonomic group can be identified in the samples by using Nanopore sequencing, but 16S rRNA gene itself would not be an accurate predictor of the actual ability of the isolates to resist radiation or desiccation [281]. If the purpose of the bioprospecting expedition is to detect specific functional activities, shotgun metagenomic data would be needed to resolve the taxonomy at the strain level [282] and to ascertain the functional potential of the different members of the microbial community according to their gene content. In this regard, it has to be noted that ONT sequencers tend to incorporate indel errors on the reads that complicate the functional prediction [283] (**see Chapter II**), and this is therefore a current limitation of the informed bioprospecting strategy we are describing in this work.

Finally, sequencing strategies show the microbiome composition based on relative abundances, which may mislead the results interpretation. For instance, if a taxon is detected in Sample 1 and in Sample 2 at 10% and 1% of relative abundance, respectively, that does not imply that Sample 1 has a higher absolute abundance of the target bacteria, since the total microbial load of the samples has not been measured. This should be taken into account when selecting the samples of interest for further exploitation.

Notwithstanding the limitations, our results clearly show that Nanopore sequencing is a powerful tool for deciphering the microbial composition of different samples during a bioprospecting expedition, and that it can contribute to optimize the sampling strategy *in situ*. With microorganisms colonizing almost any known biotope [284–286], an instrument able to resolve microbial communities inhabiting different niches is a valuable resource that can be used for targeting sample collection. Therefore, it can be envisaged a close future in microbial ecology, in which bioprospecting journeys will start with a preliminary sampling step, coupled to Nanopore-based *in situ* analysis, which will enable a second, more targeted sampling (of specific plant species, soil depths, geological substrates, salt concentration, humidity level, etc.) in a very short time lapse. This strategy will both ease further work in the lab and increase the chances of identification of the target microbial taxa and/or biomolecule of interest.

The content of this chapter has been published in Frontiers in Microbiology as: **Latorre-Pérez A**, Gimeno-Valero H, Tanner K, Pascual J, Vilanova C, Porcar M. A round trip to the desert: *in situ* Nanopore sequencing informs targeted bioprospecting. *Front. Microbiol.* 2021;12. doi:10.3389/fmicb.2021.768240 **(Publication V in Appendix C)**

Chapter II. Towards long-read metagenomics: a comparative study of assembly methods for Nanopore sequencing



Chapter II. Towards long-read metagenomics: a comparative study of assembly methods for Nanopore sequencing

Abstract:

Metagenomic sequencing has allowed for the recovery of previously unexplored microbial genomes. Whereas short-read sequencing platforms often result in highly fragmented metagenomes, Nanopore-based sequencers could lead to more contiguous assemblies due to their potential to generate long reads. Nevertheless, there is a lack of updated and systematic studies evaluating the performance of different assembly tools on Nanopore data. In this chapter, we have benchmarked the ability of different assemblers to reconstruct two different commercially-available mock communities that have been sequenced using Oxford Nanopore Technologies platforms. Among the tested tools, only metaFlye, Raven, and Canu performed well in all the datasets. These tools retrieved highly contiguous genomes (or even complete genomes) directly from the metagenomic data. Despite the intrinsic high error of Nanopore sequencing, final assemblies reached high accuracy (~99.5 to 99.8% of consensus accuracy). Polishing strategies demonstrated to be necessary for reducing the number of indels, and this had an impact on the prediction of biosynthetic gene clusters. Correction with high quality short reads did not always result in higher quality draft assemblies. Overall, Nanopore metagenomic sequencing data -adapted to MinION's current output- proved sufficient for assembling and characterizing low-complexity microbial communities.

Background

Metagenomic sequencing has become a powerful tool for recovering and studying individual genomes directly from complex microbiomes [287–289], leading to the identification and description of new -and mostly unculturable- taxa with meaningful implications [290]. Illumina has been the most widely used platform for metagenomic studies. As thoroughly discussed in the introduction to this thesis, Illumina reads are characterized by their short length and high accuracy (**Table GI.1**). When performing *de novo* assemblies, Illumina sequences often result in highly fragmented genomes, even when sequencing pure cultures [176, 183]. This is a consequence of the inability to correctly assemble genomic regions containing repetitive elements that are longer than the read length [183]. This fragmentation problem is magnified when handling metagenomic sequences due to the existence of intergenomic repeats that are shared by more than one taxon present in the microbial community [291]. It has to be noted that microbial communities often contain related species or sub-species in different -and unknown- abundances, resulting in extensive intergenomic overlaps that can hinder the assembly process [111, 292].

Third-generation sequencing (TGS) platforms have recently emerged as a solution to resolve ambiguous repetitive regions and to improve genome contiguity. Despite the considerable error associated with these technologies (**Table GI.1**), their ability to produce long reads can be used to assembly genomes with a high degree of completeness [105, 293]. Currently, the most widely used third-generation technologies are Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), both based on single molecule sequencing, and therefore, PCR-free. PacBio was the first long-read technology established in the market. On the other hand, ONT platforms are becoming increasingly popular among researchers, mainly thanks to MinION sequencers. MinION is a cost-effective, portable sequencing device, which enables real-time analysis pipelines. This platform has been broadly applied over the last few years for sequencing complete prokaryotic and eukaryotic genomes [105, 294–296], and characterizing microbial communities [73, 107].

Benchmarking is the usual way to evaluate genomic methodologies(i.e.,DNAextraction,librarypreparations, etc.) and bioinformatic tools. In the metagenomic context, benchmarking studies are frequently based on mock communities. A mock community is an artificial microbial community in which the abundance of all the microorganisms is known [297]. Mock communities can be generated in silico [298] or experimentally, as a mixture of defined DNA proportions. For *de novo* assemblies, a great effort has been made in order to benchmark all the available tools and methodologies suitable for studying microbial ecosystems via Illumina shotgun sequencing [111, 177, 299]. Nevertheless, due to the highly dynamic development of new software applicable to ONT platforms, we found that the few evaluation studies that have been focused to date on Nanopore-based metagenomic assembly did not cover the current spectrum of available assemblers [181, 300, 301].

In this chapter, we have used the data generated by Nicholls et al. [186] to comprehensively assess the current state-of-art of *de novo* assembly tools suitable for Nanopore-based, metagenomic sequencing. Original data was generated through metagenomic sequencing of two microbial communites (ZymoBIOMICS Microbial Community Standards CS and CSII) with both GridION and PromethION platforms. Overall, this work demonstrates the suitability of using Nanopore sequencing for assembling low-complexity microbial communities, and paves the way towards the standardization of bioinformatic pipelines for long-read metagenomics.

Materials and methods

Dataset description

Benchmarking datasets were extracted from Nicholls et al. [186] (Bioproject Accesion: PRJEB29504), and consisted of the high-coverage sequencing of two individual mock communities: ZymoBIOMICS Microbial Community Standards CS Even and CSII Log (Zymo Research, CA, United States, Cat. No.: ZRC190633 and ZRC190842) with both GridION and PromethION platforms. The mock communities contained the same species (eight bacteria; two yeasts), but differed in the expected proportion for each microorganism. CS mock community had a homogeneous distribution of microorganisms (12% for each bacteria and 2% for the yeasts), while the species present in CSII were distributed on a logarithmic scale, with relative abundances ranging from 89.1 to 0.000089% (**Table II.1**). Following the nomenclature from Nicholls et al. [186], the terms “Even” and “Log” were used when referring to the CS Even or the CSII Log mock communities, respectively.

Nicholls et al. [186] yielded ~14 Gbp of data on a single GridION flowcell (48 h of sequencing) and~152 Gbp on the PromethION platform (64 h of sequencing). In order to reduce the necessary computational effort, we performed an initial subsampling of this data. In particular, GridION and PromethION datasets were subsampled at two different sequencing depths (3 Gbp and 6 Gbp) to recreate MinION runs with different outputs, and the yield matched the output described in recent shotgun sequencing experiments based on MinION [140, 183, 301–305]. Subsampling was performed by selecting the top lines of the FASTQ files. Nevertheless, the most promising tools were further tested using the original GridION dataset (14 Gbp) to check their computational efficiency and general performance. All the datasets were processed with porechop (v. 0.2.4) (https://github.com/rrwick/Porechop) to remove adapters from read ends and split sequences with internal adapters.

Reference genomes

All the species included in the mock communities had an available reference genome sequenced with a combination of Illumina and Nanopore reads (available at https://doi.org/10.5281/zenodo.3935737). These assemblies, provided by Zymo Research Corporation, consisted of eight complete genomes for the bacterial strains, and two draft genomes for the yeasts.

Table II.1. Description of the microorganisms comprising the ZymoBIOMICS mock communities and their theoretical composition.

Species	Abbreviation[1]	Estimated genome size (Mbp)	Composition Even (CS)	Composition Log (CSII)
Bacillus subtilis	***B. subtilis***	4,134	12.00%	0.89%
Cryptococcus neoformans	***C. neoformans***	18,599	2.00%	0.00089%
Enterococcus faecalis	***En. faecalis***	2,965	12.00%	0.00089%
Escherichia coli	***E. coli***	5,140	12.00%	0.89 %
Lactobacillus fermentum	***L. fermentum***	2,012	12.00%	0.0089%
Listeria monocytogenes	***Li. monocytogenes***	3,008	12.00%	89.1%
Pseudomonas aeruginosa	***P. aeruginosa***	6,592	12.00%	8.9%
Saccharomyces cerevisiae	***S. cerevisiae***	11,864	2.00%	0.89%
Salmonella enterica	***Sa. enterica***	4,781	12.00%	0.89 %
Staphylococcus aureus	***St. aureus***	2,838	12.00%	0.00008%

1. Custom abbreviations were designed to facilitate the differentiation of bacterial genera whose names begin with the same letter. These abbreviations do not follow the binomial nomenclature system.

Nicholls et al. [186] sequenced and assembled each genome again from pure cultures using Illumina reads only. In the present work, however, ZymoBIOMICS genomes were used as references for carrying out the comparative analyses, due to their higher level of completeness. Although these reference genomes cannot be considered as "gold standards", Goldstein et al. [183] demonstrated that the error profile obtained through hybrid assembly (ONT+Illumina) was similar to the one obtained with MiSeq-only assembly, but the former resulted in higher contiguity. Reference genomes were gathered in a single multi-FASTA file to create a single-reference metagenome.

Evaluation of the assemblies

The assemblies were evaluated as described in **General Material and Methods**. Briefly, both metaQUAST (v. 5.0.2) [112] and minimap2 (v. 2.15) [110] were used to align the assemblies to the reference metagenome.

The presence of mismatches (SNPs) and indels was analyzed using two different strategies: minimap2 plus bcftools, and MuMmer4 plus count_SNPS_indels.pl [183]. In the former case, the in-house script indels_and_snps.py (https://doi.org/10.5281/zenodo.3935763) was applied to quantify the variants. The number of variants was normalized to the total assembly size of each metagenome in both strategies.

All the assemblers were run on the same desktop computer (CPU: AMD RYZEN 7 1700X 3.4GHZ; Cores: 8; Threads: 16; RAM: Corsair Vengeance 64 GB; SSD: Samsung 860 EVO Basic SSD 500 GB) working under Ubuntu 18.04 operative system. The time required by each tool to perform the assembly was measured with the built-in bash version of the "time" command.

Assembly polishing

Polishing was carried out as reported in **General Material and Methods:** ONT or Illumina reads were used for iteratively running 4 rounds of Racon [185], and only the draft assemblies corrected with ONT reads were further polished with Medaka (https://github.com/nanoporetech/medaka). Indels and SNPs were evaluated using the strategy based on Mummer.

Data availability

Raw data was deposited in the NCBI database under the BioProject number PRJNA564477 (https://www.ncbi.nlm.nih.gov/bioproject/564477). Raw datasets from Nicholls et al. [186] can be downloaded from the ENA (https://www.ebi.ac.uk/ena/data/view/PRJEB29504). All the code used in this study is publicly available at doi: https://doi.org/10.5281/zenodo.3935763. It includes the bash scripts designed for the automatic execution of the different bioinformatic analysis, the R code and data (CSV tables), and other in-house and third-party scripts needed to reproduce the analyses.

Results

Subsampling

Data released by Nicholls et al. [186] (ultra-deep sequencing of two different mock communities using GridION and PromethION platforms) was used in order to study the suitability of Nanopore sequencing to characterize low complex microbial communities. The mock communities were composed by the same ten microorganisms, but in different proportions (**Table II.1**). With the aim of reducing the computational resources needed for the first screening of the selected assemblers, the GridION and PromethION datasets were subsampled to obtain an output comparable with recent genomic or metagenomic studies based on MinION (approximately 3 Gbp and 6 Gbp) [140, 183, 301–305]. In general, mean read length remained the same in the subsampled datasets in comparison to the original sequencing data [186]. However, read quality was higher in the subsampled dataset. This fact suggested a bias towards higher qualities at the start of the run, since subsampling was carried out by selecting the top reads of the original files (**Table II.2**). In fact, the bottom reads which are acquired later in the sequencing run displayed the same quality than the whole dataset.

Metagenome assembly

From the selected tools, we were able to correctly install and run nine out of the ten long-read assemblers, and two out of the three short-read assemblers (**Table GMM.1**). In total, 74 assemblies were generated, 40 for the Even mock community and 34 for the Log community. Six assemblies could not be completed because miniasm and Pomoxis failed to run with the 6 Gbp Log datasets, whereas Unicycler failed to run with the 3 Gbp Log datasets. The total size of each draft assembly and the fraction of metagenome recovered from the reference genomes were evaluated for the Even datasets in order to obtain a first view of the general tool performance.

Overall, long-read assemblers resulted in a total assembly size closer to the theoretical size, and also recovered a larger metagenome fraction, with some exceptions (**Figure II.1**). Nevertheless, large differences were detected for both metrics among the assemblers. All the assemblers were far from recovering the totality of the metagenome, both in the 3 Gbp and the 6 Gbp datasets (**Figure II.1A**). It must be noted that metaQUAST and minimap2 results were consistent for the long-read assemblers, but not for the short-read assemblers, where minimap2 metric was significantly higher (**Figure II.1B**). MetaFlye (both versions) yielded the best assemblies in terms of total metagenome size and metagenome recovery except for the minimap2 metric, followed by Pomoxis, Canu and Raven (previously known as Ra). Interestingly, assembly pipelines based on the miniasm algorithm (Pomoxis, Unycicler, and miniasm itself) presented huge variations in their performance. Unicycler and miniasm performed relatively well for the 3 Gbp dataset, but when using 6 Gb, the final assembly did not improve significantly in the

Table II.2. Sequencing statistics for the original and the subsampled datasets.

ORIGINAL DATASET					
Dataset Name	**Data size (Gbp)**	**Number of reads**	**Mean read length** (bp)	**Mean read quality** (Q)	**Intended subsampling (Gbp)**
Even GridION	14.01	3,491,078	4,012.3	8.4	3
Log GridION	16.03	3,667,007	4,372.0	8.0	3
Even PromethION	146.29	36,527,376	4,005.0	7.3	3
Log PromethION	148.03	35,118,078	4,215.2	7.6	3
Even GridION	14.01	3,491,078	4,012.3	8.4	6
Log GridION	16.03	3,667,007	4,372.0	8.0	6
Even PromethION	146.29	36,527,376	4,005.0	7.3	6
Log PromethION	148.03	35,118,078	4,215.2	7.6	6
NEW DATASET					
Dataset Name	**Data size (Gbp)**	**Number of reads**	**Mean read length** (bp)	**Mean read quality** (Q)	**SRA accession number**
Even GridION	3.04	747,682	4,069.5	8.9	SRX6817349
Log GridION	3.05	685,926	4,451.0	8.7	SRX6817351
Even PromethION	2.98	748,367	3,981.0	8.2	SRX6817353
Log PromethION	2.99	711,524	4,203.3	8.3	SRX6817355
Even GridION	6.09	1,495,377	4,073.9	8.8	SRX6817350
Log GridION	6.09	1,371,820	4,442.4	8.5	SRX6817352
Even PromethION	5.97	1,496,919	3,988.8	8.2	SRX6817354
Log PromethION	5.96	1,422,918	4,185.8	8.2	SRX6817356

case of miniasm, and the general performance was highly reduced for Unicycler. This is in contrast to Pomoxis, which produced the second most complete assemblies with both dataset sizes. Although based on miniasm, it is worth highlighting that Unicycler's pipeline is designed for single isolate assembly, so reduced performance was expected for metagenomic studies. Finally, Redbean (previously known as wtdbg2) and Shasta resulted in poor assembly performance in comparison to the other long-read tools.

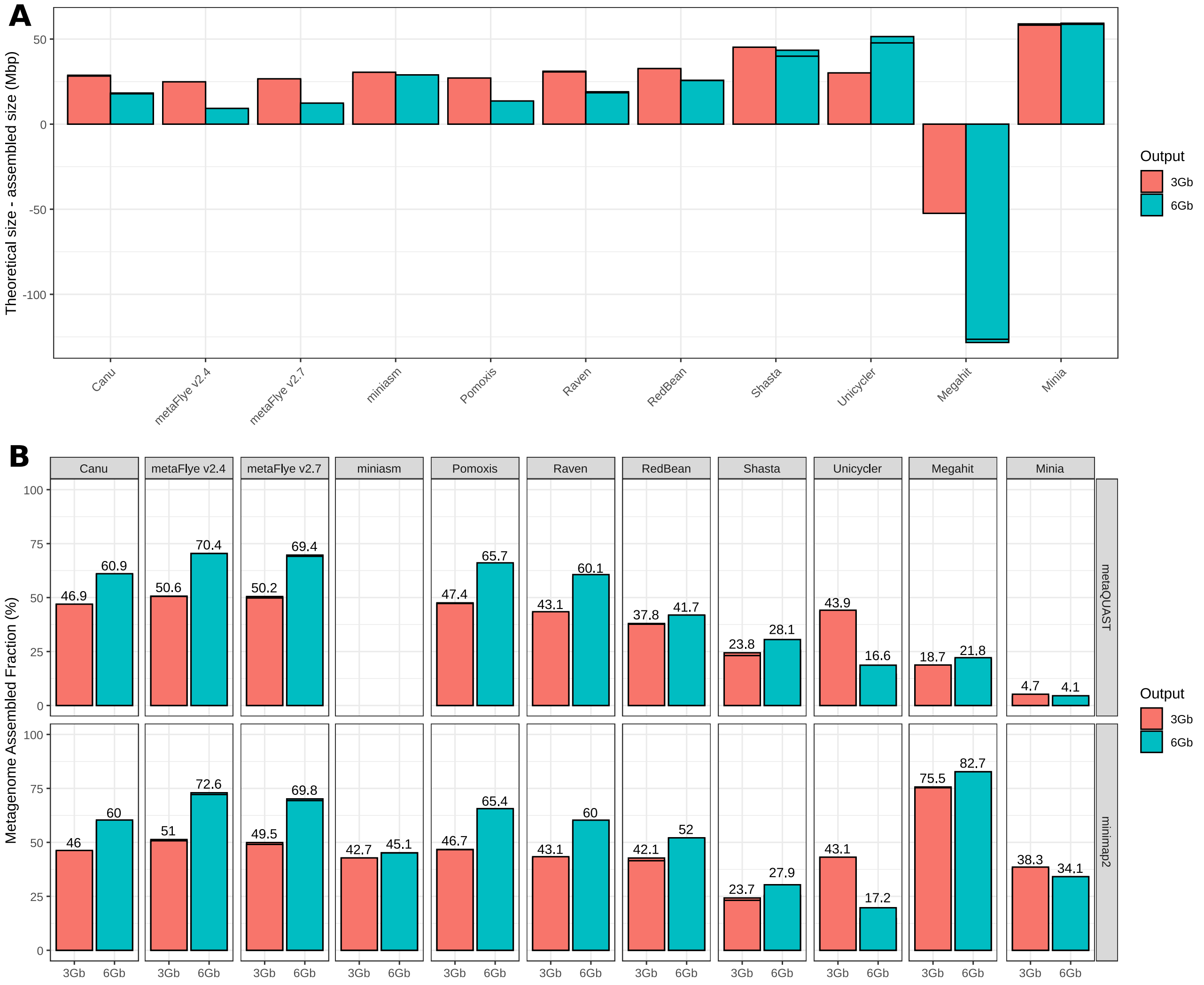


Figure II.1. Evaluation of metagenome assembly size corresponding to each tested tool for the subsampled Even datasets. (A) Total assembled size of draft assemblies with respect to the total size of the reference metagenome; (B) Fraction of the reference metagenome covered by the draft assembly, calculated by two different methods: metaQUAST (top panel) and minimap2 + BBTools (bottom panel).

MetaQUAST was used for further evaluating the degree of completeness of each individual draft genome (**Figure II.2**). As expected, yeast genomes were generally less recovered than bacterial ones, due to their lower abundance (2%) and higher size, explaining the low metagenome fraction generally recovered by all the assemblers (**Figure II.1**). In fact, the maximum average recovery fraction for the bacterial genomes was 99.92% (**Supplementary Figure II.S1**). Minia and Megahit were not able to recover any single genome with high completeness (>95% of genome coverage) in any dataset.

For the 3 Gbp dataset, metaFlye (both versions) and Unicycler recovered the eight bacterial genomes with a high completeness (>98.6%), while Pomoxis achieved lower recovery fractions for two genomes (~96.9 to 97.4%). Raven and Canu resulted in reduced recovery percentages, but still retrieved all the prokaryotic genomes with a mean covered fraction greater than 85% and 87%, respectively. Redbean and Shasta achieved particularly low fractions of genome recovery.

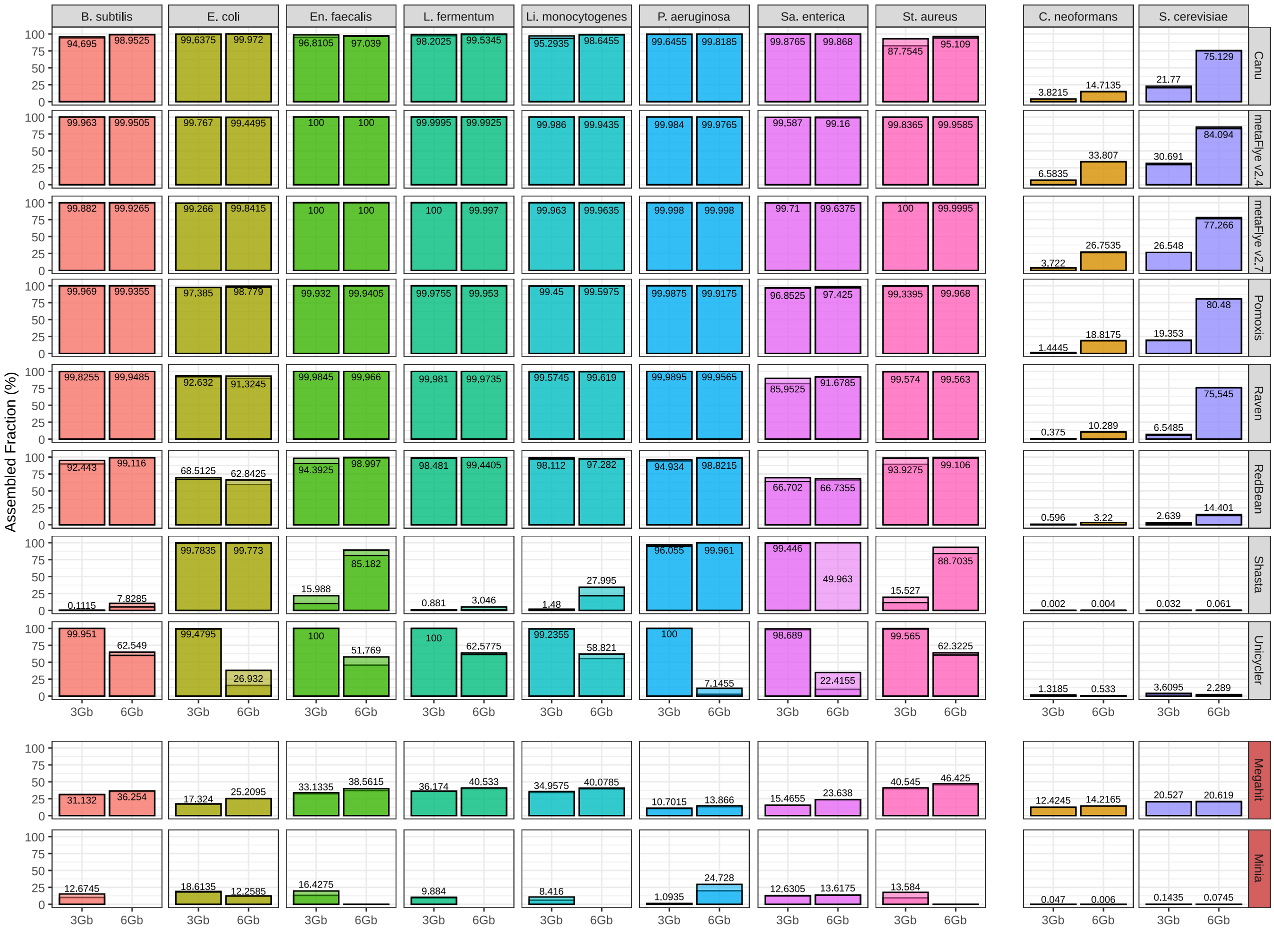


Figure II.2. Fraction of the genome covered by draft assemblies obtained with each tool, and for each individual microorganism (subsampled Even datasets). Miniasm assemblies are not shown, since it was not possible to evaluate them with metaQUAST.

For the 6 Gbp dataset, Unicycler performance decreased substantially as noted in **Figure II.1,** while Canu, Pomoxis, Raven and metaFlye achieved similar or better results. In general, metaFlye displayed the best performance on both dataset sizes in terms of genome recovery, closely followed by Pomoxis. This trend was also observed when analyzing the proportion of yeast genomes recovered by each tool. In this context, it is important to highlight that metaFlye's ability to recover eukaryotic genomes was reduced when using metaFlye v2.7. This is due to the lower number of misassemblies retrieved by this metaFlye version, indicating that the reduced fraction of genome recovery is compensated with more reliable assemblies (**Supplementary Figure II.S2**).

Results were confirmed when analyzing the Log mock community (**Supplementary Figure II.S3**). Canu, metaFlye, Raven and Pomoxis were able to recover *Listeria monocytogenes* and *Pseudomonas aeruginosa* genomes (89.1% and 8.9% of total genomic DNA in the Log mock community, respectively) with a level of completeness higher than 99%. These assemblers also recovered a significant fraction of *Bacillus subtilis* (0.89%

of total genomic DNA in the Log mock community). In fact, Raven was able to reconstruct>99% of its genome using the 6 Gbp datasets, whereas metaFlye recovered~98%. In this case, both tools outperformed Canu. Nevertheless, Raven did not recover a significant fraction of *Saccharomyces cerevisiae*, whereas Canu and metaFlye did (>8%). Pomoxis worked correctly when using the 3 Gbp datasets, but failed to run with both 6 Gbp files. The other tools based on the miniasm algorithm also failed to run the 3 Gbp (Unicycler) and/or 6 Gbp datasets (miniasm). In all cases, the error was related to memory usage and accession (segmentation violation), and could not be solved. Nevertheless, using a computer with more RAM would help to easily overcome this problem. Shasta, RedBean, Minia and Megahit performed poorly in comparison to the other tools (**Supplementary Figure II.S3**). It has to be noted that Shasta and RedBean were not originally designed to work with metagenomic data, which could result in problems when handling uneven coverage.

Regarding the time consumed by each tool, Shasta was the fastest assembler (**Figure II.3A**). This tool was able to assemble the 6 Gbp datasets in only 285 s, approximately. RedBean and miniasm were the second and third most fast software, followed by Raven (1.5–1.9 times faster than metaFlye v2.7). MetaFlye was 1.4–1.7 times faster than Pomoxis, and 3.8–5.5 times faster than Canu, which proved to be the slowest tool. These trends were also found in the Log mock community (**Supplementary Figure II.S4**), where Canu spent up to 22 h reconstructing a draft metagenome assembly from the 6 Gbp datasets. In this case, Raven was faster than metaFlye v2.7 for the 3 Gbp datasets, but not for the 6 Gbp ones.

General metagenome statistics (N50, L50, and number of contigs) were evaluated using QUAST (**Figure II.3; Supplementary Table II.S1**). It has to be stressed that the comparisons based on these metrics are difficult to analyze due to the large variation in the general performance among the different assemblers. For instance, Shasta resulted in the highest N50 and the lowest L50 values for the 6 Gbp dataset, but this tool was able to cover less than 35% of the metagenome. In fact, the total assembly size for Shasta was approximately 18–21 Mbp, in comparison to the 49–53 Mbp assembled by metaFlye.

As expected, short-read assemblers did not perform well with Nanopore data, resulting in thousands (Minia), or even hundreds of thousands of contigs (Megahit). Interestingly, long-read assemblers resulted in more fragmented draft genomes when using the 6 Gbp datasets. Except for Shasta, the other long-read assemblers also reduced their N50 and increased their L50 and number of contigs score when using 6 Gbp. Goldstein et al. [183] demonstrated that Canu assemblies improved with higher coverage when assembling bacterial isolates. This fact suggests that the loss of contiguity detected may be a direct consequence of a higher recovery rate of yeast genomes, which might be more fragmented. Indeed, assembly statistics of the Canu draft assemblies remained almost the same for the bacterial species when using 3 or 6 Gbp (**Supplementary Table II.S2**). Finally, metaFlye and Raven resulted in a more contiguous assembly with higher N50 and lower L50 in comparison to the other best performing tools (Canu and Pomoxis), for both 3 and 6 Gbp datasets (**Figure II.3; Supplementary Table II.S1**). Remarkably, metaFlye v2.7 yielded slightly better results than metaFlye v2.4 (**Figure II.3B–D**), and required less time (**Figure II.3A**).

ONT hardware, protocols and software are in constant development, leading to large improvements in short periods of time. Recently, an optimized DNA extraction and purification methodology has allowed to reach an average yield of~15.9 Gbp per flowcell [306]. For that reason, we decided to run the most promising assemblers directly on GridION's original data (Even mock community; 14 Gbp). RedBean was included because of its computational efficiency, which is a key factor for

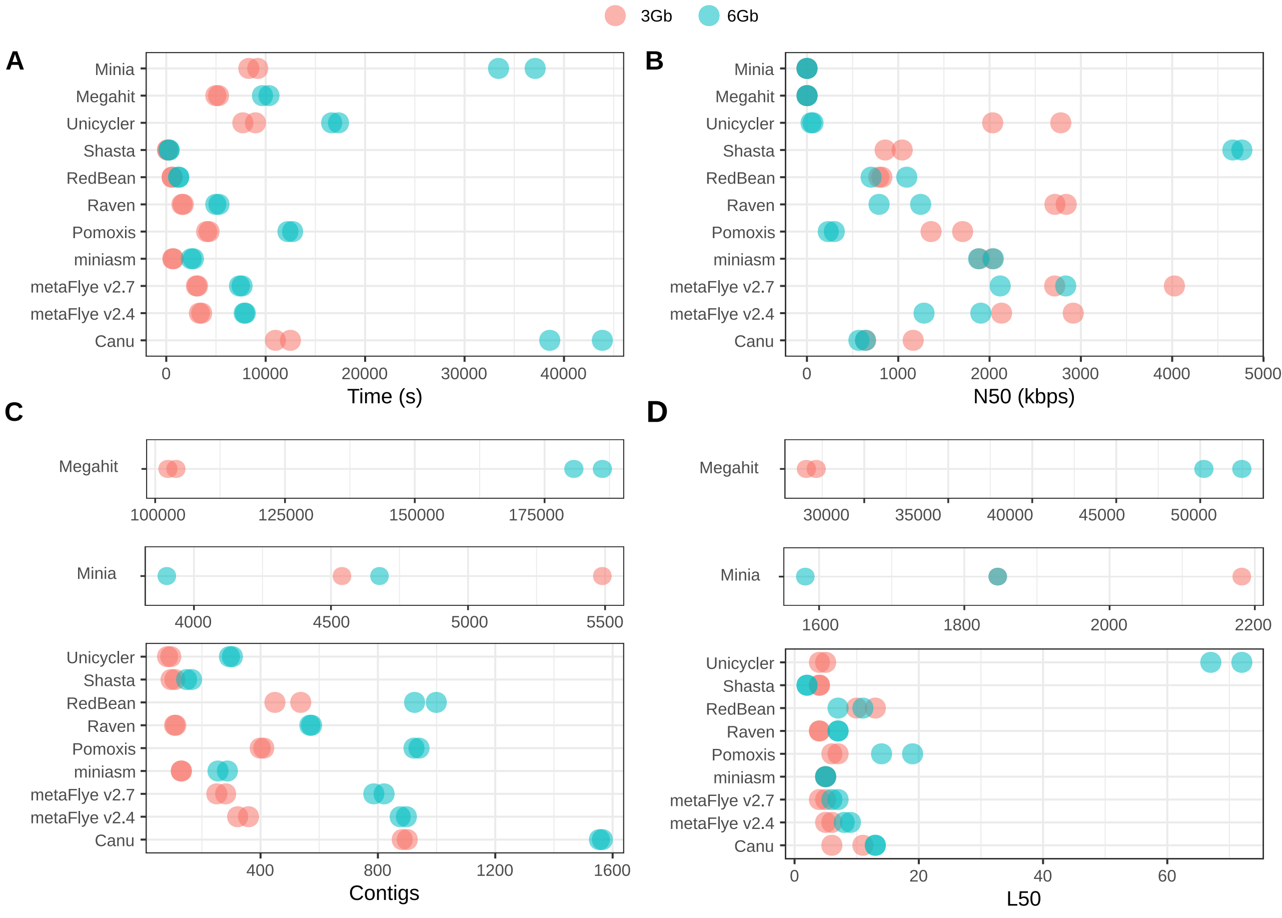


Figure II.3. General assembly performance of each tool for the subsampled Even datasets. (A) Run time; (B) N50; (C) Number of contigs; (D) L50.

the analysis of deeply sequenced microbiomes. Results were similar to those obtained for the 3 and 6 Gbp (**Figure II.4**). Canu recovered the highest proportion of bacterial genomes, closely followed by metaFlye. Raven, once again, displayed problems when reconstructing the whole *Escherichia coli* and *Salmonella enterica* genomes, an issue also detected for RedBean. MetaFlye and Raven achieved a better recovery ratio than Canu for the yeast genomes. Overall, metaFlye genomes were more complete but less contiguous than the Raven draft assemblies, which presented a lower number of contigs for all the species with the exception of *E. coli* and *S. enterica* (**Figure II.4B**). This trend was also observed for the Log datasets (**Supplementary Figure II.S4**). Remarkably, Raven was able to assemble two bacterial genomes in only one contig (*Lactobacillus fermentum* and *P. aeruginosa*), and retrieved four additional genomes in only 2–3 contigs. Finally, it was not possible to run Pomoxis on this dataset because of the unsolvable error previously described.

Assembly accuracy

Sequencing errors are the biggest drawback of TGS platforms. These errors can reach the final assemblies, resulting in lower quality draft genomes. In order to evaluate how the different assemblers handle the specific error profile of ONT platforms, we analyzed the total number of SNPs and indels present in each draft metagenome. As described in General **Material and methods**, two different and complementary strategies were used to quantify these types of errors: (1) minimap2+bcftools, and (2) MuMmer (**Figure II.5**). Both strategies relied on the alignment of the draft assemblies to the reference metagenome, composed by a

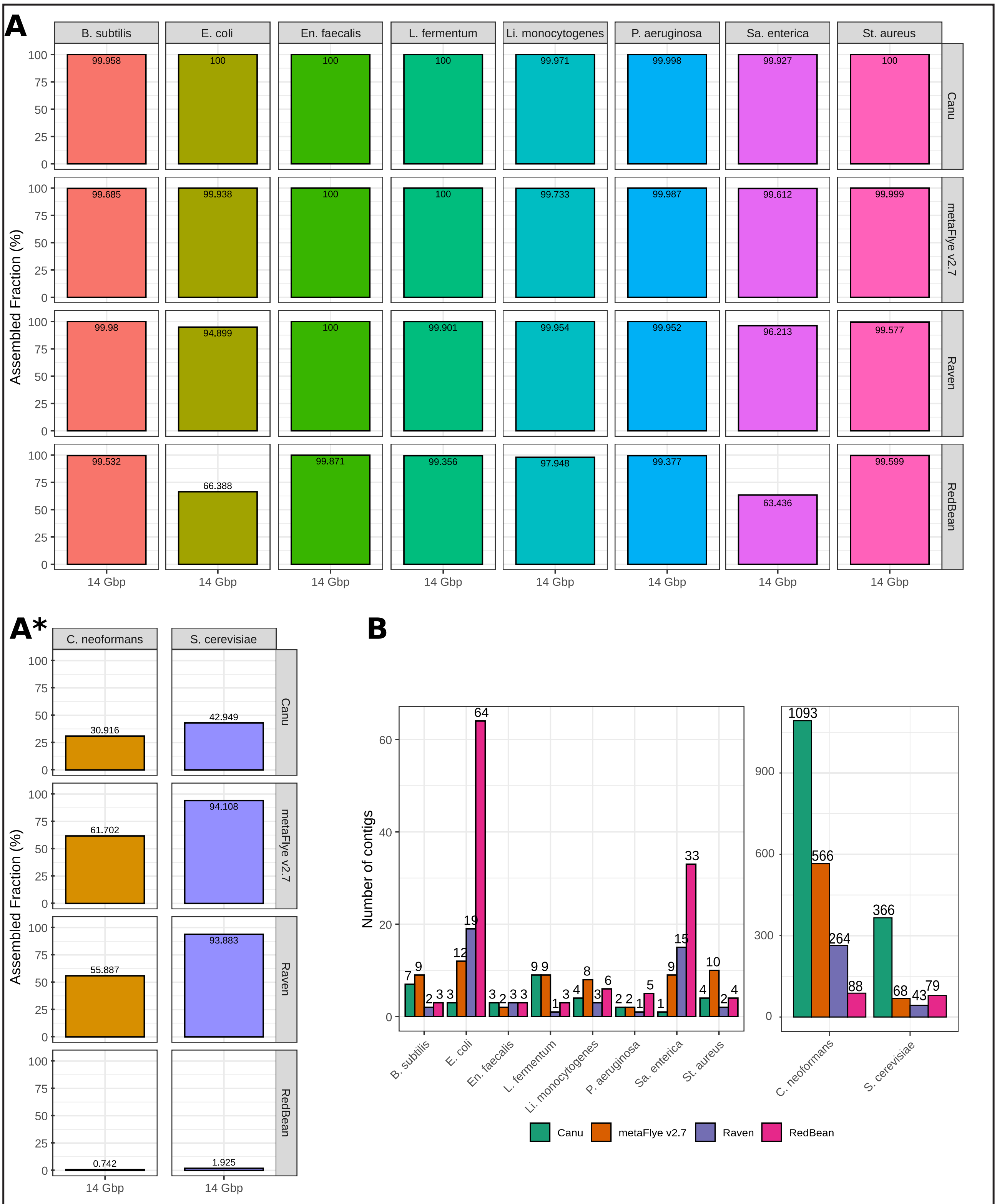


Figure II.4. Assembly evaluation of the best performing tools using the Even GridION dataset (14 Gbp). (A) Fraction of the genome covered by the draft assemblies; (B) Number of contigs for each microorganism.

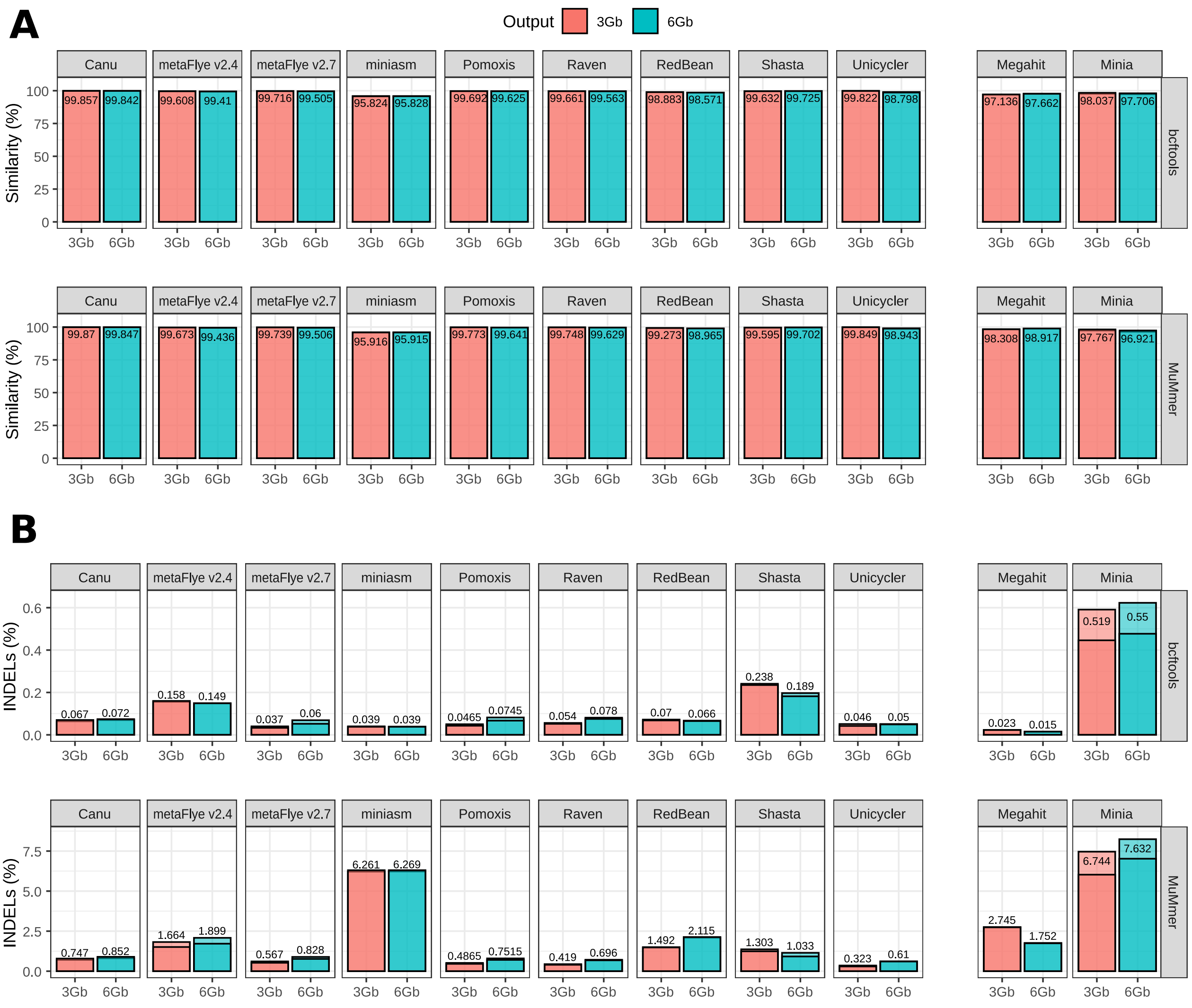


Figure II.5. Assembly accuracy of the draft assemblies (subsampled Even datasets). (A) Percentage of similarity calculated as the total number of matches normalized by the metagenome size; (B) Percentage of indels calculated as the total number of indels normalized by the metagenome size. In both cases, two different strategies were used: (top panel) alignment with minimap and evaluation with bcftools + 'indels_and_snps.py' in-house script; (bottom panel) alignment with MuMMer and evaluation with 'count_SNPS_indels.pl' script from Goldstein et al. (2019) **[183]**.

mix of all the complete genomes of each strain present in the mock community.

Results were not fully consistent between the two methodologies, especially for the indels estimation, but they still showed similar trends. All the long-read assemblers retrieved draft metagenomes with an average similarity higher than ~98.9%, with the exception of miniasm, which resulted in an approximate accuracy of only 96%. This low accuracy could explain the inability of metaQUAST to evaluate miniasm assemblies. It has to be noted that the other pipelines based on miniasm, Pomoxis and Unicycler, incorporated several rounds of polishing via Racon [185], which substantially reduced the number of SNPs and indels in the final draft assembly (see below).

Canu displayed a higher percentage of similarity for both methodologies and datasets, followed by Unicycler for the 3 Gbp dataset, and Shasta for the 6 Gbp one.

Pomoxis, metaFlye, and Raven presented similarities over 99.5%. In the case of the indel profile, Unicycler and metaFlye v2.7 clearly outperformed Canu. Raven and Pomoxis also achieved a better indel ratio than Canu, except for the 6 Gbp dataset and the bcftools metric. Redbean, miniasm, and Shasta results were inconsistent between the two methodologies tested (**Figure II.5**).

Biosynthetic gene cluster prediction

Gene prediction is highly affected by genome contiguity, completeness and accuracy. Biosynthetic gene clusters (BGCs) are especially influenced by these factors, since they are usually found in repetitive regions which are often poorly assembled. AntiSMASH was used to assess the number of clusters found in the draft assemblies retrieved by each tool in comparison to the reference metagenome with the aim of evaluating BGC prediction on Nanopore-based metagenomic assemblies (**Figure II.6**). As expected, none of the tools recovered the entire BGC profile, since metagenomes were not completely reconstructed (**Figure II.1**). Using the entire GridION dataset (14 Gbp) did not improve the number of BGCs recovered (**Supplementary Table II.S3**). Overall, when considering the total number of BGCs predicted and the similarity of the obtained profile compared to the reference profile, Raven displayed the best performance for both 3 Gbp datasets, whereas metaFlye v2.7 displayed the best performance for the 6 Gbp datasets. Pomoxis also achieved good predictions, outperforming Canu. All the predicted profiles presented an enrichment in lasso peptides (ribosomally-synthesized short peptides), which were not present in the reference profile. To further study this phenomenon, lasso peptides predicted by the different tools were searched using BLAST against the BGCs predicted in the reference metagenome. No hits were found, suggesting that these results might be prediction artefacts mainly caused by indels, which are probably introducing frameshift errors, and artificially increasing the number of short peptides being predicted (i.e., lasso peptides). In fact, metaFlye v2.7, which had a significantly lower indel ratio, retrieved fewer lasso peptides than metaFlye 2.4 (**Figure II.5**). We also corrected Pomoxis assemblies with Medaka, leading to a lower indel ratio (see the following section). Lasso

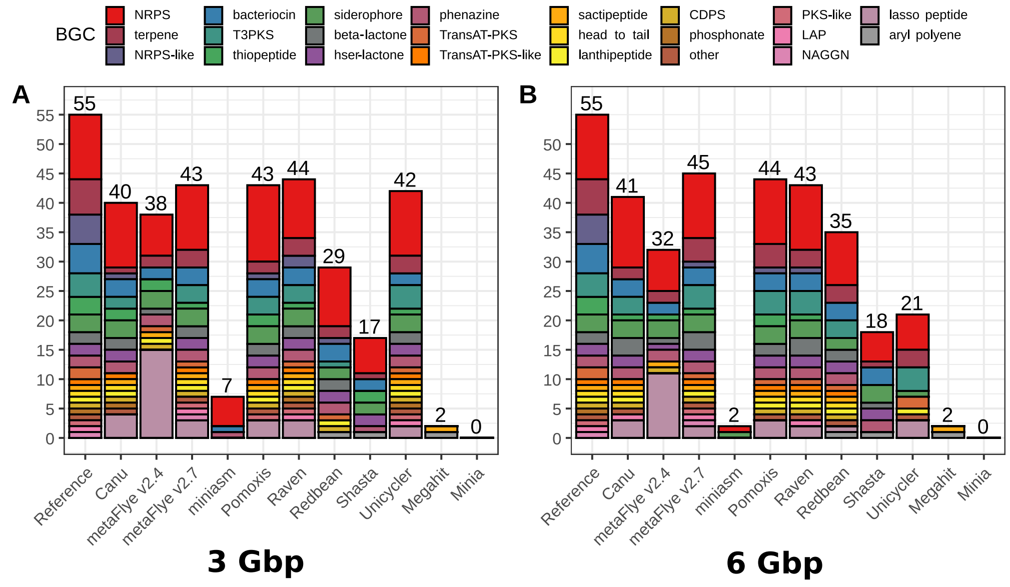


Figure II.6. Number of biosynthetic gene clusters (BGCs) predicted by antiSMASH for each draft assembly in the Even GridION datasets. (A) BGCs predicted using the 3 Gbp dataset; (B) BGCs predicted using the 6 Gbp dataset.

peptides were not detected in Pomoxis+Medaka assemblies, highlighting the importance of indel correction for functional prediction (**Supplementary Figure II.S5**).

Polishing evaluation

Polishing is the process of correcting assemblies in order to generate improved consensus sequences. Input for polishing Nanopore-based assemblies can be raw ONT reads (i.e., Racon or Medaka)[185], raw electrical signal (i.e., Nanopolish) (https://github.com/jts/nanopolish), or even high-quality short reads (i.e., Racon). The state-of-art polishing workflow for Nanopore sequencing consists of correcting the draft assemblies through several rounds of Racon (typically 2–4), followed by a single Medaka step.

Some of the tested tools automatically incorporated Racon (Raven, Pomoxis and Unicylcer) in their pipelines, whereas the others included different algorithms for correcting the reads before (Canu) or after (metaFlye and ReadBean) the assembly process. Thus, we wanted to assess how various steps of polishing could affect the SNP and indel ratio of the different assemblers. Results were highly heterogeneous (**Figure II.7; Supplementary Table II.S4**). Pomoxis and Raven drastically improved their accuracy after several rounds of polishing with the original Nanopore reads (**Supplementary Table II.S4**). In fact, accuracy with no polishing steps was close to 96%, as reported for miniasm (**Figure II.5**). Higher similarity percentages were observed after one round of Racon (1R) for Raven, and four rounds of Racon+one round of Medaka (4R+m) for Pomoxis. Redbean and metaFlye -which were run again without using their built-in polishers- also improved their accuracy after 1R or 4R+m, respectively. Canu presented a lower percentage of SNPs when no polishing steps were added to the pipeline (**Supplementary Table II.S4**). Nevertheless, all the tools drastically improved their indel ratio after 4R+m. The percentage of improvement varied between 41% (Canu) and 91% (Raven and Pomoxis) (**Figure II.7A**). It has to be highlighted that the lowest number of SNPs and indels was achieved by Canu, which is the only tool that carries out error correction before assembling the reads.

The error profiles were evaluated again to further assess whether polishing draft assemblies with high quality short reads led to improved assemblies. Albeit yielding heterogeneous results, all the tools achieved better indel ratios after four rounds of Racon correction with Illumina reads (**Supplementary Table II.S4**). In this case, all the assemblers improved their accuracy (% of similarity) after one (Canu and metaFlye) or four (Pomoxis, Raven and RedBean) Racon rounds. When comparing the highest scores obtained with Illumina-based correction to the highest scores achieved after ONT-based polishing (**Figure II.7**), the percentage of similarity was higher for metaFlye and Canu assemblies corrected with Illumina reads, and lower for Pomoxis, Raven and RedBean, where ONT polishing outperformed Illumina's. A similar trend was observed for the indel ratio. This time, Illumina correction clearly enhanced the indel correction for metaFlye and Canu. In fact, Canu+Illumina correction retrieved the lowest indel ratio. Pomoxis, Raven and RedBean achieved a better indel correction with ONT reads.

Discussion

Assembling shotgun sequencing data is often a key factor for characterizing the functional and taxonomic diversity of microbial communities. In the recent years, Nanopore-based sequencers (Oxford Nanopore Technologies; ONT) are rapidly growing in popularity due to four basic reasons: (1) low cost, (2) long-read generation, (3) portability, and (4) real-time analysis. Several bioinformatic tools have been developed to handle Nanopore sequences during the assembly process. Nevertheless, there is a lack of systematic, up-to-date, independent studies comparing the performance of the currently available tools. This work is aimed at

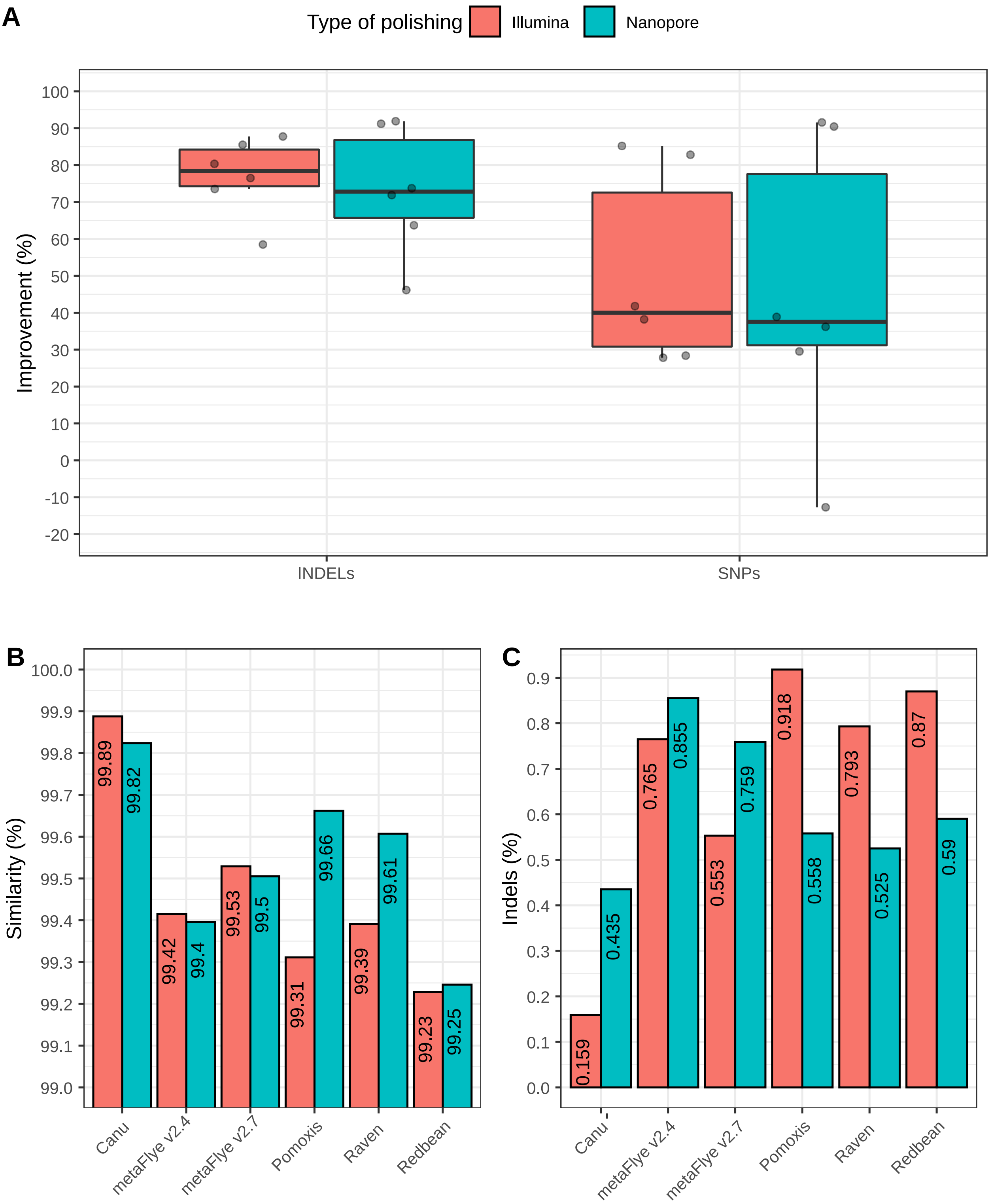


Figure II.7. Polishing evaluation. (A) Percentage of improvement, taking as a reference the number of errors prior polishing; (B) Best percentage of similarity achieved by each tool; (C) Best indel ratio achieved by each tool. Note that different number of polishing rounds were needed for achieving the highest similarity and the lowest indel ratio depending on the tool **(Supplementary Table II.S4)**.

filling this gap using the data previously published by Nicholls et al. [186], which consisted of the ultra-deep sequencing of two different mock communities (**Table II.1**) on GridION and PromethION platforms (ONT). These platforms follow the same sequencing principles as MinION, but they have a significantly higher output. For that reason, the datasets were subsampled in order to adapt their output to the current yield offered by MinION (3–6 Gbp) [140, 183, 301–305], then extending the study to higher yields comparable with other recent works [306].

Despite the relatively low complexity of the mock communities analyzed in this evaluation study, our results show that there is a huge variation in assembly results depending on the software chosen to perform the analysis. Minia and Megahit poorly reconstructed the microbial genomes (**Figures II.1 & II.2**) and produced highly fragmented draft assemblies (**Figure II.3**). This output was expected, since these assemblers are highly optimized to work on short reads, which are very different to the data generated by ONT sequencers.

Long-read assemblers (Canu, metaFlye, Unicycler, miniasm, Raven, Shasta and Readbean) also presented significant differences in the general assembly performance. This was expected too, since some of the tools were not specifically designed for assembling metagenomes (**Table GMM.1**). Overall, only metaFlye, Raven, and Canu worked well on all the tested datasets. They were able to recover the eight bacterial genomes from the Even dataset with a high degree of completeness, and also reconstructed a significant fraction of the yeast genomes. Draft assemblies were highly contiguous when using these three tools, as they were able to reconstruct bacterial genomes in only 1–19 contigs (**Figure II.4B**). Unicycler and, especially, Pomoxis, also performed well for some datasets and metrics, but failed to run in some cases (**Table GMM.1**). Both tools are pipelines based on miniasm that include further polishing steps by Racon. Miniasm alone was also unable to assemble the Log 6 Gbp dataset, indicating a lack of consistency of the algorithm for different microbial community structures. Finally, Shasta and RedBean (wtdgb2) retrieved incomplete assemblies and they did not provide any additional advantage other than computational efficiency.

Our results are in accordance with previous studies. MetaFlye has proved to outperform other tools in terms of metagenome recovery when using different mock communities [181, 300], although it must be noted that these previous studies did not include all the tools selected in the present benchmark. Canu also performed well in other studies [181], and has been proposed for increasing the contiguity of metagenome assembled genomes recovered from real samples [306]. Nevertheless, its high computational cost limits the use of Canu for bigger datasets (**Figure II.3**, **Supplementary Figure II.S4**) [61, 181]. RedBean displayed a reduced performance in comparison to other long-read assemblers [181, 300, 306, 307]. To the best of our knowledge, no other metagenome assembly benchmark has included Pomoxis, Shasta, or Raven. Wick and Holt [307] evaluated different tools for single isolate assembly (not metagenomic assembly), and reported that Shasta was more likely to produce incomplete draft assemblies, while Raven was reliable for chromosome assembly, as also seen in our work. Although Pomoxis was not included in this last benchmark, another miniasm + Racon strategy was used. This strategy, that was reported to perform robustly among different genomic datasets, is equivalent to one of the pipelines used in the present study (here referred to as Unicycler). This observed robustness is in contrast to our results, supporting the idea that the intrinsic differential coverage of metagenomic datasets could be the cause of the inconsistency detected for miniasm in this benchmark.

Although sequencing errors are one of the main drawbacks of third-generation sequencing platforms, the best performing tools (metaFlye v2.7, Canu, Raven and Pomoxis) achieved >99.5% of accuracy in the final assemblies. Indels may be especially problematic, since they can introduce frameshift errors, which hinder functional prediction. After analyzing the different BGCs profiles, metaFlye and Raven demonstrated to reach better results and they outperformed Canu. This is in accordance with the indel ratio calculated for each tool (**Figure II.5B**). It has to be highlighted that these results were obtained by using ONT configurations explicitly recommended in the manual of each tool. The use of other tools (i.e., polishers) led to assemblies with enhanced quality. The lowest number of SNPs and indels were achieved after different rounds of polishing for some assemblers (**Supplementary Table II.S4**). As a consequence, the number of polishing rounds is variable and must be carefully chosen by the user. Correction with Illumina reads is a useful strategy for reducing the number of indels and SNPs produced by metaFlye and Canu, as also reported in Moss et al. [306]. The combination of Canu with polishing tools resulted in the best accuracy, especially when using Illumina reads for the correction.

Finally, time is a crucial parameter when choosing a bioinformatic tool, even more if considering MinION's ability to generate real-time data. In this sense, metaFlye v2.7 was up to 6.7 times faster than Canu, which was the slowest tool tested on this benchmark. Raven was even faster than metaFlye, and tended to generate fewer contigs (**Figure II.3, Supplementary Figure II.S4; Supplementary Table II.S1**).

Taken together, our results show that Nanopore data (accommodated to current MinION's output) can lead to highly contiguous and accurate assemblies when using the adequate tools, with no need of complementary sequencing with Illumina. From all the tested software, metaFlye v2.7 resulted the best in terms of metagenome recovery fraction and total metagenome assembled size. Raven achieved slightly lower genome fractions than metaFlye, but was faster and generally retrieved a lower number of contigs. Canu was the most accurate tool and introduced fewer indels when combined with polishing tools, but its assembly process also demonstrated to be time consuming. Pomoxis and other miniasm-based pipelines are also promising, but their inconsistency problems should be addressed. This work may help software developers to design new bioinformatic tools optimized for Nanopore-based shotgun metagenomic sequencing, although further research is still needed in order to benchmark the different assemblers on more complex microbial communities.

The content of this chapter has been published in Scientific Reports as: **Latorre-Pérez A**, Villalba-Bermell P, Pascual J, Vilanova C. Assembly methods for Nanopore-based metagenomic sequencing: a comparative study. *Sci Rep* 2020;**10**:1–14. doi:10.1038/s41598-020-70491-3 **(Publication VI in Appendix C)**

General Discussion

1. Expanding the scope of Nanopore sequencing

Nanopore-based platforms have transformed the *status quo* of sequencing technologies by offering a set of new features (e.g., portability or real-time analysis)[15] that are beyond the reach of other systems. Despite the limitations associated with Oxford Nanopore Technologies (ONT) devices[15], the potential of Nanopore sequencing has been demonstrated in multiple branches of life sciences, including microbiology and microbiome research[16]. Nevertheless, and as expected for an emerging and rapidly evolving field, the applications of this technology require optimization and standardization. In the present work, new methodologies (**Figure GMM.1**) have been designed and implemented in order to test the suitability of Nanopore sequencing for addressing two different problems of industrial and biotechnological relevance: monitoring microbial industry-relevant processes and improving bioprospecting strategies.

1.1. A monitoring tool for microbial processes

In this thesis, Nanopore sequencing has been used to study the microbial communities associated with anaerobic digestion (AD), as reported in **Chapter IA and Appendix C**. Our work has proven that ONT devices could be applied to monitor AD in an industrial environment. This application has been primarily hampered by:

- The **economic investment** needed to acquire a conventional (NGS) sequencer, and the **technical complexity** of sequencing and bioinformatics. This process is typically simplified by submitting the collected samples to specialized sequencing facilities. Unfortunately, this makes the whole procedure significantly slower (results are typically obtained after some weeks). This limitation can be overcome with Nanopore sequencing[15], as demonstrated in the application described in **Chapter IB**.
- The **lack of characterization** of industrial microbiomes, including AD-related microbiomes, which have been traditionally considered a black box [203].

Biogas production is a very complex process that requires a deep knowledge of anaerobic digestion. The intention of this thesis was not to improve the technical aspects of AD, but to evaluate the suitability of using Nanopore sequencing to solve problems of industrial relevance. Nonetheless, the studies included in **Chapter IA** have also contributed to understand how different operational parameters can affect the microbial communities of anaerobic digesters. Specifically, we analyzed the effect of co-digesting sewage sludge with the liquid and solid fraction of grass biomass, revealing that liquid co-substrates resulted in a more effective methanogenic microbiome (dominated by *Methanosarcina*), and were associated with a higher biogas production (**Figures IA.2 & IA.4**). Bacteria involved in acidification during AD were also investigated, and the influence of different ammonia removal methods on those taxa was assessed. Despite the treatments, bacterial communities proved to be very similar in all the experiments (**Figure IA.8**). This was in line with other works, which have demonstrated the high robustness and resilience of AD microbiomes [219][17]. However, some microorganisms experimented interesting changes: the abundance of *Acholeplasma* and *Erysipelotrichaceae* UCG-004 tended to increase during acidification. These genera are predominant in microbiomes showing a robust performance under high ammonia concentrations, suggesting that the bacterial communities studied in our work progressively adapted to high ammonia levels (**Figure IA.9**).

The microorganisms highlighted above, together with other taxa proposed by previous studies (e.g., *Methanoculleus,* which is abundant in viscous sludge, or *Methanosaeta,* which is characteristic of sewage sludge[18])

[15] See subsection 4.3 of General Introduction for a full description of the advantages and drawbacks of Nanopore sequencing

[16] See section 5 of General Introduction for a report of the applications of real-time, *in situ* microbiome sequencing

[17] Indeed, to characterize this phenomenon was the main motivation of Publication IV in Appendix C.

[18] As can be seen in Study I – Chapter IA

[219], can be used as microbial markers for monitoring biogas production. Nonetheless, the knowledge about AD is far from complete. This is the main motivation of the MICRO4BIOGAS project (https://micro4biogas.eu/). In the framework of this European project, we will study hundreds of publicly available and *de novo* generated datasets to analyze the influence of several operational parameters (i.e., reactor configuration, pH, temperature, substrate, etc.) on the process. This will lead to the identification of new microbial markers that could be used to detect perturbations in the AD process, which would eventually be corrected by bioaugmentation strategies (**Figure GD.1A**). Furthermore, due to the power of sequencing analyses, we hypothesize that the future of industrial monitoring could go beyond particular microbial markers to focus on the full microbiome. In combination with the information about other operational parameters, this data could be used to train machine learning models, which could ultimately predict the performance of the reactor and suggest corrective actions.

1.2. Next-generation bioprospecting

The goal of bioprospecting is to explore different environments to find biological resources with practical applications. The main limitation of this process is that exploration (i.e., collecting samples) and searching (i.e., sample characterization) occur at different stages: the first one is carried out in the field, while the second one is performed in the laboratory and it usually takes several months to be concluded. This limitation is exacerbated in the case of microbial bioprospecting, since microorganisms are able to prosper in almost any known biotope, both natural (e.g., thermal springs or polluted rivers) [308, 309] or artificial (e.g., prosthetic devices or wasted chewing gums) [310, 311]. Microniches are especially relevant to microbial biotechnology and, for instance, solar panels display a highly diverse and dynamic biocenosis that produces commercially valuable molecules such as carotenoids [252, 312]. During a microbial bioprospecting expedition, scientists can access tens, or even hundreds of different microniches. Although the decision of focusing on some sample types over others can be guided by previous literature or the biological resources that are being searched (i.e., halophilic microorganisms are likely present in saline environments), at some point this choice becomes arbitrary. The problem is that the actual potential of the selected samples is not evaluated until the researcher returns to the laboratory, where a time-consuming screening takes place. Even in the best-case scenario, that is, the selected samples show biotechnologically-relevant activities, these microniches cannot be further exploited without repeating the whole bioprospecting expedition.

In the present thesis, we used two key advantages of Nanopore sequencing (i.e., portability and real-time analysis) to prove that sample collection and characterization can be fully integrated into the bioprospecting expedition (**Figure IB.1**). Despite the fact that ONT devices can operate out of the laboratory, as has been extensively demonstrated[19], in-field sequencing does not necessarily mean that sequencing must be actually performed in the field. In our work, a mobile laboratory was set up in an apartment near the sampling location. This allowed us to apply the same experimental protocols that we were using in the laboratory with small modifications, which was reflected in the performance of the sequencing run: the sequencing yield was substantially higher than the output reported in other on-site studies [94], and it was comparable to the yield of runs performed in fully-equipped laboratories [168, 255]. Moreover, the output that was obtained in this experiment was higher than the yield achieved in the studies included in **Chapter IA**, even though the run was stopped after only 6.5 h. This could be attributed to the update of the Ligation Sequencing Kit from version SQK-LSK108 (used in **Chapter IA**) to version SQK-LSK109 (used in **Chapter IB**), emphasizing the rapid evolution of Nanopore technology. Indeed, versions SQK-LSK110 and SQK-LSK112 of the kit are currently available[20] .

[19] See section 5.3. of General Introduction

[20] According to https://store.nanoporetech.com/eu/sample-prep.html (accessed 11 December 2021)

In the context of this dynamic change of Nanopore sequencing, the bioinformatic pipelines need to be constantly revisited and adapted. During the course of this thesis, we moved from a strategy based on BLAST (as implemented in QIIME[21]) to a method using minimap2 as the core of the analysis (i.e., Spaghetti, an analysis pipeline developed in the framework of this thesis)[22]. This shift was motivated by two main reasons:

- The BLAST + QIIME strategy showed the best results in internal tests that were performed at the beginning of this thesis (**Supplementary Figure GD.S1**). Nevertheless, as new tools were released, minimap2 proved to be a better choice based on other works and benchmarks [73, 167, 168].
- Minimap2 was substantially faster and less computationally expensive, and hence it was more suitable for in-field applications. In fact, the bioinformatic analysis during the expedition to the Tabernas Desert finished in ~3 h (**Figure IB.1**).

According to our proof of concept, Nanopore sequencing can be used for characterizing microbial communities in the field, generating the first results in only 24h after sample collection. This data is not only relevant from an ecological point of view, but it also has applied implications since it can be used to guide bioprospecting activities: microniches maximizing the diversity and abundance of radiation- and desiccation-resistant bacteria can be quickly detected and selected for further sampling (**Figure IB.3**). Despite the limitations associated with microbial culturing techniques[23], Nanopore and culturing data correlated well (**Figure IB.4**), since samples holding a greater potential at the microbiome level also yielded a more interesting set of microbial isolates, whereas samples showing less biodiversity resulted in a reduced (and redundant) set of culturable bacteria (**Figure IB.5**). This demonstrates that Nanopore sequencing can confidently be used as a screening tool for assessing the potential of samples *in situ*.

In **Chapter IB**, the analysis focused on bacteria with the potential to resist desiccation and radiation for several reasons: (i) different taxa from unrelated bacterial phyla have been described to resist such conditions [252]; (ii) these bacteria hold interest for biotechnological applications [252, 266, 312]; (iii) the Tabernas Desert is reported to harbor these taxa [37]. Similarly, sampling was focused on biocrusts and soils, since these ecosystems have proven to be a rich source of radiation- and desiccation-resistant bacteria [267–274]. In any case, this strategy could be applicable to many other bioprospecting goals, such as the search of antibiotic producing bacteria [313] or the identification of thermophilic bacteria and archaea[24] [314]. In a broader context, taxonomic information can also be exploited for detecting those microniches that harbor a higher microbial alpha diversity or that are extremely similar to other sample types based on beta diversity analyses (**Figure IB.2**) In summary, the present work lays the foundations for an improved bioprospecting strategy, in which *in situ* Nanopore sequencing can be used to inform sampling (**Figure GD.1B**). This would increase the chances of recovering the target microorganisms, although advances in culturomics are still necessary to maximize the results of this application [315].

2. The impact of Nanopore sequencing on (meta)-genomics

Genome assembly is the process of reconstructing the genetic information of an organism from shorter fragments of DNA. When assembly is applied to microbiome data, then it is called metagenome assembly. This process is hindered by several factors: (i) genomes include repetitive elements that are longer than the typical read length of NGS platforms (**Table GI.1**); (ii) metagenomes contain intergenomic repeats, that

[21] See “Pipeline 1” in General Materials and Methods

[22] See “Pipeline 2” in General Materials and Methods

[23] See Discussion in Chapter IB

[24] Using the primers described in Study I – Chapter IA.

is, sequences that are shared by more than one taxon present in the microbial community; (iii) microbial communities often contain related species or subspecies in different -and unknown- abundances. Long-read sequencing technologies, including ONT, have the potential to overcome these limitations. However, assembling long, error-prone reads represents a new algorithmic challenge [316]. This has stimulated the creation of various assembly tools that were specifically designed to work with this type of data (**Table GMM.1**), raising an important question: which assembly method is best for Nanopore sequencing?

Answering this question was the main goal of **Chapter II**. In consequence, we carried out a benchmark of the available tools, and by using data generated from two different mock communities we demonstrated that metaFlye [181] met the perfect compromise between general performance and computational efficiency. Canu [109] and Raven [173] also performed well in all the datasets tested, and they achieved the best results for some of the metrics analyzed (**Figures II.4 & II.7**). The most promising tools were further tested using publicly available Nanopore sequencing data from four additional mock communities with different levels of complexity (see **Appendix A**). These analyses helped to consolidate previous results: metaFlye showed the best overall performance for Nanopore-based metagenomic assembly and it worked especially well for recovering plasmids. Raven displayed a remarkable performance, but it did not excel in any specific parameter. Finally, Canu proved not to be sufficiently scalable for some datasets, while other tools did not introduce any substantial improvement to metaFlye or Raven.

Similar results were obtained by Wick & Holt [307] when assembling individual bacterial genomes. In this benchmark, Flye[25] demonstrated to be robust, to make the smallest sequence errors and to recover plasmids at a broad range of size and sequencing depth. This tool also performed well for reconstructing metagenomes from complex samples. For instance, metaFlye was able to recover several circular genomes from human [306] or canine faeces [317]. Altogether, Flye/metaFlye would be the preferred choice for general assembly applications. Nevertheless, as each assembler has its advantages and drawbacks [307], combining various tools for analyzing the same dataset can lead to improved results [318]. Analogously, using different polishing tools can increase the accuracy of the assemblies [319], as anticipated in our work (**Supplementary Table II.4; Appendix A**).

In general, our results confirmed that metagenomes assembled from Nanopore sequences were highly contiguous (**Figures II.3 and II.4**), which was in line with previous genomic and metagenomic studies [183, 306, 307, 317]. On the other hand, Nanopore-based assemblies tend to accumulate errors, especially indels (**Figure II.5**) [183, 306]. Hybrid assembly strategies combine the best of both worlds: the accuracy of NGS and the contiguity of TGS. Therefore, this approach may be considered the current gold standard for genome and metagenome assembly [183, 318, 320, 321]. However, using two different sequencing technologies is not ideal, since it increases the cost of the analysis, and it requires twice the effort. Assemblies relying only on Nanopore data can reach an accuracy of >99.99% with an appropriate sequencing coverage [318, 319] thanks to improvements in basecalling models and polishing tools. More recently, the introduction of the Q20+ chemistry and the 'Duplex' mode has increased the accuracy of raw reads up to 99.3% and 99.8%, respectively [98, 99]. The Q20+ chemistry includes a new motor protein which improves the translocation of DNA molecules across the nanopore, thus increasing the basecalling accuracy. The 'Duplex' mode refers to the ability of nanopores to read one DNA strand and then its complementary chain. This phenomenon is not size dependent and it occurs naturally ~1-4% of the times, but ONT claims to have increased this rate up to ~40% with optimizations in library preparation kits [98]. The impact of these advances in genomics and metagenomics has not been

[25] Flye and metaFlye are the same tool, but metaFlye is the preferred name when used for metagenomics.

evaluated yet, although it can be predicted that they will improve the consensus accuracy of the assemblies. The effect will be more noticeable on low-coverage genomes, in which raw accuracy is key, since errors cannot be compensated with sequencing depth. Considering the advantages of Nanopore sequencing[26] , this technology has the potential to become the standard for (meta)-genome assembly, but systematic evaluations are still necessary to demonstrate the benefits of ONT platforms over Illumina, PacBio (especially in Consensus Circular Sequencing -CCS- or HiFi mode) and/or hybrid approaches.

Apart from improving assembly, Nanopore metagenomic sequencing could be integrated in the on-site applications described in previous sections. This would overcome some of the limitations associated with metataxonomics (**Table GI.2**), and it would allow to increase the power of the analyses in many aspects. For instance, monitoring could go beyond prokaryotic organisms, since metagenomics allows the detection of eukaryotes, prokaryotes and viruses in the same experiment. With this approach, species- and strain-level resolution would be generally possible, thus increasing the specificity and sensitivity of the applications. Finally, both industrial monitoring and *in situ* bioprospecting would benefit from functional metagenomics, since it could be used to track specific genes or metabolic pathways (e.g., biosynthetic gene clusters) and to identify new genetic resources (e.g., highly divergent enzymes). As metagenomic sequencing also has disadvantages (**Table GI.2**), future works should help to decide which strategy (metagenomics or metataxonomics) fits better with the objective, logistics and budget of each analysis.

3. Epilogue: the future of Nanopore-based applications

In the last few years, ONT devices have dramatically evolved from prototypes to one of the most promising alternatives to traditional NGS platforms. Although this technology is still under development and requires further optimizations, it has already played a crucial role in the genomic surveillance of SARS-CoV-2 [123, 322], especially in low-income countries with limited access to large sequencing facilities [323, 324] . Indeed, the low start-up cost, the portability and the potential of real-time analyses make Nanopore sequencing a great solution for outbreak control and many other clinical applications[27]. Beyond problems related to human health, it has been demonstrated throughout this work that Nanopore sequencing could also contribute to monitor and improve processes of industrial and/or biotechnological relevance, such as anaerobic digestion or bioprospecting. This is likely just the beginning, as novel applications of ONT platforms will continuously emerge in the future. Based on the experience acquired during the present thesis, the most ambitious applications may gravitate towards two concepts, which I hypothesize here:

- **End-to-end vs. swiss-army-knife solutions.** Sooner or later, Nanopore sequencing will cross the borders of scientific research and reach other areas, such as food safety, forensics or agriculture. This is in line with the main slogan of ONT, which is “to enable the analysis of anything, by anyone, anywhere”. To fulfil this goal, sample preparation and data analysis protocols should be substantially simplified, since the staff in charge of routine analyses in the industry will probably lack the technical knowledge of a molecular biologist or bioinformatician. ONT is making progress in that direction, and the VolTRAX, a programmable lab-on-a-chip device for automatic library preparation, is just an example[28]. At the bioinformatic level, user-friendly software is also being developed (e.g., EPI2ME by ONT), but it is unlikely that a generalist computer program could cover all the possible uses of Nanopore sequencing. Therefore, future designs should focus on achieving an end-to-end platform combining software and hardware improvements for

[26] See subsection 4.3 of General Introduction

[27] See subsection 5.1. of General Introduction.

[28] See https://nanoporetech.com/products/voltrax (accessed 21 December 2021)

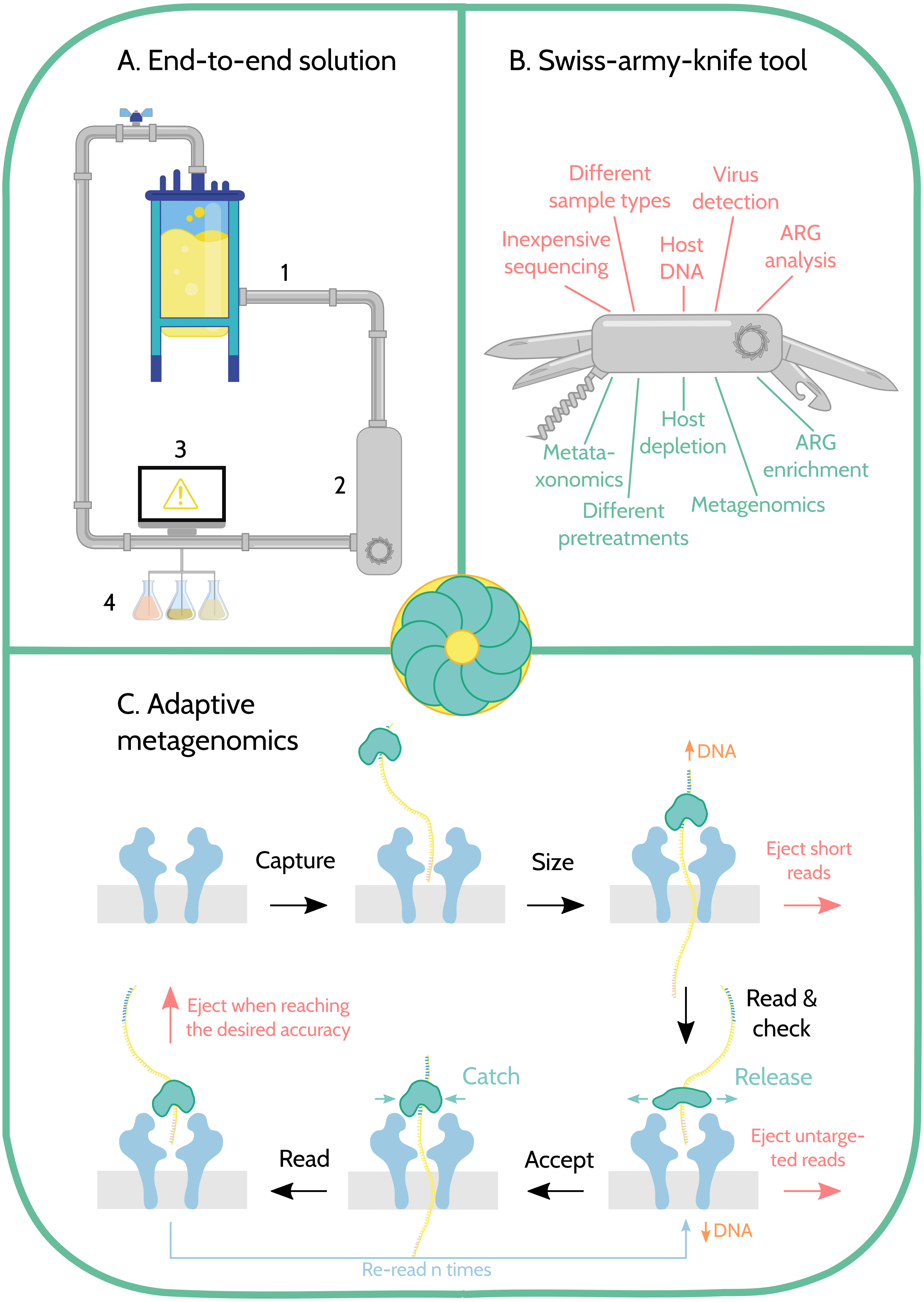

A. End-to-end solution
1
2
3
4
B. Swiss-army-knife tool
Different sample types
Virus detection
Inexpensive sequencing
Host DNA
ARG analysis
Host depletion
Metata-xonomics
ARG enrichment
Different pretreatments
Metagenomics
C. Adaptive metagenomics
Capture
Size
DNA
Eject short reads
Read & check
Eject when reaching the desired accuracy
Catch
Release
Read
Accept
Eject untarge-ted reads
DNA
Re-read n times

Figure GD.1. **(A)** Nanopore sequencing as an end-to-end solution for industrial monitoring. Samples collected from the anaerobic digester (1) are processed with minimal human intervention (e.g., VolTRAX for DNA extraction and library preparation). Then, *in situ* sequencing is used to characterize the microbiome (2). Sequencing results are combined with data from other chemical parameters (pH, temperature, methane production, etc.), and the efficiency of the process is evaluated by a software (3). Corrective actions (i.e., bioaugmentation with optimized microbial strains) are automatically applied (4). Monitoring is periodically repeated. **(B)** Nanopore sequencing as a multipurpose tool for bioprospecting. A versatile laboratory toolkit, including a portable sequencer and adaptable bioinformatic tools, can be used by experts for solving several technical problems during a sampling expedition. **(C)** Schematic representation of adaptive metagenomics. A DNA molecule (yellow) coupled to a motor protein (green) is captured by an empty nanopore (blue). The molecule passes through the pore until the motor protein stops the translocation. This first pass across the membrane allows to estimate the length of the molecule. Therefore, short reads can be ejected (adaptive length). Accepted fragments are read in "Outy" mode (see **Supplementary Figure GD.S2**). Sequences are compared to a database to decide if they are relevant or not (adaptive sampling). Relevant sequences (targets) are read several times to reduce sequencing errors (adaptive accuracy). After reaching the desired accuracy threshold (e.g., 99.99%), the DNA molecule is ejected and the nanopore is free to read another strand. Panel C is inspired by Clive Brown's presentation at the Nanopore Community Meeting 2021 (NCM21) **[331]**.

each specific application. Continuing with the example of industrial monitoring introduced in previous sections, the ideal application would start with a plant operator collecting a sample from the digester, which would be analyzed with minimal human intervention. After several hours of sequencing, the system should indicate the status of the reactor and suggest interventions to improve the efficiency of the process (i.e., bioaugmentation with optimized microbial strains) (**Figure GD.1A**).

Nevertheless, other applications of Nanopore sequencing would need a totally different approach. In this case, versatility of the platform will be preferred over simplicity. For instance, in a bioprospecting expedition, various sample types are usually collected (i.e. water, soil, plant matter, etc.). These samples need different and complex pretreatments (i.e., filtration for water samples, washing steps for clay samples, etc.). Moreover, the bioprospecting goals may change depending on the sample type (i.e., the microorganisms of interest for water samples may not be of interest for soil samples). Hence, a universal end-to-end solution is not possible in this context. Instead, a flexible bioprospecting and data analysis toolkit (e.g., Bento Lab[29] + sampling tools + MinION + laptop) should be prepared and used by technical personnel, which could adapt the protocol to every possible scenario in order to maximize the effectiveness of the *in situ* application (**Figure GD.1B**).

- **Adaptive metagenomics.** In Nanopore sequencing, raw electrical signals become completely available for analysis even when the DNA strand is still being read. A rapid exploration of this data allows, theoretically, to decide whether a particular sequence is interesting depending on the specific goal of the sequencing. Indeed, non-relevant reads can be ejected from the nanopores that are reading

[29] Bento Lab is a portable DNA analysis laboratory.

them by reversing the polarity of the voltage across the specified pore [325]. Considering that DNA is read at 450 bp/s (250 bp/s with the new Q20 chemistry [326]), ~2.6h are required to sequence a 4.2 Mbp molecule (the longest single molecule that has been ever read). Therefore, to accept or reject this molecule after only a few seconds could help to use all the sequencing capacity of the device to read only the desired targets, thus enriching these fragments without including any extra sample preparation step (i.e., PCR, hybridization…) [327]. Adaptive metagenomics is a novel idea that combines three different sequencing modes that are compatible with Nanopore sequencing (**Figure GD.1C**):

1. **Adaptive length**[30] . When a DNA molecule passes through a nanopore, its length can be effectively measured. After being sized, the DNA fragment can be discarded if it is too short, thus enriching the sequencing results towards long reads. Polynucleotide sizing was the first application of Nanopore sequencing to be published[31] [81]. This option is currently unavailable in ONT devices, but it is under active development [98].
2. **Adaptive sampling**. This concept refers to the ability of Nanopore sequencing to reject off-target sequences before they are completely read. Sequencing data generated while reading the DNA molecule can be mapped against a database of desired or undesired targets to decide if the molecule is interesting or not. Mapping can be performed with both raw electrical signals [327] or basecalled sequences [328]. In a metagenomic context, adaptive sampling could be used to deplete undesirable DNA (i.e. eukaryotic DNA from the host) [329], and to enrich specific genes (e.g., antibiotic resistance genes or ARGs) [330] or genomes (i.e., specific microbial species)[32].
3. **Adaptive accuracy**. Motor proteins can be modified to allow a single DNA molecule to be read several times by the same nanopore. Similarly to PacBio CCS mode (**Figure GI.3**), errors can be corrected with sequencing coverage. When a certain threshold of accuracy has been reached, the molecule can be ejected from the nanopore and another DNA fragment would be read.

In adaptive metagenomics, Nanopore sequencing would start as usual, although adaptive length could be activated if ultra-long reads are desired. Once the most abundant members of the microbial community have been characterized (i.e., their genomes have been assembled), their genetic information could be stored in a database that would be used by the adaptive sampling software to deplete DNA fragments coming from these microbial species. In consequence, the sequences corresponding to rare microorganisms could be dynamically enriched, as has already been demonstrated [328]. Finally, adaptive accuracy would allow the generation of high-accuracy long reads, which would reduce the overall sequencing coverage needed to obtain an error-free genome for each microbial strain. In summary, adaptive metagenomics holds potential to explore the so-called microbial “dark matter” without the need of time-consuming sample preparation (i.e., enrichment methods) or expensive ultra-deep sequencing [331].

Besides these directions, the full potential of Nanopore sequencing seems far from explored, and its influence on the sequencing market is thus expected to steadily increase in the following years, especially if ONT devices keep evolving at the same rate. At this point, the future of sequencing technologies is uncertain, since it is not possible to forecast which sequencing platform, if any, will prevail. Nevertheless, to reach the same level of robustness, standardization and adoption as NGS,

[30] Adaptative length and adaptative accuracy modes require a change in the way that DNA molecules are currently read. See Supplementary Figure GD.S2.

[31] See section 4 of the General Introduction

[32] See Publication VII in Appendix C for a short review of applications of adaptative sampling (previously known as ‘Read until’)

further efforts addressing all the aspects of usage lifecycle for TGS technologies are still required. In particular -and in line with the spirit of this thesis-, improving usability in real-life scenarios, if successful, will allow the ONT statement on its technology designed “to enable the analysis of anything, by anyone, anywhere”, to evolve from a mere commercial slogan to an overwhelming reality.

Conclusions

In the present thesis, different metataxonomic and metagenomic strategies based on Nanopore sequencing have been designed, implemented, and tested. These approaches have been effectively used for characterizing ecosystems of industrial and/or biotechnological relevance. The general conclusions derived from this work are listed below:

- Nanopore sequencing can be applied as a monitoring tool to study bacterial and archaeal communities related to anaerobic digestion (AD). Our analyses allowed to:
  - Detect microbial markers associated with an improved production of methane from sewage sludge (i.e., *Methanosarcina*)
  - Measure the impact of different parameters (i.e., substrate for co-digestion or pretreatment for ammonia removal) on the AD microbiome.
  - Propose a nanopore-based strategy for monitoring industrial bioprocesses, after identifying the current strengths and limitations of the technique.
- *In situ* sequencing can be a powerful tool for deciphering the microbial composition of different samples during a bioprospecting expedition, as demonstrated in a proof-of-concept study in the Tabernas Desert (Almería, Spain). Nanopore and culturing data correlated well, and samples holding a greater potential at the microbiome level also yielded a more interesting set of microbial isolates. Therefore, this approach can be used to inform decision-making during sampling, thus increasing the chances of achieving the bioprospecting goals (i.e., identification of biotechnologically-relevant microbial taxa and/or biomolecules).
- Metataxonomic protocols have been implemented and updated in parallel to the advances of the technology. This includes Spaghetti, a custom pipeline for automatic bioinformatic analysis of Nanopore sequencing data that proved to be adequate for *in situ* applications.
- Nanopore metagenomic sequencing allowed to recover extremely contiguous genomes directly from microbial communities. The results from the benchmark showed that metaFlye was the best-performing assembler, although other tools such as Raven were also promising. Our evaluation of metagenomic assembly was the first to include all the tools designed to handle Nanopore data and it paved the way towards the standardization of bioinformatic pipelines for Nanopore-based metagenomic sequencing.

Resumen en Castellano

Introducción

La tecnología ha jugado un papel clave a lo largo de la evolución del ser humano. Aunque el término "biotecnología" no fue introducido hasta 1919 por Karl Ereky, los humanos llevaban milenios buscando, seleccionando y utilizando otras especies en su propio beneficio. En este sentido, la fermentación de alimentos para mejorar su conservación podría considerarse el primer ejemplo de biotecnología microbiana y, pese a que el conocimiento científico ha progresado enormemente desde entonces, el objetivo de esta disciplina sigue siendo el mismo: explorar los recursos biológicos disponibles para encontrar sus aplicaciones prácticas. El microscopio y las técnicas de aislamiento y cultivo de microorganismos son algunas de las herramientas más usadas para este fin y, de hecho, contribuyeron en gran medida a los hallazgos más célebres de la microbiología durante el siglo XX (p. ej., descripción de patógenos, desarrollo de antibióticos...). Sin embargo, el descubrimiento de que la mayor parte de los microorganismos son incultivables puso en evidencia la necesidad de desarrollar nuevos métodos para caracterizar toda la diversidad microbiana que no había sido explorada hasta la fecha.

La secuenciación del ADN se ha convertido en una herramienta esencial desde entonces. Aunque hubo varios intentos previos de desarrollar técnicas de secuenciación de ácidos nucleicos, el **método de Sanger** (o secuenciación de Sanger, en honor a su principal inventor) fue el primero en ser adoptado de forma masiva. Este método se basa en el uso de didesoxinucleótidos (ddNTPs), ADN polimerasa, cebadores (*primers*) y desoxirribonucleótidos (dNTPs). La extensión del ADN empieza de forma normal usando dNTPs hasta que un ddNTP es aleatoriamente incorporado a la cadena, causando la parada de la reacción. Al final del proceso se obtiene una mezcla de oligonucleótidos de diferentes tamaños que pueden ser resueltos mediante electroforesis, utilizando un canal diferente para cada ddNTP. Actualmente, se usan ddNTPs marcados por fluorescencia, lo que permite realizar la lectura de la secuencia de ADN en un solo canal. Además, la separación se realiza mediante electroforesis capilar en instrumentos automatizados que permiten varias reacciones en paralelo. El método de Sanger ha sido usado en diferentes proyectos, destacando el Proyecto Genoma Humano (PGH). A nivel técnico, esta metodología se caracteriza por generar secuencias de gran calidad (≈0.1% de error) y de tamaño moderado (≈1.000 pb), a costa de tener un bajo rendimiento (pocas secuencias por carrera de secuenciación) y un elevado precio por nucleótido leído. De hecho, pese al éxito de los primeros proyectos de secuenciación genómica, pronto se hizo evidente que la cantidad de infraestructura, tiempo, dinero y personal necesario para obtener una secuencia de ADN completa a partir de un único organismo usando esta técnica no era sostenible.

En consecuencia, se desarrollaron nuevos métodos que se caracterizaban por la paralelización masiva de las reacciones de secuenciación, incrementando así el rendimiento del proceso. Estas tecnologías son comúnmente conocidas como plataformas de secuenciación de siguiente generación (**NGS**, por sus siglas en inglés). **Illumina** es la plataforma NGS más utilizada en la actualidad. Su método de secuenciación es parecido al ideado por Sanger, ya que usa dNTPs modificados que impiden la elongación de las cadenas de ADN. Sin embargo, la reacción es reversible en este caso: los grupos químicos que bloquean el extremo 3'-OH de la ribosa pueden ser eliminados. En Illumina, los cuatro dNTPs bloqueados y marcados por fluorescencia son proporcionados al mismo tiempo. Uno de ellos será añadido a la cadena que se está sintetizando. Tras la eliminación del resto de dNTPs, la fluorescencia es medida y se puede leer el dNTP que ha sido incorporado. Por último, el extremo 3'-OH es regenerado, dando inicio a otro ciclo de elongación y lectura. Esta reacción ocurre simultáneamente en millones de fragmentos de ADN diferentes (*clusters*). Cabe resaltar que la secuenciación no se produce a partir de una única molécula, sino que cada *cluster* está constituido por miles de clones de un mismo fragmento. Al proceso utilizado para conseguir los clones de cada molécula de ADN original se le denomina amplificación puente. La secuenciación Illumina se caracteriza por resultar en lecturas cortas (25 – 300 pb) y de alta calidad (0.5 – 0.1% de error).

En los últimos años, han aparecido tecnologías de **secuenciación de tercera generación** (TGS). Las plataformas TGS se caracterizan por producir secuencias largas (>10 Kpb) en tiempo real, a partir de moléculas únicas de ADN (sin amplificación previa). Estos secuenciadores han sido desarrollados por dos compañías, Pacific Biosciences (PacBio) y Oxford Nanopore Technologies (ONT), y se basan en metodologías completamente diferentes.

La **secuenciación PacBio** utiliza ADN polimerasas modificadas que se unen a fragmentos de ADN previamente circularizados. Este complejo es inmovilizado en el fondo de unos micropocillos llamados ZMWs. En concreto, cada pocillo contiene una única molécula de ADN. A continuación, la ADN polimerasa empieza a incorporar dNTPs marcados, que son detectados en tiempo real, es decir, sin detener la reacción tras la adición del dNTP. El fluoróforo unido al dNTP es liberado por la propia polimerasa, permitiendo que la señal fluorescente se desvanezca antes de leer el siguiente dNTP. PacBio ofrece dos modos de secuenciación diferentes: CCS y CLR. En el modo CCS, fragmentos de 10-25Kbp son leídos varias veces por la misma polimerasa, permitiendo obtener un porcentaje de error inferior al 0.1%. En el modo CLR, la molécula es leída una sola vez por la polimerasa. Esto permite secuenciar fragmentos de ADN más grandes (>50 Kpb), pero provoca un aumento en la tasa de error (8 – 13%).

La **secuenciación ONT** o **secuenciación por nanoporos** se fundamenta en hacer pasar una molécula de ADN (o ARN) por un canal (nanoporo), provocando un cambio en la corriente iónica. El bloqueo de la corriente es específico a la estructura química de los dNTPs. Leyendo los cambios en la corriente eléctrica se puede deducir la secuencia nucleotídica del fragmento de ADN. Los elementos básicos de los secuenciadores ONT son:

- **Nanoporos**: proteínas que forman los canales transmembranales por los que pasan las moléculas.
- **Celdas de flujo (*flow cells*)**: superficie física donde se sitúan los nanoporos y que incluye los electrodos y circuitos que controlan y detectan los cambios en la corriente eléctrica.
- **Proteínas motoras**: controlan la velocidad a la que las moléculas atraviesan los nanoporos.
- ***Basecallers***: programa informático que convierte las señales eléctricas en secuencias de nucleótidos. Se basan en inteligencia artificial.
- **Componentes computacionales**: discos duros, memorias RAM, procesadores y tarjetas gráficas.

En general, las plataformas TGS (ONT + PacBio) ofrecen una serie de ventajas con respecto a las tecnologías NGS, en general, y a Illumina, en particular:

- **Tamaño de lectura.** El tamaño de lectura máximo en Illumina se sitúa sobre las 300 pb x 2 (se lee la secuencia en sentido directo e inverso), mientras que tanto en PacBio como en ONT se superan de largo las 10 Kpb.
- **No se necesita amplificación**. La secuenciación se basa en moléculas únicas, lo que evita los sesgos introducidos por la PCR.
- **Tiempo de secuenciación reducido**. En TGS, la reacción no se pausa tras la incorporación de cada dNTP, lo que permite acelerar la obtención de resultados.
- **Análisis epigenético directo**. Las plataformas

TGS pueden detectar nucleótidos modificados sin necesidad de realizar transformaciones químicas durante la preparación de librerías.

Además, la secuenciación ONT, que juega un papel fundamental en esta tesis, posee una serie de particularidades con respecto al resto de plataformas:

- **Portabilidad**. La secuenciación por nanoporos no implica elementos ópticos, sino que se basa en componentes electrónicos que pueden ser miniaturizados. Prueba de ello es el secuenciador portátil MinION de ONT.
- **Análisis en tiempo real**. Las secuencias se generan inmediatamente después de que la molécula de ADN atraviese el nanoporo, o incluso de manera simultánea a la translocación. Estas lecturas pueden analizarse sin necesidad de esperar a que la carrera de secuenciación termine.
- **Tamaño de lectura**. En teoría, la secuenciación ONT puede producir lecturas de cualquier tamaño. De hecho, se ha conseguido leer fragmentos de hasta 4.2 Mpb de longitud con esta tecnología.
- **Precisión de la secuenciación**. Es la principal desventaja de esta plataforma. Actualmente, el error de secuenciación varía desde el 13% al 2%, dependiendo de varios factores (tipo de secuenciación, tipo de muestra, modelo de nanoporo y *basecaller* usado…). Sin embargo, los nuevos kits de secuenciación de ONT prometen una precisión mayor al 99.3% (o 99.8% si se usa el modo de secuenciación 'Duplex').
- **Coste**. El MinION es, de lejos, el secuenciador más económico del mercado, pero el coste por nucleótido leído de este dispositivo es alto en comparación con otros secuenciadores (p. ej., PromethION de ONT o NovaSeq 6000 de Illumina).
- **Secuenciación directa de ARN.** Las plataformas ONT son las únicas que permiten leer el ARN directamente, sin conversión previa a ADN.

Pese a todas estas características, la tecnología Illumina sigue siendo la manera más económica de conseguir lecturas de alta calidad, dado el elevado rendimiento de sus secuenciadores y la alta implantación de la compañía en el mercado.

Todas las técnicas de secuenciación descritas hasta el momento han sido aplicadas de una u otra manera para estudiar comunidades microbianas o 'microbiomas'. El término microbioma define el conjunto de microorganismos (microbiota) que habitan un biotopo, así como los elementos genéticos y biológicos asociados a estos microbios (p. ej., genes, genomas, proteínas, metabolitos…). Los microbiomas se pueden estudiar mediante múltiples técnicas moleculares, aunque las más comunes son:

- **Metataxonomía**. Se basa en la amplificación por PCR de genes marcadores que son característicos de un grupo taxonómico determinado (p. ej., gen 16S ribosomal para procariotas o 18S ribosomal para eucariotas).
  - <u>Ventajas</u>: técnica económica, análisis sencillo y bases de datos (BBDD) más completas.
  - <u>Desventajas</u>: solo se detecta un grupo taxonómico, no hay gen universal para virus, resolución taxonómica limitada (p. ej., nivel de género) y no aporta información sobre funciones.
- **Metagenómica**. Consiste en extraer el ADN total de una muestra y secuenciar los genomas completos. Principalmente, el análisis de los datos metagenómicos se puede realizar de dos formas: (1) asignando directamente la taxonomía y la función de las lecturas generadas o (2) ensamblando estas lecturas para formar fragmentos de ADN más grandes (*contigs*) y reconstruir el metagenoma (conjunto de genomas presentes en una muestra).

  - Ventajas: se detectan todos los grupos taxonómicos a la vez, aporta información funcional, permite la recuperación de genomas de microorganismos individuales y tiene una mayor resolución taxonómica.
  - Desventajas: más cara, BBDD menos completas y difícil de aplicar en muestras con escasa biomasa.
- **Metatrancriptómica**. Se basa en la extracción del ARN total de una muestra y la secuenciación del ARN mensajero (ARNm).
  - Ventajas: se detectan todos los grupos taxonómicos a la vez y aporta información sobre las funciones activas.
  - Desventajas: más cara, el ARN es inestable, la concentración de ARN ribosomal (ARNr) es mucho mayor que la de ARNm y es difícil de aplicar en muestras con escasa biomasa.

Todas estas técnicas pueden ser llevadas a cabo mediante secuenciación ONT. Las ventajas de esta tecnología (p. ej., portabilidad o análisis en tiempo real), favorecen el desarrollo de aplicaciones novedosas que pueden abrir nuevos mercados. Dado que la presente tesis se llevó a cabo en el marco de un doctorado industrial, esta perspectiva adquiere una especial relevancia. Los secuenciadores ONT han sido aplicados para una gran variedad de fines, destacando las **aplicaciones clínicas** (p. ej., vigilancia genómica de patógenos emergentes o detección de genes de resistencia a antibióticos), **industriales** (p. ej., análisis de calidad de aguas residuales o detección de patógenos en cultivos agrícolas) y **ecológicas** (p. ej., secuenciación de microbiomas en la Antártida o la estación espacial internacional). De entre todas las posibles aplicaciones de la tecnología en el marco de la biotecnología microbiana, dos de ellas despiertan un gran interés:

- **Monitorización de procesos de digestión anaerobia**. La digestión anaerobia (DA) implica la conversión de sustratos complejos, típicamente residuos (restos agrícolas, estiércol...), en biogás (mezcla de metano y dióxido de carbono) mediante transformaciones microbianas. La DA ocurre en cuatro fases: **hidrólisis, acidogénesis, acetogénesis y metanogénesis**. Todas estas fases se pueden desarrollar en un solo reactor (o digestor) o en varios. En este último caso, la hidrólisis y la acidogénesis ocurre en un digestor (**fase de acidificación**) y la acetogénesis y la metanogénesis en otro (**fase de metanización**). El resultado de la fase de acidificación es una mezcla de ácidos grasos de cadena corta (ácidos grasos volátiles) que pueden ser usados para producir biogás u otras sustancias de interés industrial. La DA es altamente variable, ya que depende de muchos factores (pH, temperatura, tipo de reactor...). Estos factores también afectan a los microorganismos responsables de la producción de biogás y pueden influir en la eficiencia del proceso. Por ello, en la presente tesis se ha desarrollado una aplicación basada en secuenciación por nanoporos para el monitoreo de comunidades bacterianas y arqueanas.
- **Mejora del proceso de bioprospección**. La bioprospección microbiana es la búsqueda de microorganismos y productos biológicos relevantes desde el punto de vista biotecnológico. Las expediciones de bioprospección suelen desarrollarse en lugares inexplorados y recónditos e implican la toma de distintas muestras que son analizadas posteriormente en el laboratorio. Estos análisis pueden prolongarse durante meses. Una vez conocidos los resultados, la explotación de los recursos naturales de más relevancia (según los estudios) queda supeditada a una gran limitación: el número de muestras obtenido durante la bioprospección. En este sentido, obtener más muestras implica repetir la expedición de muestreo, lo cual no siempre es posible. En la presente tesis se ha evaluado la idoneidad de la secuenciación ONT para caracterizar muestras *in situ*, es decir, durante la propia expedición.

Pese al enorme potencial de las plataformas ONT, las herramientas experimentales y bioinformáticas disponibles para caracterizar microbiomas mediante está técnica eran muy limitadas cuando se inició esta tesis. Por ello, la **motivación principal** del presente trabajo fue identificar, optimizar y validar nuevas metodologías metataxonómicas y metagenómicas basadas en secuenciación por nanoporos, así como diseñar e implementar aplicaciones novedosas de esta tecnología para resolver problemas de relevancia industrial o biotecnológica.

Objetivos

Esta tesis fue diseñada con la finalidad de optimizar la secuenciación ONT para estudiar comunidades microbianas de interés industrial o biotecnológico. Para ello, se definieron los siguientes objetivos:

- Desarrollar protocolos experimentales y bioinformáticos para el análisis metataxonómico de comunidades de bacterias y arqueas usando la tecnología ONT y:
  - Aplicar estos nuevos protocolos para caracterizar microbiomas asociados con la producción de biogás.
  - Evaluar el potencial de las plataformas portátiles de ONT para mejorar la estrategia de muestreo durante una expedición de bioprospección.
- Comprobar el rendimiento de diferentes métodos de ensamblaje de lecturas generadas a partir de la secuenciación metagenómica por nanoporos y definir las ventajas y limitaciones de esta tecnología en comparación con el estado del arte del ensamblaje metagenómico.

Metodología

Secuenciación y análisis metataxonómico

El protocolo básico para los estudios metataxonómicos consistió en los siguientes pasos:

- **Extracción y cuantificación de ADN**. Se realizó usando dos kits comerciales diferentes: DNeasy Power Soil Kit (QIAGEN, Alemania) y FastDNA Spin Kit for Soil (MP Biomedicals GmbH, Alemania). Las muestras procedentes de digestores anaerobios fueron lavadas con tampón fosfato salino (PBS) para reducir la concentración de sustancias contaminantes. El ADN fue cuantificado preferentemente por fluorescencia, usando el kit Qubit x1 dsDNA High-Sensitivity Assay kit (Qubit 2.0 Fluorometer, Thermo Fisher, EE.UU.), aunque también se usó el equipo Nanodrop 1000 Spectrophotometer (Thermo Scientific, EE.UU.)
- **Amplificación del gen 16S ARNr por PCR**. Se eligió este gen por ser específico de los microorganismos procariotas, que eran el objetivo de los análisis. Como la secuenciación ONT es capaz de producir lecturas largas, se amplificó el gen ribosomal completo (≈1.500 pb). Para ello, se usaron dos parejas de cebadores diferentes, una para amplificar selectivamente las arqueas (Arch8F y Arch1492R) y otra para amplificar las bacterias (S-D-Bact-0008-a-S-16 y S-D-Bact-1492-a-A-16). Ambos cebadores incluían los adaptadores universales de ONT. Los amplicones fueron purificados usando el kit Agencourt AMPure XP beads (Beckman Coulter, EE.UU.) o el kit NucleoMag (Macherey-Nagel, Alemania) y el ADN fue cuantificado con el kit de Qubit antes mencionado.
- ***Barcoding.*** Los *barcodes* (etiquetas para reconocer diferentes muestras) se añadieron usando los kits PCR Barcoding Expansion 1-12 y PCR Barcoding Expansion 1-96 (ONT, Reino Unido). El ADN fue purificado y cuantificado como se ha descrito anteriormente.
- **Preparación de librerías**. Se usaron dos versiones diferentes del Ligation Sequencing Kit (ONT, R.U.): SQK-LSK108 y SQK-LSK109. Luego se repararon los extremos del ADN con el kit NEBNext FFPE DNA Repair Mix (New

England Biolabs, EE.UU.) y se purificaron y cuantificaron los amplicones como ya se ha descrito.

- **Secuenciación ONT**. Todos los experimentos se llevaron a cabo usando las *flow cells* R9.4.1 (ONT, R.U.) y el *basecalling* se realizó mediante MinKNOW usando los *basecallers* más actualizados en el momento de la secuenciación (Albacore o Guppy). Las lecturas con una calidad de Phred menor a 7 fueron descartadas.
- **Análisis bioinformático**. Las secuencias fueron analizadas siguiendo dos protocolos diferentes desarrollados en el marco de este trabajo. El primer protocolo (o *pipeline*) se basaba en BLAST, tal y como se implementa en la herramienta QIIME (v. 1.9.1.), para realizar la asignación taxonómica de las lecturas. Al segundo protocolo se le denominó 'Spaghetti'. Este usaba la herramienta minimap2 (v. 2.17) para identificar la taxonomía de las secuencias usando una base de datos de referencia. Ambos *pipelines* incluían un paso previo de pretratamiento de las secuencias que consistió en: (1) eliminación de adaptadores con Porechop; (2) eliminación de secuencias demasiado cortas (<700–1.200 pb) o demasiado largas (>1700-1800 pb) con Nanofilt; (3) comprobación de la calidad de las lecturas con NanoStat; (4) eliminación de secuencias quiméricas con YACRD. Como BBDD de referencia se usaron GreenGenes (v. 13.8) y SILVA (v. 132 o v. 138). El resultado de la asignación taxonómica fue modificado mediante *scripts* de QIIME o propios, según el caso, obteniéndose tablas de abundancia absoluta y relativa.
- **Análisis estadístico y visualización de datos**. Todos los análisis estadísticos se realizaron con el paquete phyloseq (v. < 1.30.0), dentro de R. Para la diversidad alfa, todas las muestras fueron rarefactadas para obtener un número de secuencias uniforme para cada muestra estudiada. Para la diversidad beta, se realizaron Análisis de Coordenadas Principales (PcoA) usando la métrica Bray-Curtis para calcular la disimilitud entre muestras. Los *heatmaps*, fueron creados con ampvis2 (v. < 3.3.1), mientras que las figuras interactivas se construyeron con plotly (v. 4.9.2.1). Los diagramas de Venn se generaron usando una herramienta web[33]. Los análisis de abundancia diferencial se realizaron con DESeq2 (v. < 1.26.0). Cuando se llevaron a cabo pruebas múltiples, los p-valores se corrigieron con el método de Benjamini-Hochberg. Cabe resaltar que algunos de los análisis descritos fueron incluidos en Spaghetti.

Análisis metagenómico

Para evaluar las características del ensamblaje metagenómico de datos ONT, se seleccionaron varias herramientas diferentes y se realizaron distintas pruebas de uso. La metodología de este estudio consistió en los siguientes pasos:

- **Selección de herramientas para el ensamblaje *de novo***. Las herramientas se eligieron siguiendo tres normas básicas: (1) el software debía ser gratuito, (2) incluir una guía de usuario y (3) haber sido ampliamente usado por la comunidad científica. En total, se seleccionaron tres ensambladores metagenómicos desarrollados para secuencias cortas (metaSPAdes, Megahit, Minia) y diez herramientas para ensamblar lecturas de plataformas TGS (Canu, HINGE, miniasm, metaFlye v2.4, metaFlye v.2.7, Pomoxis, Raven, Readbean, Shasta, Unicycler), aunque algunas de ellas no se pudieron instalar o ejecutar correctamente[34] .
- **Evaluación de los ensamblajes**. La integridad y contigüidad de los ensamblajes fueron evaluados con QUAST (v. 5.0.2). Los ensamblajes *de novo* se compararon con los metagenomas de referencia usando metaQUAST (v. 5.0.2) o

[33] http://bioinformatics.psb.ugent.be/webtools/Venn/

[34] Los comandos utilizado para ejecutar cada herramienta se pueden consultar en https://zenodo.org/record/3935763

minimap2 (v. 2.15). La precisión del ensamblaje se comprobó mediante dos estrategias: (1) minimap2 y (2) MuMmer4 (v. 3.23). Los *clusters* de genes biosintéticos (BGCs), que son estructuras difíciles de ensamblar por ser largas y contener material repetitivo, se predijeron usando AntiSMASH (v 5.0).

- **Corrección (*polishing*) de los ensamblajes**. Los ensamblajes se corrigieron mediante dos herramientas: Racon (v. 1.4.3) y Medaka (v. 2.2.2) usando secuencias generadas por ONT e Illumina. Solo los ensamblajes corregidos con lecturas ONT fueron sometidos a la corrección con Medaka. La precisión de los ensamblajes fue determinada después de cada ronda de corrección.

Resultados y Discusión

Desenmarañando los microbiomas industriales: el caso de la digestión anaerobia

Tradicionalmente, la digestión anaerobia (DA) ha sido considerada un sistema de caja negra, ya que no se conocen en profundidad los procesos microbianos que propician la conversión de materia orgánica diversa en biogás u otras sustancias de valor añadido. En este estudio, los protocolos para el análisis metataxonómico de secuencias ONT descritos anteriormente fueron puestos en práctica para evaluar el impacto de distintos parámetros operacionales (tipo de sustrato, pretratamientos, presencia de sustancias nocivas...) en las bacterias y arqueas características de la DA.

En un **primer trabajo**, las fracciones líquidas y solidas resultantes del licuado de biomasa herbácea fueron usadas como co-sustratos para la co-digestión anaerobia de lodos activos provenientes de aguas residuales. La comunidad de arqueas metanógenas fue analizada a lo largo del tiempo mediante secuenciación ONT, en paralelo a la cuantificación de otros parámetros químicos de interés (pH, concentración de metano, concentración de ácidos grasos volátiles...). Los reactores alimentados con los lodos y la fracción herbácea líquida desarrollaron un microbioma más estable, enriquecido en el género *Methanosarcina,* resultando también en un mayor rendimiento en la producción de metano. Por el contrario, los digestores alimentados con los lodos y la fracción herbácea sólida mostraron un microbioma más inestable, con una producción de biogás menor. En este caso, el género mayoritario fue *Methanosaeta*, que es un taxón arqueano típico de lodos activos. En otras palabras, la co-digestión de lodos activos con biomasa herbácea sólida no propició un cambio en el microbioma de los lodos. Dicho cambio sí que se consiguió al usar biomasa líquida, mejorando el rendimiento del proceso.

El **segundo trabajo** consistió en examinar el impacto de distintos métodos de eliminación de nitrógeno (NH-3 *stripping* y precipitación MAP) sobre el proceso de DA, usando estiércol de gallina (rico en amoníaco) como sustrato. El amoníaco puede causar la inhibición de las reacciones de metanogénesis y, por tanto, la reducción de su concentración en el sustrato de la DA es vital para aumentar la eficiencia del proceso. Durante el estudio, se caracterizaron las comunidades bacterianas presentes en varios reactores de acidificación, donde se realiza la hidrólisis y la acidogénesis de la materia orgánica en los procesos de DA de dos fases. Para ello, se aplicó secuenciación metataxonómica basada en nanoporos a lo largo del tiempo, obteniéndose muestras antes y después de los tratamientos de eliminación. Cada método de eliminación se implementó y evalúo por separado. A pesar de los tratamientos, los microbiomas resultaron ser muy similares en todos los experimentos, lo que demostró la gran robustez y resiliencia de los microorganismos que participan en la DA. En cualquier caso, la precipitación MAP tuvo un impacto más notable, provocando cambios en las abundancias de un mayor número de taxones. Además, la abundancia de *Acholeplasma* y *Erysipelotrichaceae* UCG-004 aumentó progresivamente en todos los experimentos. Estos géneros son predominantes en procesos de AD sometido a altas concentraciones de amoniaco, lo que sugirió que los microbiomas iniciales evolucionan hacia comunidades bacterianas mejor adaptadas a las condiciones.

Aparte de estos trabajos, la tecnología ONT se aplicó en dos estudios más para: (1) analizar las interacciones entre bacterias durante el proceso de DA mediante la aplicación del modelo de Lotka-Volterra; (2) estudiar la composición taxonómica y las características funcionales de un biofilm bacteriano formado en la cara interna de un reactor transparente expuesto a la luz natural. En este último caso, la secuenciación metataxonómica por nanoporos fue complementada con secuenciación metagenómica por Illumina.

En definitiva, a lo largo de la presente tesis se ha demostrado que la secuenciación ONT puede ser aplicada con éxito para caracterizar comunidades bacterianas y arqueanas relacionadas con la DA. Nuestros estudios han demostrado que los microbiomas responsables de la producción de biogás son complejos y están influenciados por múltiples factores. Pese a ello, existen distintos marcadores microbianos, como por ejemplo *Methanosarcina*, que se asocian con una mayor eficiencia del proceso. Este hecho allana el camino para un objetivo más ambicioso: usar las plataformas portátiles de ONT para monitorizar cualquier transformación microbiana a escala industrial. En este sentido, distintos marcadores podrían ser examinados *in situ* para determinar el rendimiento de los procesos, usando estrategias de bioaumentación con cepas optimizadas en caso de que sea necesario corregir el comportamiento del sistema. Más allá de centrarse en marcadores individuales, las abundancias de todos los microorganismos podrían combinarse con datos químicos procedentes de otros análisis (pH, temperatura, producción de metano...) para alimentar algoritmos de inteligencia artificial que puedan modelizar el estado del sistema con mayor exactitud.

Secuenciación de microbioma in situ para guiar el proceso de bioprospección

El proceso de bioprospección microbiana se inicia con la toma de muestras (trabajo de campo) y finaliza con la búsqueda de recursos biológicos, es decir, microorganismos y sus funciones, con potencial biotecnológico (trabajo de laboratorio). En este trabajo, se evalúo la idoneidad de usar la secuenciación ONT para predecir el potencial de las distintas muestras *in situ*, sin necesidad de volver al laboratorio. Para ello, se preparó una expedición al Desierto de Tabernas (Almería, España). Como prueba de concepto, se acotó el objetivo de la bioprospección a la búsqueda de microorganismos que, según la bibliografía previa, son capaces de resistir a condiciones extremas de desecación y radiación. El foco se puso en estas bacterias por varios motivos: (1) se trata de un grupo amplio de microorganismos de afiliación taxonómica diversa; (2) estas bacterias pueden tener aplicaciones biotecnológicas importantes, por ejemplo, en la industria cosmética (protección solar); (3) el Desierto de Tabernas alberga una gran riqueza de microorganismos de este tipo.

Durante la primera jornada de muestreo, se tomaron catorce muestras diferentes, principalmente costras biológicas (*biocrusts*) y suelos. Estos tipos de muestras se eligieron porque suelen ser ricos en microorganismos resistentes a las condiciones deseadas. Sobre estas muestras se aplicó la secuenciación metataxonómica, tal y como se ha descrito anteriormente. Cabe resaltar que la secuenciación no se realizó directamente en el campo, sino que se produjo en un laboratorio móvil instalado en un apartamento a 15 km del desierto. Esto permitió poder aplicar los mismos protocolos que se llevan a cabo en un laboratorio convencional, lo que originó un aumento de la eficiencia de la secuenciación en comparación con otros estudios de campo. El análisis de datos fue realizado con la herramienta Spaghetti. Esta herramienta fue específicamente desarrollada para este estudio por dos motivos: (1) el protocolo anterior (basado en BLAST) presentaba limitaciones computacionales que demoraban la obtención de resultados. Con la nueva herramienta, todos los análisis, incluyendo la visualización de datos y la toma de decisiones, terminaron en menos de tres horas; (2) Spaghetti se basa en minimap2, que demostró realizar una asignación taxonómica más precisa de acuerdo con otros estudios publicados meses antes a la preparación de la expedición. Los resultados del protocolo bioinformático permitieron categorizar las muestras en base a la diversidad y proporción de bacterias resistentes que contenían.

Seguidamente, se realizó una segunda jornada de muestreo, que se centró en obtener muestras relevantes, según los análisis. En concreto, se seleccionó (1) la muestra que presentaba una mayor abundancia y diversidad de bacterias resistentes (*biocrust*), (2) la muestra que presentaba una menor proporción y abundancia (*biocrust;* control 'negativo' del proceso) y (3) una muestra de suelo (control 'negativo' frente a las muestras de *biocrust*). A continuación, se tomaron réplicas biológicas de estas muestras. De vuelta en el laboratorio, se estableció una colección de cultivos a partir de todas las muestras. En total, se obtuvieron 166 cepas diferentes afiliadas a 50 géneros distintos.

Pese a las limitaciones asociadas con las técnicas de cultivo (p. ej., la mayor parte de los microorganismos son incultivables), los datos metataxonómicos y los datos de la colección de cultivos correlacionaron bien, es decir, las bacterias aisladas habían sido detectadas durante la expedición de muestreo a través de la secuenciación *in situ*. Además, la muestra que mostraba un mayor potencial a nivel de microbioma también dio lugar a un conjunto de aislados más interesante, que incluía un mayor número de cepas pertenecientes a géneros típicamente resistentes a la radiación y a la desecación. Por su parte, la muestra que mostraba menor diversidad de estas bacterias resultó en un conjunto de cultivos redundante (con respecto a las otras muestras) y menos interesante. De este modo, se anticipa que la secuenciación portátil de ONT puede ser usada para mejorar la toma de decisiones durante las expediciones de bioprospección. En otras palabras, los investigadores pueden usar esta herramienta para decidir qué muestras son más interesantes según sus objetivos, centrando la estrategia de muestreo en explotar intensivamente esos tipos de muestras frente a otros.

Hacia la metagenómica basada en secuencias largas: estudio comparativo de métodos de ensamblaje para secuenciación por nanoporos

La metagenómica ha permitido recuperar genomas de microorganismos previamente desconocidos, gracias a la capacidad de esta técnica para estudiar especies no cultivables. La secuenciación con plataformas NGS (lecturas cortas) suele dar lugar a genomas y metagenomas muy fragmentados, dado que los microorganismos suelen contener elementos genéticos repetitivos de mayor tamaño que las lecturas obtenidas. Las plataformas TGS son capaces de generar secuencias largas o ultra largas, que tienen el potencial de dar lugar a ensamblajes más contiguos. No obstante, los errores asociados a estas tecnologías dificultan la reconstrucción de los metagenomas y suponen un nuevo reto algorítmico.

En este trabajo, se evaluó de forma exhaustiva el rendimiento de diferentes herramientas de ensamblaje para reconstruir metagenomas simples y conocidos (comunidades *mock*) secuenciados mediante plataformas ONT. De todas las herramientas probadas, tan solo se obtuvieron resultados robustos con metaFlye, Raven y Canu, aunque este último ensamblador demostró ser ineficiente desde el punto de vista computacional. Estas tres herramientas dieron lugar a genomas muy contiguos (pocos *contigs*) e íntegros a partir del ensamblaje metagenómico. A pesar de la baja precisión de la tecnología ONT, los ensamblajes finales tuvieron un error menor al 0.5%-0.2%, dependiendo del ensamblador.

La corrección de los metagenomas (*polishing*) fue necesaria para revertir los errores de tipo inserción/deleción (indel), que pueden originar fallos en la predicción de genes codificantes, ya que introducen cambios en el marco abierto de lectura. De hecho, se demostró que la corrección daba lugar a una mejor predicción de los *clusters* de genes biosintéticos (BGCs), que son estructuras complejas de ensamblar. El *polishing* se realizó usando los propios datos de ONT o secuencias generadas con otra plataforma. En este sentido, aunque las secuencias de Illumina tienen una mayor calidad, la corrección de los ensamblajes usando estos datos no siempre da lugar a un aumento en la precisión de los mismos, sino que este proceso depende de la herramienta elegida para reconstruir el metagenoma.

En general, se probó que la secuenciación metagenómica por nanoporos es suficiente para caracterizar comunidades microbianas poco complejas. Estos resultados están en la misma línea que los obtenidos por otros grupos de investigación, que además han demostrado que la secuenciación ONT aumenta la contigüidad del ensamblaje en muestras reales y complejas, permitiendo la obtención de genomas bacterianos circulares directamente desde el metagenoma. Aunque la mayor parte de los errores asociados con esta tecnología se corrigen durante el proceso de ensamblaje, los indels representan un problema para la predicción génica. Por ello, los ensamblajes híbridos usando datos provenientes de plataformas NGS y TGS podrían considerarse el estándar actual para el ensamblaje, ya que combinan lo mejor de los dos mundos: la contigüidad e integridad de TGS y la precisión de NGS. No obstante, usar dos secuenciadores diferentes implica un mayor coste económico y una elevada carga de trabajo. Por ello, es probable que la tendencia evolucione hacia usar tan solo una plataforma. Con esta perspectiva, ONT está mejorando la calidad de sus secuencias, logrando una precisión de hasta el 99.8% cuando se combina el nuevo kit Q20+ con el modo de secuenciación 'Duplex'. Estas mejoras todavía siguen en la fase de implementación y, en consecuencia, su impacto sobre el ensamblaje genómico o metagenómico no ha sido evaluado todavía. Sin embargo, se puede predecir que estas nuevas características mejorarán la calidad de los ensamblajes y que el efecto será más notable en aquellos genomas con menor cobertura de lecturas, puesto que en estos casos los errores no se pueden corregir con la profundidad de secuenciación. Considerando estos avances, la secuenciación ONT puede convertirse en la referencia para el ensamblaje (meta)genómico en un futuro próximo, aunque todavía se necesitan evaluaciones sistemáticas que demuestren los beneficios de esta tecnología frente a otras metodologías (Illumina, PacBio o ensamblaje híbrido).

Aparte de mejorar el ensamblaje, la secuenciación metagenómica por nanoporos puede ser integrada en las aplicaciones *in situ* descritas anteriormente. Esto ayudaría a superar algunas de las limitaciones de la metataxonomía y a incrementar la potencia del análisis en muchos aspectos. Por ejemplo, la monitorización industrial podría ir más allá de los organismos procariotas, ya que la metagenómica permite la detección de eucariotas, procariotas y virus en el mismo experimento. Con esta aproximación, la resolución taxonómica aumentaría desde el nivel de género típicamente obtenido en metataxonomía, hasta el nivel de especie o incluso cepa. Esto incrementaría la sensibilidad y especificidad de las aplicaciones. Por último, tanto la monitorización como la bioprospección se beneficiarían de la capacidad de la secuenciación metagenómica para aportar datos sobre las funciones de las comunidades microbianas. Estos análisis podrían utilizarse para rastrear genes y rutas metabólicas de interés (p. ej., *clusters* de genes biosintéticos) o para identificar nuevos recursos genéticos (p. ej., enzimas altamente divergentes). Como la metagenómica también tiene sus propios inconvenientes, los trabajos futuros deben ayudar a decidir qué estrategia (metagenómica o metataxonomía) encaja mejor con el objetivo, la logística y el presupuesto de cada aplicación.

El futuro de las aplicaciones basadas en secuenciación por nanoporos

Pese a que la secuenciación ONT sigue, básicamente, en fase de desarrollo, esta tecnología ha demostrado ser ya de una gran utilidad, tal y como se deriva de los resultados de esta tesis. En un contexto más amplio, esta metodología ha jugado un papel esencial en la vigilancia genómica del SARS-CoV-2, especialmente en países en vías de desarrollo que no tienen acceso a otro tipo de secuenciadores mucho más costosos. Por tanto, la expansión de la secuenciación por nanoporos se prevé imparable y, en base a la experiencia adquirida durante esta tesis, se pueden hipotetizar varias líneas de investigación ambiciosas que pueden revolucionar la biotecnología microbiana en los años venideros:

- **Solución completa (*end-to-end*)**. Antes o después la secuenciación ONT cruzará la barrera de la investigación, expandiéndose a otros ámbitos como la seguridad alimentaria, la agricultura o la ciencia forense. Para ello, es estrictamente necesario simplificar los procesos de secuenciación y análisis de resultados, ya que el personal encargado de realizar las comprobaciones rutinarias en un ámbito no académico no dispondrán de los conocimientos técnicos de un investigador en biología molecular. ONT está invirtiendo recursos en esta dirección y el desarrollo de chips para la preparación de librerías de forma automatizada (VolTRAX) o la creación de programas bioinformáticos interactivos (EPI2ME) son solo algunos ejemplos. Poniendo como ejemplo los procesos de DA, importantes en el presente trabajo, la aplicación de monitorización ideal empezaría con un operario de la planta de biogás tomando una muestra del reactor, que sería procesada con una intervención humana mínima. Tras varias horas de secuenciación ONT, el sistema indicaría el estado del proceso y sugeriría acciones correctivas para mejorar la eficiencia, que serían implementadas por los técnicos de la planta.
- **Solución versátil (navaja suiza)**. En otras situaciones, como por ejemplo las expediciones de bioprospección, la estrategia debería ser totalmente diferente. En este caso, el personal responsable del análisis *in situ* sí que suele tener aptitudes técnicas de biología y bioinformática. Por tanto, las herramientas de secuenciación y de preparación de las muestras deberían ser fácilmente adaptables para proveer una solución a cada posible problema. Para ello, se propone el uso de laboratorios móviles (Bento Lab + MinION + ordenadores portátiles) que puedan ser usados para distintos fines: extraer ADN de cualquier tipo de muestras, realizar secuenciación metataxonómica (amplificando genes marcadores por PCR) o metagenómica, preprocesar las muestras que lo requieran (filtrar muestras de agua, quitar contaminantes de muestras de arcilla…), etc.
- **Metagenómica adaptativa**. La secuenciación ONT tiene la particularidad de que los datos se generan en el momento en el que la molécula de ADN empieza a atravesar el nanoporo. Estos datos pueden ser utilizados al instante, sin necesidad de esperar a que el nanoporo lea completamente el fragmento de ADN. Por tanto, transcurridos unos pocos segundos de la translocación (que ocurre a menos de 450 pb/s), ya se puede evaluar si el fragmento es relevante o no. Las moléculas no relevantes pueden ser eyectadas del nanoporo, que quedaría libre para leer otro fragmento. La metagenómica adaptativa es un concepto propuesto en esta tesis y que se basa en tres características básicas de la secuenciación ONT implementadas o propuestas por otros autores: (1) Tamaño adaptativo. El tamaño de la molécula de ADN puede determinarse antes de que esta empiece a ser leída. Por tanto, se pueden descartar los fragmentos por debajo de un determinado umbral de longitud. (2) Muestreo adaptativo. Esta funcionalidad

ya está disponible y consiste en comparar la secuencia que está siendo leída contra una base de datos preconfigurada para determinar si el fragmento es de interés o no. Esto permite enriquecer fragmentos de relevancia (p. ej., genes de resistencia a antibióticos) o descartar fragmentos no deseados (p. ej., secuencias provenientes del hospedador). (3) Precisión adaptativa. La misma molécula de ADN puede ser leída una y otra vez para mejorar la precisión de la secuencia.

El proceso de metagenómica adaptativa empezaría con la identificación del tamaño del fragmento. Las moléculas que superen un cierto tamaño serían leídas. Tras los primeros segundos de secuenciación, se determinaría si el fragmento es de interés o no en base a su secuencia nucleotídica. Los fragmentos relevantes se secuenciarían varias veces hasta reducir los errores por debajo de un umbral preestablecido (p. ej., 0.01%).

En un espectro más general, el futuro de las plataformas de secuenciación es tremendamente incierto y no es posible pronosticar qué tecnología se impondrá a las demás, si es que esto llega a suceder en algún momento. En cualquier caso, para que la secuenciación por nanoporos alcance el mismo estado de estandarización y robustez que goza la tecnología Illumina, todavía son necesarios muchos esfuerzos. En línea con la idea general de esta tesis, mejorar la usabilidad en escenarios "reales" (fuera del laboratorio) es estrictamente necesario para que el eslogan de ONT ("permitiendo el análisis de cualquier cosa, en cualquier lugar y por cualquier persona") deje de ser un mero reclamo publicitario para convertirse en una realidad arrolladora.

Conclusiones

A lo largo del presente trabajo, diferentes aplicaciones metagenómicas y metataxonómicas basadas en secuenciación ONT han sido diseñadas, implementadas y evaluadas. Estas aproximaciones han sido usadas con éxito para caracterizar ecosistemas de relevancia industrial y/o biotecnológica. Las conclusiones generales derivadas de esta tesis son:

- La secuenciación ONT puede ser aplicada para caracterizar las comunidades microbianas responsables de los procesos de digestión anaerobia (DA). En este sentido, nuestros análisis permitieron:
  - Detectar marcadores microbianos (bacterias y arqueas), como por ejemplo *Methanosarcina*, asociados con una mejora de la eficiencia en la producción de metano a partir de lodos activos provenientes del tratamiento de aguas residuales.
  - Medir el impacto de diferentes parámetros (sustrato de la co-digestión o pretratamiento para la eliminación de amoníaco) en el microbioma característico de la DA.
  - Proponer una estrategia basada en secuenciación ONT para monitorizar procesos industriales basados en transformaciones microbianas, después de haber identificado las ventajas y limitaciones actuales de la técnica.
- Los secuenciadores portátiles de ONT son herramientas valiosas para descifrar la composición taxonómica de las muestras tomadas durante una expedición de bioprospección, tal y como se demostró en una prueba de concepto desarrollada en el Desierto de Tabernas (Almería, España). Los datos generados por las plataformas ONT y los datos resultantes del aislamiento de microorganismos en el laboratorio correlacionaron bien, ya que a partir de las muestras que mostraban un mayor potencial a nivel de microbioma, se aislaron los microorganismos más interesantes. Por lo tanto, esta estrategia puede ser utilizada para informar la toma de decisiones durante el muestreo, aumentando así las posibilidades de cumplir los objetivos de la bioprospección (p. ej., identificación de microorganismos y biomoléculas relevantes desde una perspectiva biotecnológica).
- Varios protocolos para el análisis metataxonómico de muestras de interés han sido implementados y actualizados de acuerdo con los avances de la tecnología ONT. Entre ellos destaca Spaghetti, un programa bioinformático que permite el análisis automático de datos ONT y que demostró ser especialmente efectivo para estudios *in situ*.
- La secuenciación metagenómica por nanoporos permitió recuperar genomas extremadamente contiguos e íntegros a partir de distintos metagenomas. Los resultados de la evaluación realizada en esta tesis demuestran que metaFlye es el ensamblador más eficiente para datos ONT. Otras herramientas, como Raven, también son prometedoras. Nuestro estudio fue el primero en incluir todas las herramientas bioinformáticas diseñadas específicamente para esta tecnología, marcando el camino hacia la estandarización de los protocolos para el análisis de secuencias ONT.

References

1 Ambrose SH. Paleolithic Technology and Human Evolution. *Science* 2001;**291**:1748–53. doi:10.1126/science.1059487

2 Bud R. History of Biotechnology. *Nature* 1989;**337**:10. doi:10.1038/npg.els.0003086

3 Sibbesson E. Reclaiming the Rotten: Understanding Food Fermentation in the Neolithic and Beyond. *Environ Archaeol* 2019;**4103**. doi: 10.1080/14614103.2018.1563374

4 Achtman M. How old are bacterial pathogens? *Proc Royal Soc B* 2016;**283**:0–1. doi:10.1098/rspb.2016.0990

5 Muñoz-Ramirez ZY, Pascoe B, Mendez-Tenorio A, *et al.* A 500-year tale of co-evolution, adaptation, and virulence: Helicobacter pylori in the Americas. *ISME J* 2021;**15**:78–92. doi:10.1038/s41396-020-00758-0

6 Buchholz K, Collins J. The roots - A short history of industrial microbiology and biotechnology. *Appl Microbiol Biotechnol* 2013;**97**:3747–62. doi:10.1007/s00253-013-4768-2

7 Pariente N. A field is born. 2019. https://www.nature.com/articles/d42859-019-00006-2 (accessed 3 Oct 2021).

8 Wheelis ML. *Principles of Modern Microbiology*. London: Jones and Barlett Publishers 2007.

9 Puspita ID, Kamagata Y, Tanaka M, *et al.* Are uncultivated bacteria really uncultivable? *Microbes Environ* 2012;**27**:356–66. doi:10.1264/jsme2.ME12092

10 Louca S, Mazel F, Doebeli M, *et al.* A census-based estimate of earth's bacterial and archaeal diversity. *PLoS Biol* 2019;**17**:1–30. doi:10.1371/journal.pbio.3000106

11 Houpikian P, Raoult D. Traditional and molecular techniques for the study of emerging bacterial diseases: One laboratory's perspective. *Emerg Infect Dis* 2002;**8**:122–31. doi:10.3201/eid0802.010141

12 Mitchell PS, Persing DH. Current Trends in Molecular Microbiology. *Lab Med* 1999;**30**:263–70. doi:10.1093/labmed/30.4.263

13 Burki F, Roger AJ, Brown MW, *et al.* The New Tree of Eukaryotes. *Trends Ecol Evol* 2020;**35**:43–55. doi:10.1016/j.tree.2019.08.008

14 Parks DH, Chuvochina M, Chaumeil PA, *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* 2020;**38**:1079–86. doi:10.1038/s41587-020-0501-8

15 Cross KL, Campbell JH, Balachandran M, *et al.* Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat Biotechnol* 2019;**37**:1314–21. doi:10.1038/s41587-019-0260-6

16 Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics* 2016;**107**:1–8. doi:10.1016/j.ygeno.2015.11.003

17 Holley RW, Madison JT, Zamir A. A new method for sequence determination of large oligonucleotides. *Biochem Biophys Res Commun* 1964;**17**:389–94. doi:10.1016/0006-291X(64)90017-8

18 Sanger F, Brownlee GG, Barrell BG. A two-dimensional fractionation procedure for radioactive nucleotides. *J Mol Biol* 1965;**13**:373-IN4. doi:10.1016/S0022-2836(65)80104-8

19 Holley RW, Apgar J, Everett GA, *et al.* structure of a ribonucleic acid. *Science* 1965;**147**:1462–5. doi:10.1126/science.147.3664.1462

20 Brownlee GG, Sanger F. Nucleotide sequences from the low molecular weight ribosomal RNA of Escherichia coli. *J Mol Biol* 1967;**23**:337-IN9. doi:10.1016/S0022-2836(67)80109-8

21 Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 1975;**94**:441–8. doi:10.1016/0022-2836(75)90213-2

22 Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci* 1977;**74**:560–4. doi:10.1073/pnas.74.2.560

23 Sanger F, Nicklen S, Coulson AR. DNA

sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* 1977;**74**:5463-7. doi:10.1073/pnas.74.12.5463

24 Balch WE, Magrum LJ, Fox GE, *et al.* An ancient divergence among the bacteria. *J Mol Evol* 1977;**9**:305–11. doi:10.1007/BF01796092

25 Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci* 1977;**74**:5088–90. doi:10.1073/pnas.74.11.5088

26 Ragan MA, Bernard G, Chan CX. Molecular phylogenetics before sequences :Oligonucleotide catalogs as k-mer spectra. *RNA Biol* 2014;**11**:176–85. doi:10.4161/rna.27505

27 Bruijns B, Tiggelaar R, Gardeniers H. Massively parallel sequencing techniques for forensics: A review. *Electrophoresis* 2018;**39**:2642–54. doi:10.1002/elps.201800082

28 Clarridge JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* 2004;**17**:840–62. doi:10.1128/CMR.17.4.840-862.2004

29 Pace NR, Stahl DA, Lane DJ, *et al.* The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. ***Adv Microb ecol*** 1986;:1–55. doi:10.1007/978-1-4757-0611-6_1

30 Pace NR. A molecular view of microbial diversity and the biosphere. *Science* 1997;**276**:734–40. doi:10.1126/science.276.5313.734

31 Blattner FR, Plunkett G, Bloch CA, *et al.* The complete genome sequence of Escherichia coli K-12. *Science* 1997;**277**:1453–62. doi:10.1126/science.277.5331.1453

32 Fleischmann RD, Adams MD, White O, *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 1995;**269**:496–512. doi:10.1126/science.7542800

33 Bult CJ, White O, Olsen GJ, *et al.* Complete genome sequence of the Methanogenic archaeon, Methanococcus jannaschii. *Science* 1996;**273**:1058–73. doi:10.1126/science.273.5278.1058

34 Goffeau A, Barrell BG, Bussey H, *et al.* Life with 6000 Genes *Science (80-)* 1996;**274**:546–67.

35 Venter JC, Adams MD, Myers EW, *et al.* The Sequence of the Human Genome. *Science* 2001;**291**:1–49

36 Lander ES, Linton LM, Birren B, *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001;**412**:565–6. doi:10.1038/35087627

37 Molina-Menor E, Gimeno-Valero H, Pascual J, *et al.* High Culturable Bacterial Diversity From a European Desert: The Tabernas Desert. *Front Microbiol* 2021;**11**:1–15. doi:10.3389/fmicb.2020.583120

38 Victoria Wang X, Blades N, Ding J, *et al.* Estimation of sequencing error rates in short reads. *BMC Bioinformatics* 2012;**13**:1–12. doi:10.1186/1471-2105-13-185

39 Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45. doi:10.1038/nbt1486

40 Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods* 2008;**5**:16–8. doi:10.1038/nmeth1156

41 Wetterstrand KA. The Cost of Sequencing a Human Genome. 2020. https://bit.ly/3FgMhN2 (accessed 12 Oct 2021).

42 Lightbody G, Haberland V, Browne F, *et al.* Review of applications of high-throughput sequencing in personalized medicine: Barriers and facilitators of future progress in research and clinical application. *Brief Bioinform* 2019;**20**:1795–811. doi:10.1093/bib/bby051

43 Preston J, VanZeeland A, Peiffer DA. Innovation at Illumina: The road to the $600 human genome. 2021. https://go.nature.com/338oYrc (accessed 12 Oct 2021)

44 Hoy MA. DNA Sequencing and the Evolution of the "-Omics". In: *Insect Molecular Genetics*. 2013. 251–305. doi:10.1016/B978-0-12-415874-0.00007-X

45 Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51. doi:10.1038/nrg.2016.49

46 Reuter JA, Spacek D V., Snyder MP. High-Throughput Sequencing Technologies. *Mol Cell* 2015;**58**:586–97. doi:10.1016/j.molcel.2015.05.004

47 Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma* 2021;**3**:1–9. doi:10.1093/nargab/lqab019

48 Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* 2011;**12**:R112. doi:10.1186/gb-2011-12-11-r112

49 Nakamura K, Oshima T, Morimoto T, *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 2011;**39**. doi:10.1093/nar/gkr344

50 Tilak MK, Botero-Castro F, Galtier N, *et al.* Illumina Library Preparation for Sequencing the GC-Rich Fraction of Heterogeneous Genomic DNA. *Genome Biol Evol* 2018;**10**:616–22. doi:10.1093/gbe/evy022

51 Levene HJ, Korlach J, Turner SW, *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 2003;**299**:682–6. doi:10.1126/science.1079700

52 Buermans HPJ, den Dunnen JT. Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta - Mol Basis Dis* 2014;**1842**:1932–41. doi:10.1016/j.bbadis.2014.06.015

53 Hon T, Mars K, Young G, *et al.* Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* 2020;**7**:1–11. doi:10.1038/s41597-020-00743-4

54 Wenger AM, Peluso P, Rowell WJ, *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**:1155–62. doi:10.1038/s41587-019-0217-9

55 Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet* 2020;**21**:597–614. doi:10.1038/s41576-020-0236-x

56 Pfeiffer F, Gröber C, Blank M, *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep* 2018;**8**:1–14. doi:10.1038/s41598-018-29325-6

57 van Dijk EL, Jaszczyszyn Y, Naquin D, *et al.* The Third Revolution in Sequencing Technology. *Trends Genet* 2018;**34**:666–81. doi:10.1016/j.tig.2018.05.008

58 Tang AD, Soulette CM, van Baren MJ, *et al.* Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* 2020;**11**:1–12. doi:10.1038/s41467-020-15171-6

59 Dougherty ML, Underwood JG, Nelson BJ, *et al.* Transcriptional fates of human-specific segmental duplications in brain. *Genome Res* 2018;**28**:1566–76. doi:10.1101/gr.237610.118

60 Ishiura H, Shibata S, Yoshimura J, *et al.* Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat Genet* 2019;**51**:1222–32. doi:10.1038/s41588-019-0458-z

61 Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* 2020;**38**. doi:10.1038/s41587-020-0422-6

62 Tourancheau A, Mead EA, Zhang XS, *et al.* Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat Methods* 2021;**18**:491–8. doi:10.1038/s41592-021-01109-3

63 Grünberger F, Knüppel R, Jüttner M, *et al.* Nanopore-based native RNA sequencing provides insights into prokaryotic transcription, operon structures, rRNA maturation and modifications. *bioRxiv* Published Online First: 2020. doi:10.1101/2019.12.18.880849

64 Frank M, Prenzler A, Eils R, *et al.* Genome sequencing: A systematic review of health economic evidence. *Health Econ Rev* 2013;**3**:1–8. doi:10.1186/2191-1991-3-29

65 Whipps J, Lewis K, Cooke R. Mycoparasitism and plant disease control. *Fungi Biol Control Syst* 1988;:161–87.

66 Berg G, Rybakova D, Fischer D, *et al.* Microbiome definition re-visited: old concepts and new challenges. *Microbiome* 2020;**8**:1–22. doi:10.1186/s40168-020-00875-0

67 Whon TW, Shin NR, Kim JY, *et al.* Omics in gut microbiome analysis. *J Microbiol* 2021;**59**:292–7. doi:10.1007/s12275-021-1004-0

68 Porter TM, Hajibabaei M. Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. *Front Ecol Evol* 2020;**8**:1–15. doi:10.3389/fevo.2020.00248

69 Bartolo AG, Zammit G, Peters AF, *et al.* The current state of DNA barcoding of macroalgae in the Mediterranean Sea: Presently lacking but urgently required. *Bot Mar* 2020;**63**:253–72. doi:10.1515/bot-2019-0041

70 Johnson JS, Spakowicz DJ, Hong BY, *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 2019;**10**:1–11. doi:10.1038/s41467-019-13036-1

71 Winand R, Bogaerts B, Hoffman S, *et al.* Targeting the 16s rRNA gene for bacterial identification in complex mixed samples: Comparative evaluation of second (illumina) and third (oxford nanopore technologies) generation sequencing technologies. *Int J Mol Sci* 2020;**21**:1–22. doi:10.3390/ijms21010298

72 Earl JP, Adappa ND, Krol J, *et al.* Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. *Microbiome* 2018;**6**:190. doi:10.1186/s40168-018-0569-2

73 Cuscó A, Catozzi C, Viñes J, *et al.* Microbiota profiling with long amplicons using nanopore sequencing: Full-length 16s rRNA gene and whole rrn operon [version 1; referees: 2 approved, 3 approved with reservations]. *F1000Research* 2018;**7**:1–29. doi:10.12688/f1000research.16817.1

74 Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. *Microbiome* 2015;**3**:1–3. doi:10.1186/s40168-015-0094-5

75 Kim D, Song L, Breitwieser FP, *et al.* Centrifuge: rapid and accurate classificaton of metagenomic sequences, version *Genome Res* 2016;**26**:054965. doi:10.2202/gr.210641.116

76 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. doi:10.1093/bioinformatics/btv638

77 Hilton SK, Castro-Nallar E, Pérez-Losada M, *et al.* Metataxonomic and metagenomic approaches vs. culture-based techniques for clinical pathology. *Front Microbiol* 2016;**7**:1–12. doi:10.3389/fmicb.2016.00484

78 Petrova OE, Garcia-Alcalde F, Zampaloni C, *et al.* Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. *Sci Rep* 2017;**7**:1–15. doi:10.1038/srep41114

79 Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 2018;**20**:1125–39. doi:10.1093/bib/bbx120

80 Oxford Nanopore Technologies. Company history. 2021.https://nanoporetech.com/about-us/history (accessed 24 Oct 2021).

81 Kasianowicz JJ, Brandin E, Branton D, *et al.* Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci* 1996;**93**:13770–3. doi:10.1073/pnas.93.24.13770

82 Howorka S, Cheley S, Bayley H. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nat Biotechnol* 2001;**19**:636–9. doi:10.1038/90236

83 Van der Verren SE, Van Gerven N, Jonckheere W, *et al.* A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity. *Nat Biotechnol* 2020;**38**:1415–20. doi:10.1038/s41587-020-0570-8

84 Goyal P, Krasteva P V., Van Gerven N, *et al.* Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature* 2014;**516**:250–3. doi:10.1038/nature13768

85 Oxford Nanopore Technologies. R10.3: the newest nanopore for high accuracy nanopore sequencing. 2020.https://bit.ly/3GE6oqc (accessed 1 Nov 2021).

86 Chapman MR, Robinson LS, Pinkner JS, *et al.* Role of Escherichia coli Curli Operons in Directing Amyloid Fiber Formation. *Science* 2002;**295**:851–5. doi:10.1126/science.1067484

87 Vlaams Instituut voor Biotechnologie. Engineering protein nanopores to improve DNA sequencing. 2020.https://bit.ly/3BKQF5f (accessed 1 Nov 2020).

88 Oxford Nanopore Technologies. Flow cells and nanopores. 2021.https://bit.ly/3my9pjN (accessed 1 Nov 2021).

89 Amarasinghe SL, Su S, Dong X, *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;**21**:1–16. doi:10.1186/s13059-020-1935-5

90 Oxfrod Nanopore Technologies. How basecalling works. 2021.https://bit.ly/3EByBMs (accessed 1 Nov 2021).

91 Oxford Nanopore Technologies. MinION Mk1B IT requirements. 2021.https://bit.ly/3EDRhLs (accessed 1 Nov 2021).

92 Payne A, Holmes N, Rakyan V, *et al.* Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* 2019;**35**:2193–8. doi:10.1093/bioinformatics/bty841

93 Oxford Nanopore Technologies. At NCM, announcements include single-read accuracy of 99.1% on new chemistry and sequencing a record 10 Tb in a single PromethION run. 2020. https://bit.ly/3BA8SSO (accessed 1 Nov 2021).

94 Latorre-Pérez A, Pascual J, Porcar M, *et al.* A lab in the field: applications of real-time, in situ metagenomic sequencing. *Biol Methods Protoc* 2020;**5**:1–11.doi:10.1093/biomethods/bpaa016

95 Oxford Nanopore Technologies. SmidgION. 2021.https://bit.ly/3BD8O4J (accessed 2 Nov 2021).

96 Taxt AM, Avershina E, Frye SA, *et al.* Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing. *Sci Rep* 2020;**10**:1–11. doi:10.1038/s41598-020-64616-x

97 Jeck WR, Lee J, Robinson H, *et al.* A Nanopore Sequencing–Based Assay for Rapid Detection of Gene Fusions. *J Mol Diagnostics* 2019;**21**:58–69. doi:10.1016/j.jmoldx.2018.08.003

98 Oxford Nanopore Technologies. Oxford Nanopore Tech Update: new Duplex method for Q30 nanopore single molecule reads, PromethION 2, and more. 2021.https://bit.ly/3mBwO3W (accessed 2 Nov 2021).

99 Oxford Nanopore Technologies. Accuracy. 2021.https://bit.ly/3ECmGOG (accessed 20 Dec 2021).

100 Depledge DP, Srinivas KP, Sadaoka T, *et al.* Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun* 2019;**10**. doi:10.1038/s41467-019-08734-9

101 Xie S, Leung AW-S, Zheng Z, *et al.* Applications and potentials of nanopore sequencing in the (epi)genome and (epi) transcriptome era. *Innov* 2021;**2**:100153. doi:10.1016/j.xinn.2021.100153

102 Oxford Nanopore Technologies. Epigenetics and methylation analysis. 2021.https://bit.ly/3nSHsT9 (accessed 2 Nov 2021).

103 Kilianski A, Haas JL, Corriveau EJ, *et al.* Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *Gigascience* 2015;**4**. doi:10.1186/s13742-015-0051-z

104 Shin J, Lee S, Go MJ, *et al.* Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Sci Rep* 2016;**6**:1–10. doi:10.1038/srep29681

105 Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015;**12**:733–5. doi:10.1038/nmeth.3444

106 Myers EW, Sutton GG, Delcher AL, *et al.* A whole-genome assembly of Drosophila. *Science* 2000;**287**:2196–204. doi:10.1126/science.287.5461.2196

107 Benítez-Páez A, Sanz Y. Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinIONTM portable nanopore sequencer. *Gigascience* 2017;**6**:1–12. doi:10.1093/gigascience/gix043

108 Mitsuhashi S, Kryukov K, Nakagawa S, *et al.* A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Sci Rep* 2017;**7**:1–9. doi:10.1038/s41598-017-05772-5

109 Koren S, Walenz BP, Berlin K, *et al.* Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res* 2017;**27**:722–36. doi:10.1101/gr.215087.116

110 Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. doi:10.1093/bioinformatics/bty191

111 Sczyrba A, Hofmann P, Belmann P, *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat Methods* 2017;**14**:1063–71. doi:10.1038/nmeth.4458

112 Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics* 2016;**32**:1088–90. doi:10.1093/bioinformatics/btv697

113 McIntyre ABR, Ounit R, Afshinnekoo E, *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol* 2017;**18**:1–19. doi:10.1186/s13059-017-1299-7

114 Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci Rep* 2016;**6**:1–14. doi:10.1038/srep19233

115 McGovern E, Waters SM, Blackshields G, *et al.* Evaluating established methods for Rumen 16S rRNA amplicon sequencing with mock microbial populations. *Front Microbiol* 2018;**9**:1–14. doi:10.3389/fmicb.2018.01365

116 Matias Rodrigues JF, Schmidt TSB, Tackmann J, *et al.* MAPseq: Highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics* 2017;**33**:3808–10. doi:10.1093/bioinformatics/btx517

117 Almeida A, Mitchell AL, Tarkowska A, *et al.* Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *Gigascience* 2018;**7**:1–10. doi:10.1093/gigascience/giy054

118 Claesson MJ, Clooney AG, O'Toole PW. A clinician's guide to microbiome analysis. *Nat Rev Gastroenterol Hepatol* 2017;**14**:585–95. doi:10.1038/nrgastro.2017.97

119 Knight R, Vrbanac A, Taylor BC, *et al.* Best practices for analysing microbiomes. *Nat Rev Microbiol* 2018;**16**:410–22. doi:10.1038/s41579-018-0029-9

120 Beghini F, McIver LJ, Blanco-Míguez A, *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* 2021;**10**:1–29. doi:10.7554/eLife.65088

121 Quick J, Loman NJ, Duraffour S, *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* 2016;**530**:228–32. doi:10.1038/nature16996

122 Quick J, Ashton P, Calus S, *et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol* 2015;**16**:1–14. doi:10.1186/s13059-015-0677-2

123 Bull RA, Adikari TN, Ferguson JM, *et al.* Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun* 2020;**11**:1–8. doi:10.1038/s41467-020-20075-6

124 Peto L, Rodger G, Carter DP, *et al.* Diagnosis of SARS-CoV-2 infection with LamPORE, a high-throughput platform combining loop-mediated isothermal amplification and nanopore sequencing. *J Clin Microbiol* 2021;**59**. doi:10.1128/JCM.03271-20

125 Ashikawa S, Tarumoto N, Imai K, *et al.* Rapid identification of pathogens from positive blood culture bottles with the MinION nanopore sequencer. *J Med Microbiol* 2018;**67**:1589–95. doi:10.1099/jmm.0.000855

126 Phelan JE, O'Sullivan DM, Machado D, *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med* 2019;**11**:1–7. doi:10.1186/s13073-019-0650-x

127 Lemon JK, Khil PP, Frank KM, *et al.* Rapid nanopore sequencing of plasmids and resistance gene detection in clinical isolates. *J Clin Microbiol* 2017;**55**:3530–43. doi:10.1128/JCM.01069-17

128 Greninger AL, Naccache SN, Federman S, *et al.* Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med* 2015;**7**:1–13. doi:10.1186/s13073-015-0220-9

129 Batovska J, Lynch SE, Rodoni BC, *et al.* Metagenomic arbovirus detection using MinION nanopore sequencing. *J Virol Methods* 2017;**249**:79–84. doi:10.1016/j.jviromet.2017.08.019

130 Kafetzopoulou LE, Efthymiadis K, Lewandowski K, *et al.* Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Eurosurveillance* 2018;**23**:1–13. doi:10.2807/1560-7917.ES.2018.23.50.1800228

131 Tanaka H, Matsuo Y, Nakagawa S, *et al.* Real-time diagnostic analysis of MinION^{TM}-based metagenomic sequencing in clinical microbiology evaluation: a case report. *JA Clin Reports* 2019;**5**:5–6. doi:10.1186/s40981-019-0244-z

132 Pendleton KM, Erb-Downward JR, Bao Y, *et al.* Rapid pathogen identification in bacterial pneumonia using real-time metagenomics. *Am J Respir Crit Care Med* 2017;**196**:1610–2. doi:10.1164/rccm.201703-0537LE

133 Schmidt K, Mwaigwisya S, Crossman LC, *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J Antimicrob Chemother* 2017;**72**:104–14. doi:10.1093/jac/dkw397

134 Cheng J, Hu H, Kang Y, *et al.* Identification of pathogens in culture-negative infective endocarditis cases by metagenomic analysis. *Ann Clin Microbiol Antimicrob* 2018;**17**:1–11. doi:10.1186/s12941-018-0294-5

135 Grumaz C, Hoffmann A, Vainshtein Y, *et al.* Rapid Next-Generation Sequencing–Based Diagnostics of Bacteremia in Septic Patients. *J Mol Diagnostics* 2020;**22**:405–18. doi:10.1016/j.jmoldx.2019.12.006

136 Leggett RM, Alcon-Giner C, Heavens D, *et al.* Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. *Nat Microbiol* 2020;**5**:430–42. doi:10.1038/s41564-019-0626-z

137 Charalampous T, Kay GL, Richardson H, *et al.* Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat Biotechnol* 2019;**37**:783–92. doi:10.1038/s41587-019-0156-5

138 Mallapaty S. How sewage could reveal true scale of coronavirus outbreak. *Nature* 2020;**580**:176–7. doi:10.1038/d41586-020-00973-x

139 Hu YOO, Ndegwa N, Alneberg J, *et al.* Stationary and portable sequencing-based approaches for tracing wastewater contamination in urban stormwater systems. *Sci Rep* 2018;**8**:1–13. doi:10.1038/s41598-018-29920-7

140 Che Y, Xia Y, Liu L, *et al.* Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome* 2019;**7**:1–13. doi:10.1186/s40168-019-0663-0

141 Hu Y, Green GS, Milgate AW, *et al.* Pathogen detection and microbiome analysis of infected wheat using a portable DNA sequencer. *Phytobiomes J* 2019;**3**:92–101. doi:10.1094/PBIOMES-01-19-0004-R

142 Boykin L, Ghalab A, Marchi BR De, *et al.* Real time portable genome sequencing for global food security [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research* 2018;**7**:1–12. doi:10.12688/F1000RESEARCH.15507.1

143 Gonçalves AT, Collipal-Matamal R, Valenzuela-Muñoz V, *et al.* Nanopore sequencing of microbial communities reveals the potential role of sea lice as a reservoir for fish pathogens. *Sci Rep* 2020;**10**:1–12. doi:10.1038/s41598-020-59747-0

144 Voorhuijzen-Harink MM, Hagelaar R, van Dijk JP, *et al.* Toward on-site food authentication using nanopore sequencing. *Food Chem X* 2019;**2**:100035. doi:10.1016/j.fochx.2019.100035

145 Chakraborty D, Karthikeyan OP, Selvam A, *et al.* Two-phase anaerobic digestion of food waste: Effect of semi-continuous feeding on acidogenesis and methane production. *Bioresour Technol* 2021;:126396. doi:10.1016/j.biortech.2021.126396

146 Patel A, Mahboubi A, Horváth IS, *et al.* Volatile Fatty Acids (VFAs) Generated by Anaerobic Digestion Serve as Feedstock for Freshwater and Marine Oleaginous Microorganisms to Produce Biodiesel and Added-Value Compounds. *Front Microbiol* 2021;**12**:1–17. doi:10.3389/fmicb.2021.614612

147 Prajapati KK, Pareek N, Vivekanand V. Pretreatment and multi-feed anaerobic co-digestion of agro-industrial residual biomass for improved biomethanation and kinetic analysis. *Front Energy Res* 2018;**6**:1–18. doi:10.3389/fenrg.2018.00111

148 Pomerantz A, Peñafiel N, Arteaga A, *et al.* Real-time DNA barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* 2018;**7**:1–14. doi:10.1093/gigascience/giy033

149 Menegon M, Cantaloni C, Rodriguez-Prieto A, *et al.* On site DNA barcoding by nanopore sequencing. *PLoS One* 2017;**12**:1–18. doi:10.1371/journal.pone.0184741

150 Edwards A, Debbonaire A, Nicholls S, *et al.* In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. *bioRxiv* 2016;:073965. doi:10.1101/073965

151 Goordial J, Altshuler I, Hindson K, *et al.* In situ field sequencing and life detection in remote (79°26'N) Canadian high arctic permafrost ice wedge microbial communities. *Front Microbiol* 2017;**8**:1–14. doi:10.3389/fmicb.2017.02594

152 Johnson SS, Zaikova E, Goerlitz DS, *et al.* Real-time DNA sequencing in the antarctic dry valleys using the Oxford nanopore sequencer. *J Biomol Tech* 2017;**28**:2–7. doi:10.7171/jbt.17-2801-009

153 Gowers G-OF, Vince O, Charles J-H, *et al.* Entirely Off-Grid and Solar-Powered DNA Sequencing of Microbial Communities during an Ice Cap Traverse Expedition. *Genes (Basel)* 2019;**10**:902. doi:10.3390/genes10110902

154 Edwards A, Soares A, Rassner S, *et al.* Deep Sequencing: Intra-Terrestrial Metagenomics Illustrates The Potential Of Off-Grid Nanopore DNA Sequencing. *bioRxiv* Published Online First: 2017. doi:10.1101/133413

155 Castro-Wallace SL, Chiu CY, John KK, *et al.* Nanopore DNA Sequencing and Genome Assembly on the International Space Station. *Sci Rep* 2017;**7**:1–12. doi:10.1038/s41598-017-18364-0

156 Burton AS, Stahl SE, John KK, *et al.* Off earth identification of bacterial populations using 16S rDNA nanopore sequencing. *Genes (Basel)* 2020;**11**:1–10. doi:10.3390/genes11010076

157 Carr CE, Bryan NC, Saboda KN, *et al.* Nanopore Sequencing at Mars, Europa and Microgravity Conditions. *bioRxiv* 2020;:2020.01.09.899716. doi:10.1101/2020.01.09.899716

158 Bergmann I, Mundt K, Sontag M, *et al.* Influence of DNA isolation on Q-PCR-based quantification of methanogenic Archaea in biogas fermenters. *Syst Appl Microbiol* 2010;**33**:78–84. doi:10.1016/j.syapm.2009.11.004

159 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**:1–11. doi:10.1093/nar/gks808

160 Oxford nanopore Technologies. New basecaller now performs 'raw basecalling', for improved sequencing accuracy. 2017.https://bit.ly/3lHoH4P (accessed 8 Dec 2021).

161 Oxford Nanopore Technologies. Analysis solutions for nanopore sequencing data. 2021. https://bit.ly/3dy3BS1 (accessed 8 Dec 2021).

162 Caporaso JG, Kuczynski J, Stombaugh J, *et al.* QIIME allows analysis of high- throughput community sequencing data. *Nat Publ Gr* 2010;**7**:335–6. doi:10.1038/nmeth0510-335

163 De Coster W, D'Hert S, Schultz DT, *et al.* NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* 2018;**34**:2666–9. doi:10.1093/bioinformatics/bty149

164 Marijon P, Chikhi R, Varré J-S. yacrd and fpa: upstream tools for long-read genome assembly. *Bioinformatics* 2020;**36**:3894–6. doi:10.1093/bioinformatics/btaa262

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

166 Quast C, Pruesse E, Yilmaz P, *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:590–6. doi:10.1093/nar/gks1219

167 Santos A, van Aerle R, Barrientos L, *et al.* Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Comput Struct Biotechnol J* 2020;**18**:296–305. doi:10.1016/j.csbj.2020.01.005

168 Urban L, Holzer A, Baronas JJ, *et al.* Freshwater monitoring by nanopore sequencing. *Elife* 2021;**10**:1–27. doi:10.7554/eLife.61504

169 Gamaarachchi H, Parameswaran S, Smith MA. Featherweight long read alignment using partitioned reference indexes. *Sci Rep* 2019;**9**:1–12. doi:10.1038/s41598-019-40739-8

170 McMurdie PJ, Holmes S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* 2013;**8**. doi:10.1371/journal.pone.0061217

171 Andersen KS, Kirkegaard RH, Karst SM, *et al.* ampvis2: An R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv* 2018;:10–1. doi:10.1101/299537

172 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:1–21. doi:10.1186/s13059-014-0550-8

173 Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci* 2021;**1**:332–6.doi:10.1038/s43588-021-00073-4

174 Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 2020;**17**:155–8. doi:10.1038/s41592-019-0669-3

175 Shafin K, Pesout T, Lorig-Roach R, *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* 2020;**38**:1044–53. doi:10.1038/s41587-020-0503-6

176 Wick RR, Judd LM, Gorrie CL, *et al.* Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;**13**:1–22. doi:10.1371/journal.pcbi.1005595

177 Nurk S, Meleshko D, Korobeynikov A, *et al.* MetaSPAdes: A new versatile metagenomic assembler. *Genome Res* 2017;**27**:824–34. doi:10.1101/gr.213959.116

178 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. doi:10.1093/bioinformatics/btv033

179 Kamath GM, Shomorony I, Xia F, *et al.* HINGE: Long-read assembly achieves optimal repeat resolution. *Genome Res* 2017;**27**:747–56. doi:10.1101/gr.216465.116

180 Li H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 2016;**32**:2103–10. doi:10.1093/bioinformatics/btw152

181 Kolmogorov M, Bickhart DM, Behsaz B, *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;**17**:1103–10. doi:10.1038/s41592-020-00971-x

182 Gurevich A, Saveliev V, Vyahhi N, *et al.* QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**:1072–5. doi:10.1093/bioinformatics/btt086

183 Goldstein S, Beka L, Graf J, *et al.* Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics* 2019;**20**:23. doi:10.1186/s12864-018-5381-7

184 Blin K, Shaw S, Steinke K, *et al.* AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;**47**:W81–7. doi:10.1093/nar/gkz310

185 Vaser R, Sović I, Nagarajan N, *et al.* Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017;**27**:737–46. doi:10.1101/gr.214270.116

186 Nicholls SM, Quick JC, Tang S, *et al.* Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* 2019;**8**:1–9. doi:10.1093/gigascience/giz043

187 Abendroth C, Hahnke S, Simeonov C, *et al.* Microbial communities involved in biogas production exhibit high resilience to heat shocks. *Bioresour Technol* 2018;**249**:1074–9. doi:10.1016/j.biortech.2017.10.093

188 Abendroth C, Vilanova C, Günther T, *et al.* Eubacteria and archaea communities in seven mesophile anaerobic digester plants in Germany. *Biotechnol Biofuels* 2015;**8**:1–10. doi:10.1186/s13068-015-0271-6

189 Doloman A, Soboh Y, Walters AJ, *et al.* Qualitative Analysis of Microbial Dynamics during Anaerobic Digestion of Microalgal Biomass in a UASB Reactor. *Int J Microbiol* 2017;**2017**. doi:10.1155/2017/5291283

190 Yang Z, Koh SK, Ng WC, *et al.* Potential application of gasification to recycle food waste and rehabilitate acidic soil from secondary forests on degraded land in Southeast Asia. *J Environ Manage* 2016;**172**:40–8. doi:10.1016/j.jenvman.2016.02.020

191 Blades L, Morgan K, Douglas R, *et al.* Circular Biogas-Based Economy in a Rural Agricultural Setting. *Energy Procedia* 2017;**123**:89–96. doi:10.1016/j.egypro.2017.07.255

192 European Biogas Association. EBA Statistical Report 2017. 2017.https://bit.ly/3wLxYNM (accessed 14 Nov 2021).

193 European Biogas Association. Biogas sector launches 1st overview on European biogas and gasification technologies. 2021.https://bit.ly/30mF5Ra (accessed 14 Nov 2021).

194 Stams AJM, Plugge CM. Electron transfer in syntrophic communities of anaerobic bacteria and archaea. *Nat Rev Microbiol* 2009;**7**:568–77. doi:10.1038/nrmicro2166

195 Thauer RK, Kaster AK, Seedorf H, *et al.* Methanogenic archaea: Ecologically relevant differences in energy conservation. *Nat Rev Microbiol* 2008;**6**:579–91. doi:10.1038/nrmicro1931

196 Sundberg C, Al-Soud WA, Larsson M, *et al.* 454 Pyrosequencing Analyses of Bacterial and Archaeal Richness in 21 Full-Scale Biogas Digesters. *FEMS Microbiol Ecol* 2013;**85**:612–26. doi:10.1111/1574-6941.12148

197 Hanreich A, Schimpf U, Zakrzewski M, *et al.* Metagenome and metaproteome analyses of microbial communities in mesophilic biogas-producing anaerobic batch fermentations indicate concerted plant carbohydrate degradation. *Syst Appl Microbiol* 2013;**36**:330–8. doi:10.1016/j.syapm.2013.03.006

198 Banach A, Ciesielski S, Bacza T, *et al.* Microbial community composition and methanogens' biodiversity during a temperature shift in a methane fermentation chamber. *Environ Technol (United Kingdom)* 2019;**40**:3252–63. doi:10.1080/09593330.2018.1468490

199 Gaby JC, Zamanzadeh M, Horn SJ. The effect of temperature and retention time on methane production and microbial community composition in staged anaerobic digesters fed with food waste. *Biotechnol Biofuels* 2017;**10**:1–13. doi:10.1186/s13068-017-0989-4

200 De Vrieze J, Christiaens MER, Walraedt D, *et al.* Microbial community redundancy in anaerobic digestion drives process recovery after salinity exposure. *Water Res* 2017;**111**:109–17. doi:10.1016/j.watres.2016.12.042

201 Zhou J, Zhang R, Liu F, *et al.* Biogas production and microbial community shift through neutral pH control during the anaerobic digestion of pig manure. *Bioresour Technol* 2016;**217**:44–9. doi:10.1016/j.biortech.2016.02.077

202 Ciotola RJ, Martin JF, Tamkin A, *et al.* The influence of loading rate and variable temperatures on microbial communities in anaerobic digesters. *Energies* 2014;**7**:785–803. doi:10.3390/en7020785

203 Tzun-Wen Shaw G, Liu AC, Weng CY, *et al.* Inferring microbial interactions in thermophilic and mesophilic anaerobic digestion of HOG waste. *PLoS One* 2017;**12**:1–22. doi:10.1371/journal.pone.0181395

204 Abendroth C, Simeonov C, Peretó J, *et al.* From grass to gas: Microbiome dynamics of grass biomass acidification under mesophilic and thermophilic temperatures. *Biotechnol Biofuels* 2017;**10**:1–12. doi:10.1186/s13068-017-0859-0

205 Bonmatí A, Flotats X. Air stripping of ammonia from pig slurry: Characterisation and feasibility as a pre- or post-treatment to mesophilic anaerobic digestion. *Waste Manag* 2003;**23**:261–72. doi:10.1016/S0956-053X(02)00144-7

206 Yao Y, Yu L, Ghogare R, *et al.* Simultaneous ammonia stripping and anaerobic digestion for efficient thermophilic conversion of dairy manure at high solids concentration. *Energy* 2017;**141**:179–88. doi:10.1016/j.energy.2017.09.086

207 Huang H, Liu J, Ding L. Recovery of phosphate and ammonia nitrogen from the anaerobic digestion supernatant of activated sludge by chemical precipitation. *J Clean Prod* 2015;**102**:437–46. doi:10.1016/j.jclepro.2015.04.117

208 Sürmeli R, Bayrakdar A, Çalli B. Removal and recovery of ammonia from chicken manure. *Water Sci Technol* 2017;**75**:2811–7. doi:10.2166/wst.2017.116

209 Abendroth C, Wünsche E, Luschnig O, *et al.* Producing high-strength liquor from mesophilic batch acidification of chicken manure. *Waste Manag Res* 2015;**33**:291–4. doi:10.1177/0734242X14568536

210 Liao Q, Chang J, Hermann C, *et al. Bioreactors for Microbial Biomass and Energy Conversion*. Singapore: Springer Singapore 2018.

211 Klocke M, Nettmann E, Bergmann I, *et al.* Characterization of the methanogenic Archaea within two-phase biogas reactor systems operated with plant biomass. *Syst Appl Microbiol* 2008;**31**:190–205. doi:10.1016/j.syapm.2008.02.003

212 Zhao H, Li J, Li J, *et al.* Organic loading rate shock impact on operation and microbial communities in different anaerobic fixed-bed reactors. *Bioresour Technol* 2013;**140**:211–9. doi:10.1016/j.biortech.2013.04.027

213 Rajesh Banu J, Sugitha S, Kannah RY, *et al.* Marsilea spp.—A novel source of lignocellulosic biomass: Effect of solubilized lignin on anaerobic biodegradability and cost of energy products. *Bioresour Technol* 2018;**255**:220–8. doi:10.1016/j.biortech.2018.01.103

214 Guštin S, Marinšek-Logar R. Effect of pH, temperature and air flow rate on the continuous ammonia stripping of the anaerobic digestion effluent. *Process Saf Environ Prot* 2011;**89**:61–6. doi:10.1016/j.psep.2010.11.001

215 Kataki S, West H, Clarke M, *et al.* Phosphorus recovery as struvite: Recent concerns for use of seed, alternative Mg source, nitrogen conservation and fertilizer potential. *Resour Conserv Recycl* 2016;**107**:142–56. doi:10.1016/j.resconrec.2015.12.009

216 Herrmann C, Idler C, Heiermann M. Improving aerobic stability and biogas production of maize silage using silage additives. *Bioresour Technol* 2015;**197**:393–403. doi:10.1016/j.biortech.2015.08.114

217 Elbeshbishy E, Dhar BR, Nakhla G, *et al.* A critical review on inhibition of dark biohydrogen fermentation. *Renew Sustain Energy Rev* 2017;**79**:656–68. doi:10.1016/j.rser.2017.05.075

218 Niu Q, Hojo T, Qiao W, *et al.* Characterization of methanogenesis, acidogenesis and hydrolysis in thermophilic methane fermentation of chicken manure. *Chem Eng J* 2014;**244**:587–96. doi:10.1016/j.cej.2013.11.074

219 Abendroth C. *Paving the crossroad of biorefinery*. 2017. PhD Thesis. University of Valencia. Valencia.

220 Solli L, Håvelsrud OE, Horn SJ, *et al.* A metagenomic study of the microbial communities in four parallel biogas reactors. *Biotechnol Biofuels* 2014;**7**:1–15. doi:10.1186/s13068-014-0146-2

221 Wirth R, Pap B, Dudits D, *et al.* Genome-centric investigation of anaerobic digestion using sustainable second and third generation substrates. *J Biotechnol* 2021;**339**:53–64. doi:10.1016/j.jbiotec.2021.08.002

222 Singh A, Müller B, Schnürer A. Profiling temporal dynamics of acetogenic communities in anaerobic digesters using next-generation sequencing and T-RFLP. *Sci Rep* 2021;**11**:1–14. doi:10.1038/s41598-021-92658-2

223 Shang Y, Kumar S, Oakley B, *et al.* Chicken gut microbiota: Importance and detection technology. *Front Vet Sci* 2018;**5**. doi:10.3389/fvets.2018.00254

224 Carrasco JMD, Casanova NA, Miyakawa MEF. Microbiota, gut health and chicken productivity: What is the connection? *Microorganisms* 2019;**7**:1–15. doi:10.3390/microorganisms7100374

225 Calusinska M, Goux X, Fossépré M, *et al.* A year of monitoring 20 mesophilic full-scale bioreactors reveals the existence of stable but different core microbiomes in bio-waste and wastewater anaerobic digestion systems. *Biotechnol Biofuels* 2018;**11**:1–19. doi:10.1186/s13068-018-1195-8

226 Hassa J, Klang J, Benndorf D, *et al.* Indicative marker microbiome structures deduced from the taxonomic inventory of 67 full-scale anaerobic digesters of 49 agricultural biogas plants. *Microorganisms* 2021;**9**:1–18. doi:10.3390/microorganisms9071457

227 Hahnke S, Langer T, Koeck DE, *et al.* Description of Proteiniphilum saccharofermentans sp. nov., Petrimonas mucosa sp. nov. and Fermentimonas caenicola gen. nov., sp. nov., isolated from mesophilic laboratory-scale biogas reactors, and emended description of the genus Proteiniphilum. *Int J Syst Evol Microbiol* 2016;**66**:1466–75. doi:10.1099/ijsem.0.000902

228 Heitkamp K, Latorre-Pérez A, Nefigmann S, *et al.* Monitoring of seven industrial anaerobic digesters supplied with biochar. *Biotechnol Biofuels* 2021;**14**:1–14. doi:10.1186/s13068-021-02034-5

229 Kollarcikova M, Faldynova M, Matiasovicova J, *et al.* Different bacteroides species colonise human and chicken intestinal tract. *Microorganisms* 2020;**8**:1–14. doi:10.3390/microorganisms8101483

230 Ao T, Chen L, Chen Y, *et al.* The screening of early warning indicators and microbial community of chicken manure thermophilic digestion at high organic loading rate. *Energy* 2021;**224**:120201. doi:10.1016/j.energy.2021.120201

231 Fischer MA, Ulbricht A, Neulinger SC, *et al.* Immediate Effects of Ammonia Shock on Transcription and Composition of a Biogas Reactor Microbiome. *Front Microbiol* 2019;**10**. doi:10.3389/fmicb.2019.02064

232 Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A* 2016;**113**:5970–5. doi:10.1073/pnas.1521291113

233 Bull AT, Goodfellow M. Dark, rare and inspirational microbial matter in the extremobiosphere: 16 000 m of bioprospecting campaigns. *Microbiol (United Kingdom)* 2019;**165**:1252–64. doi:10.1099/mic.0.000822

234 Lauber CL, Hamady M, Knight R, *et al.* Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 2009;**75**:5111–20. doi:10.1128/AEM.00335-09

235 DiGiulio DB, Callahan BJ, McMurdie PJ, *et al.* Temporal and spatial variation of the human microbiota during pregnancy. *Proc Natl Acad Sci* 2015;**112**:11060–5. doi:10.1073/pnas.1502875112

236 Bhattacharjee A, Velickovic D, Wietsma TW, *et al.* Visualizing Microbial Community Dynamics via a Controllable Soil Environment. *mSystems* 2020;**5**:1–10. doi:10.1128/msystems.00645-19

237 Palatnick A, Zhou B, Ghedin E, *et al.* IGenomics: Comprehensive DNA sequence analysis on your Smartphone. *Gigascience* 2021;**9**:1–12. doi:10.1093/gigascience/giaa138

238 Chan WS, Au CH, Lam HY, *et al.* Evaluation on the use of Nanopore sequencing for direct characterization of coronaviruses from respiratory specimens, and a study on emerging missense mutations in partial RdRP gene of SARS-CoV-2. *Virol J* 2020;**17**:1–13. doi:10.1186/s12985-020-01454-3

239 Tytgat O, Gansemans Y, Weymaere J, *et al.* Nanopore sequencing of a forensic STR multiplex reveals loci suitable for single-contributor STR profiling. *Genes (Basel)* 2020;**11**. doi:10.3390/genes11040381

240 Vasiljevic N, Lim M, Humble E, *et al.* Developmental validation of Oxford Nanopore Technology MinION sequence data and the NGSpeciesID bioinformatic pipeline for forensic genetic species identification. *Forensic Sci Int Genet* 2021;**53**:102493. doi:10.1016/j.fsigen.2021.102493

241 McHugh AJ, Yap M, Crispie F, *et al.* Microbiome-based environmental monitoring of a dairy processing facility highlights the challenges associated with low microbial-load samples. *npj Sci Food* 2021;**5**:1–13. doi:10.1038/s41538-021-00087-2

242 Hardegen J, Latorre-Pérez A, Vilanova C, *et al.* Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass. *Bioresour Technol* 2018;**265**. doi:10.1016/j.biortech.2018.06.005

243 Edwards U, Rogall T, Blöcker H, *et al.* Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res* 1989;**17**:7843–53. doi:10.1093/nar/17.19.7843

244 Stackebrandt E, Liesack W. Nucleic acids and classification. In: Goodfellow M, O'Donnell AG, eds. *Handbook of new bacterial systematics*. London: : London Academic Press 1993. 152–89.

245 Okonechnikov K, Golosova O, Fursov M, *et al.* Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* 2012;**28**:1166–7. doi:10.1093/bioinformatics/bts091

246 Montero-Calasanz M del C, Göker M, Broughton WJ, *et al.* Geodermatophilus tzadiensis sp. nov., a UV radiation-resistant bacterium isolated from sand of the Saharan desert. *Syst Appl Microbiol* 2013;**36**:177–82. doi:10.1016/j.syapm.2012.12.005

247 Yu LZH, Luo XS, Liu M, *et al.* Diversity of ionizing radiation-resistant bacteria obtained from the Taklimakan Desert. *J Basic Microbiol* 2015;**55**:135–40. doi:10.1002/jobm.201300390

248 Deng W, Yang Y, Gao P, *et al.* Radiation-Resistant Micrococcus luteus SC1204 and Its Proteomics Change Upon Gamma Irradiation. *Curr Microbiol* 2016;**72**:767–75. doi:10.1007/s00284-016-1015-y

249 Etemadifar Z, Gholami M, Derikvand P. UV-Resistant Bacteria with Multiple-Stress Tolerance Isolated from Desert Areas in Iran. *Geomicrobiol J* 2016;**33**:1–7. doi:10.1080/01490451.2015.1063025

250 Paulino-Lima IG, Fujishima K, Navarrete JU, *et al.* Extremely high UV-C radiation resistant microorganisms from desert environments with different manganese concentrations. *J Photochem Photobiol B Biol* 2016;**163**:327–36. doi:10.1016/j.jphotobiol.2016.08.017

251 Golinska P, Montero-Calasanz M del C, Świecimska M, *et al.* Modestobacter excelsi sp. nov., a novel actinobacterium isolated from a high altitude Atacama Desert soil. *Syst Appl Microbiol* 2020;**43**:0–9. doi:10.1016/j.syapm.2019.126051

252 Tanner K. Life under the sun: microbial ecology and applications of the solar panel microbiota. 2020. PhD thesis. University of Valencia. Valencia.

253 Chun J, Oren A, Ventosa A, *et al.* Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* 2018;**68**:461–6. doi:10.1099/ijsem.0.002516

254 Ciuffreda L, Rodríguez-Pérez H, Flores C. Nanopore sequencing and its application to the study of microbial communities. *Comput Struct Biotechnol J* 2021;**19**:1497–511. doi:10.1016/j.csbj.2021.02.020

255 Nygaard AB, Tunsjø HS, Meisal R, *et al.* A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. *Sci Rep* 2020;**10**:3209. doi:10.1038/s41598-020-59771-0

256 Edwards A, Debbonaire AR, Nicholls SM, *et al.* In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. *bioRxiv* 2019;:073965. doi:10.1101/073965

257 Williams L, Loewen-Schneider K, Maier S, *et al.* Cyanobacterial diversity of western European biological soil crusts along a latitudinal gradient. *FEMS Microbiol Ecol* 2016;**92**:1–9. doi:10.1093/femsec/fiw157

258 Machado-de-Lima NM, Fernandes VMC, Roush D, *et al.* The Compositionally Distinct Cyanobacterial Biocrusts From Brazilian Savanna and Their Environmental Drivers of Community Diversity. *Front Microbiol* 2019;**10**:1–10. doi:10.3389/fmicb.2019.02798

259 Buckley DH, Huangyutitham V, Nelson TA, *et al.* Diversity of Planctomycetes in soil in relation to soil history and environmental heterogeneity. *Appl Environ Microbiol* 2006;**72**:4522–31. doi:10.1128/AEM.00149-06

260 Spain AM, Krumholz LR, Elshahed MS. Abundance, composition, diversity and novelty of soil Proteobacteria. *ISME J* 2009;**3**:992–1000. doi:10.1038/ismej.2009.43

261 Bergmann GT, Bates ST, Eilers KG, *et al.* The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol Biochem* 2011;**43**:1450–5. doi:10.1016/j.soilbio.2011.03.012

262 DeBruyn JM, Nixon LT, Fawaz MN, *et al.* Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil. *Appl Environ Microbiol* 2011;**77**:6295–300. doi:10.1128/AEM.05005-11

263 Zhang B, Wu X, Tai X, *et al.* Variation in Actinobacterial Community Composition and Potential Function in Different Soil Ecosystems Belonging to the Arid Heihe River Basin of Northwest China. *Front Microbiol* 2019;**10**:1–11. doi:10.3389/fmicb.2019.02209

264 Kalam S, Basu A, Ahmad I, *et al.* Recent Understanding of Soil Acidobacteria and Their Ecological Significance: A Critical Review. *Front Microbiol* 2020;**11**. doi:10.3389/fmicb.2020.580024

265 Larsbrink J, McKee LS. Bacteroidetes bacteria in the soil: Glycan acquisition, enzyme secretion, and gliding motility. 2020. 63–98. doi:10.1016/bs.aambs.2019.11.001

266 Gabani P, Singh O V. Radiation-resistant extremophiles and their potential in biotechnology and therapeutics. *Appl Microbiol Biotechnol* 2013;**97**:993–1004. doi:10.1007/s00253-012-4642-7

267 Holmes AJ, Bowyer J, Holley MP, *et al.* Diverse, yet-to-be-cultured members of the Rubrobacter subdivision of the Actinobacteria are widespread in Australian arid soils. *FEMS Microbiol Ecol* 2000;**33**:111–20. doi:10.1016/S0168-6496(00)00051-9

268 Rainey FA, Ray K, Ferreira M, *et al.* Extensive diversity of ionizing-radiation-resistant bacteria recovered from Sonoran Desert soil and description of nine new species of the genus Deinococcus obtained from a single soil sample. *Appl Environ Microbiol* 2005;**71**:5225–35. doi:10.1128/AEM.71.9.5225-5235.2005

269 Zhang Q, Liu C, Tang Y, *et al.* Hymenobacter xinjiangensis sp. nov., a radiation-resistant bacterium isolated from the desert of Xinjiang, China. *Int J Syst Evol Microbiol* 2007;**57**:1752–6. doi:10.1099/ijs.0.65033-0

270 Abed RMM, Al Kharusi S, Schramm A, *et al.* Bacterial diversity, pigments and nitrogen fixation of biological desert crusts from the Sultanate of Oman. *FEMS Microbiol Ecol* 2010;**72**:418–28. doi:10.1111/j.1574-6941.2010.00854.x

271 Amin A, Ahmed I, Habib N, *et al.* Microvirga pakistanensis sp. nov., a novel bacterium isolated from desert soil of Cholistan, Pakistan. *Arch Microbiol* 2016;**198**:933–9. doi:10.1007/s00203-016-1251-3

272 Hu QW, Chu X, Xiao M, *et al.* Arthrobacter deserti sp. Nov., isolated from a desert soil sample. *Int J Syst Evol Microbiol* 2016;**66**:2035–40. doi:10.1099/ijsem.0.000986

273 Wübbeler JH, Oppermann-Sanio FB, Ockenfels A, *et al.* Sphingomonas jeddahensis sp. nov., isolated from Saudi Arabian desert soil. *Int J Syst Evol Microbiol* 2017;**67**:4057–63. doi:10.1099/ijsem.0.002249

274 Liang Y, Tang K, Wang Y, *et al.* Hymenobacter crusticola sp. nov., isolated from biological soil crust. *Int J Syst Evol Microbiol* 2019;**69**:547–51. doi:10.1099/ijsem.0.003196

275 da Rocha UN, Cadillo-Quiroz H, Karaoz U, *et al.* Isolation of a significant fraction of non-phototroph diversity from a desert biological soil crust. *Front Microbiol* 2015;**6**:1–14. doi:10.3389/fmicb.2015.00277

276 Heikema AP, Horst-Kreft D, Boers SA, *et al.* Comparison of illumina versus nanopore 16s rRNA gene sequencing of the human nasal microbiota. *Genes (Basel)* 2020;**11**:1–17. doi:10.3390/genes11091105

277 Matsuo Y, Komiya S, Yasumizu Y, *et al.* Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. *BMC Microbiol* 2021;**21**:35. doi:10.1186/s12866-021-02094-5

278 Curry KD, Wang Q, Nute MG, *et al.* Emu: Species-Level Microbial Community Profiling for Full-Length Nanopore 16S Reads. *bioRxiv* 2021;:1–14.https://www.biorxiv.org/content/early/2021/05/03/2021.05.02.442339

279 Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics* 2021;**37**:1600–1. doi:10.1093/bioinformatics/btaa900

280 Xu Z, Mai Y, Liu D, *et al.* Fast-Bonito : A Faster Basecaller for Nanopore Sequencing. *bioRxiv* 2020.

281 Steen AD, Crits-Christoph A, Carini P, *et al.* High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J* 2019;**13**:3126–30. doi:10.1038/s41396-019-0484-y

282 Dilthey AT, Jain C, Koren S, *et al.* Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat Commun* 2019;**10**. doi:10.1038/s41467-019-10934-2

283 Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol* 2019;**37**:124–6. doi:10.1038/s41587-018-0004-z

284 Archer SDJ, Lee KC, Caruso T, *et al.* Airborne microbial transport limitation to isolated Antarctic soil habitats. *Nat Microbiol* 2019;**4**:925–32.doi:10.1038/s41564-019-0370-4

285 Sielaff AC, Urbaniak C, Babu G, *et al.* Characterization of the total and viable bacterial and fungal communities associated with the International Space Station surfaces. *Microbiome* 2019;**7**:50.

286 Tanner K, Molina-Menor E, Latorre-Pérez A, *et al.* Extremophilic microbial communities on photovoltaic panel surfaces: a two-year study. *Microb Biotechnol* 2020;:1751-7915.13620. doi:10.1111/1751-7915.13620

287 Hug LA, Baker BJ, Anantharaman K, *et al.* A new view of the tree of life. *Nat Microbiol* 2016;**1**:1–6. doi:10.1038/nmicrobiol.2016.48

288 Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* 2018;**5**:1–8. doi:10.1038/sdata.2017.203

289 Nayfach S, Shi ZJ, Seshadri R, *et al.* New insights from uncultivated genomes of the global human gut microbiome. *Nature* 2019;**568**:505–10. doi:10.1038/s41586-019-1058-x

290 Fettweis JM, Serrano MG, Brooks JP, *et al.* The vaginal microbiome and preterm birth. *Nat Med* 2019;**25**:1012–21. doi:10.1038/s41591-019-0450-2

291 Olson ND, Treangen TJ, Hill CM, *et al.* Metagenomic assembly through the lens of validation: Recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief Bioinform* 2018;**20**:1140–50. doi:10.1093/bib/bbx098

292 Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. *Brief Bioinform* 2020;**21**:584–94. doi:10.1093/bib/bbz020

293 Jayakumar V, Sakakibara Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief Bioinform* 2017;**20**:866–76. doi:10.1093/bib/bbx147

294 Wick RR, Judd LM, Gorrie CL, *et al.* Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genomics* 2017;**3**:0–6. doi:10.1099/mgen.0.000132

295 Deschamps S, Zhang Y, Llaca V, *et al.* A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat Commun* 2018;**9**. doi:10.1038/s41467-018-07271-1

296 Jain M, Koren S, Miga KH, *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 2018;**36**:338–45. doi:10.1038/nbt.4060

297 Bokulich NA, Rideout JR, Mercurio WG, *et al.* mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking. *mSystems* 2016;**1**. doi:10.1128/msystems.00062-16

298 Fritz A, Hofmann P, Majda S, *et al.* CAMISIM: Simulating metagenomes and microbial communities. *Microbiome* 2019;**7**:1–12. doi:10.1186/s40168-019-0633-6

299 Vollmers J, Wiegand S, Kaster AK. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! *Plos One* 2017. doi:10.1371/journal.pone.0169662

300 Hu Y, Fang L, Nicholson C, *et al.* Implications of Error-Prone Long-Read Whole-Genome Shotgun Sequencing on Characterizing Reference Microbiomes. *iScience* 2020;**23**. doi:10.1016/j.isci.2020.101223

301 Sevim V, Lee J, Egan R, *et al.* Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Sci data* 2019;**6**:285. doi:10.1038/s41597-019-0287-z

302 Parajuli P, Deimel LP, Verma NK, *et al.* Genome Analysis of Shigella flexneri Serotype 3b Strain SFL1520 Reveals Significant Horizontal Gene Acquisitions Including a Multidrug Resistance Cassette. *Genome Biol Evol* 2019;**11**:776–85. doi:10.1093/gbe/evz026

303 Dhar R, Seethy A, Pethusamy K, *et al.* De novo assembly of the Indian blue peacock (Pavo cristatus) genome using Oxford Nanopore technology and Illumina sequencing. *Gigascience* 2019;**8**:1–13. doi:10.1093/gigascience/giz038

304 Leidenfrost RM, Pöther DC, Jäckel U, *et al.* Benchmarking the MinION: Evaluating long reads for microbial profiling. *Sci Rep* 2020;**10**:1–11. doi:10.1038/s41598-020-61989-x

305 Hamner S, Brown BL, Hasan NA, *et al.* Metagenomic profiling of microbial pathogens in the little bighorn river, Montana. *Int J Environ Res Public Health* 2019;**16**. doi:10.3390/ijerph16071097

306 Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* Published Online First: 2020. doi:10.1038/s41587-020-0422-6

307 Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research* 2019;**8**:1–22. doi:10.12688/f1000research.21782.1

308 Pedron R, Esposito A, Bianconi I, *et al.* Genomic and metagenomic insights into the microbial community of a thermal spring. *Microbiome* 2019;**7**:1–13. doi:10.1186/s40168-019-0625-6

309 Sánchez-Andrea I, Rodríguez N, Amils R, *et al.* Microbial diversity in anaerobic sediments at Río Tinto, a naturally acidic environment with a high heavy metal content. *Appl Environ Microbiol* 2011;**77**:6085–93. doi:10.1128/AEM.00654-11

310 Martinez RM, Bowen TR, Foltzer MA. Prosthetic Device Infections. *Microbiol Spectr* 2016;**4**:197–210. doi:10.1128/microbiolspec.DMIH2-0004-2015

311 Satari L, Guillén A, Vidal-Verdú À, *et al.* The wasted chewing gum bacteriome. *Sci Rep* 2020;**10**:1–10. doi:10.1038/s41598-020-73913-4

312 Dorado-Morales P, Vilanova C, Peretó J, *et al.* A highly diverse, desert-like microbial biocenosis on solar panels in a Mediterranean city. *Sci Rep* 2016;**6**:1–9. doi:10.1038/srep29235

313 Hashmi MZ, Strezov V, Varma A. *A*ntibiotics and Antibiotics Resistance Genes in Soils: Monitoring, Toxicity, Risk Assessment and Management. ***Soil Biology*** 2017.

314 Zeldes BM, Keller MW, Loder AJ, *et al.* Extremely thermophilic microorganisms as metabolic engineering platforms for production of fuels and industrial chemicals. *Front Microbiol* 2015;**6**:1–17. doi:10.3389/fmicb.2015.01209

315 Vilanova C, Porcar M. Are multi-omics enough? *Nat Microbiol* 2016;**1**:1–2. doi:10.1038/nmicrobiol.2016.101

316 Kolmogorov M, Yuan J, Lin Y, *et al.* Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;**37**:540–6. doi:10.1038/s41587-019-0072-8

317 Cuscó A, Pérez D, Viñes J, *et al.* Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces. *BMC Genomics* 2021;**22**:1–15. doi:10.1186/s12864-021-07607-0

318 Wick RR, Judd LM, Cerdeira LT, *et al.* Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol* 2021;**22**:1–17. doi:10.1186/s13059-021-02483-z

319 Huang YT, Liu PY, Shih PW. Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. *Genome Biol* 2021;**22**:1–17. doi:10.1186/s13059-021-02282-6

320 Wick RR, Judd LM, Gorrie CL, *et al.* Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genomics* 2017;**3**:1–7. doi:10.1099/mgen.0.000132

321 Bertrand D, Shaw J, Kalathiyappan M, *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol* 2019;**37**:937–44. doi:10.1038/s41587-019-0191-2

322 Baker DJ, Aydin A, Le-Viet T, *et al.* CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes. *Genome Med* 2021;**13**:1–

11. doi:10.1186/s13073-021-00839-5

323 Onwuamah CK, Kanteh A, Abimbola BS, *et al.* SARS-CoV-2 sequencing collaboration in west Africa shows best practices. *Lancet Glob Heal* 2021;**9**:e1499–500. doi:10.1016/S2214-109X(21)00389-2

324 Butera Y, Mukantwari E, Artesi M, *et al.* Genomic sequencing of SARS-CoV-2 in Rwanda reveals the importance of incoming travelers on lineage diversity. *Nat Commun* 2021;**12**:1–11. doi:10.1038/s41467-021-25985-7

325 Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. *Nat Methods* 2016;**13**:751–4. doi:10.1038/nmeth.3930

326 Oxford Nanopore Technologies. Q20+ Chemistry for single molecule accuracy of 99% and higher. 2021.https://bit.ly/3sw9VCi (accessed 22 Dec 2021).

327 Kovaka S, Fan Y, Ni B, *et al.* Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol* 2021;**39**:431–41. doi:10.1038/s41587-020-0731-9

328 Payne A, Holmes N, Clarke T, *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol* 2021;**39**:442–50. doi:10.1038/s41587-020-00746-x

329 Gan M, Wu B, Yan G, *et al.* Combined nanopore adaptive sequencing and enzyme-based host depletion efficiently enriched microbial sequences and identified missing respiratory pathogens. *BMC Genomics* 2021;**22**:1–11. doi:10.1186/s12864-021-08023-0

330 Viehweger A, Marquet M, Hölzer M, *et al.* Adaptive Nanopore Sequencing on Miniature Flow Cell detects extensive antimicrobial. *bioRxiv* 2021.

331 Brown C. NCM update 2021. 2021 https://bit.ly/3mIhkuI (accessed 20 Dec 2021).

Appendix A

Consolidating and revisiting: extended evaluation of assembly methods for Nanopore-based metagenomics

Adriel Latorre-Pérez[1*], Morgane Blanot[1*], Manuel Porcar[1,2] and Cristina Vilanova[1]

[1]Darwin Bioprospecting Excellence S.L., Calle Catedrático Agustín Escardino 9, Paterna, 46980, Spain.

[2]Institute for Integrative Systems Biology I2SysBio (University of València-CSIC), Catedrático José Beltrán 2, Paterna, 46980, Spain.

*Equal contributions

[Preliminary and unpublished data]

Background

Mock communities are artificial and defined microbial communities in which the abundance of all the microorganisms is known. Sequencing data generated from mock communities is central for benchmarking metagenomic tools, since the results obtained after data analysis can be compared to the theoretical composition of the community to evaluate the performance of each bioinformatic pipeline. In the work reported in **Chapter II** (**Publication VI in Appendix C**) we analyzed two different mock communities, the ZymoBIOMICS Microbial Community Standards CS Even and the CSII Log (Zymo Research, United States, Cat. No.: ZRC190633 and ZRC190842), which were sequenced by Nicholls et al. [1] using GridION and Promethion (Oxford Nanopore Technologies -ONT-). In parallel to our study, other research groups published Nanopore data produced from metagenomic sequencing of other mock communities. Therefore, the main goal of this work was to assess the performance of several metagenomic assemblers on this data in order to consolidate and revisit the conclusions of **Chapter II**.

Brief experimental procedures

Four different datasets were selected and downloaded:

- **BenchEV (Leidenfrost et al. [2])**. Mock community comprising 12 type strains of gram-positive and gram-negative bacteria with varying GC. Homogeneous distribution of microorganisms.
- **BenchHE (Leidenfrost et al. [2]).** Same as BenchEV. Logarithmic distribution of the microorganisms.
- **BMock12 (Sevim et al. [3]).** Mixture of 12 bacterial strains, whose genomes are particularly challenging to reconstruct due to GC content, genome size and/or repeats. Heterogeneous distribution (the abundance of microorganisms ranged from 1.6% to 16.2%).
- **MSA2006 (Moss et al. [4]).** Mixture of 12 bacterial strains from the human gut microbiome. Homogeneous distribution.

All these mock communities were sequenced on MinION (ONT) using R.9.4.1 flow cells (ONT, UK, Cat. No.: FLO-MIN106D). Reference genomes were available for all the microorganisms included in each community.

Only the assembly tools that showed a good performance on **Chapter II** (Canu [5], Raven [6], Flye -or metaFlye- [7] and Pomoxis) were originally considered for evaluation. Necat [8], which was released after the publication of the first benchmark, was also included. Finally, RedBean [9] was assessed due to its high computational efficiency. Tools were used as indicated in Table 1. All the analyses were run on the same desktop computer (CPU: AMD RYZEN 7 1700X 3.4GHZ; Cores: 8; Threads: 16; RAM: Corsair Vengeance 64 GB; SSD: Samsung 860 EVO Basic SSD 500GB; HDD: x2 Toshiba Canvio Basics 2Tb; Operating System: Ubuntu 18.04).

Table 1. Commands used for running each assembler. $filepath: path to input reads (FASTQ); $size: estimated metagenome size.

Assembler (version)	Command
Flye (v. 2.7)	flye –nano-raw $filepath –out-dir $results –genome-size $size –threads 16 –meta –plasmids
Canu (v. 2.0)	canu -p assembly -d $results genomeSize=$size corOutCoverage=10000 corMhapSensitivity=high corMinCoverage=0 redMemory=32 oeaMemory=32 batThreads=16 batMemory=60 -nanopore $filepath
Pomoxis (v. 0.3.2)	mini_assemble -i $filepath -o $results -p assembly -l $size -t 16
Raven (v. 1.1.5)	raven –threads 16 $filepath > assembly.fa
Redbean (v. 2.5)	wtdbg2 -x ont -t 16 -g $size -i $filepath -fo step1 wtpoa-cns -t 16 -i step1.ctg.lay.gz -fo assembly.ctg.fa
Necat (v. 0.01)	necat.pl bridge config.txt

Draft assemblies were polished using one round of Racon [10], and one round of Medaka (https://github.com/nanoporetech/medaka), as described in **Chapter II**. In both cases, only Nanopore reads were used for polishing. Assembly metrics were calculated with MetaQUAST [11], which was also applied to compare the draft assemblies to the reference metagenomes. Thus, mismatches (or SNPs) and indels were detected and quantified.

Results and discussion

In total, 21 draft assemblies were obtained after assembling four different mock communities (BenchEV, BenchHE, BMock12 and MSA2006) using six different metagenomic assemblers. The assembly of MSA2006 could not be completed with Pomoxis (https://nanoporetech.github.io/pomoxis/) due to the same memory issue reported in **Chapter II** (i.e., "segmentation violation"), which confirmed the lack of consistency of this tool. The MSA2006 dataset was extremely large (30.3 Gbp), which may have caused the RAM issues. Additionally, Canu was stopped for both the BMock12 and the MSA2006 datasets after five days of execution. It must be noted that the rest of the tools spent less than 5 h in assembling these metagenomes (**Figure 1A**). Therefore, our work proved the computational inefficiency of Canu, which may hamper the application of this tool to study complex datasets.

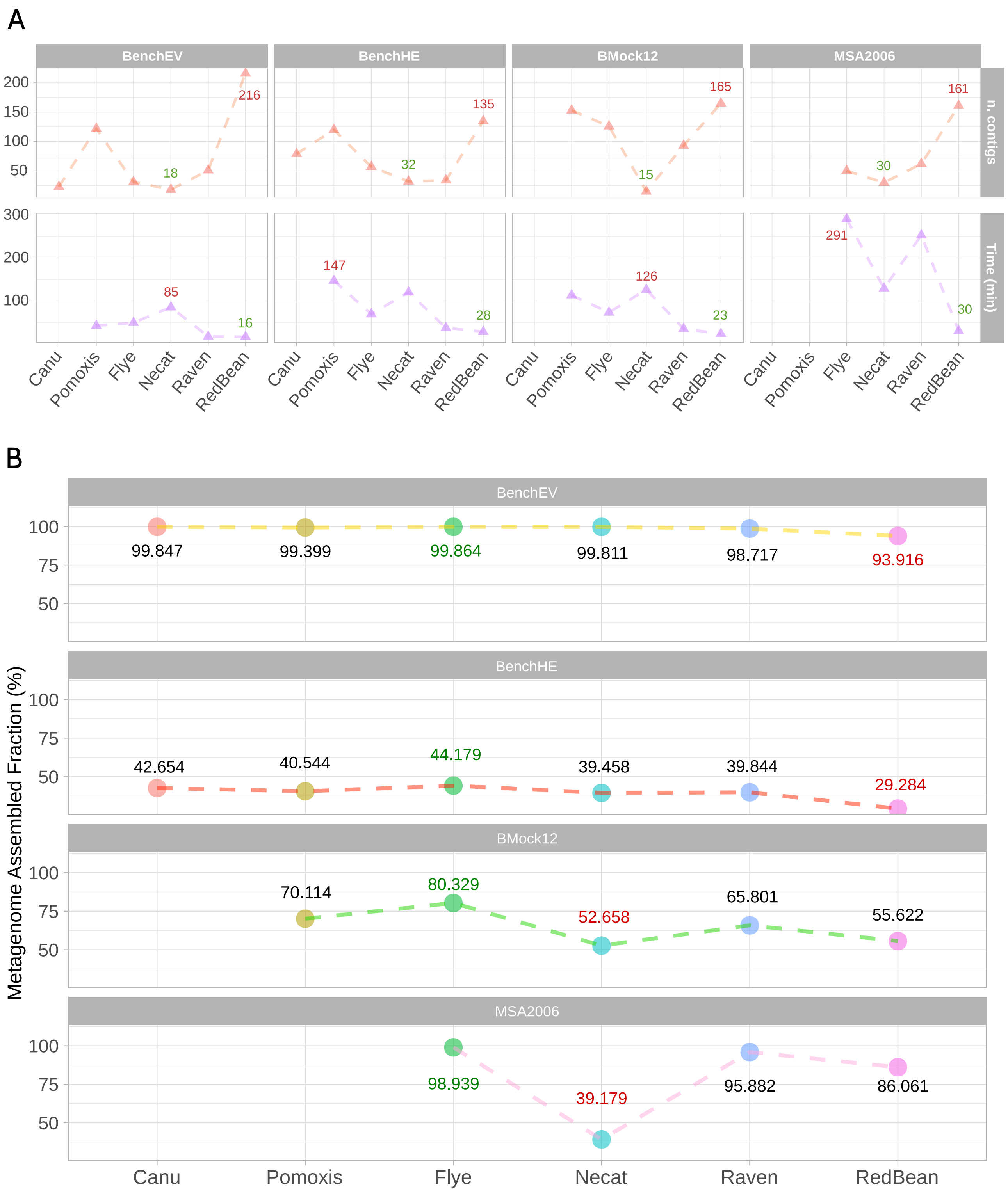
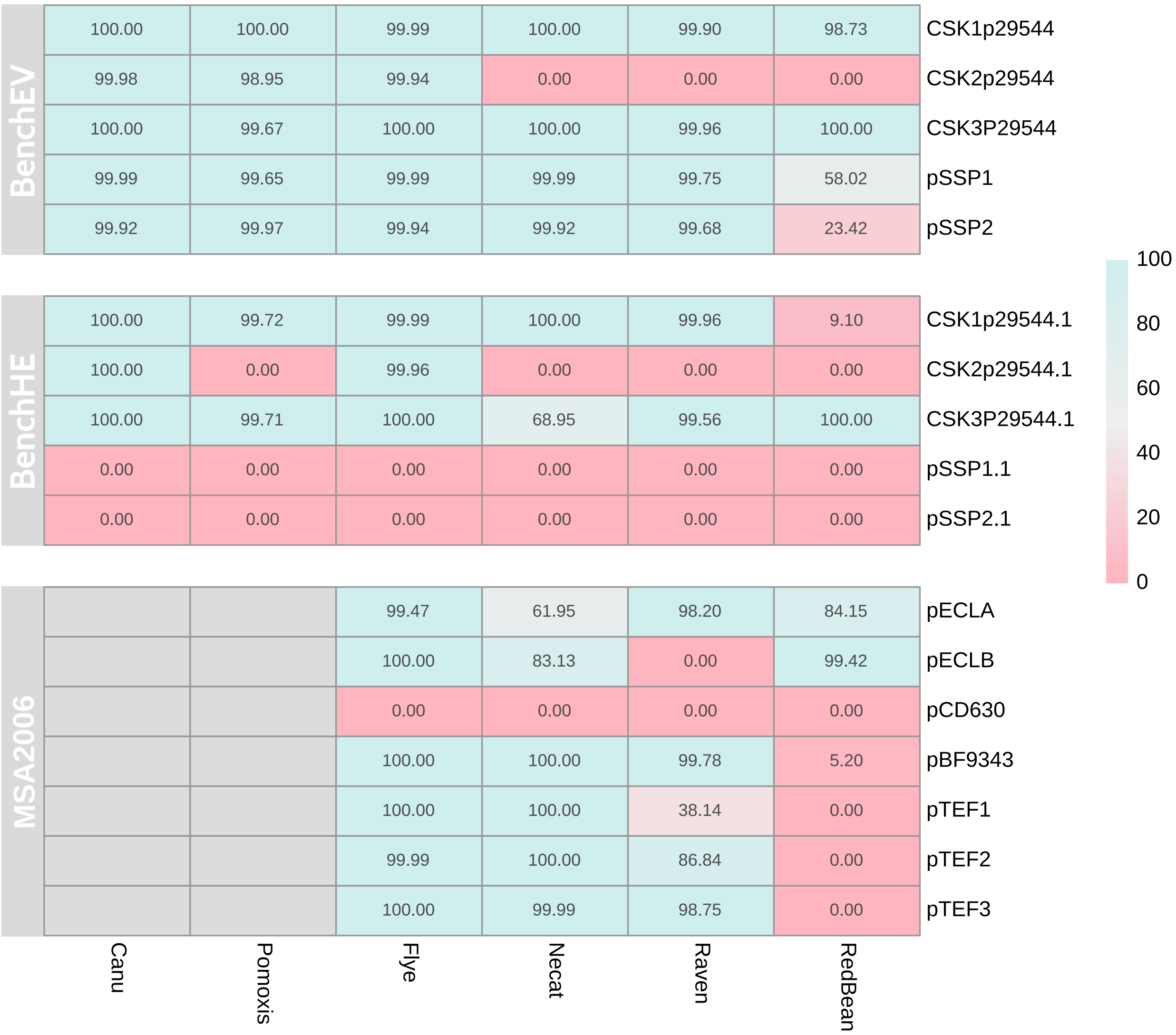


Figure 1. General performance of the different assembly tools on each mock community. (A) Number of contigs retrieved by each tool (top panel) and time spent to finish the assembly (bottom panel). Times for Canu are not shown (out of scale). (B) Fraction of the metagenome assembled by each tool. Missing points indicate that this assembly was not completed.

In agreement with the results of **Chapter II**, Flye recovered the highest metagenome fraction in all cases (**Figure 1B**). By contrast, the lowest recovery ratio was obtained by RedBean, for the BenchEV and BenchHE communities, and by Necat, for the BMock12 and MSA2006 datasets. Raven was robust in all the experiments, while Canu and Pomoxis also showed a good performance on the datasets that could be assembled.



		Canu	Pomoxis	Flye	Necat	Raven	RedBean
BenchEV	CSK1p29544	100.00	100.00	99.99	100.00	99.90	98.73
	CSK2p29544	99.98	98.95	99.94	0.00	0.00	0.00
	CSK3P29544	100.00	99.67	100.00	100.00	99.96	100.00
	pSSP1	99.99	99.65	99.99	99.99	99.75	58.02
	pSSP2	99.92	99.97	99.94	99.92	99.68	23.42
BenchHE	CSK1p29544.1	100.00	99.72	99.99	100.00	99.96	9.10
	CSK2p29544.1	100.00	0.00	99.96	0.00	0.00	0.00
	CSK3P29544.1	100.00	99.71	100.00	68.95	99.56	100.00
	pSSP1.1	0.00	0.00	0.00	0.00	0.00	0.00
	pSSP2.1	0.00	0.00	0.00	0.00	0.00	0.00
MSA2006	pECLA			99.47	61.95	98.20	84.15
	pECLB			100.00	83.13	0.00	99.42
	pCD630			0.00	0.00	0.00	0.00
	pBF9343			100.00	100.00	99.78	5.20
	pTEF1			100.00	100.00	38.14	0.00
	pTEF2			99.99	100.00	86.84	0.00
	pTEF3			100.00	99.99	98.75	0.00

Figure 2. Fraction of the plasmids (y axis) recovered by each tool (x axis) in each experiment (panels). Grey squares correspond to non-available data (assemblies were not finished).

Mock communities based on the homogeneous distribution of microorganisms (BenchEV and MSA2006) were generally better reconstructed, although Necat had problems assembling the MSA2006 dataset (high sequencing depth). Results for BMock12 were heterogeneous, and the metagenome recovery ratio ranged from ~52.7% (Necat) to 80.3% (Flye).

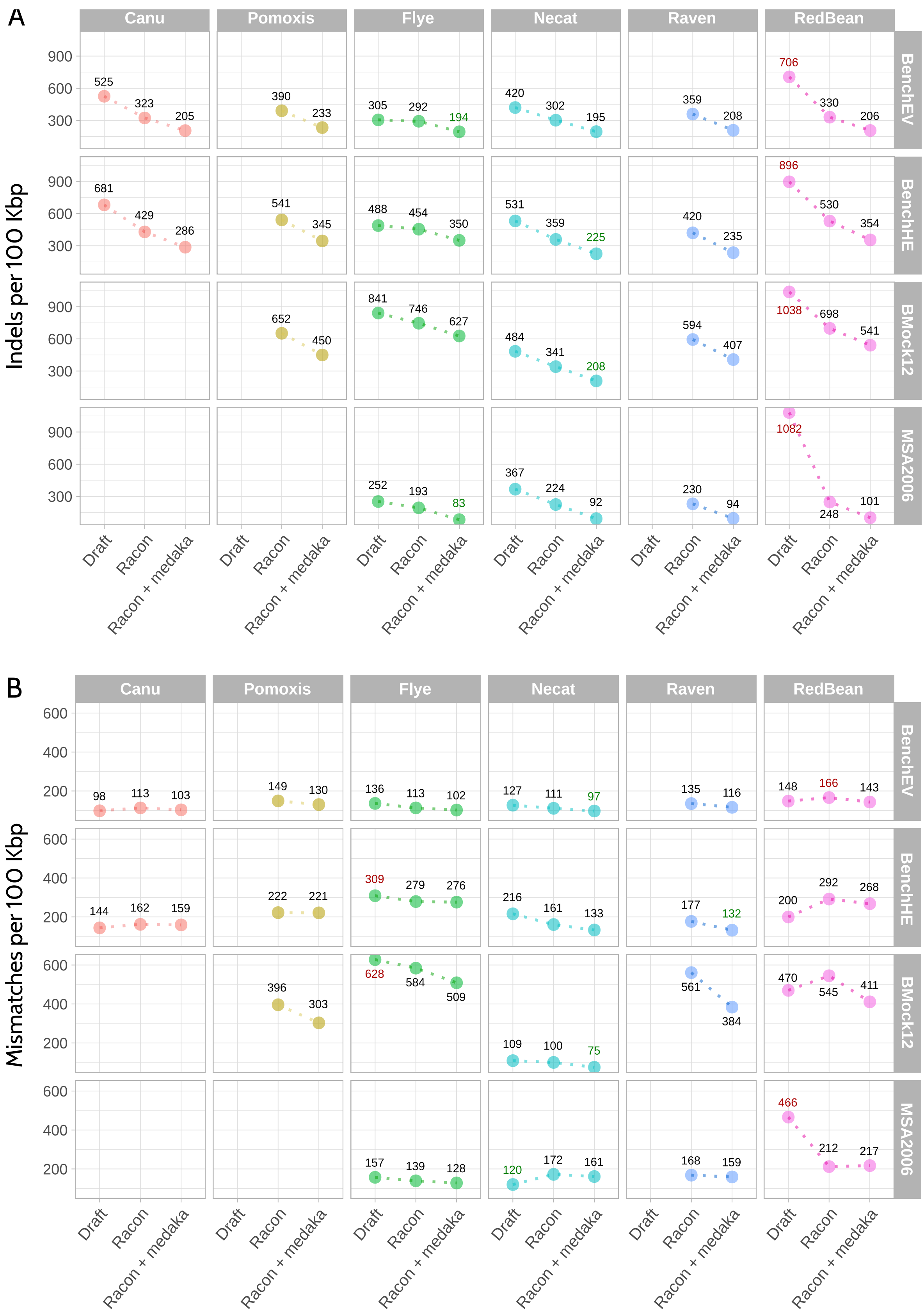


Figure 3. Evaluation of the accuracy of the assemblies before (draft) and after polishing with Racon and Medaka. (A) Indels per 100 Kbp. (B) Mismatches per 100 Kbp. By default, Raven and Pomoxis include a round of Racon polishing. Therefore, errors before polishing with Racon were not calculated for these assemblers.

Finally, BenchHE was poorly reconstructed by all tools. This was expected, as this mock community was created using a logarithmic distribution of bacteria. In consequence, sequencing coverage was insufficient to recover and assemble low-abundant microorganisms (**Figure 1B**).

In terms of contiguity, Necat and RedBean achieved the lowest and highest number of contigs in all the experiments, respectively (**Figure 1A**). The performance of Raven and Flye was similar to that of Necat, except for the BMock12 dataset. Nevertheless, this data was clearly biased, as Necat recovered a lower fraction of this metagenome (**Figure 1B**). Regarding computational efficiency, RedBean was the fastest tool and, in general, Raven was faster than Flye, Pomoxis and Necat (except for the MSA2006 dataset) (**Figure 1A**).

Plasmid recovery was also evaluated for the BenchEV, BenchHE and MSA2006 mock communities, showing that both Canu and Flye were particularly good in reconstructing these extrachromosomal DNA fragments compared to Raven, Necat and Pomoxis. Once again, the worst results were obtained with RedBean (**Figure 2**).

All the draft assemblies were polished with Racon and Medaka. As anticipated in **Chapter II**, the ratio of indels decreased substantially in all cases after polishing (**Figure 3A**). Results for mismatch correction were heterogeneous. For instance, polishing tools corrected mismatches in the assemblies generated by Raven and Flye, but the ratio of mismatches increased in the case of Canu (**Figure 3B**). Necat, Flye and Raven showed a comparable performance on BenchEV and MSA2006 in terms of both indels and mismatches. Nevertheless, the accuracy of Flye was reduced in the BenchHE and BMock12 mock communities. To further investigate this phenomenon, we calculated the ratio of mismatches for each genome included in the BenchHE and BMock12 after metagenomic assembly with Flye and Necat (**Figure 4**). This analysis clearly demonstrated that accuracy was similar for the genomes that were equally recovered by both tools. However, Flye is less conservative than Necat, and it recovered genomes that were less complete. These genomes tended to accumulate more errors, thus biasing the general accuracy of the metagenome.

Conclusions

In this work, several long-read metagenomic assemblers were used for reconstructing four different mock communities, which were sequenced with MinION (Oxford Nanopore Technologies). Flye showed the best overall performance for Nanopore-based metagenomic assembly and it worked especially well for recovering plasmids. Raven displayed a remarkable performance, but it did not excel in any specific parameter. Finally, Canu proved not to be sufficiently scalable for some datasets, while other tools (i.e., Pomoxis or Necat) did not introduce any substantial improvement to metaFlye or Raven. Although very fast, RedBean is not recommended for metagenome assembly. Polishing with Racon (one round) and Medaka (one round) reduced the number of indels, which may be key for functional prediction. Finally, polishing tools also minimized the number of mismatches in the metagenomes assembled with Racon and Flye. In general, the present benchmark helped to consolidate the results from **Chapter II**.

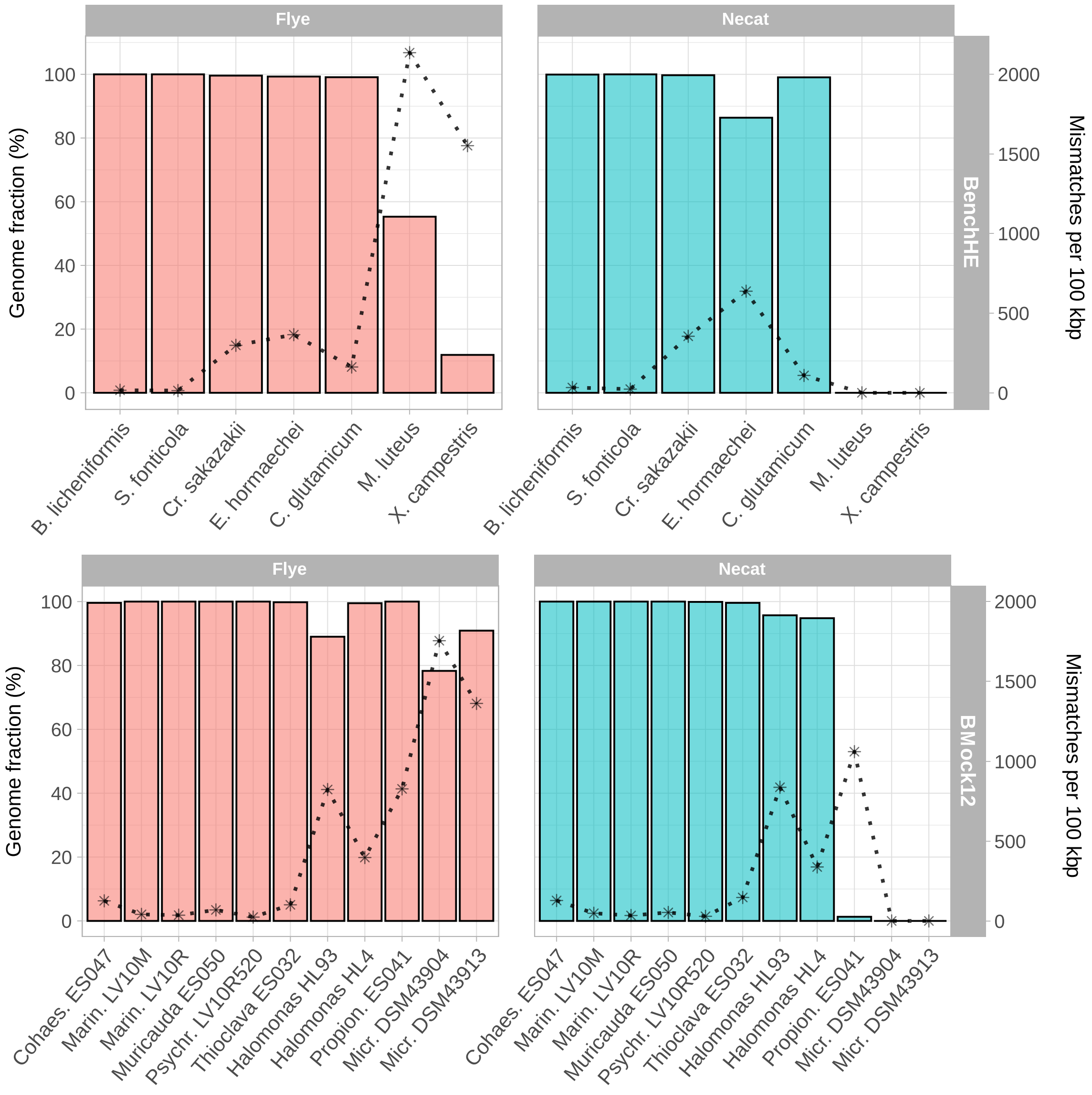


Figure 4. Accuracy calculated for each genome included in the BenchHE (top panel) and BMock12 (bottom panel) mock communities. Only Flye and Necat assemblies (after polishing) are shown. Bars display the fraction of the genome assembled by each tool. Points ("*") show mismatches per 100 Kbp. Incomplete genomes recovered by Flye (but not by Necat) tended to accumulate more errors.

References

1 Nicholls SM, Quick JC, Tang S, *et al.* Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* 2019;**8**:1–9. doi:10.1093/gigascience/giz043

2 Leidenfrost RM, Pöther DC, Jäckel U, *et al.* Benchmarking the MinION: Evaluating long reads for microbial profiling. *Sci Rep* 2020;**10**:1–11. doi:10.1038/s41598-020-61989-x

3 Sevim V, Lee J, Egan R, *et al.* Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Sci data* 2019;**6**:285. doi:10.1038/s41597-019-0287-z

4 Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* Published Online First: 2020. doi:10.1038/s41587-020-0422-6

5 Koren S, Walenz BP, Berlin K, *et al.* Canu: Scalable and accurate long-read assembly via adaptive ϰ-mer weighting and repeat separation. *Genome Res* 2017;**27**:722–36. doi:10.1101/gr.215087.116

6 Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci* 2021;**1**:332–6. doi:10.1038/s43588-021-00073-4

7 Kolmogorov M, Bickhart DM, Behsaz B, *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;**17**:1103–10. doi:10.1038/s41592-020-00971-x

8 Chen Y, Nie F, Xie SQ, *et al.* Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun* 2021;**12**:1–10. doi:10.1038/s41467-020-20236-7

9 Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 2020;**17**:155–8. doi:10.1038/s41592-019-0669-3

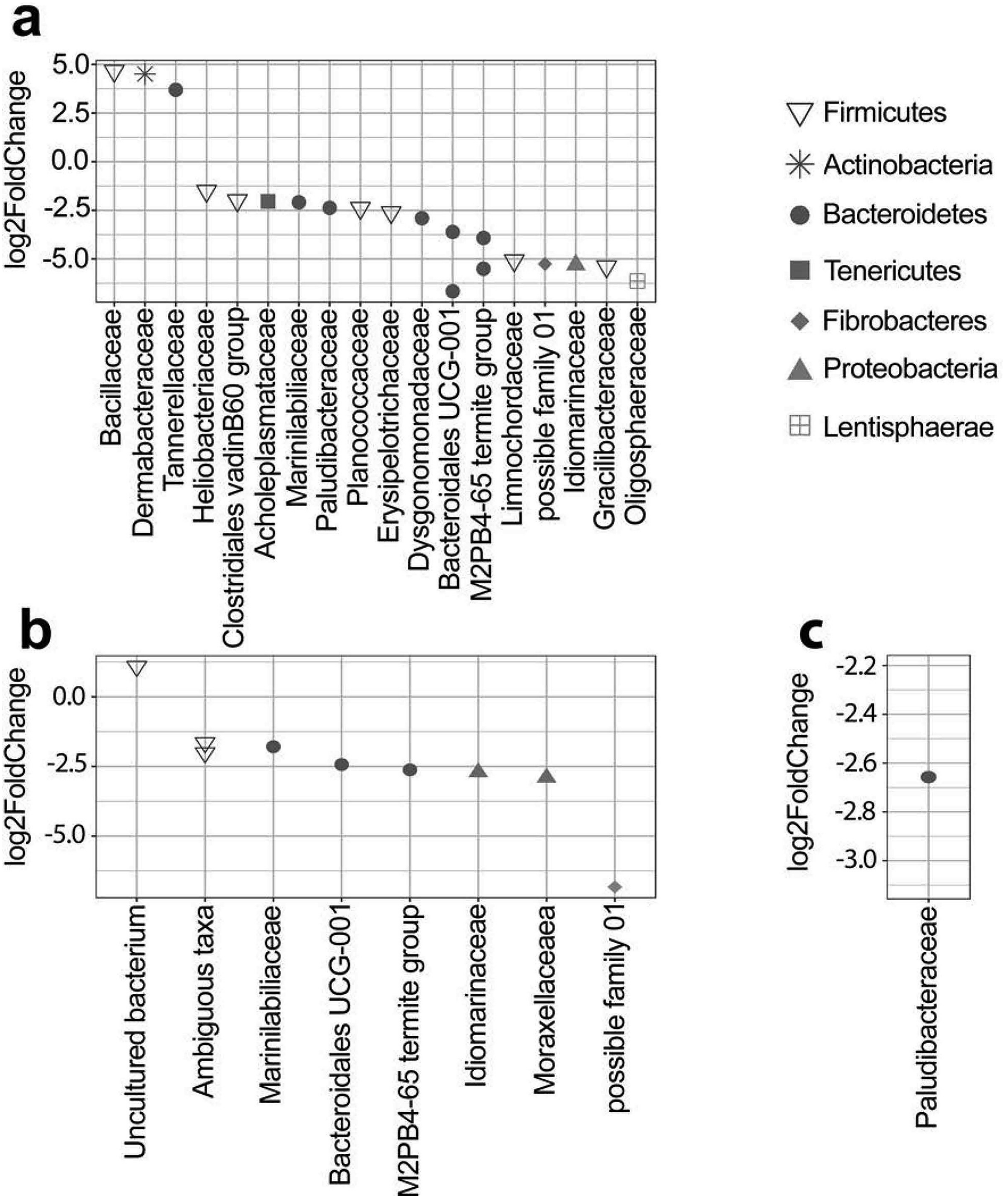
10 Vaser R, Sović I, Nagarajan N, *et al.* Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017;**27**:737–46. doi:10.1101/gr.214270.116

11 Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics* 2016;**32**:1088–90. doi:10.1093/bioinformatics/btv697

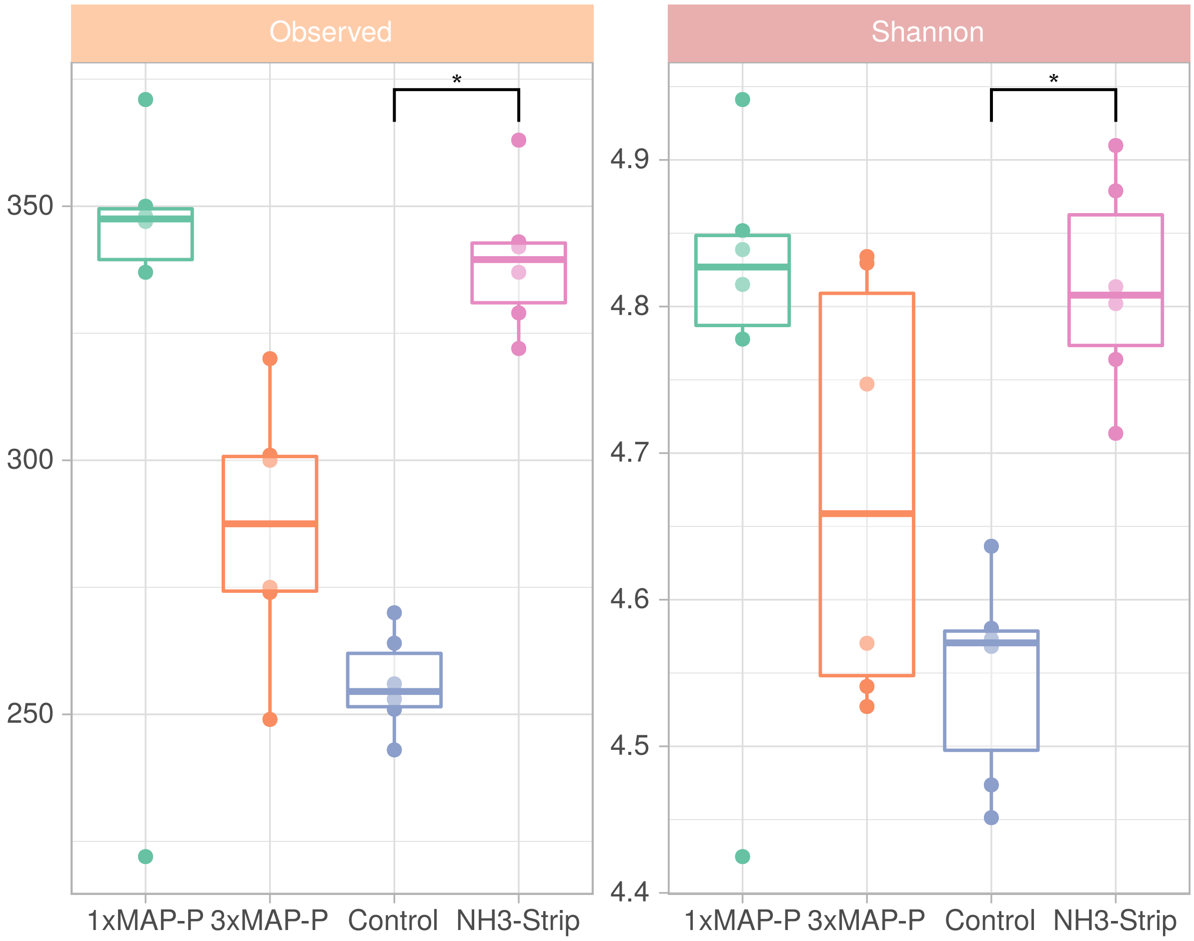
Appendix B

Supplementary material

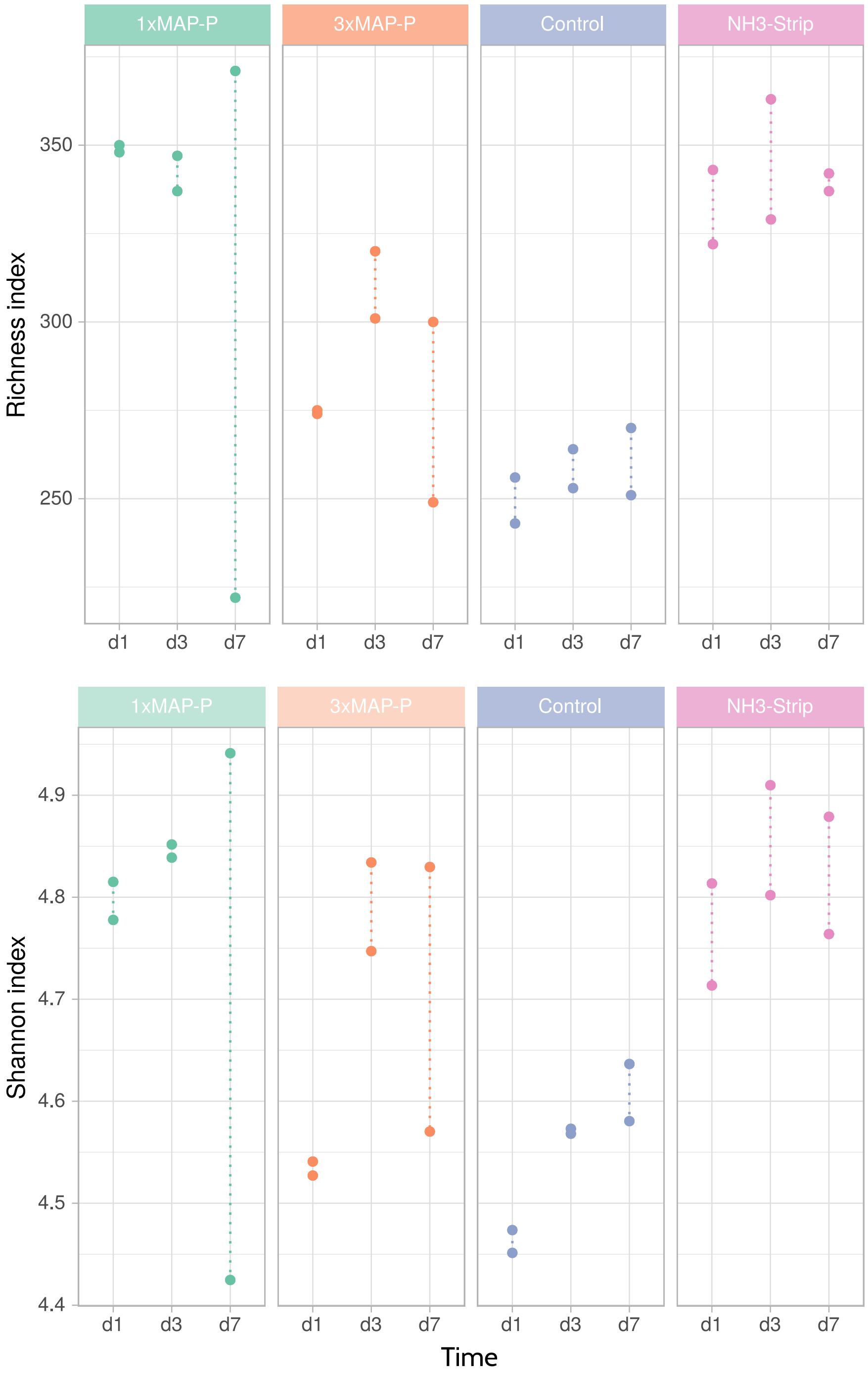
Chapter IA



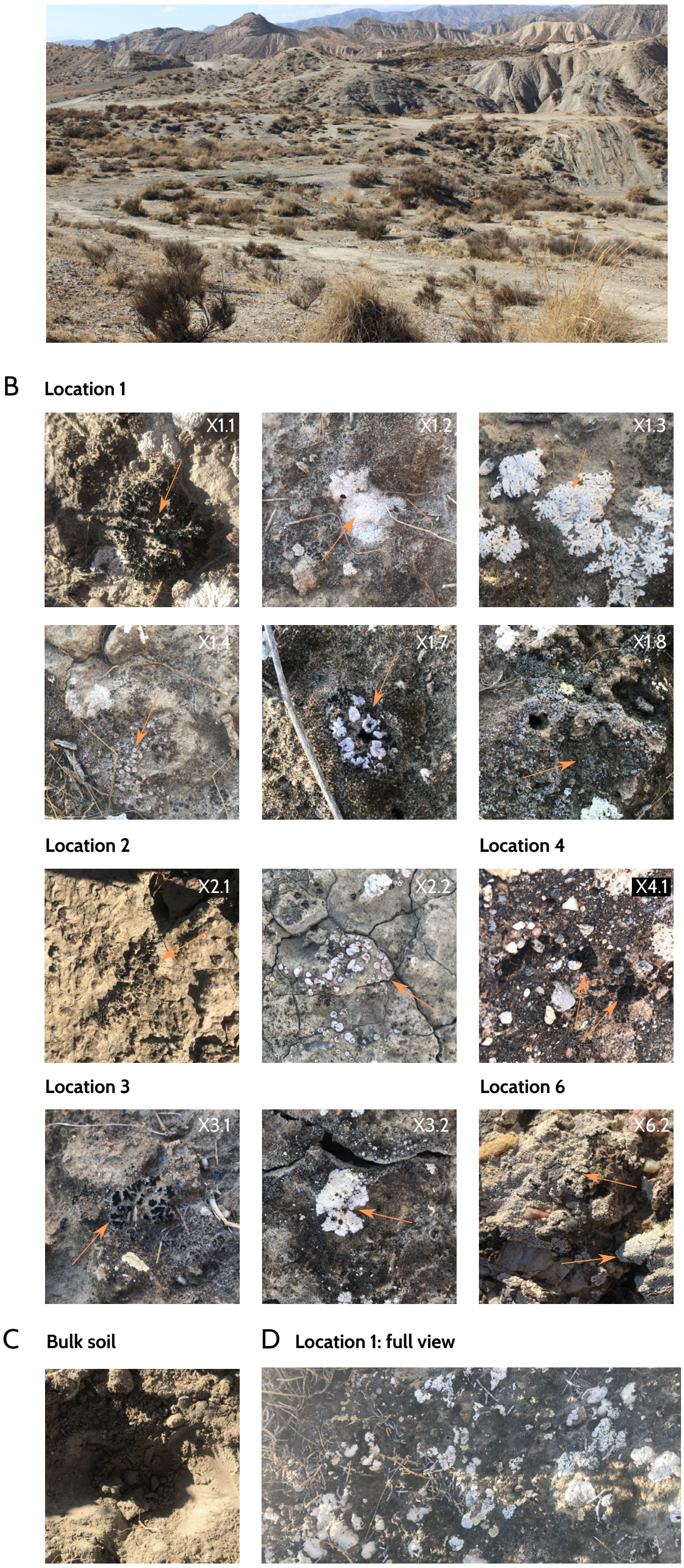
Supplementary Figure IA.S1. Differential abundance analysis (DESeq2 test) based on Nanopore sequencing data. Differences between the different treatments and the control reactors were calculated at day 7. Genera showing significant differences (FDR-adjusted p-value < 0.05) were grouped by family (x axis) and phylum (point shape). (a) 3x MAP precipitation, (b) 1x MAP precipitation and (c) NH3-stripping vs. control (no treatment). Positive log2FoldChange values mean that the taxon is overrepresented in the treatment, while negative values mean that the taxon is overrepresented in the control.



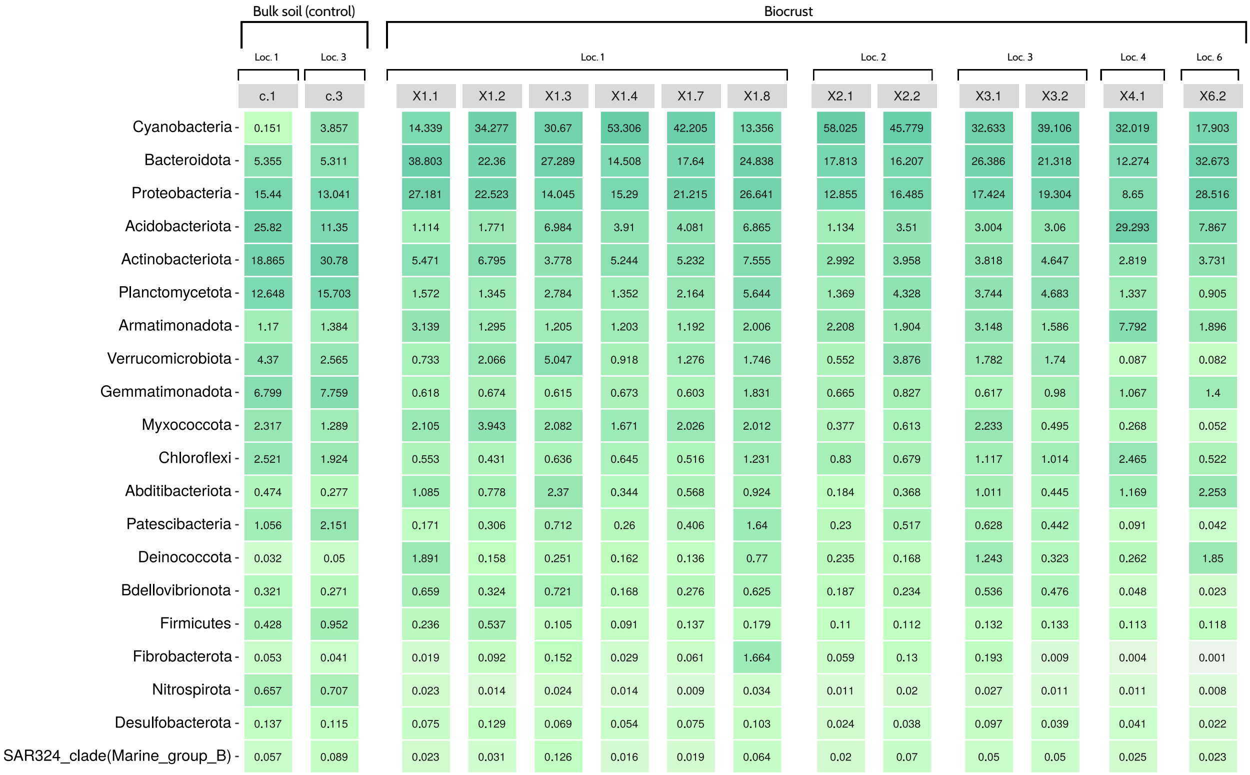
Supplementary Figure IA.S2. Alpha diversity analysis. Richness (“Observed”; left panel) and Shannon (right panel) metrics grouped by treatment and considering all the samples from days 1, 3 and 7. “*” indicates a significant difference (FDR-adjusted p-value < 0.05; Wilcoxon rank-sum test).



Supplementary Figure IA.S3. Alpha diversity analysis. Richness (top panel) and Shannon (bottom panel) metrics grouped by treatment vs. day of acidification ('d').

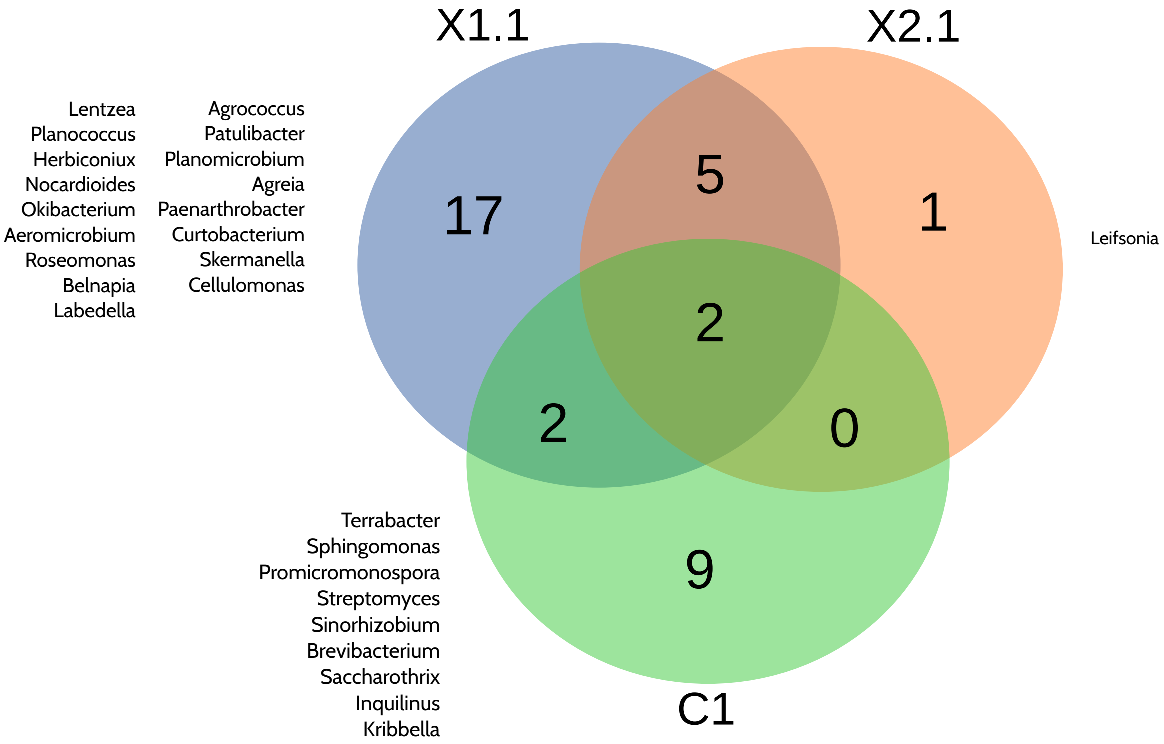


Supplementary Figure IB.S1. Sampling overview. (A) Tabernas Desert landscape. (B) Pictures of the biocrust samples. (C) Picture of a bulk soil sample. (D) Full view of Location 1. The diversity of biocrusts can be observed. GPS coordinates of the samples are available upon reasonable request, since the natural resources of the Tabernas Desert are protected by regional rules.



	Bulk soil (control)		Biocrust											
	Loc. 1	Loc. 3	Loc. 1						Loc. 2		Loc. 3		Loc. 4	Loc. 6
	c.1	c.3	X1.1	X1.2	X1.3	X1.4	X1.7	X1.8	X2.1	X2.2	X3.1	X3.2	X4.1	X6.2
Cyanobacteria	0.151	3.857	14.339	34.277	30.67	53.306	42.205	13.356	58.025	45.779	32.633	39.106	32.019	17.903
Bacteroidota	5.355	5.311	38.803	22.36	27.289	14.508	17.64	24.838	17.813	16.207	26.386	21.318	12.274	32.673
Proteobacteria	15.44	13.041	27.181	22.523	14.045	15.29	21.215	26.641	12.855	16.485	17.424	19.304	8.65	28.516
Acidobacteriota	25.82	11.35	1.114	1.771	6.984	3.91	4.081	6.865	1.134	3.51	3.004	3.06	29.293	7.867
Actinobacteriota	18.865	30.78	5.471	6.795	3.778	5.244	5.232	7.555	2.992	3.958	3.818	4.647	2.819	3.731
Planctomycetota	12.648	15.703	1.572	1.345	2.784	1.352	2.164	5.644	1.369	4.328	3.744	4.683	1.337	0.905
Armatimonadota	1.17	1.384	3.139	1.295	1.205	1.203	1.192	2.006	2.208	1.904	3.148	1.586	7.792	1.896
Verrucomicrobiota	4.37	2.565	0.733	2.066	5.047	0.918	1.276	1.746	0.552	3.876	1.782	1.74	0.087	0.082
Gemmatimonadota	6.799	7.759	0.618	0.674	0.615	0.673	0.603	1.831	0.665	0.827	0.617	0.98	1.067	1.4
Myxococcota	2.317	1.289	2.105	3.943	2.082	1.671	2.026	2.012	0.377	0.613	2.233	0.495	0.268	0.052
Chloroflexi	2.521	1.924	0.553	0.431	0.636	0.645	0.516	1.231	0.83	0.679	1.117	1.014	2.465	0.522
Abditibacteriota	0.474	0.277	1.085	0.778	2.37	0.344	0.568	0.924	0.184	0.368	1.011	0.445	1.169	2.253
Patescibacteria	1.056	2.151	0.171	0.306	0.712	0.26	0.406	1.64	0.23	0.517	0.628	0.442	0.091	0.042
Deinococcota	0.032	0.05	1.891	0.158	0.251	0.162	0.136	0.77	0.235	0.168	1.243	0.323	0.262	1.85
Bdellovibrionota	0.321	0.271	0.659	0.324	0.721	0.168	0.276	0.625	0.187	0.234	0.536	0.476	0.048	0.023
Firmicutes	0.428	0.952	0.236	0.537	0.105	0.091	0.137	0.179	0.11	0.112	0.132	0.133	0.113	0.118
Fibrobacterota	0.053	0.041	0.019	0.092	0.152	0.029	0.061	1.664	0.059	0.13	0.193	0.009	0.004	0.001
Nitrospirota	0.657	0.707	0.023	0.014	0.024	0.014	0.009	0.034	0.011	0.02	0.027	0.011	0.011	0.008
Desulfobacterota	0.137	0.115	0.075	0.129	0.069	0.054	0.075	0.103	0.024	0.038	0.097	0.039	0.041	0.022
SAR324_clade(Marine_group_B)	0.057	0.089	0.023	0.031	0.126	0.016	0.019	0.064	0.02	0.07	0.05	0.05	0.025	0.023

Supplementary Figure IB.S2. Heatmap showing the top 20 phyla detected in the samples and their relative abundances. Loc. = Location.



Supplementary Figure IB.S3. Venn diagram showing the genera isolated from any replicate of each sample

Supplementary Figure IB.S4. Barplot displaying the cumulative relative abundances (x axis) of the radiation- and desiccation-resistant genera.

This figure could not be integrated into this thesis, but it is available on:

https://doi.org/10.5281/zenodo.5771104

Supplementary Table IB.S1. Mobile laboratory setup. List of main reagents, fungible and equipment used.

Instrument	Description	Units
Laptop	MSI GF63 Thin 9SC-047XES laptop (CPU: Intel Corei7-9750H, 6 core, 12 threads; RAM: 16GB; SSD: 512 Gb; Graphics Card: GeForce GTX 1650)	2
MinION	MinION Mk1B (ONT, Oxford, UK)	1
Flow cell	R9.4.1 MinION flow cell (ONT, Oxford, UK, Cat. No.: FLO-MIN106D)	2
Thermoblock	24 tubes (1.5 mL) thermoblock Labnet 596111 (Labnet, Madrid, Spain)	1
Horizontal vortex	Horizontal vortex for 24 tubes (Selecta J.P., Barcelona, Spain)	1
Microcentrifuge	12 tubes (1.5 mL); 13400 rpm max.; Minispin eppendorf F45-12-11 (Eppendorf, Hamburg, Germany)	1
Qubit	Qubit™ 2.0 Flex Fluorometer (Thermo Fisher, Waltham, United States, Cat. No.: Q33327)	1
Thermocycler	Mastercycler Eppendorf 5332 (Eppendorf, Hamburg, Germany)	1
Pipettes	10, 100 and 1000 ㎕ pipettes	1
Bunsen burner	-	1
Fungible & reagents	**Description**	**Units**
Qubit DNA Kit	x1 dsDNA High-Sensitivity Assay kit (Thermo Fisher, Waltham, United States, Cat. No.: Q33230)	1
Pipette tips	10, 100 and 1000 ㎕ x96 pipette tips	2
DNA extraction kit	DNEasy Power Soil Kit (QIAGEN, Germany, Cat. No.: 12888)	1
16S primers	S-D-Bact-0008-a-S-16 and S-D-Bact-1492-a-A-16. Modified with ONT Universal tags.	-
PCR mix	NZYTaq II 2x Green Master Mix (NZYTech, Lisboa, Portugal, Cat. No.: MB358)	1
Purification kit	NucleoMag kit for PCR clean up with magnetic beads (Macherey-Nagel, Germany, Cat. No.: 744100.4).	1
ONT barcoding kit	PCR Barcoding Expansion Pack 1-96 (ONT, Oxford, UK, Cat. No.: EXP-PBC096)	1
ONT ligation kit	Ligation Sequencing Kit (ONT, Oxford, UK, Cat. No.: SQK-LSK109)	1
End-prep kit	NEBNext FFPE DNA Repair Mix (New England Biolabs, Ipswich, US, Cat. No.: M6630)	1
Flow cell wash kit	Flow Cell Wash Kit (ONT, Oxford, UK, Cat. No.: EXP-WSH004)	1

Supplementary Table IB.S2. Summary of the microbial culture collection. Only dereplicated strains have been included.

This table could not be integrated into this thesis, but it is available on:

https://doi.org/10.5281/zenodo.5771104

Supplementary Table IB.S3. Top 20 phyla and their relative abundances in the different samples.

Phylum	X1.1	X1.2	X1.3	X1.4	X1.7	X1.8	X2.1	X2.2	X3.1	X3.2	X4.1	X6.2	Biocrust Avg	c.1	c.3	Soil Avg	Average (Avg)
Cyanobacteria	14.34	34.28	30.67	53.31	42.21	13.36	58.03	45.78	32.63	39.11	32.02	17.90	34.47	0.15	3.86	2.00	29.83
Bacteroidota	38.80	22.36	27.29	14.51	17.64	24.84	17.81	16.21	26.39	21.32	12.27	32.67	22.68	5.35	5.31	5.33	20.20
Proteobacteria	27.18	22.52	14.05	15.29	21.22	26.64	12.86	16.49	17.42	19.30	8.65	28.52	19.18	15.44	13.04	14.24	18.47
Acidobacteriota	1.11	1.77	6.98	3.91	4.08	6.87	1.13	3.51	3.00	3.06	29.29	7.87	6.05	25.82	11.35	18.59	7.84
Actinobacteriota	5.47	6.80	3.78	5.24	5.23	7.55	2.99	3.96	3.82	4.65	2.82	3.73	4.67	18.86	30.78	24.82	7.55
Planctomycetota	1.57	1.34	2.78	1.35	2.16	5.64	1.37	4.33	3.74	4.68	1.34	0.90	2.60	12.65	15.70	14.18	4.26
Armatimonadota	3.14	1.30	1.20	1.20	1.19	2.01	2.21	1.90	3.15	1.59	7.79	1.90	2.38	1.17	1.38	1.28	2.22
Verrucomicrobiota	0.73	2.07	5.05	0.92	1.28	1.75	0.55	3.88	1.78	1.74	0.09	0.08	1.66	4.37	2.56	3.47	1.92
Gemmatimonadota	0.62	0.67	0.62	0.67	0.60	1.83	0.66	0.83	0.62	0.98	1.07	1.40	0.88	6.80	7.76	7.28	1.79
Myxococcota	2.10	3.94	2.08	1.67	2.03	2.01	0.38	0.61	2.23	0.50	0.27	0.05	1.49	2.32	1.29	1.80	1.53
Chloroflexi	0.55	0.43	0.64	0.64	0.52	1.23	0.83	0.68	1.12	1.01	2.46	0.52	0.89	2.52	1.92	2.22	1.08
Abditibacteriota	1.09	0.78	2.37	0.34	0.57	0.92	0.18	0.37	1.01	0.44	1.17	2.25	0.96	0.47	0.28	0.38	0.87
Patescibacteria	0.17	0.31	0.71	0.26	0.41	1.64	0.23	0.52	0.63	0.44	0.09	0.04	0.45	1.06	2.15	1.60	0.62
Deinococcota	1.89	0.16	0.25	0.16	0.14	0.77	0.24	0.17	1.24	0.32	0.26	1.85	0.62	0.03	0.05	0.04	0.54
Bdellovibrionota	0.66	0.32	0.72	0.17	0.28	0.62	0.19	0.23	0.54	0.48	0.05	0.02	0.36	0.32	0.27	0.30	0.35
Firmicutes	0.24	0.54	0.10	0.09	0.14	0.18	0.11	0.11	0.13	0.13	0.11	0.12	0.17	0.43	0.95	0.69	0.24
Fibrobacterota	0.02	0.09	0.15	0.03	0.06	1.66	0.06	0.13	0.19	0.01	0.00	0.00	0.20	0.05	0.04	0.05	0.18
Nitrospirota	0.02	0.01	0.02	0.01	0.01	0.03	0.01	0.02	0.03	0.01	0.01	0.01	0.02	0.66	0.71	0.68	0.11
Desulfobacterota	0.08	0.13	0.07	0.05	0.07	0.10	0.02	0.04	0.10	0.04	0.04	0.02	0.06	0.14	0.12	0.13	0.07
Other	0.21	0.18	0.46	0.16	0.18	0.34	0.14	0.25	0.23	0.19	0.19	0.14	0.22	1.39	0.47	0.93	0.32

Supplementary Table IB.S4. Full genus-level relative abundance table.

This table could not be integrated into this thesis, but it is available on:

https://doi.org/10.5281/zenodo.5771104

Supplementary Table IB.S5. Genera isolated from each combination of samples.

Samples	Number	Genus
C1 & X1.1 & X2.1	2	*Arthrobacter* *Pseudarthrobacter*
X1.1 & X2.1	5	*Kocuria* *Clavibacter* *Methylorubrum* *Mycolicibacterium* *Blastococcus*
C1 & X1.1	2	*Staphylococcus* *Bacillus*
X1.1	17	*Lentzea* *Planococcus* *Herbiconiux* *Nocardioides* *Okibacterium* *Aeromicrobium* *Roseomonas* *Belnapia* *Labedella* *Agrococcus* *Patulibacter* *Planomicrobium* *Agreia* *Paenarthrobacter* *Curtobacterium* *Skermanella* *Cellulomonas*
X2.1	1	*Leifsonia*
C1	9	*Terrabacter* *Sphingomonas* *Promicromonospora* *Streptomyces* *Sinorhizobium* *Brevibacterium* *Saccharothrix* *Inquilinus* *Kribbella*

Supplementary Table IB.S6. Radiation- and desiccation-resistant genera isolated from each combination of samples.

Samples	Number	Genus
C1 & X1.1 & X2.1	1	*Arthrobacter*
X1.1 & X2.1	2	*Methylorubrum* *Kocuria*
C1 & X1.1	1	*Bacillus*
X1.1	4	*Planomicrobium* *Cellulomonas* *Nocardioides* *Curtobacterium*
C1	1	*Sphingomonas*

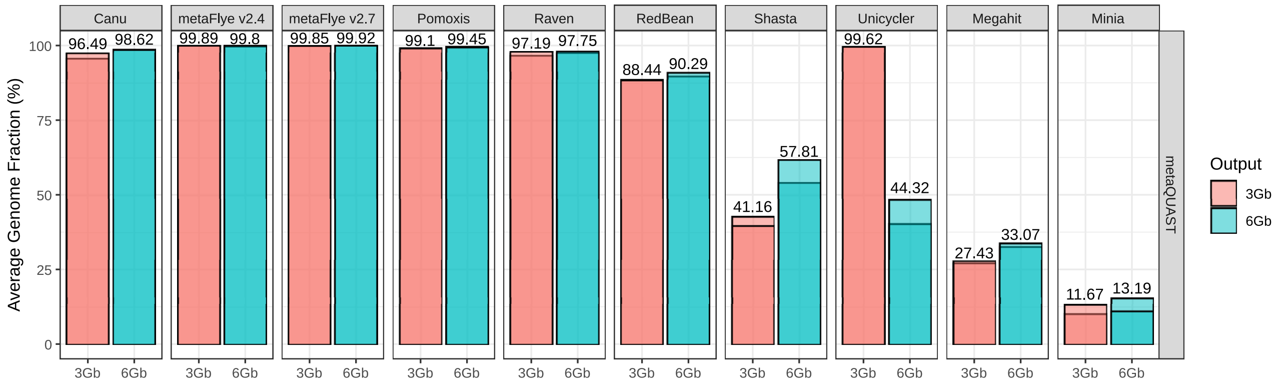
Supplementary Table IB.S7. Genera isolated from each combination of culture conditions.

Culture Media	Number	Genus
TSA	27	*Lentzea* *Klenkia* *Planococcus* *Herbiconiux* *Nocardioides* *Kocuria* *Okibacterium* *Terrabacter* *Clavibacter* *Mycetocola* *Cellulosimicrobium* *Promicromonospora* *Cryobacterium* *Diaminobutyricimonas* *Kineococcus* *Agrococcus* *Friedmanniella* *Patulibacter* *Planomicrobium* *Agreia* *Saccharothrix* *Paenarthrobacter* *Staphylococcus* *Cellulomonas* *Leifsonia* *Microbacterium* *Kribbella*
SSE/HD	4	*Aureimonas* *Rhodococcus* *Frondihabitans* *Methylobacterium*
SSE/HD + light	4	*Aeromicrobium* *Sphingomonas* *Roseomonas* *Brevibacterium*
TSA & SSE/HD	6	*Methylorubrum* *Streptomyces* *Labedella* *Skermanella* *Curtobacterium* *Bacillus*
SSE/HD + Light & TSA	2	*Mycolicibacterium* *Pseudarthrobacter*
SSE/HD & SSE/HD + Light	4	*Belnapia* *Sinorhizobium* *Pseudomonas* *Inquilinus*
SSE/HD & SSE/HD + Light & TSA	3	*Arthrobacter* *Blastococcus* *Modestobacter*

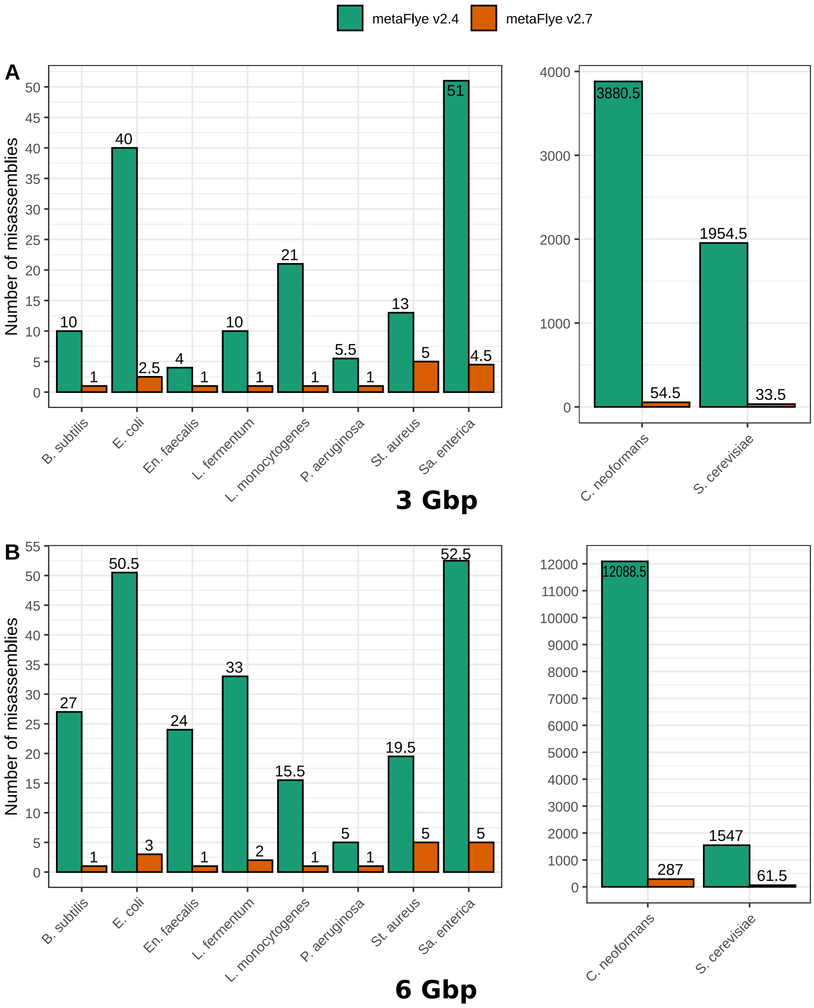
Supplementary File IB.S1. Raw Spaghetti output (interactive HTML file).

This file could not be integrated into this thesis, but it is available on:

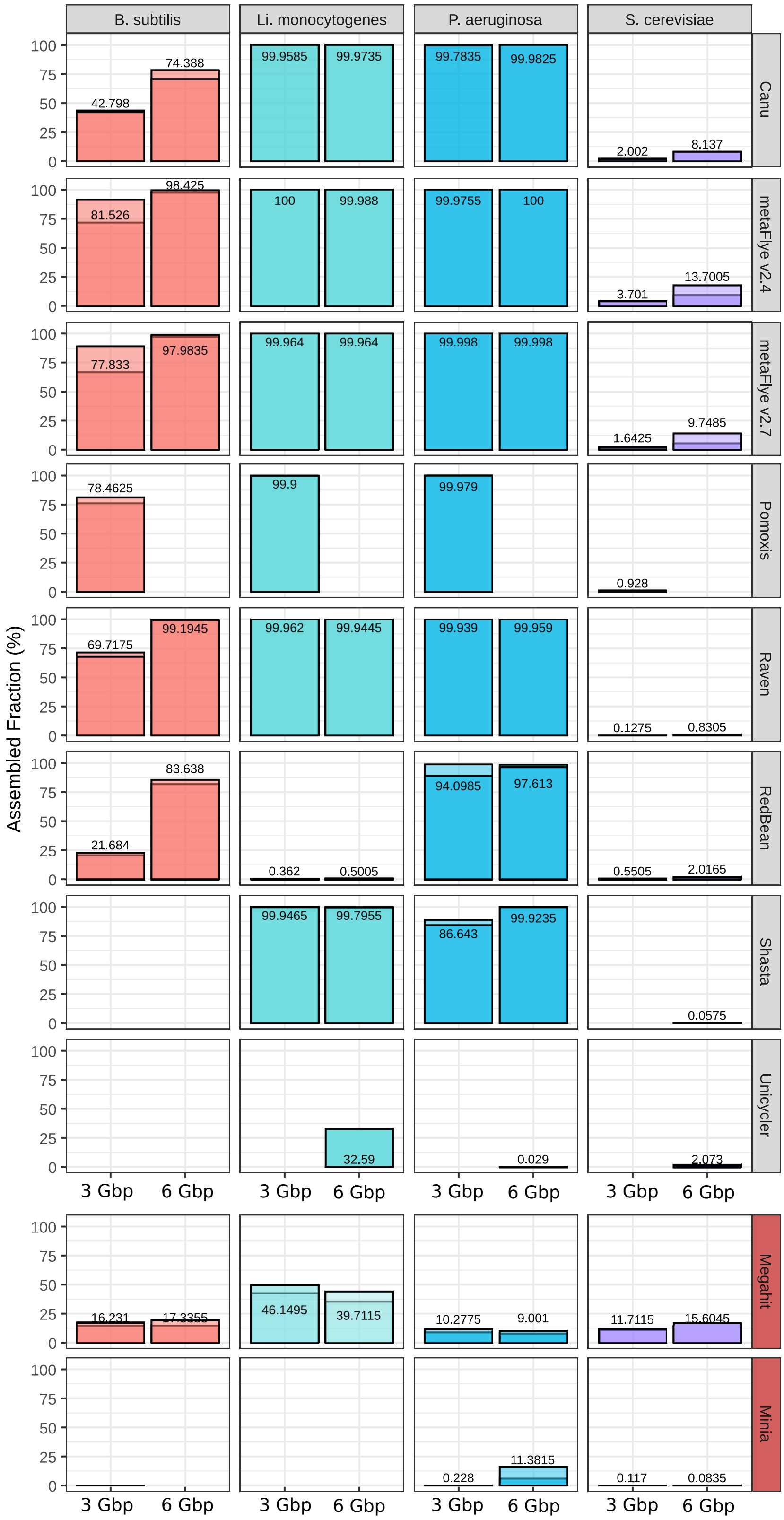
https://doi.org/10.5281/zenodo.5771104



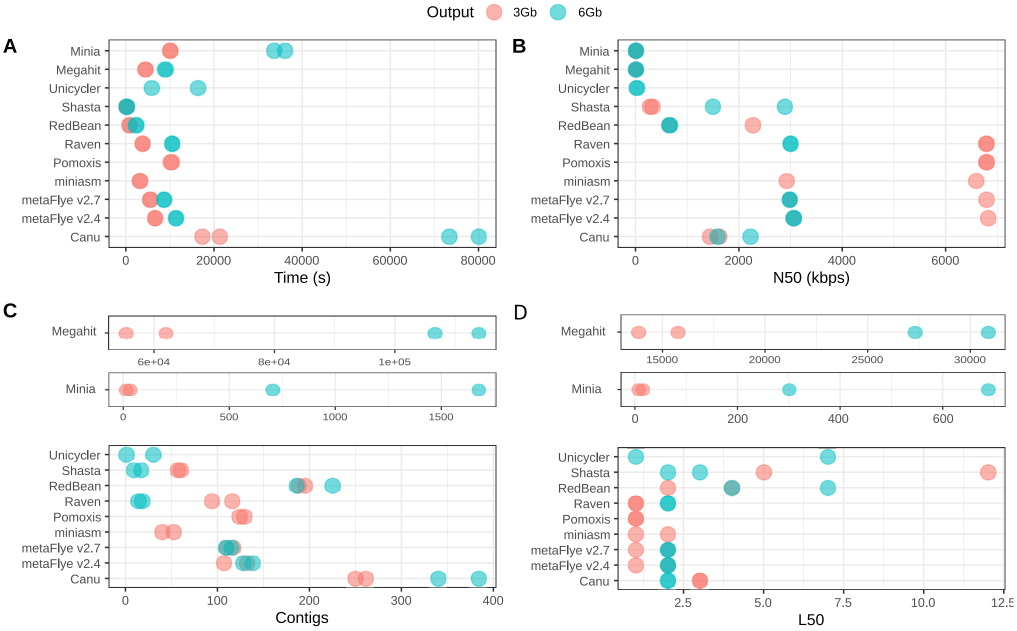
Supplementary Figure II.S1. Average recovery fraction for the bacterial genomes.



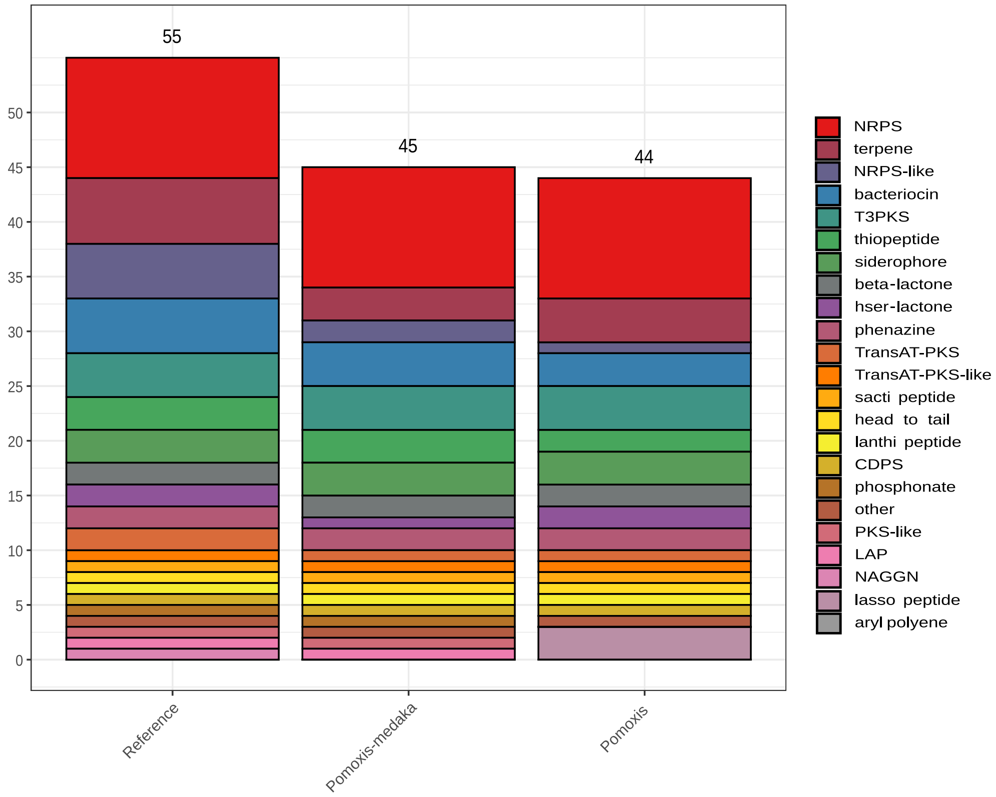
Supplementary Figure II.S2. Number of missassemblies detected in metaFlye v2.4 vs metaFlye v2.7.



Supplementary Figure II.S3. Fraction of the genome covered by the draft assemblies obtained with each tool, and for each individual microorganism (Log datasets). Minimap2 + miniasm assemblies are not shown, since it was not possible to evaluate them with metaQUAST. Only microorganisms with >1% genome fraction recovered for at least one long-read assembler are shown.



Supplementary Figure II.S4. General assembly performance of each tool for the subsampled Log datasets. (A) Run time; (B) N50; (C) Number of contigs; (D) L50.



Supplementary Figure II.S5. Number of biosynthetic gene clusters (BGCs) predicted by antiSMASH for Pomoxis and Pomoxis + one round of Medaka polishing.

Supplementary Table II. S1. General assembly metrics for the subsampled Even datasets.

	Even 3 Gbp				Even 6 Gbp			
Assembler	**Time**	**Contigs**	**N50 (bp)**	**L50**	**Time**	**Contigs**	**N50 (bp)**	**L50**
Minia (GridION)	9197.737	4540	643	1846	37099.649	3901	654	1581
Minia (PromethION	8304.526	5490	659	2182	33416.486	4677	679	1846
Megahit (G)	5253.507	103985	1216	27139	10327.801	186134	1105	52475
Megahit (P)	4989.36	102456	1253	26557	9674.742	180636	1139	50217
Canu (G)	10978.687	883	1163820	6	38568.393	1555	640396	13
Canu (P)	12483.032	900	642230	11	43858.387	1565	568807	13
Unicycler (G)	7700.913	94	2779880	4	16631.639	304	45271	72
Unicycler (P)	8968.157	83	2034125	5	17319.445	294	66707	67
miniasm (G)	654.505	130	1885531	5	2510.74	287	1880034	5
miniasm (P)	708.913	130	2033537	5	2729.335	255	2041461	5
metaFlye v2.4 (G)	3559.961	359	2916346	5	7950.856	897	1282129	9
metaFlye v2.4 (P)	3330.583	322	2131086	6	7838.849	876	1904318	8
metaFlye v2.7 (G)	3027.983	281	2713249	5	7362.655	821	2116196	7
metaFlye v2.7 (P)	3145.585	251	4024367	4	7623.861	786	2833748	6
Raven (G)	1575.824	107	2838930	4	5295.183	574	1245136	7
Raven (P)	1709.401	111	2716031	4	4979.566	568	789908	7
RedBean (G)	581.017	537	820672	10	1238.877	999	1093208	7
RedBean (P)	595.181	449	787102	13	1259.508	925	702782	11
Shasta (G)	127.911	108	855983	4	284.967	165	4762717	2
Shasta (P)	143.778	95	1043084	4	285.085	148	4664589	2
Pomoxis (G)	7594.56	411	1359179	6	17591.121	940	233266	19
Pomoxis (P)	6928.873	399	1705012	7	17020.362	923	299631	14

Supplementary Table II. S2. Assembly metrics for the subsampled GridION datasets assembled with Canu

		3Gb			**6Gb**	
	Contigs	**N50 (bp)**	**L50**	**Contigs**	**N50 (bp)**	**L50**
Bacillus subtilis	34	298,071	5	17	655,353	3
Enterococcus faecalis	13	388,478	3	10	747,976	3
Escherichia coli	5	2,669,962	1	6	4,941,166	1
Lactobacillus fermentum	13	402,806	2	14	4,941,166	1
Listeria monocytogenes	18	4,942,769	1	14	2,747,940	2
Pseudomonas aeruginosa	3	5,593,153	1	4	2,747,940	2
Salmonella enterica	3	4,942,769	1	11	2,075,612	2
Staphylococcus aureus	17	769,443	2	17	640,396	3
Cryptococcus neoformans	199	3,722	51	807	5,9140	189
Saccharomyces cerevisiae	596	6,062	142	684	18,390	145

Supplementary Table II.S3. BGC profiles for the GridION datasets.

This table could not be integrated into this thesis, but it is available on:

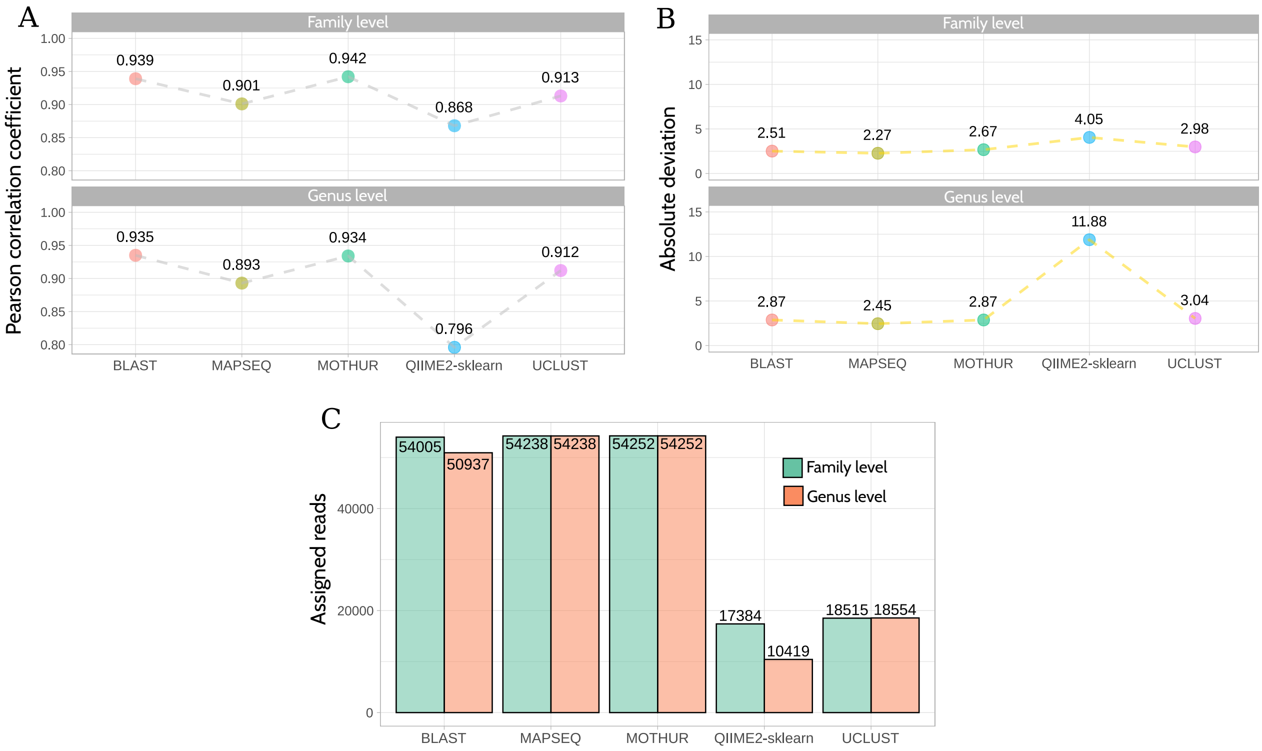
https://doi.org/10.5281/zenodo.5832041

Supplementary Table II.S4. Accuracy metrics after each round of polishing.

This table could not be integrated into this thesis, but it is available on:

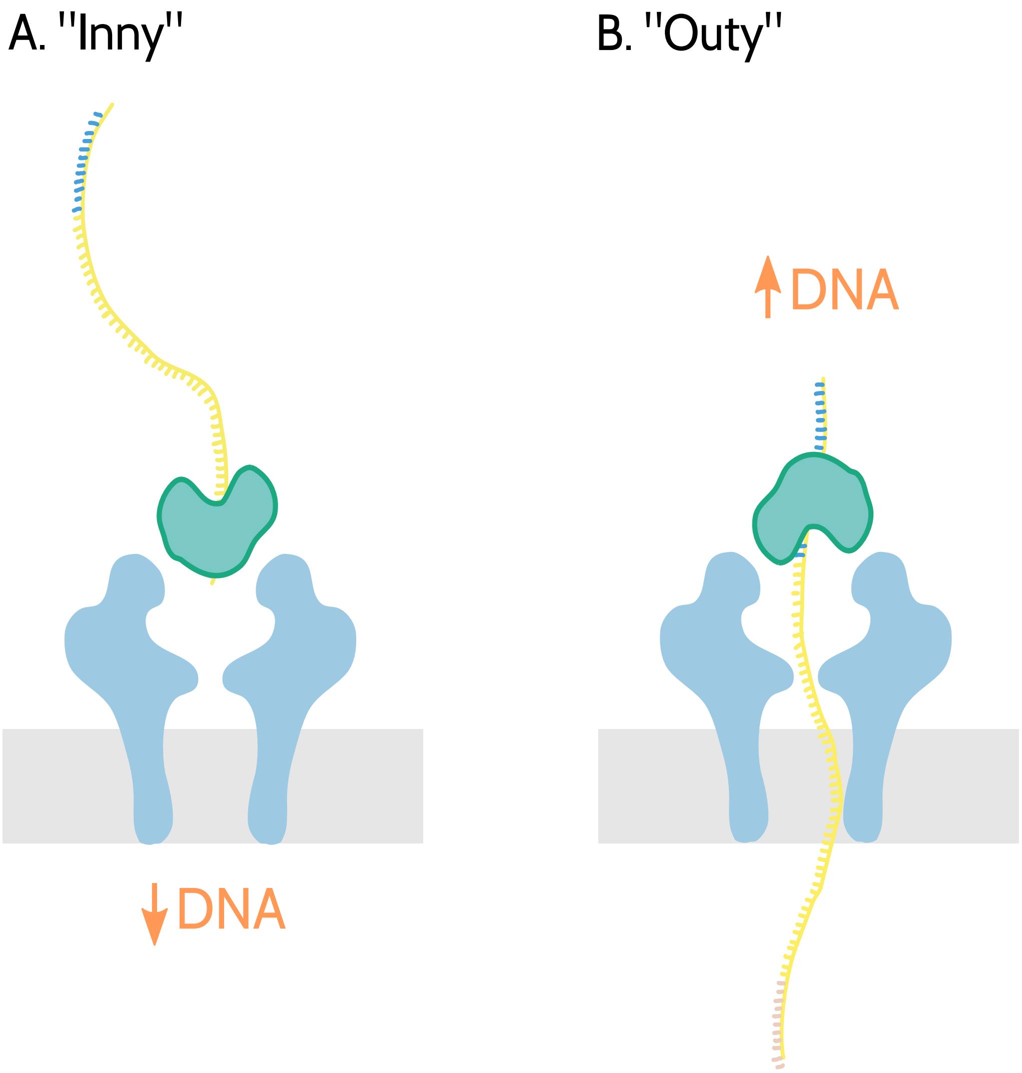
https://doi.org/10.5281/zenodo.5832041

General Discussion.



Supplementary Figure GD.S1. Benchmark of different bioinformatic tools tested on the ZymoBIOMICS Microbial Community Standards CS (described in **Chapter II**). (A) Pearson correlation coefficient (rho) values from comparing the theoretical composition to the relative abundances detected by each tool. (B) Absolute deviation from the theoretical composition. (C) Total number of sequences assigned. 16S rRNA gene amplification and sequencing was performed as reported for **Chapter IB**. Data preprocessing was carried out as described in Pipeline 1 (see **subsection 1.6. of General Materials and Methdos**). Pearson correlation was calculated with the cor.test R package. Deviations on the theoretical composition were calculated by applying the following formula: Σ | log2 (detected abundance / theoretical abundance) |.

Bioinformatic tools: BLAST and UCLUST were used within QIIME (http://qiime.org/); the classify-sklearn naïve Bayes taxonomy classifier was used with QIIME 2 (https://qiime2.org/); MOTHUR (https://mothur.org/) and MAPseq (https://github.com/jfmrod/MAPseq) were used with default parameters.



Supplementary Figure GD.S2. ‘Inny’ vs. ‘Outy’ sequencing mode. In the ‘Inny’ mode (currently implemented in Oxford Nanopore Technologies devices), the motor protein directly moves the DNA molecule across the membrane. In the ‘Outy’ mode, the DNA molecule passes through the nanopore until the motor protein stops the translocation. Translocation occurs rapidly, as the process is not controlled by the motor protein. Therefore, changes in the electrical signal cannot be converted into nucleotide sequences, but can be used to size the DNA fragment. Then, the motor protein moves the DNA molecule across the membrane in the opposite direction, and the fragment can be read.

Appendix C

Publications related to this thesis

Publication I

Bioresource Technology 265 (2018) 275–281

Contents lists available at ScienceDirect

Bioresource Technology

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

Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass

Justus Hardegen[a], Adriel Latorre-Pérez[b], Cristina Vilanova[b], Thomas Günther[c], Manuel Porcar[b,d], Olaf Luschnig[e], Claudia Simeonov[a], Christian Abendroth[a,d,f,*]

[a] Robert Boyle Institut e.V., Jena, Germany
[b] Darwin Bioprospecting Excellence, S.L., Paterna, Valencia, Spain
[c] Eurofins Umwelt Ost GmbH, Jena, Germany
[d] Institute for Integrative Systems Biology (I2SysBio), Paterna, Valencia, Spain
[e] Bio H2 Umwelt GmbH, Jena, Germany
[f] Technische Universität Dresden, Chair of Waste Management, Pratzschwitzer Str. 15, Pirna, Germany

ARTICLE INFO

Keywords:
Sewage sludge
Sewage sludge digestion
Nanopore sequencing
Co-factor F420 quantification
Wastewater treatment plants
Anaerobic digestion

ABSTRACT

In this work, liquid and solid fractions of grass biomass were used as co-substrates for anaerobic co-digestion of sewage sludge. The input of grass biomass was increased gradually, and the underlying methanogenic microbiome was assessed by means of microscopy-based cell counting and full-length 16S rRNA gene high-throughput sequencing, proving for the first time the suitability of nanopore-based portable sequencers as a monitoring tool for anaerobic digestion systems. In both cases co-fermentation resulted in an increased number of bacteria and methanogenic archaea. Interestingly, the microbial communities were highly different between solid and liquid-fed batches. Liquid-fed batches developed a more stable microbiome, enriched in *Methanosarcina* spp., and resulted in higher methanogenic yield. In contrast, solid-fed batches were highly unstable at higher substrate concentrations, and kept *Methanosaeta* spp. – typically associated to sewage sludge – as the majoritary methanogenic archaea.

1. Introduction

Anaerobic digestion is a well-known technology that allows microbial conversion of biomass into methane and carbon dioxide. Often the process is combined with physical, chemical or biological pretreatments, which facilitate the subsequent biomass degradation (Rani et al., 2012; Kannah et al., 2017; Kavitha et al., 2017). Basically, anaerobic fermentation consists of four phases (Bischofsberger et al., 2005): hydrolysis (biomass fragmentation), acidogenesis (formation of organic acids, alcohols, carbon dioxide, and hydrogen), acetogenesis (formation of acetic acid), and methanogenesis (last phase of the process, in which acetic acid, hydrogen, and carbon dioxide are the main substrates for the formation of methane). The key microorganisms in the methanogenesis phase are methanogenic archaea, whose composition depends on the operation conditions and strongly changes when co-digestion with additional substrates occurs (Sundberg et al., 2013). Mesophilic methanogens that can be found in especially high abundances belong to the genus *Methanosaeta, Methanoculleus*, and *Methanosarcina* (Abendroth et al., 2015, 2017a).

A strong microbial shift can be observed under thermophilic conditions, in which methanogens such as *Methanothermobacter* or *Methanobacterium* show an increased abundance (Maus et al., 2016; Lin et al., 2017; Xiao et al., 2018). Besides temperature, methanogens are also very sensitive to the organic loading rate. For example, under mesophilic conditions digestion processes with high amounts of chemical oxygen demand (COD) tend to have high amounts of *Methanosarcina* and *Methanoculleus*. On the other hand, sewage sludge, which has typically lower amounts of COD compared to typical industrial co-digesters, tends to have higher amounts of archaea corresponding to the genus *Methanosaeta* (Abendroth et al., 2015, 2017a).

Even though there is a basic understanding of the distribution of methanogenic genera under certain digestive conditions, there are still some gaps remaining. For example, the gradual increase of COD in sewage sludge by means of co-digestion with other substrates has not been sufficiently characterized. However, as the application of co-substrates in sewage mono digesters could lead to a substrate overload, it is important to characterize the transition of a typical sewage microbiome into a high-performance microbiome. According to the current state of art, the mentioned microbial transition is of high interest for the scientific community in this field, as this process is of crucial importance

[*] Corresponding author at: Robert Boyle Institut e.V., Jena, Germany.
E-mail address: abendroth.ca@gmail.com (C. Abendroth).

https://doi.org/10.1016/j.biortech.2018.06.005
Received 28 March 2018; Received in revised form 30 May 2018; Accepted 4 June 2018
Available online 06 June 2018

for an efficient transition from anaerobic sewage sludge mono-digestion to co-digestion. With aims to reach the climate objectives of the European Union for the 21th century, researchers are continuously investigating new technologies and methodologies that might help to build a green and self-sustainable economy (European Commission, 2007). In this context, a very promising approach is to increase the efficiency of wastewater treatment plants by using existing sewage sludge digesters, in which co-digestion is implemented. Such a technological upgrade would allow efficient and local usage of organic waste sources, which are produced by surrounding communities and industries. In addition, it would be a further step towards powering self-sufficient wastewater treatment plants. Based on this idea, a number of works is showing the possibility to upgrade anaerobic sewage sludge digesters by using co-substrates. The proposed co-substrates include grass biomass (Hidaka et al., 2016; Abendroth et al., 2017a,b), food waste (Zahan et al., 2016), municipal solid waste (Cabbai et al., 2016), glycerol (Jensen et al., 2014), microalgae (Mahdy et al., 2015), or pear residues (Arhoun et al., 2013). To facilitate the transition from anaerobic sewage sludge mono-digestion to co-digestion, and to meet the high standards of wastewater treatment plants regarding process stability, a better understanding of the microbial changes occurring during the transition from typical sewage digestion to high-load digestion processes is still needed. This work aimed to investigate the impact of slowly increasing concentrations of COD on the underlying microbiome of sewage sludge digesters. The impact of different feeding strategies (feeding with liquids or solids) was also analysed. On the one hand, lignocellulose from fresh grass biomass was mechanically treated, separated from liquids, and used for the solid feeding strategy. On the other hand, grass liquor (after separation from the solid fraction), was used for the liquid feeding strategy.

A powerful tool to investigate microbiome changes is 16S rRNA gene amplicon high-throughput sequencing or shotgun metagenomic approaches (e.g. Vanwonterghem et al., 2014; Abendroth et al., 2017b), since they enable the detection of thousands of species in one single experiment, and can also yield information on the metabolic pathways underlying the biogas production process. Within the present work, it was aimed aimed to apply the recently developed ONT™MinION sequencing platform, as this technology could have the potential to become a suitable monitoring tool for anaerobic digestion plants. The use of metagenomic sequencing as a monitoring tool for industrial processes (i.e. fermentations) has not been sufficiently explored to date.

One of the main reasons for this is the economic investment needed to acquire a sequencer, as well as the technical complexity of the sequencing process and the ulterior bioinformatic analysis. This process is typically simplified by submitting samples to specialized sequencing facilities. Unfortunately, this makes the whole procedure significantly slower (results are typically obtained after some weeks). However, the launch of new generation, portable sequencers opens up a new scenario for real-life sequencing applications. Therefore, the features of this technology (user-friendly operation, real-time analysis, and portability) prove the unprecedented impact in the clinical (Quick et al., 2017), biosecurity (Pritchard et al., 2016), and environmental (Brown et al., 2017) fields. This work assesses for the first time the suitability of the ONT™MinION platform as a monitoring tool for anaerobic digestion systems, and uses this technology to follow up the changes in the archaeal methane-producing community at nearly real time.

2. Material and Methods

2.1. Substrate

Fresh grass biomass was chosen to be used as substrate (Gramineae). In was collected from the front garden of the Robert Boyle Institute and stored at 0° after pretreatment. To separate lquids from solids a conventional juicer (Angel Juicer 8500 s, Angel Co. LTD., Korea) was used for pre-treatment of grass biomass. The solid fraction contained a COD of 366 mg O_2 per g of substrate (according to the German guideline DIN 3814-S9), and the liquid fraction had a COD of 82 mg O_2 per g of substrate (according to the German guideline DIN 38409-H41). The produced liquid fraction contained 11.63% of total solids (TS) with 76.63% of volatile solids (VS, percentage of TS). The remaining solid fraction contained 55.80% of TS and 90.16% of VS.

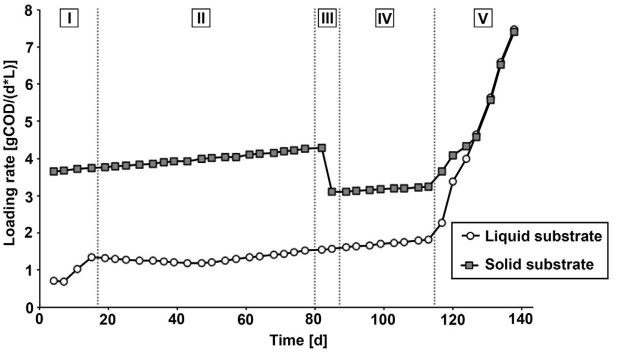


Fig. 1. Substrate load per day and litre in time. The addition of liquid (batch reactions B and C) and solid (batch reactions D and E) substrate was performed in five different phases, as indicated in roman numbers. Phase I: The COD input concentration was adjusted to a value in which similar amounts of biogas (methane) were produced in batches fed with liquid or solid substrates. Phase II: A stable period of feeding occurred. Phase III: solid-fed batches (reactors D and E) reached an extremely high viscosity, and small amounts of water were added to enable stirring. Phase IV: Another phase of stable conditions followed. Phase V: in order to drastically increase the organic loading rate, the substrate was changed to molasses in all the reactors.

2.2. Seed sludge

As seed sludge of choice sewage sludge from a mesophilic anaerobic digester of a municipal wastewater treatment plant in Jena (Germany) was used. After collection the sludge was stored for one week. The sludge contained 3.52% of TS and 56.30% of VS.

2.3. Semicontinuous batch digestion

Five batch reactors (A–E) were set up according to the German guideline VDI 4630. Feeding occurred semi-continuously with gradually increasing loading rates as shown in Fig. 1. Incubation occurred at 37 °C without stirring and incubation bottles were agitated manually before biogas measurements or sampling. Reactor A was used as a negative control (without substrate input); reactors B and C were fed with liquid substrate (liquids separated from grass biomass); and reactors D and E were fed with solid substrate input (remaining solids after liquid separation). Each reactor was filled with 300 mL of sewage sludge. The reactors were opened every 3–4 days – twice a week – to take samples and to add substrate. Afterwards, the reactors were closed and flushed with nitrogen to ensure an anaerobic atmosphere. Gas was collected in a liquid displacement device (eudiometer) and the volume of biogas, as well as the ratio of CO_2 and CH_4 was measured daily. Produced gas was analysed using the “COMBIMASS 99 GA-m” gas measurement device (Binder, Germany) to determine the ratio of CO_2 and CH_4.

The amount of total volatile fatty acids (TVFA) and the solubilisation of COD were monitored using conventional photometer-based assays (Nanocolor CSB15000 and Nanocolor organische Säuren 3000, Macherey-Nagel, Germany).

In the beginning (Fig. 1, phase I), the loading rate was adjusted in such a way that liquid- and solid-fed reactors produced similar amounts of methane. After running the reactors with a constant input (phase II), solid-fed batches (reactors D and E) reached an extremely high viscosity, and 100 mL of water were added to enable stirring (phase III).

After another phase of stable conditions (phase IV) and from rom day 113 onwards, the input of both liquid and solid substrates was reduced by 25% each cycle, reaching a reduction of 100% at day 124. At the same time, the grass biomass was replaced with molasses to induce a shock loading. The molasses input was increased by 0.5 g per cycle (phase V).

2.4. Fluorescent microscopy

Prokaryotes were quantified after staining with 4′,6-diamidino-2-phenylindole (DAPI) using a epifluorescent microscope (Axio Lab.A1, Carls Zeiss, Germany). Firstly, Teflon-coated slides with 10 wells (Carl Roth, Germany) were covered with a gelatine membrane. To do this, 10 μL of gelatine solution (0.1% gelatine und 0.01% $CrK(SO4)_2$) were pipetted on each well and dried at 50 °C for 10 min. Depending on the density of cells, samples were diluted 1:200 with PBS buffer. Each well was filled with 10 μL of the diluted sample and dried at 50 °C for 10 min. Finally, 2.5 μL of fluorescent mounting solution (RotiR-Mount Aqua, Carl-Roth, Germany) were applied. Quantification was performed under 400 × magnification and 450 ms exposure time. For each time point, 32 pictures ware taken and evaluated using the ImageJ software. Excitation occurred with wavelengths ranging from 360 to 375 nm and only emitted light with wavelengths above 400 nm was collected.

Methanogenic archaea were quantified using the same microscope, but with a different set of optical filters and an excitation wavelength adjusted to the quantification of the cofactor F420 (which is associated with methanogenic archaea). Excitation occurred with wavelengths ranging from 400 to 440 nm and only emitted light with wavelengths between 500 nm and 550 nm was collected. Samples were diluted 1:2 with a mounting solution (10 μL each) (RotiR-Mount FluorCare, Carl-Roth, Germany) and 3 μL of the suspension were applied between the cover slip and the slide. Pictures were taken with 400 × magnification and 126 ms exposure time. For each time point, 48 pictures ware taken and evaluated using the ImageJ software.

2.5. DNA isolation

In order to reduce the amount of inhibiting substances, biomass samples were sedimented by centrifugation (10 min at 20,000*g*) and washed several times with sterile PBS buffer until a clear supernatant was observed. Then, metagenomic DNA was isolated using the Power Soil DNA Isolation kit (MO BIO Laboratories) following the manufacturers instructions. The quantity and quality of the DNA was determined on a 1.5% agarose gel and with a Nanodrop-1000 Spectrophotometer (Thermo Scientific, Wilmington, DE).

2.6. 16S rRNA gene amplification and barcoding

The full-length 16S rRNA gene of archaea was PCR-amplified using universal primers Arch8F (5′-TCCGGTTGATCCTGCC-3′) and Arch1492R (5′-GGCTACCTTGTTACGACTT-3′), for which specificity had been previously reported (Klindworth et al., 2013). Primer sequences were tailored to add the ONT™ Universal Tags (5′-TTTCTGTT GGTGCTGATATTGC-3′ for the forward primer and 5′-ACTTGCCTGTC GCTCTATCTTC-3′ for the reverse primer) to their 5′ ends. These universal tags allowed the barcoding of the amplicons in the second PCR using the ONT^{TM} PCR Barcoding kit (EXP-175 PBC001).

For the first PCR, the mixture consisted of 1 × Taq Polymerase Buffer, 200 μM dNTPs, 200 nM primers, 1 U of Taq DNA polymerase (VWR), and 10 ng of DNA template in a final volume of 50 μL. PCR conditions were an initial denaturation step at 94 °C for 1 min, followed by 35 cycles of amplification (denaturing, 1 min at 95 °C; annealing, 1 min at 49 °C; extension, 2 min at 72 °C), with a final extension at 72 °C for 10 min. To assess possible reagent contamination, each PCR reaction included a negative control without template DNA, which did not amplify. A purification step using Agencourt AMPure XP beads (Beckman Coulter) at 0.5 × concentration was performed to remove primer-dimers and non-specific amplicons, and the resulting DNA was recovered and assessed by Qubit quantification.

In the second PCR, the mixture contained 0.5 nM of the first PCR product, 1 × Taq Polymerase Buffer, 200 M of dNTPs, 1 U of Taq DNA polymerase (VWR), and the corresponding specific barcode (EXP-PBC001) as recommended in the ONT protocol 1D PCR barcoding amplicons (SQK-LSK108). The PCR conditions consisted of an initial denaturation step for 30 s at 98 °C, followed by 15 cycles at 98 °C for 15 s, 15 s at 62 °C for annealing, 45 s at 72 °C for extension, and a final extension step for 7 min at 72 °C. A clean-up step using AMPure XP beads at 0.5 × concentration was used again to discard short fragments as recommended by the manufacturer. Finally, an equimolar pool of amplicons was prepared for the subsequent library construction.

2.7. Library construction and sequencing

The Ligation Sequencing Kit 1D (SQK-LSK108) was used to prepare the amplicon library to load into the MinION™ following the instructions of the 1D PCR barcoding amplicon protocol of ONT. The barcoded pool of amplicons (1 μg) was used as input DNA. The DNA was processed for end repair and dA-tailing using the NEBNext End Repair/dA-tailing Module (New England Biolabs), and the resulting DNA was purified using Agencourt AM206 Pure XP beads (Beckman Coulter) and assessed by Qubit quantification. For the adapter ligation step, a total of 0.2 pmol of the end-prepped DNA was added to a mix containing 50 μL of Blunt/TA ligase master mix (New England Biolabs) and 20 μL of Adapter Mix (SQK-LSK108), and was incubated at room temperature for 10 min. DNA was purified again with the Agencourt AMPure XP beads (Beckman Coulter) and the Adapter Bead Binding buffer provided on SQK-LSK108 kit to finally obtain the DNA library.

The flow cell (R9.4, FLO-MIN106) was primed and then loaded as indicated in the ONT^{TM} protocols. Sequencing was performed during 12 h using the standard sequencing protocol implemented in the MinKNOW™ software.

2.8. Metagenomic data analysis

Reads were basecalled using the Metrichor™ agent, and sequencing statistics were followed in real time using the EPI2ME debarcoding workflow. The fast5 files obtained were converted to fastq files using poRe (Watson et al., 2015) and adapters were trimmed using Porechop (https://github.com/rrwick/Porechop). The resulting sequences were analyzed with the QIIME software. Briefly, reads were aligned, and then identified through BLAST searches against the latest version of the GreenGenes database (13 8). Data was then further analyzed and represented with custom scripts. In order to detect significant changes in the relative abundance of particular taxa (*Methanosaeta* spp. and *Methanosarcina* spp.), a Welchs *t*-test for unequal variances was performed.

3. Results and discussion

3.1. Reactor performance: liquid vs. solid feeding

Two different strategies for the repowering of sewage sludge involving co-digestion were compared: (1) using a liquid co-substrate with very low percentage of total solids (TS) and, therefore, with low amounts of lignocellulose (batch reactions B and C); and (2) using a solid co-substrate with a very high percentage of TS and, therefore, with high amounts of lignocellulose (batch reactions D and E). Both co-substrates were obtained from fresh grass (Graminidae) biomass. Additionally, a control digester was kept without co-substrate input (batch reaction A). The loading rate of both experimental approaches was adjusted in such a way that both systems produced similar volumes

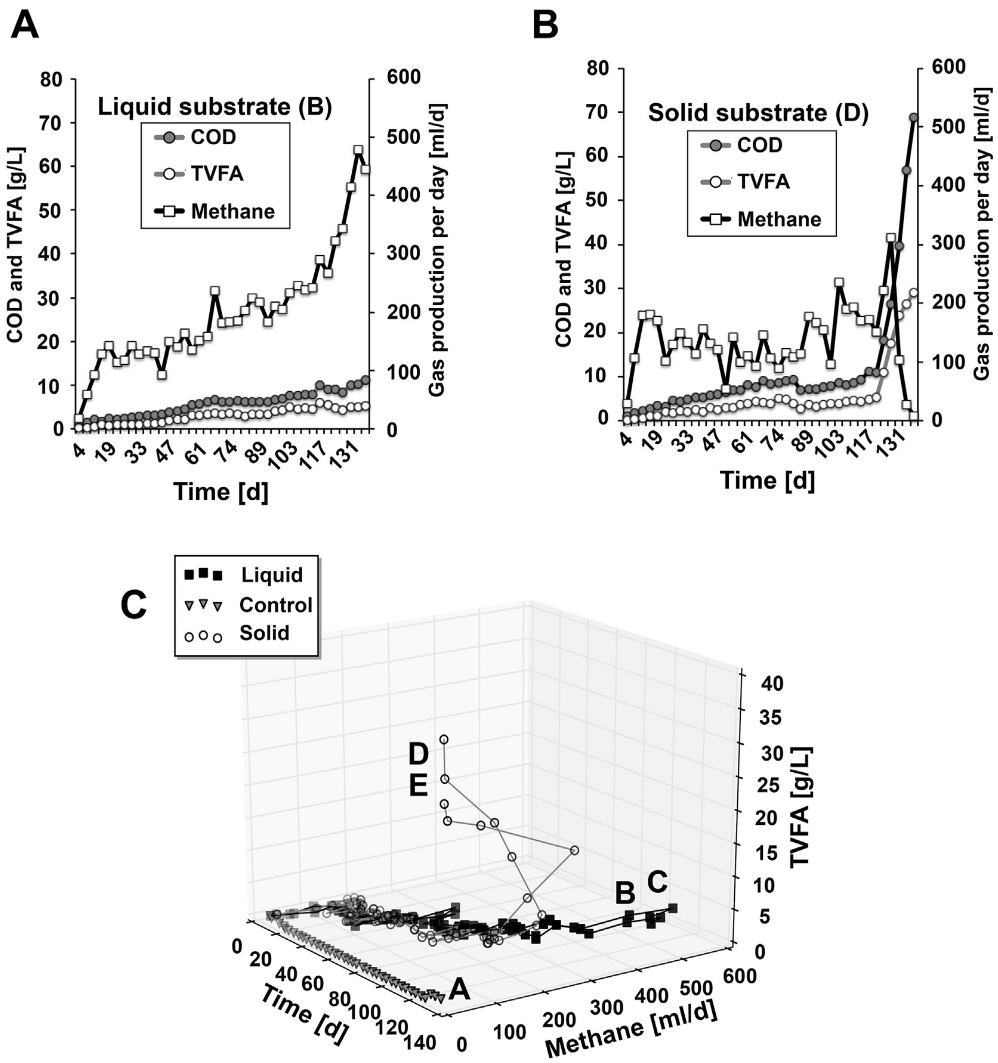


Fig. 2. Chemical parameters measured for the different experimental set-ups. Data on COD, TVFA and methane production are represented for one of the liquid fed reactors (A) and one of the solid fed reactors (B). A tridimensional representation of the evolution of all the reactors is also shown (C).

of methane per working volume, as described in Section 2 (Figs. 1 and 2).

In both co-digestion strategies, the amount of biogas ranged between approximately 100 and 200 mL of methane per day during phase I and II (day 1–82) (Fig. 2). By the end of phase II the liquid fed batches reached a solubilized COD of 5.9 ± 0.2 g COD/L, and a TVFA concentration of 2.81 ± 0 g TVFA/L. Although the solubilized COD and TVFA of the solid fed system were higher at that time (12.02 ± 2.98 g COD/L, and 3.98 ± 0.02 g TVFA/L), the produced amount of methane was slightly higher in the liquid fed system (Fig. 2A and B). Moreover, methane production within the liquid fed system proved more stable in time. By the end of phase II, the digestion sludge in the solid fed system reached such high viscosity that no stirring was possible. In order to ensure a better substrate distribution and to facilitate to movement of bubbles, a small amount of water was added to the solid fed batch systems D and E (100 mL). Due to the dilution, the loading rate of the solid fed batches was slightly reduced during phase IV (Fig. 1).

Since the high viscosity of the solid fed batch prevented any further increase of the loading rate, the substrate was changed stepwise to molasses for both digestion experiments (liquid and solid fed batches), starting at day 117 (phase V). In parallel, the loading rate of the other substrates (liquid and solid grass biomass) was lowered stepwise until day, when both experimental set-ups were fed exclusively with molasses.

The solubilized COD increased drastically in both digestion approaches, indicating a substrate overload. However, from day 132 onwards (Fig. 1, phase V), the produced amount of methane became drastically reduced in the solid fed batches (Fig. 2B and C). In contrast, the liquid fed batch systems displayed higher stability, with continuously increasing levels of methane production. Moreover, a sudden acid shock was detected in the concentration of TVFAs in the solid fed batches, reaching more than 20 g TVFA/L at day 132 (Fig. 2B). The liquid fed batches remained with a lower concentration of TVFAs, reaching 5.52 ± 0.54 g TVFA/L at day 32, indicating a much more efficient conversion of TVFAs into methane.

3.2. Changes in the abundance of prokaryotes and methanogenic archaea

Fluorescent microscopy was performed to complement the chemical analysis during the experiments. As described in Section 2, DAPI staining was used to count the total number of prokaryotes, and

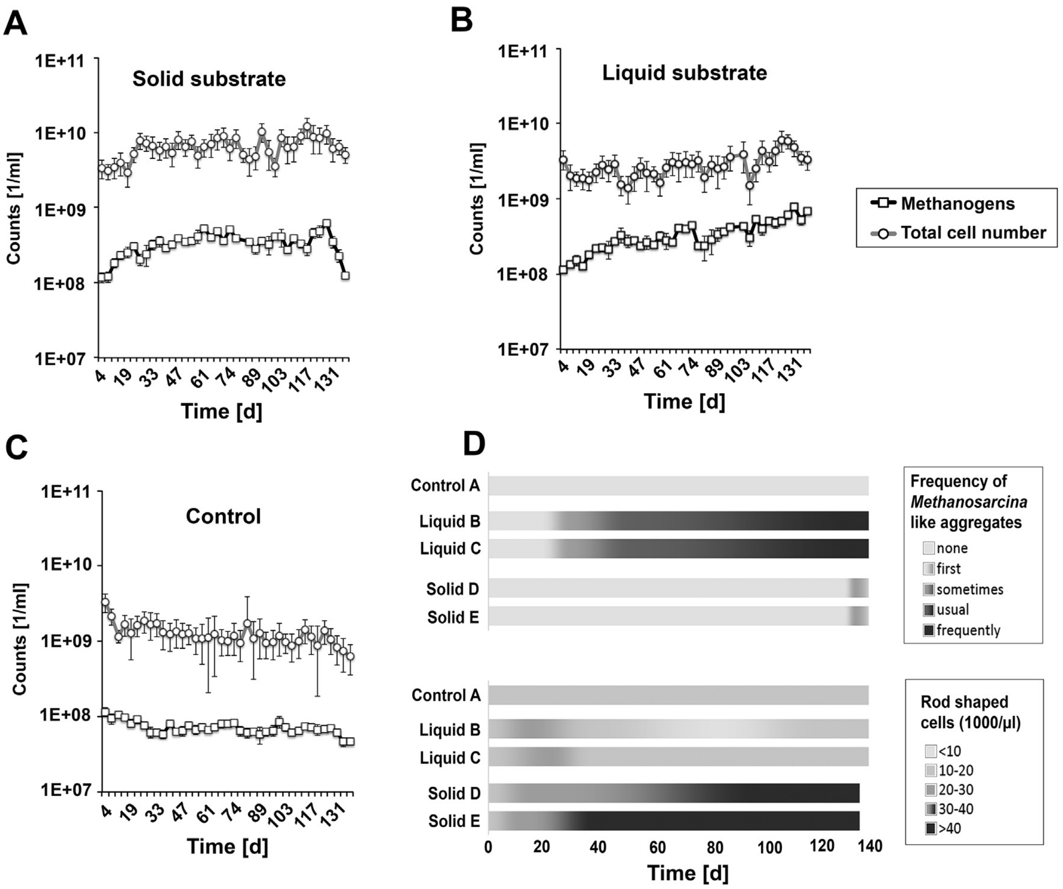


Fig. 3. Microscopic analysis of prokaryotic and methanogenic communities. Methanogens were screened by quantifying the co-factor F420, whereas total Prokaryota were stained with DAPI. Quantified F420- and DAPI signals are shown for a liquid-fed reactor (A), a solid-fed reactor (B), and the unfed control (C). Additionally, *Methanosarcina* spp. like cell aggregates and rod shaped F420-signals were analysed semi-quantitatively (D). High amounts of rod shaped F420-signals were used as indicators for high concentrations of *Methanosaeta*, which is typical for sewage plants and sludges with low COD content.

fluorescence associated to the cofactor F420 was used to quantify methanogenic archaea (Fig. 3). The number of prokaryotes varied between 1E + 09 and 1E + 10 per mL, which is comparable with previous studies (Nettmann et al., 2010). The feeding events in both liquid- and solid-fed batches did not cause a noticeable increase in the number of prokaryotes. The substrate overload at the end of the experiment (Fig. 1, phase 4) did not cause a shift in the number of prokaryotes, indicating a high stability of the underlying bacterial community. However, starvation in the unfed control (batch reaction A) caused a decrease in the number of prokaryotes below 1E + 09. A similar influence of starvation was observed recently in another study, where the effect of starvation was analysed based on fluorescent in-situ microscopy, comparing a wide variety of anaerobic digesters, which were fed with a wide variety of substrates (Abendroth et al., 2016).

The number of methanogens increased continuously from 1E + 08 to 1E + 09 per mL, which is in accordance with other studies (Nettmann et al., 2010). However, in the unfed control the number of methanogens remained unexpectedly stable, indicating a high resistance against starvation. At day 132, the number of methanogens decreased back to 1E + 08 cells in the solid-fed batches, but remained stable in the liquid-fed batches. This is in accordance with the halt in methane production observed in the solid fed batches at day 132 (Fig. 2). Taken together, these results suggest the presence of a better-adjusted microbiome in the liquid fed batches.

3.3. Changes in the composition of the methanogenic microbiome

Changes in the relative abundance of the main genera involved in methane production were followed up by means of microscopy and confirmed by full-length 16S rRNA gene high-throughput sequencing, using specific primers targeting archaea. As shown in Fig. 3D, microscopic analysis revealed that number of rod-shaped methanogens in solid-fed batches kept high throughout the experiment, indicating a high number of *Methanosaeta*, a typical genus observed in sewage sludge (Abendroth et al., 2015, 2017a). Interestingly, liquid-fed batches showed decreasing numbers of rod-shaped methanogens and increasing numbers of *Methanosarcina*-like complexes. Microscopic analysis made it difficult to quantify *Methanosarcina* species, as they tend to form large complexes, which prevent the identification of single cells. *Methanosarcina* is a typical genus found in co-digesters with higher loading rate then sewage sludge (Abendroth et al., 2015).

In parallel, the archaeal communities present in the different reactor configurations were analysed by means of full-length archaeal 16S rRNA gene high throughput analysis using ONT™ MinION sequencing. In accordance with microscopic data, the principal component analysis performed with the archaeal profile of each sample showed a different microbial evolution for solid and liquid-fed reactors (Fig. 4A). Fig. 4B shows the evolution of the taxonomic composition in each experimental condition. At day 29, *Methanosaeta* was the most abundant genus in all cases, accounting for 23–75% of sequences. *Methanosaeta* remained the majoritary genus throughout the experiment in solid-fed batches, as

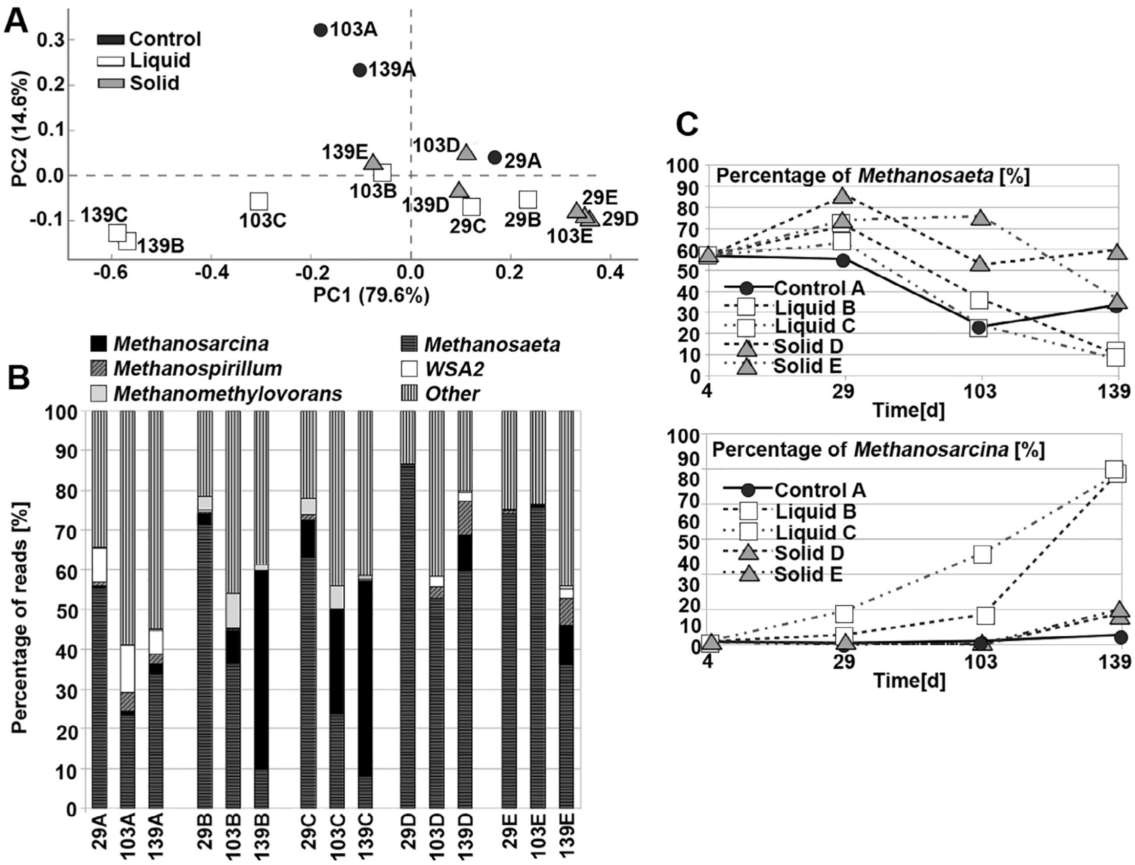


Fig. 4. Taxonomic analysis of the archaeal communities based on full-length 16S rRNA gene high-throughput sequencing. (A) Principal Component Analysis (PCA) based on the taxonomic profile of the samples. (B) Relative frequency of the most abundant taxa in each sample. Numbers in the X axis indicate the day in which samples were taken. Letters indicate the type of feeding (A: unfed control; B and C: liquid-fed batches; D and E: solid-fed batches). (C) Evolution of the frequency of *Methanosaeta* spp. (up) and *Methanosarcina* spp. (down) in the different reactors.

well as in the control digester. However, the abundance of *Methanosarcina* spp. increased in liquid-fed batches, with it becoming the majoritary genus at days 103 (batch C) and 139 (batch B and C). Fig. 4B and C show the decline of *Methanosaeta* spp., and the enrichment of *Methanosarcina spp*. Moreover, the increase in *Methanosarcina* was accompanied by a transient increase of *Methanomethylovorans* at day 29 and 103 (Fig. 4B). Consistently with these findings, earlier work reported a high abundance of *Methanomethylovorans* in sewage digesters which received co-ferments from biodiesel production (Abendroth et al., 2015).

It has to be noted that *Methanosarcina* spp. are methanogens which can be found in co-digesters with high loading rates, especially in the leachate of leach-bed systems (Klocke et al., 2008; Zhao et al., 2013; Abendroth et al., 2015). Therefore, the observed results indicate that the methanogenic microbiome of liquid-fed systems can be successfully shaped and adapted to higher loading rates. This is in contrast with solid-fed systems, which remained enriched in *Methanosaeta* spp., a typical genera from sewage sludge digesters, which are usually operated at lower loading rates (Abendroth et al., 2015).

The presented results show that the successful adaption of microbiomes of sewage sludge to higher loading rates depends on the feeding strategy and is a time consuming process. To ensure stable process conditions, it is recommended to intensively screen the taxonomic profile of industrial sewage digesters, when increasing the loading rate due to the application of co-substrates. As lignocellulose enriched substrates are problematic during the digestion process, it is recommended to use preferred liquid separates and to remove the lignocellulolytic fraction. Alternatively, it is also possible to liquefy the lignocellulolytic fraction prior to the digestion process, as discussed recently by Rajesh Banu et al. (2018).

4. Conclusions

Following the presented results, the microbiome of liquid-fed reactors was re-shaped, changing from a *Methanosaeta* spp.-dominated community, to a *Methanosarcina* spp.-enriched community. For the first time in this field, 16S rRNA gene high-throughput sequencing was performed with an ONT^{TM} MinION sequencer, allowing the tracking of changes in taxonomic data from full-length 16S rRNA gene sequencing. This work reports that the addition of liquid co-substrates resulted in a more effective methanogenic microbiome, and allowed higher biogas production. Altogether, the presented results confirm the high potential for increasing the efficiency of sewage sludge digesters from wastewater treatment plants.

Acknowledgements

We are grateful for proofreading by Kristie Tanner. We are also thankful for funding of the work by the German Ministry of Economic Affairs and Energy (grant nos. 03KB110A, 16KN041331 and 16KN070128).

Conflict of interest

The authors declare no conflict of interest.

References

Abendroth, C., Vilanova, C., Günther, T., Luschnig, O., Porcar, M., 2015. Eubacteria and Archaea communities in seven mesophile anaerobic digester plants. Biotechnol. Biofuels 8, 87.

Abendroth, C., Hahnke, S., Klocke, M., Luschnig, O., 2016. Potential pitfalls of FISH microscopy as assessment method for anaerobic digesters. bioRxiv. http://dx.doi.org/10.1101/054999.

Abendroth, C., Simeonov, C., Peret, J., Antnez, O., Gavidia, R., Luschnig, O., Porcar, M., 2017a. From grass to gas: microbiome dynamics of grass biomass acidification under mesophilic and thermophilic temperatures. Biotechnol. Biofuels 10, 171.

Abendroth, C., Hahnke, S., Simeonov, C., Klocke, M., Casani-Miravalls, M., Ramm, P., Bürger, C., Luschnig, O., Porcar, M., 2017b. Microbial communities involved in biogas production exhibit high resilience to heat shocks (. Bioresour. Technol in press).

Arhoun, B., Bakkali, A., El Mail, R., Rodriguez-Maroto, J.M., Garcia-Herruzo, F., 2013. Biogas production from pear residues using sludge from a wastewater treatment plant digester. Influence of the feed delivery procedure. Bioresour. Technol. 127, 242–247.

Bischofsberger, W., Dichtl, N., Rosenwinkel, K.H., Seyfried, C.F., Böhnke, B., 2005. In: Anaerobtechnik. Springer, Heidelberg, pp. 39–43.

Brown, B., Watson, M., Minot, S., Rivera, M., Franklin, R., 2017. MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience 6 (3), 1–10.

Cabbai, V., De Bortoli, N., Goi, D., 2016. Pilot plant experience on anaerobic codigestion of source selected OFMSW and sewage sludge. Waste Manage. 49, 47–54.

DIN 38409-H41: German Standard Methods for Examination of Water, Waste Water and Sludge; Summary Action and Material Characteristic Parameters (Group H); Determination of the Chemical Oxygen Demand (COD) in the Range over 15 mg/l (H41).

DIN 38414-S9: German Standard Methods for the Examination of Water, Waste Water and Sludge; Sludge and Sediments (Group S); Determination of the Chemical Oxygen Demand (COD) (S 9).

European Commission, 2007. Renewable Energy Road Map Renewable Energies in the 21st Century: Building a More Sustainable Future. COM (2006) 848 final. European Commission, Brussels.

Jensen, P.D., Astals, S., Lu, Y., Devadas, M., Batstone, D.J., 2014. Anaerobic codigestion of sewage sludge and glycerol, focusing on process kinetics, microbial dynamics and sludge dewaterability. Water Res. 67, 355–366.

Kannah, R.Y., Kavitha, S., Rajesh Banu, J., Yeom, I.T., Johnson, M., 2017. Synergetic effect of combined pretreatment for energy efficient biogas generation. Bioresour. Technol. 232, 235–246.

Kavitha, S., Yukesh Kannah, R., Rajesh Banu, J., Kaliappan, S., Johnson, M., 2017. Biological disintegration of microalgae for biomethane recovery-prediction of biodegradability and computation of energy balance. Bioresour. Technol. 244 (Pt 2), 1367–1375.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glöckner, F.O., 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 41, e1.

Klocke, M., Nettmann, E., Bergmann, I., Mundt, K., Souidi, K., Mumme, J., Linke, B., 2008. Characterization of the methanogenic Archaea within two-phase biogas reactor systems operated with plant biomass. Syst. Appl. Microbiol. 31, 190205.

Lin, L., Yu, Z., Li, Y., 2017. Sequential batch thermophilic solid-state anaerobic digestion of lignocellulosic biomass via recirculating digestate as inoculum – Part II: microbial diversity and succession. Bioresour. Technol. 241, 1027–1035.

Mahdy, A., Mendez, L., Ballesteros, M., Gonzlez-Fernndez, C., 2015. Algaculture integration in conventional wastewater treatment plants: anaerobic digestion comparison of primary and secondary sludge with microalgae biomass. Bioresour. Technol. 184, 236–244.

Maus, I., Koeck, D.E., Cibis, K.G., Hahnke, S., Kim, Y.S., Langer, T., Kreubel, J., Erhard, M., Bremges, A., Off, S., Stolze, Y., Jaenicke, S., Goesmann, A., Sczyrba, A., Scherer, P., König, H., Schwarz, W.H., Zverlov, V.V., Liebl, W., Pühler, A., Schlüter, A., Klocke, M., 2016. Unravelling the microbiome of a thermophilic biogas plant by metagenome and metatranscriptome analysis complemented by characterization of bacterial and archaeal isolates. Biotechnol. Biofuels 11 (9), 171.

Nettmann, E., Bergmann, I., Pramschfüer, S., Mundt, K., Plogsties, V., Herrmann, C., Klocke, M., 2010. Polyphasic analyses of methanogenic archaeal communities in agricultural biogas plants. Appl. Environ. Microbiol. 76, 2540–2548.

Pritchard, L., Glover, R., Humphris, S., Elphinstone, J., Toth, I., 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal. Methods 8 (1), 12–24.

Quick, J., Grubaugh, N., Pullan, S., Claro, I., Smith, A., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T.F., Beutler, N.A., Burton, D.R., Lewis-Ximenez, L.L., Goes de Jesus, J., Giovanetti, M., Hill, S.C., Black, A., Bedford, T., Caroll, M.W., Nunez, M., Alcantara Jr, L.C., Sabino, E.C., Baylis, S.A., Faria, N.R., Loose, M., Simpson, J.T., Pybus, O.G., Andersen, K.G., Loman, N.J., 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat. Prot. 12 (6), 1261–1276.

Rajesh Banu, J., Sugitha, S., Kannah, R.Y., Kavitha, S., Yeom, I.T., 2018. Marsilea spp.-a novel source of lignocellulosic biomass: effect of solubilized lignin on anaerobic biodegradability and cost of energy products. Bioresour. Technol. 255, 220–228.

Rani, R.U., Kumar, S.A., Kaliappan, S., Yeom, I.T., Banu, J.R., 2012. Low temperature thermo-chemical pretreatment of dairy waste activated sludge for anaerobic digestion process. Bioresour. Technol. 103 (1), 415–424.

Sundberg, C., Al-Soud, W.A., Larsson, M., Alm, E., Yekta, S.S., Svensson, B.H., Sørensen, S.J., Karlsson, A., 2013. 454 pyrosequencing analyses of bacterial and archaeal richness in 21 full-scale biogas digesters. FEMS Microbiol. Ecol. 85 (3), 612–626.

Vanwonterghem, I., Jensen, P.D., Dennis, P.G., Hugenholtz, P., Rabaey, K., Tyson, G.W., 2014. Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. ISME 8, 2015–2018.

Xiao, Z., Lin, M., Fan, J., Chen, Y., Zhao, C., Liu, B., 2018. Anaerobic digestion of spent mushroom substrate under thermophilic conditions: performance and microbial community analysis. Appl. Microbiol. Biotechnol. 102 (1), 499–507.

Zahan, Z., Othman, M.Z., Rajendram, W., 2016. Anaerobic codigestion of municipal wastewater treatment plant sludge with food waste: a case study. BioMed Res. Int. 2016, 8462928 Epub 2016 Sep 5.

Zhao, H., Li, J., Li, J., Yuan, X., Piao, R., Zhu, W., Li, H., Wang, X., Cui, Z., 2013. Organic loading rate shock impact on operation and microbial communities in different anaerobic fixed-bed reactors. Bioresour. Technol. 140, 211–219.

Publication II

Renewable Energy 146 (2020) 1021–1030

Contents lists available at ScienceDirect

Renewable Energy

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

Ammonia removal during leach-bed acidification leads to optimized organic acid production from chicken manure

Patrice Ramm [1, a, b], Christian Abendroth [1, c, d, *], Adriel Latorre-Pérez [e], Christiane Herrmann [a], Stefan Sebök [a], Anne Geißler [c], Cristina Vilanova [e], Manuel Porcar [e, f], Christina Dornack [c], Christoph Bürger [g], Hannah Schwarz [d], Olaf Luschnig [g]

[a] Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB), Bioengineering, Potsdam, Germany
[b] Institute of Agricultural and Urban Ecological Projects affiliated to Berlin Humboldt University (IASP), Berlin, Germany
[c] Technische Universität Dresden, Institute of Waste Management and Circular Economy, Pirna, Germany
[d] Robert Boyle Institut e.V., Jena, Germany
[e] Darwin Bioprospecting Excellence, S.L. Parc Cientific Universitat de Valencia, Paterna, Valencia, Spain
[f] Institute for Integrative Systems Biology (I2SysBio), University of Valencia-CSIC, Paterna, Valencia, Spain
[g] Bio H2 Umwelt GmbH, Jena, Germany

ARTICLE INFO

Article history:
Received 13 December 2018
Received in revised form
19 May 2019
Accepted 3 July 2019
Available online 4 July 2019

Keywords:
Biogas production
Anaerobic digestion
Chicken manure
Acidification
MAP-Precipitation
NH_3-Stripping
16S-rRNA sequencing

ABSTRACT

This work demonstrates the suitability of nitrogen removal during anaerobic acidification in batch configuration for a more efficient pre-treatment of chicken manure prior to anaerobic digestion. High loading rates corresponding to a total nitrogen input between 6.3 and 9.5 $g\,L^{-1}$ allowed successful suppression of methanogenic archaea. To eliminate nitrogen, NH_3-stripping and MAP (magnesium ammonium phosphate hexahydrate) precipitation were compared. In spite of decreased cell quantities detected using qPCR, removal of nitrogen caused an increase in volatile fatty acid (VFA) formation from 13 to 19%. The highest nitrogen removal during acidification (up to 29%) was achieved with three consecutives MAP precipitation steps, however, conductivity values were affected too, reaching 53.3 and 53.1 $mS\,cm^{-1}$ after the three consecutive MAP precipitations. Additionally, MAP-precipitation reduced the concentration of important trace elements and 16S-rRNA amplicon sequencing revealed an altered taxonomic pattern, in which especially the bacterial families Marinilabiliaceae, Bacteroidales UCG-001, M2PB4-65 termite group and Idiomarinaceae were impaired. However, in spite of these inhibitory effects, nitrogen removal proved able to prevent unwanted methanogenesis and to enhance the yield of VFAs, and this strategy thus holds great potential for the optimized production of biogas in a two-phase system.

1. Introduction

Poultry farming is one of the most important activities in livestock and a global production of 92.5 million tons of broiler meat is expected in 2018 [1]. These activities in poultry livestock also result in high amounts of poultry manure, which can be problematic due to the high contents of nitrogen, mainly in the form of uric acid and proteins [2]. Poultry manure poses a threat for the environment and it can cause health problems to humans and animals [3]. One of the main routes for the disposal of this toxic waste is through anaerobic digestion [4], a technology that allows conversion of multiple types of biomass into methane [5]. However, despite the robustness of anaerobic microbiomes against process disturbances due to their high microbial redundancy [6], mono-digestion of poultry manure is problematic. During the hydrolysis of complex compounds, the initial step of the anaerobic digestion process, organically-bound nitrogen is released in the form of ammonia, which causes the inhibition of anaerobic microorganisms involved in biomethanation. Contents of total ammoniacal nitrogen above 3–5 $g\,L^{-1}$ in the fermentation liquid are generally reported to cause

[*] Corresponding author. Technische Universität Dresden, Institute of Waste Management and Circular Economy, Pirna, Germany.
E-mail address: abendroth.ca@gmail.com (C. Abendroth).
[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.renene.2019.07.021
0960-1481/© 2019 Elsevier Ltd. All rights reserved.

inhibition [7], but this effect strongly depends on process parameters such as pH, temperature, and the acclimation of the anaerobic microbiome [8]. In order to avoid the formation of high concentrations of toxic ammonia during anaerobic digestion of poultry litter, co-digestion of carbon-rich feedstocks with only marginal shares of poultry manure between 4% and 10% of fresh input substrate is recommended [9].

To reduce the toxic effects of poultry manure, different experimental approaches have been investigated. Several reports described the possibility of adaptive evolution, in order to acclimatize involved microorganisms to high ammonia levels [10,11]. However, microbial adaptation to high ammonia concentrations is a time-consuming process, which might need up to several months [12]. Another option is the application of physico-chemical methods for ammonia removal, such as ammonia stripping [13–15] or precipitation of ammonia and phosphate together with magnesium salts [16]. Stripping of ammonia involves blowing air or biogas through the fermentation liquid, resulting in the transfer of free ammonia into the gas phase. The removal of ammonia is facilitated at high pH and temperature, since ammonic nitrogen is present in the shape of free ammonia under these process conditions [17]. The ammonia can subsequently be captured and recovered from the gas phase by absorption, e.g. with sulphuric acid, to yield ammonium sulphate. MAP (magnesium ammonium phosphate hexahydrate) precipitation, also known as struvite precipitation, is based on the ability of ammonia to bind with magnesium and phosphate [18]. Under weakly alkaline conditions, and if phosphate and magnesium salts are added at required amounts, magnesium ammonium phosphate crystallises, and it can be withdrawn from the liquid phase for further utilisation as fertilizer. Ammonia stripping and MAP precipitation methods are frequently used for wastewater treatment, and have also successfully been applied for the removal and recovery of ammonia from poultry manure [2,19], poultry litter leachate [20] and effluent from anaerobic digestion of poultry manure [21–23] with removal efficiencies above 90%.

An alternative possibility to facilitate anaerobic digestion of poultry manure is the biological pre-treatment in a separated acidification stage [24–26]. Separated acidification allows a rapid formation and, due to inhibition of methanogenic archaea, the accumulation of organic acids becomes possible. The resulting high-strength liquor can be used as energy storage, when stored under anaerobic conditions and, therefore, facilitates energy production on demand [27]. The produced high-strength liquor could also be used to extract volatile fatty acids, which are useful intermediates for biorefinery platforms [28]. Moreover, very high input concentrations can be used during separated acidification. For example, Gijzen et al. used loading rates up to 25.8 g L^{-1} [29].

Another advantage when using solid feedstock as input material is that an acidic high-strength liquor can be produced and separated from solids prior to methanation, for example by using leach bed reactors (LBR). This has recently been demonstrated in an experiment, where solids from grass acidification were successfully removed prior to methanisation [30]. This facilitates pumping, and allows the application of a high-performance methane reactor, such as an anaerobic filter (AF), upflow anaerobic sludge blanket (UASB) or expanded granular sludge blanket (EGSB) for efficient methanation in a second stage [31]. Indeed, the separation of the methanogenic process in two different stages improves process stability, since microbial populations in biofilms or granules of AF, UASB or EGSB reactors are more robust against unfavourable conditions and might tolerate higher inhibitor concentrations [32]. In addition, producing a high-strength liquor with low content of total solids (TS) in a separated hydrolysis and acidification stage can also be combined with ammonia removal strategies such as NH_3-stripping or MAP-precipitation, since these methods usually require low TS levels for efficient application [2]. Based on this idea, the present study aims at demonstrating the possibility of combining leach-bed acidification of poultry manure with ammonia stripping or ammonia precipitation. The objectives of the present work were:

- to increase the production of organic acids and decrease the production of methane from poultry manure in a 7-day leach-bed process coupled with ammonia stripping or MAP precipitation
- to chemically characterize the resulting high-strength liquor, and
- to investigate the influence of the procedure on the underlying microbiome, by means of 16S-rRNA high-throughput sequencing

2. Material and methods

2.1. Raw materials

The solid part of separated digestate was obtained from a local biogas plant in Woltow (Germany) and used as seed sludge. This biogas plant was linked to an egg-producing poultry farm. Due to its high TS content, the separated digestate was suitable to be used for the high-solids fermentation tests in the LBR in the present study. Chicken manure from laying hens was used as the main substrate, which was collected from the egg-producing poultry farm NEN Marth GmbH, Woltow 19, Selpin-Woltow, Germany. Additionally, wheat straw was deployed as the packing material for the leach bed. The applied raw materials were chemically characterized (Table 1).

2.2. Experimental set-up and procedure

2.2.1. Batch acidification of chicken manure

To compare methane production during acidification, preliminary batch acidification experiments were performed in 2 L glass bottles, as described by Abendroth et al. [29]. The produced gas was collected in a eudiometer (liquid displacement systems) and each bottle was flushed with nitrogen gas prior to the

Table 1
Characterisation of the seed sludge and substrates used in this study (TS – total solids, FM – fresh matter; VS – volatile solids; COD – chemical oxygen demand).

Function	Digestate	Chicken Manure	Wheat straw
	Seed sludge	Substrate	Packing material
TS (% of FM)	18.03 ± 1.04	35.71 ± 1.14	88.81 ± 2.40
VS (% of TS)	70.25 ± 2.28	51.30 ± 2.45	94.71 ± 2.73
Total nitrogen (Kjeldahl, g kg^{-1} of FM)	3.10 ± 0.29	8.31 ± 0.82	5.89 ± 1.20
COD (g kg^{-1} of FM)	183.09 ± 15.86	257.18 ± 10.26	1074.92 ± 17.87
pH	8.81 ± 0.10	7.77 ± 0.24	6.91 ± 0.15
Conductivity (mS cm^{-1})	27.41 ± 11.79	54.40 ± 5.19	8.08 ± 3.59

experimentation to ensure an anaerobic atmosphere. The bottle was continuously shaken at 60 rpm to achieve a homogeneous mixture. The experiments were performed without seed sludge in order to inhibit methane production. Different substrate input concentrations up to 396 g L^{-1} were compared, resulting in total nitrogen concentrations ranging from 1.6 to 9.5 g N L^{-1}.

2.2.2. *Acidification of chicken manure in leach-bed configuration*

Leach-bed acidification was performed in custom-made, stainless-steel reactors with a total volume of 100 L (number 1 in Fig. 1a) [32]. The leach-bed, a mixture of 5 kg seed sludge and 5 kg chicken manure, was incubated at mesophilic temperature (37 °C) for 7 days. Seven days were chosen as treatment time in the LBR since this duration was previously found suitable to achieve high volatile fatty acid concentrations and to prevent conversion of acids into methane (data not shown). As packing material for the leach-bed, 0.5 kg wheat straw was used and retained in a strainer. A custom-made storage tank with a total volume of 60 L (number 2 in Fig. 1a) was filled with 20 L of fresh tap water at the beginning of the experiment. During the incubation period of 7 days, the liquor was circulated between the storage tank and the leach-bed reactor with a flow rate of 193.8 mL s^{-1} resulting in a high-strength liquor (leachate). Circulation was achieved by using a pump (Wilden Model P1 Plastic/P1/PPPPP/WFS/WF/KWF"; Wilden Pumps Grand Terrace, CA, U.S.A.).

Liquor was pumped from the storage tank and percolated over the leach-bed every 3 h. After passing the leach-bed, the liquor was returned to the storage tank. Percolation was adjusted to achieve a complete cycle of circulation of the liquor each 24 h. The storage tank was also temperated at 37 °C.

The volume of produced gas was measured in a multi-chamber rotor gas meter (TG 05/5 model, Dr.-Ing. Ritter Apparatebau GmbH & Co. KG, Bochum, Germany) and collected in a gas bag (Tesseraux Spezialverpackungen GmbH, Bürstadt, Germany), where the gas was kept until analysis. The measured biogas volume was normalized to standard conditions (temperature 273.15 K, pressure 1013.25 mbar).

Leach-bed acidification was performed without nitrogen removal (control), and with nitrogen removal by either NH$_3$-stripping or MAP precipitation. Acidification experiments were repeated twice for each configuration (Duplicate A and B).

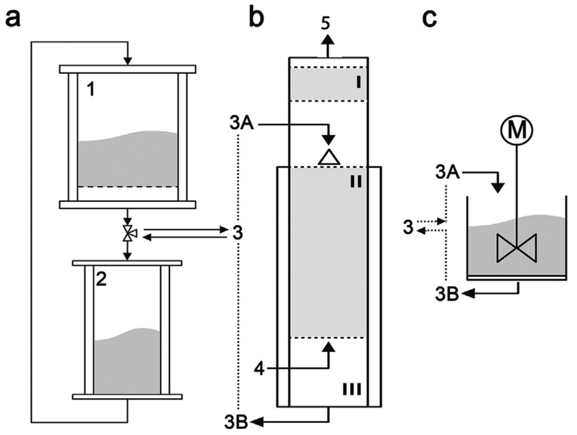


Fig. 1. Experimental set-up: Acidification of chicken manure in a leach-bed configuration (a). Leachate was collected in a storage tank (2), and continuously recirculated into the leach-bed reactor (1). To apply NH$_3$-stripping and MAP precipitation, liquor was removed from leach-bed system and then returned back after nitrogen removal (3); NH$_3$-stripping occurred in a stripping column (b). For this, leachate was pumped into segment I of the column (3A), which was equipped with a droplet separator. From there, the leachate was transferred through a packed bed in segment II. In opposite direction to the leachate flow, atmospheric air was pumped into the column in section III through the packed bed (4), which was used as trickling filter. The treated leachate was collected in the lower part of the column, from where it was removed (3B) and, finally, refilled into the leach-bed acidification (A). NH3-enriched gas escaped through the upper part of the column (5). MAP precipitation occurred in an open tank coupled to a laboratory stirrer (c).

2.2.3. *NH3-stripping*

For NH$_3$-stripping a custom-made stripping column was used [32]. The column had an inner diameter of 0.168 m and a total length of 2.05 m. It was divided into three segments: segment I = leachate feed, segment II = packed bed, and segment III = gas carrier feed (Fig. 1b). The packed bed of the column contained a loose bulk of plastic rings (Pall-Ring 15; Raschig GmbH, Ludwigshafen, Germany) showing a diameter of 0.015 m, a specific surface of 350 m^2 m^{-3}, and a porosity of 0.88 m^3 m^{-3}. The bulk height was 0.81 m. The column was heated with a water bath (Lauda Alpha RA8; Dr. R. Wobser GmbH & Co. KG, Lauda-Königshofen) up to 90 °C.

The column was loaded with approximately 0.02 m^3 h^{-1} of leachate from the storage tank of the leach-bed acidification (number 3 in Fig. 1a) using a pump (Watson-Marlow 323du Drive 400 RPM EU; Spirax-Sarco Engineering group, Falmouth Cornwall, England). To facilitate ammonia release, the pH of the leachate was adjusted to 11.6 using approximately 50 g L^{-1} of sodium hydroxide (32%).

Released ammonium was removed from the column by a constant air stream, which was used as gas carrier. The gas carrier was pumped at a rate of 20 m^3 h^{-1} through the stripping column, using an air ventilator (RL65-21/14; ebm-pabst, St. Georgen, Hungary). After passing through the stripping column, leachate was taken from the bottom of the column and returned back into the storage tank, the pH was re-adjusted to its former value using approximately 12.3 g L^{-1} of sulphuric acid (75%).

NH$_3$-stripping was performed only once during the acidification of chicken manure (the treatment was applied after three days of the 7-day leach-bed incubation time). In the applied treatment procedure, the leachate was passed two times through the stripping column to achieve an appropriate removal of ammonia.

2.2.4. *MAP precipitation*

For the MAP precipitation (precipitation of $MgNH_4PO_4 \bullet 6H_2O$), an open tank coupled to a laboratory stirrer for continuous mixing was used (Fig. 1c). The working volume of the tank was 10 L. The stirrer was a paddle agitator, which was powered by a Eurostar 40 digital stirrer drive (IKA Werke GmbH & Co. KG, Staufen, Germany). The reaction was conducted at a constant temperature of 22 °C.

For each MAP precipitation step, 30.5 g L^{-1} of potassium hydrogenphosphate and 89 g L^{-1} of magnesium chloride hexahydrate (both previously dissolved in deionised water) were applied to the leachate taken from the storage tank of the leach-bed acidification (number 3 in Fig. 1a). After 30 min of incubation with agitation (100 rpm), a sedimentation step of 60 min without agitation was carried out in order to achieve a separation between the precipitated MAP crystals and the leachate. After sedimentation, the supernatant was released from the open tank by a lateral outlet and returned back into the storage tank. The MAP sludge was discarded. The dry weight of the MAP sludge was approximately 4.7 w/w-% of the treated leachate.

Two different MAP precipitation experiments were performed: in one of them, MAP precipitation was applied only once, after three days of the 7-day leach-bed incubation time (further referred as 1xMAP-P); while in the other, MAP precipitations were performed after days 1, 2 and 4 (further referred as 3xMAP-P).

2.3. Chemical analysis

pH and conductivity were analysed using the WTW pH/Cond 340i device (WTW, Weilheim, Germany), equipped with the WTW pH Electrode SenTix41 and the WTW TetraCon 325 electrodes. The TS, VS, NH_4–N, total nitrogen (Kjeldahl) and trace elements were analysed according to VDLUFA guideline (1983/2006) [33]. The COD was analysed according to the guideline DIN 38409–41:1980–12. Volatile fatty acids (VFA), including acetic, propionic, butyric, iso-butyric, valeric, iso-valeric, and caproic acid, were analysed as described by Ref. [34].

The composition of the produced biogas was analysed on a daily basis using the portable gas analyser GA 2000 (Geotechnical Instruments (UK) Ltd., Warwickshire, England).

2.4. DNA-analysis

2.4.1. DNA isolation

DNA isolation was performed in accordance to the proceedings established in the labs of the ATB, as described by Nettmann et al. [35] and Bergmann et al. [36]. For each DNA isolation, 500 μL of the samples were used. A FastDNA Spin Kit for Soil (MP Biomedicals GmbH, Eschwege, Germany) and a FastPrep 24 instrument (MP Biomedicals GmbH, Eschwege, Germany) were used for the isolation of DNA from the sampled liquor. For the elution of the DNA, 100 μL of DES buffer were used.

2.4.2. qPCR

The applied q-PCR method was performed in accordance to Nettmann et al. [35] and Bergmann et al. [37]. The chosen primer and TaqMan sets were BAC fw 5′-ACT CCT ACG GGA GGC AG-3′, BAC rev 5′-GAC TAC CAG GGT ATC TAA TCC-3′, and BAC TaqMan 5′-TGC CAG CAG CCG CGG TAA TAC-3′ for bacteria; and ARC fw 5′-ATT AGA TAC CCS BGT AGT CC-3′, ARC rev 5′-GCC ATG CAC CWC CTC T-3′, and ARC TaqMan 5′-AGG AAT TGG CGG GGG AGC AC-3′ for archaea.

The q-PCR-method was conducted utilizing the 2x qPCR Probe Mix (Bioenzym Scientific GmbH, Hess. Oldendorf, Germany), TaqMan probes, primer-fw/rev (biomers.net GmbH, Ulm, Germany) and a CFX96 Real-Time System (Bio-Rad Laboratories GmbH, München, Germany). The PCR mixture consisted of 10 μL 2x pPCR Probe Mix, 1.8 μL primer-fw, 1.8 μL primer-rev, 0.4 μL TaqMan probe, 1000 pg of sampled DNA (2 μL of the isolated DNA with a concentration of 500 pg μL^{-1}) and 4 μL PCR grade water.

The PCR program for bacteria started with a 7 min step 95 °C, followed by 45 cycles with 15 s at 95 °C for denaturation, 30 s at 57 °C for primer annealing, and 60 s at 60 °C for elongation. The thermocycler program for archaea started with 7 min at 95 °C too and was followed by 40 cycles with 15 s at 95 °C for denaturation, and 60 s at 60 °C for primer annealing and DNA elongation.

Evaluation of the PCR results was supported by the program CFX Manager 3.1 (Bio-Rad Laboratories GmbH, München, Germany).

2.4.3. 16S rRNA gene amplification and barcoding

The bacterial full-length 16S rRNA gene was amplified via PCR using the primers S-D-Bact-0008-a-S-16 (5′AGRGTTYGATY MTGGCTCAG3′) and S-D-Bact-1492-a-A-16 (5′TACCTTGTTA YGACTT3′), which were recommended by Klindworth et al. [38]. For barcoding, ONT™ Universal Tags were added to the 5′ end of the forward primer (5′ TTTCTGTTGGTGCTGATATTGC-3′) and to the reverse primer (5′-ACTTGCCTGTCGCTCTATCTTC-3′), using the ONT™ PCR Barcoding kit (EXP-175 PBC001). A mixture of 1 × Taq Polymerase Buffer, 200 μM dNTPs, 200 nM primers, 1 U of Taq DNA polymerase (VWR), and 10 ng of DNA template was used for the primary PCR (final volume: 50 μL). The PCR started with an initial denaturation step at 94 °C for 1 min, followed by 35 cycles of amplification (denaturing, 1 min at 95 °C; annealing, 1 min at 49 °C; extension, 2 min at 72 °C); and with a final extension step at 72 °C for 10 min. A negative control without DNA template was included. Primer-dimers and non-specific amplicons were removed using the Agencourt AMPure XP beads (Beckman Coulter) at 0.5 × concentration. The purified DNA was recovered and assessed using Qubit (Qubit® 2.0 Fluorometer, Thermofisher, Waltham, USA).

In the secondary PCR, amplicons from the primary PCR were used as template with a concentration of 0.5 nM, and mixed with 1X Taq Polymerase Buffer, 200M of dNTPs, 1 U of Taq DNA polymerase (VWR), and the corresponding specific barcode (EXP-PBC001 kit) as recommended in the ONT protocol for 1D PCR barcoding amplicons (SQK-LSK108). The PCR conditions consisted of an initial denaturation step at 98 °C for 30 s, followed by 15 cycles at 98 °C for 15 s, 15 s at 62 °C for annealing, 45 s at 72 °C for extension, and a final extension step at 72 °C for 7 min. AMPure XP beads at 0.5X concentration were again used to discard short fragments as recommended by the manufacturer. Finally, an equimolar pool of amplicons was prepared for the subsequent library construction. Library construction and sequencing was performed as recently described by Hardegen et al. [39].

Three MinION™ runs of 4 h each were carried out to sequence all barcoded samples, using the same flow cell for all of them (R9.4, FLO-MIN106). Recommended ONT™ protocols were followed for priming, loading and washing of the flow cell. The sequencing process was controlled using the MinKNOW™ software (version 1.13.1; standard sequencing protocol).

2.4.4. Bioinformatic and metagenomic analysis

Reads were basecalled in real time using the MinKNOW™ software (version 1.13.1), and sequencing statistics were followed in real time using the EPI2ME debarcoding workflow. Porechop (https://github.com/rrwick/Porechop) was applied for demultiplexing the barcodes and removing the adaptors. The resulting sequences were analysed using the QIIME software. In QIIME, reads were aligned, and then identified through BLAST searches against the latest version (132) of the SILVA database [40].

The taxonomic results were further analysed with R using the phyloseq package (version 1.22.3) [41]. As recommended by the authors of the phyloseq package, taxa which were not detected more than 3 times in at least 20% of the samples were removed, and abundances were standardized to the median sequencing depth in order to correct different library sizes. A Principal Coordinates Analysis (PCoA) was carried out based on Bray-Curtis distances of the different microbial samples. Additionally, a comparative analysis of the microbial profiles observed at day 7 was carried out using DESeq2 package [42].

3. Results and discussion

3.1. Determination of the optimal input concentration for chicken manure

Preliminary acidification experiments with different nitrogen concentrations were performed in batch configuration to assess the undesired production of methane (Fig. 2). Usually, during acidification, no methane is produced due to high loading rates. High loading rates often result in low pH-values, which inhibits methane formation. However, high concentrations of ammonia can buffer the pH in a range where methanogenesis can occur. Because of this, all batch acidifications produced small amounts of methane, which is in concordance with a recent work by Abendroth et al. [29]. However, when exceeding a total nitrogen input of 6.3 g L^{-1}, high concentrations of solubilized COD were achieved and the methane formation was strongly inhibited (Fig. 2), even though the pH

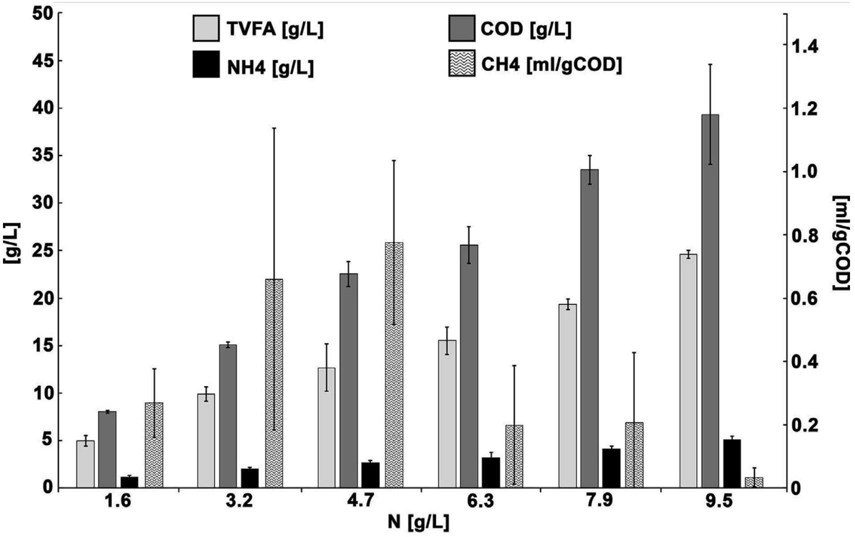


Fig. 2. Chemical characterization of the high-strength liquor produced: Total volatile fatty acids (TVFA), as well as solubilised chemical oxygen demand (COD) and NH_4 are given in g L^{-1}. Methane formation is given in ml CH_4 per g of input COD. Different input concentrations of substrate with a wide diversity of total nitrogen contents (Kjeldahl) are shown.

remained above 6.0 in all cases (data not shown). The inhibition of methanogens might be explained by a high concentration of solubilized substances such as ammonia or TVFA (Fig. 2).

To reach nitrogen input concentrations of 6.3–9.5 g L^{-1}, an input of fresh substrate higher than 0.5 kg L^{-1} was necessary. Since that feeding rate did not allow a proper homogenization in the applied batch configuration, further experiments were performed in a leach-bed configuration, which ensured a very low total solids (TS) content of 3.2% ± 1.4% in the generated high-strength liquor (leachate) (Fig. 3a).

3.2. N-removal during acidification

Applying a leach-bed configuration for further acidification experiments, NH_3-stripping and MAP-precipitation were compared to a control reaction, which was acidified without any nitrogen removal. Usually, nitrogen was removed at day 3, in order to ensure that a significant fraction of nitrogen was converted into ammonia. Additionally, a fourth experiment was performed, where the MAP-precipitation was repeated three times (3x MAP-P). The resulting high-strength liquors were chemically characterized, and the results are shown in Fig. 3.

After 7 days of acidification, the treated liquors always showed lower nitrogen concentrations than the untreated control. The highest nitrogen removal was achieved with the triple MAP-precipitation, where the total nitrogen (Kjeldahl) concentration was about 29% lower compared to the control. The concentration of NH_4—N was even reduced by 38% (Fig. 3b). To enable the comparison of the different removal procedures, the treated process liquids were directly analysed directly before and after treatment (data not shown). Since it is only possible to separate nitrogen in the form of NH_4—N from liquid by the applied procedures, particular attention was paid on the NH4—N removal efficiency. We found NH_4—N removal efficiencies of approximately 55% for NH_3-stripping (initial NH_4—N concentration 2349.6 ± 243.0 mg kg^{-1}, ammonia removal 1313.1 ± 401.4 mg kg^{-1}) and 53% for single MAP precipitation (initial NH_4—N concentration 2078.0 ± 97.4 mg kg^{-1}, NH_4—N-removal 1088.9 ± 21.1 mg kg^{-1}) if applied to the process liquor during acidification (results not shown). The three treatments conducted in 3x MAP-precipitation experiment showed efficiencies of 42% (treatment after 1 day of acidification, initial NH_4—N concentration 1661.5 ± 241.1 mg kg^{-1}, NH_4—N-removal 682.9 ± 129.9 mg kg^{-1}), 45% (treatment after 2 days of acidification, initial NH_4—N concentration 1789.0 ± 82.0 mg kg^{-1}, NH_4—N-removal 806.9 ± 3.0 mg kg^{-1}) and 35% (treatment after 4 days of acidification, initial NH_4—N concentration 1888.0 ± 26.9 mg kg^{-1}, NH_4—N-removal 658.0 ± 343.6 mg kg^{-1}), respectively. Therefore, the efficiency of nitrogen removal was clearly influenced by the initial NH_4—N concentration of the treated process liquor. In general, (as expected for the third treatment in the 3x MAP-precipitation, where the duplicates showed high differences) a higher initial NH_4—N concentration of the treated process liquor caused higher removal efficiencies. The high fluctuation of the process liquid in spite of identical operation of the leach-bed system at all experiments and the use of the same raw materials might be attributed to natural variations in microbiology and chemical composition of the raw material. This is typical if complex natural materials are used. The fluctuation in chemical composition of the used raw material is shown in Table 1. Nevertheless, the amount of removed NH_4—N was in a comparable range for the different applied techniques. In spite of showing the lowest efficiencies, the triple application of MAP precipitation in the 3x MAP-precipitation experiment lead to the lowest concentration of nitrogen and NH_4—N in process liquor in the end of the 7 days of acidification. A continuous release of NH_4—N during acidification was observed. This effect caused the difference between the values of NH_4—N concentrations of the process liquid directly after treatment and in the end of the acidification. To give here another example: During the NH_3-stripping, the NH_4—N concentration of the process liquor was approximately 1037 mg kg^{-1} directly after treatment, but 1964 mg kg^{-1} upon finishing the acidification at day 7.

For enhanced comparability of the determined N-removal efficiencies during the conducted experiment with complex and natural raw materials, additionally, single experiments with artificial mixtures consisting only of deionized water and ammonia carbonate were conducted. In these experiments, ammonia removal rates of up to 76.4% and 96.2% were achieved for NH_3-stripping and MAP-precipitation (results not shown). With both techniques, a maximum amount of approximately 1.8 g NH_4—N could be separated from the treated artificial liquor. N removal capacities of

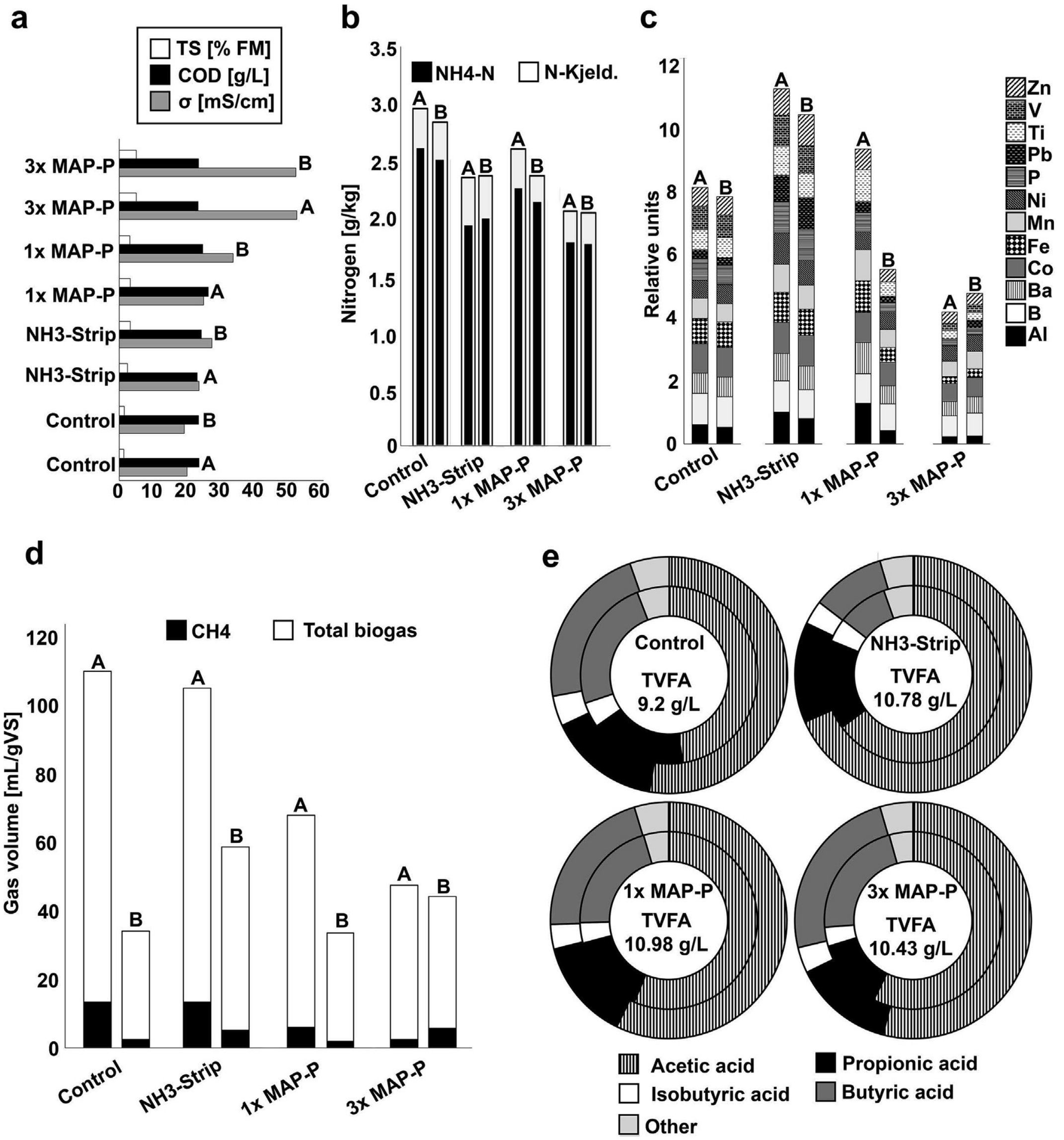


Fig. 3. Characterization of the high-strength liquor produced with the different treatments: total solids (TS), chemical oxygen demand (COD) and conductivity (σ) measurements are shown (a), as well as the total nitrogen content (N-Kjeld.) and nitrogen content corresponding to ammonia (b); to facilitate the comparison of trace elements and heavy metals, their amount is normalized to a value between 0 and 1 (with 1 representing the highest observed abundance for each element). Only the elements showing a decrease after repeated ammonia precipitation (3x MAP) are shown (c); biogas and methane production (d), and the ratio of organic acids are shown in relative units (%), and the total amount of volatile fatty acids (TVFAs) produced is shown in the centre of each pie diagram (e).

approximately 90% have been reported if applying NH_3-stripping and MAP-precipitation for treatment of digestate or wastewater, however, these techniques are quite demanding [43–45]. The efficiency of NH_3-stripping is influenced by pH, temperature, air flow, treatment duration and concentration of VFAs [43,44]. For MAP-precipitation, important influencing factors are suspended solids, dissolved solids, ionic strength, pH and concentrations of calcium and magnesium [45]. In the current study, the practical implementation of the applied techniques “NH_3-stripping” and “MAP-precipitation” represents an appropriate compromise between efficiency and cost.

It has to be stressed that the content of trace elements of the resulting process liquor was also affected, especially in the case of the triple MAP-precipitation (Fig. 3c). The biogas produced strongly varied between duplicates, this observation can also be attributed to the natural fluctuations of the used raw materials as described above. Particularly low biogas productions were detected when three MAP precipitations were performed (Fig. 3d).

Taken together, our results indicate that MAP-precipitation is an effective treatment to prevent methane production during acidogenesis. This could be a consequence of the reduced amount of trace elements during this treatment, which are probably co-precipitated with ammonia (Fig. 3c). During MAP-precipitation, the addition of salt for ammonia precipitation resulted in a high conductivity too (25.4 and 34.2 mS cm^{-1} in the simple treatment, and 53.3 and 53.1 mS cm^{-1} in the triple treatment), which could also explain the reduced methane production, since conductivity values higher than 35 mS cm^{-1} have been previously been reported to inhibit methanogenesis [46]. The conductivity during the other treatment remained below 35 mS cm^{-1}).

The content of TVFAs increased between 13 and 19% when N-removal methods were applied. High ammonia concentrations can inhibit hydrolysis and acidification, and thus the formation of VFAs. Methanogenesis is more susceptible to high ammonia concentrations, but hydrolysis and acidification have also been found to be negatively affected by high ammonia concentrations. Excess ammonia nitrogen may inhibit bacterial growth due to its influence on intracellular pH, and the inhibition of enzyme activities [47]. For example, Niu et al. [48] noticed a 50% inhibition of hydrolysis and acidogenesis at TAN concentrations of 5305 and 5707 mg/L, respectively, when digesting chicken manure under thermophilic conditions. However, values for inhibition levels exhibit a large variation in the scientific literature, and tolerable ammonia concentrations also depend on operating conditions such as the pH and temperature, and acclimatization of the microorganisms [47]. In the present study, ammonia removal during the acidification process correlated with enhanced acidification at high nitrogen input concentrations. TVFA concentrations were similar in all three approaches for N-removal (averaged 10.73 ± 0.28 gTVFAs L^{-1}), but the ratio between different kinds of VFAs differed (Fig. 3e). While in the untreated process liquid the share of acetic acid amounted 59% of TVFAs, the process liquid treated with MAP precipitation and NH_3-stripping showed values of approximately 65% and 74%, respectively. The removal of nitrogen obviously enhanced the formation of acetic acid, which is convenient for a subsequent use of the process liquor in methanisation. The lower share of acetic acid in experiments with MAP-precipitation in comparison to NH_3-stripping might be linked to a reduced amount of trace elements and the increase in conductivity leading to an inhibition of microbial activity.

3.3. Analysing microbial quantities

The absolute abundance of bacteria and archaea was analysed on the last day of each leach-bed experiment by means of qPCR (Fig. 4a and b). The number of bacteria and archaea found in the control is in concordance with other studies [34,37]. However, all experiments, except the control, showed a strong decrease in the number of archaea. This indicates that the harsh N-removal method might contribute to prevent contamination with methanogenic microorganisms. This result is in concordance with a previous report on the acidification of chicken manure [29].

In the case of the triple treatment with MAP-precipitation, qPCR results show that the bacterial community was strongly affected too. Surprisingly, this reduction in the bacterial cell number did not result in an apparent reduction of TVFAs production. However, it has to be noted that TVFAs content reached values of 8.9 and 8.6 g L^{-1} at day 3 (data not shown), indicating that most of the TVFAs were produced before the second MAP-precipitation was conducted.

3.4. Characterization of taxonomic profiles

Besides qPCR, high-throughput sequencing of the 16S-rRNA gene was performed with a MinION (Oxford Nanopore) sequencer. This is the first study in which this device is used to analyse the

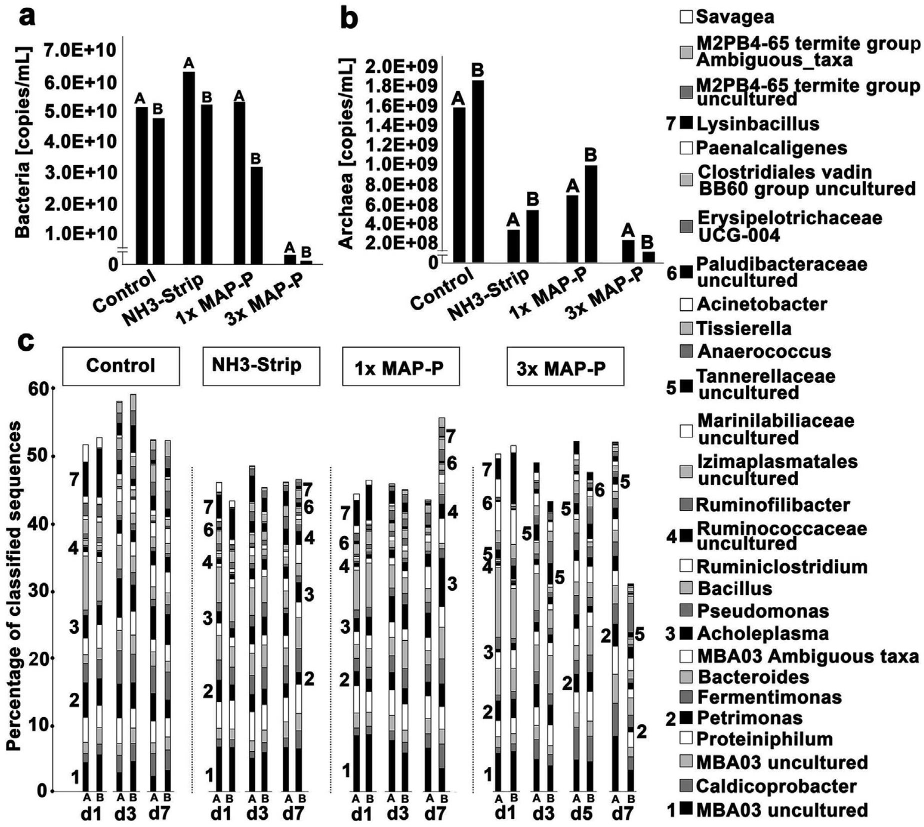


Fig. 4. Microbial characterisation of the high-strength liquor produced with the different treatments. The copy numbers of 16S-rRNA genes of bacteria (a) and archaea (b) were analysed based on qPCR. Taxonomic profiles on genus level were analysed for bacteria based on 16S-rRNA high-throughput sequencing. Only the most abundant genera are shown (c).

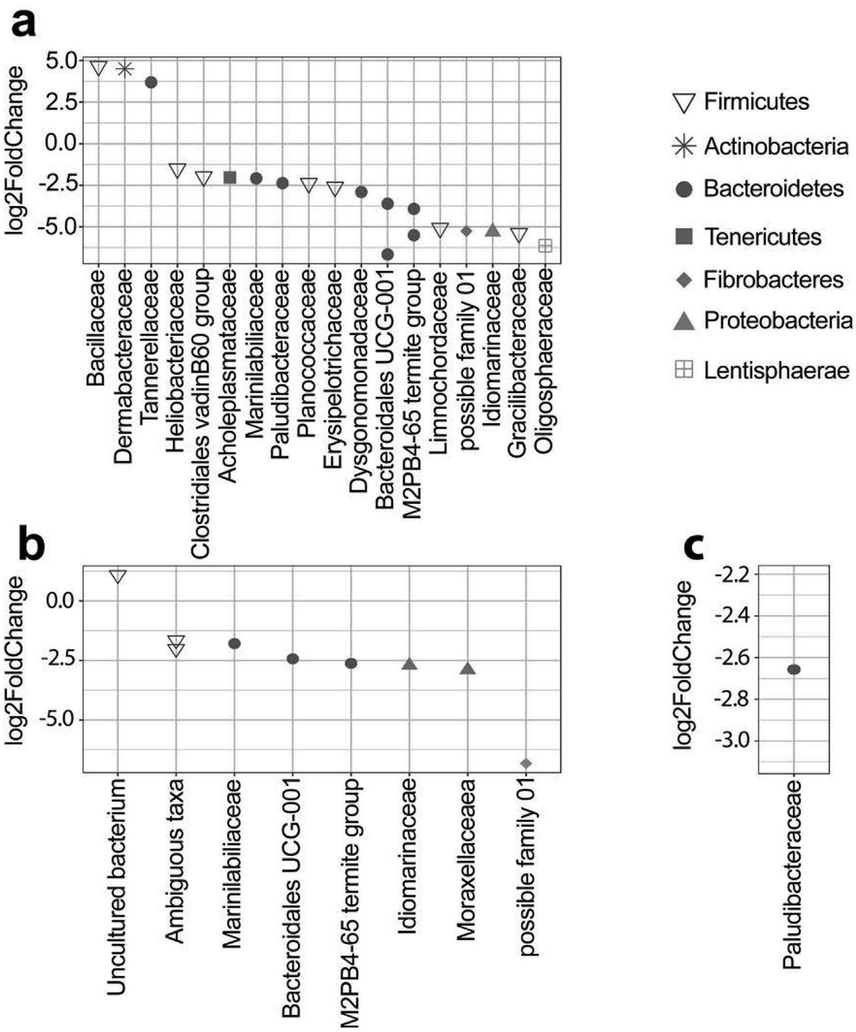


Fig. 5. Microbial taxa that are different within taxonomic profiles associated to nitrogen removal: taxonomic profiles (family level) corresponding to the last day of the triple MAP precipitation (a), single MAP precipitation (b) and NH3-stripping (c) experiments were compared to the control. Only statistically-significant results ($p < 0.05$) are shown.

bacterial fraction of anaerobic digestion experiments. This technology was recently applied for the first time to the archaeal population of anaerobic digestion experiments [39], where MinION sequencing resulted in extra-long amplicon sequences, which were covering the entire 16S-rRNA gene. Due to this advantage, a large fraction of bacterial sequences could be assigned to the genus level. In this study, the most abundant genera detected by sequencing are shown in Fig. 4c. Control experiments as well as NH_3-stripping and MAP-precipitation experiments showed similar microbial profile patterns. However, triple MAP-precipitation resulted in a different pattern. To further investigate the influence of nitrogen removal on the underlying biocenosis, a statistical comparison was performed using the microbial profiles corresponding to the last day of each experiment (Fig. 5). The microbial communities were only slightly affected by NH3-stripping, with a significant reduction in the abundance of only one family (Fig. 5c). The communities proved much more affected by MAP-precipitation, which resulted in a significant change in eight different families (Fig. 5b). In concordance with qPCR results, the most aggressive treatment in terms of bacterial community modification was the triple MAP-precipitation. This denitrification process affected eighteen different families (Fig. 5a). A multivariate analysis performed with the microbial profiles also confirmed that the triple MAP precipitation treatment had the highest impact on the microbiome (Fig. S1).” The families Marinilabiliaceae, Bacteroidales UCG-001, M2PB4-65 termite group and Idiomarinaceae seem to be particularly altered by this denitrification process, as they were altered in both, single and triple MAP-precipitation. Therefore, further studies are recommended to prove the suitability of these families as microbial markers for process disturbance due to nitrogen removal.

4. Conclusions

The presented work shows, for the first time, the suitability of using NH_3-stripping or MAP-precipitation during an acidification treatment of chicken manure in a leach-bed configuration.

In general, all of the investigated procedures showed comparable efficiencies for removal of nitrogen. A triple MAP-precipitation treatment proved the best method; allowing the removal of almost one third of the total nitrogen during acidification, and, due to low TS concentrations, resulted in a high-strength liquor that might be suitable for methanisation in an anaerobic filter, provided that the liquor is supplemented with trace elements (e.g. cobalt or nickel), which are important for subsequent methanisation. The removal of nitrogen lead to an up to 19% increase of production of VFA, however, highest production of acetic acid (8 g per L process liquid) favourable for methanisation was observed at NH_3-stripping.

Based on the promising results, further investigations aiming at the methanisation of a high-strength liquor generated by means of the described techniques are needed.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

We are thankful for the funding received from the German Ministry of Economic Affairs and Energy (grant no. 16KN041338, 16KN041342, 16KN041339 and 16KN041341). We are also thankful to Claudia Wahl, Paul Moritz Wander and the molecular biology lab as well as the chemical analysis lab from the Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB) for their ambitious technical support. Finally, we are thanking Kay Beinersdorf and Andreas Underberg for support with the experimental design.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.renene.2019.07.021.

References

[1] United States Department of Agriculture, Livestock and Poultry: World Markets and Trade. Foreign Agricultural Service, 2018. http://usda.mannlib.cornell.edu/usda/fas/livestock-poultry-ma//2010s/2018/livestock-poultry-ma-04-10-2018.pdf.

[2] F. Abouelenien, W. Fujiwara, Y. Namba, M. Kosseva, N. Nishio, Y. Nakashimada, Improved methane fermentation of chicken manure via ammonia removal by biogas recycle, Bioresour. Technol. 101 (2010) 6368–6373.

[3] A. Nowak, K. Matusiak, S. Borowski, T. Bakuła, S. Opaliński, R. Kołacz, B. Gutarowska, Cytotoxicity of odorous compounds from poultry manure, Int. J. Environ. Res. Public Health 13 (2016) 1064.

[4] B.P. Kelleher, J.J. Leahy, A.M. Henihan, T.F. O'Dwyer, D. Sutton, M.J. Leahy, Advances in poultry litter disposal technology – a review, Bioresour. Technol. 83 (2002) 27–36.

[5] F. Cecchi, C. Cavinato, Anaerobic digestion of bio-waste: a mini-review focusing on territorial and environmental aspects, Waste Manag. Res. 33 (2015) 429–438.

[6] C.M. Spirito, S.E. Daly, J.J. Werner, L.T. Angenent, Redundancy of anaerobic digestion microbiomes during disturbances by the antibiotic monensin, Appl. Environ. Microbiol. (2018), 02692-17.

[7] B. Drosg, Process monitoring in biogas plants, in: Technical brochure, IEA bioenergy task 37, 2013.

[8] Y. Chen, J.J. Cheng, K.S. Creamer, Inhibition of anaerobic digestion process: a review, Bioresour. Technol. 99 (2008) 4044–4064.

[9] G. Bujoczek, J. Oleszkiewicz, R. Sparling, S. Cenkowski, High solid anaerobic digestion of chicken manure, J. Agric. Eng. Res. 101 (2000) 6368–6373.

[10] I.W. Koster, G. Lettinga, Anaerobic digestion at extreme ammonia concentrations, Biol. Wastes 25 (1988) 51–59.

[11] H. Tian, I.A. Fotidis, E. Mancini, I. Angelidaki, Different cultivation methods to acclimatise ammonia-tolerant methanogenic consortia, Bioresour. Technol. 232 (2017) 1–9.

[12] L.P. Wilson, L.H. Loetscher, S.E. Sharvelle, S.K. De Long, Microbial community acclimation enhances waste hydrolysis rates under elevated ammonia and salinity conditions, Bioresour. Technol. 146 (2013) 15–22.

[13] Y. Yao, L. Yu, R. Ghogare, A. Dunsmoor, M. Davaritouchaee, S. Chen, Simultaneous ammonia stripping and anaerobic digestion for efficient thermophilic conversion of dairy manure at high solids concentration, Energy 141 (2017) 179–188.

[14] A. Bonmatí, X. Flotats, Air stripping of ammonia from pig slurry: characterisation and feasibility as a pre-or post-treatment to mesophilic anaerobic digestion, Waste Manag. 23 (2003) 261–272.

[15] N. Krakat, B. Demirel, R. Anjum, D. Dietz, Methods of ammonia removal in anaerobic digestion: a review, Water Sci. Technol. 76 (2017) 1925–1938.

[16] H. Huang, J. Liu, L. Ding, Recovery of phosphate and ammonia nitrogen from the anaerobic digestion supernatant of activated sludge by chemical precipitation, J. Clean. Prod. 102 (2015) 437–446.

[17] S. Guštin, R. Marinšek-Logar, Effect of pH, temperature and air flow rate on the continuous ammonia stripping of the anaerobic digestion effluent, Process Saf. Environ. 89 (2011) 61–66.

[18] S. Kataki, H. West, M. Clarke, D.C. Baruah, Phosphorus recovery as struvite from farm, municipal and industrial waste: feedstock suitability, methods and pre-treatments, Waste Manag. 49 (2016) 437–454.

[19] G. Markou, Improved anaerobic digestion performance and biogas production from poultry litter after lowering its nitrogen content, Bioresour. Technol. 196 (2015) 726–730.

[20] A. Gangagni Rao, T. Sasi Kanth Reddy, S. Surya Prakash, J. Vanajakshi, J. Joseph, A. Jetty, A. Rajashekhara Reddy, P.N. Sarma, Biomethanation of poultry litter leachate in UASB reactor coupled with ammonia stripper for enhancement of overall performance, Bioresour. Technol. 99 (2008) 8679–8684.

[21] A. Muhmood, S. Wu, J. Lu, Z. Ajmal, H. Luo, R. Dong, Nutrient recovery from anaerobically digested chicken slurry via struvite: performance optimization and interactions with heavy metals and pathogens, Sci. Total Environ. 635 (2018) 1–9.

[22] S. Wu, P. Ni, J. Li, H. Sun, Y. Wang, H. Luo, J. Dach, R. Dong, Integrated approach to sustain biogas production in anaerobic digestion of chicken manure under recycled utilization of liquid digestate: dynamics of ammonium accumulation and mitigation control, Bioresour. Technol. 205 (2016) 75–81.

[23] K. Yetilmezsoy, Z. Sapci-Zengin, Recovery of ammonium nitrogen from the effluent of UASB treating poultry manure wastewater by MAP precipitation as a slow release fertilizer, J. Hazard Mater. 166 (2009) 260–269.

[24] A.R. Webb, F.R. Hawkes, The anaerobic digestion of poultry manure: variation of gas yield with influent concentration and ammonium-nitrogen levels, Agric. Wastes 14 (1985) 135–156.

[25] C. Abendroth, E. Wünsche, O. Luschnig, C. Bürger, T. Günther, Producing high-strength liquor from mesophilic batch acidification of chicken manure, Waste Manag. Res. 33 (2015) 291–294.

[26] R.Ö. Sürmeli, A. Bayrakdar, B. Çalli, Removal and recovery of ammonia from chicken manure, Water Sci. Technol. 75 (2017) 2811–2817.

[27] C. Abendroth, C. Simeonov, J. Peretó, O. Antúnez, R. Gavidia, O. Luschnig, M. Porcar, From grass to gas: microbiome dynamics of grass biomass acidification under mesophilic and thermophilic temperatures, Biotechnol. Biofuels 10 (2017) 171.

[28] G. Strazzera, F. Battista, N.H. Garcia, N. Frison, D. Bolzonella, Volatile fatty acids production from food wastes for biorefinery platforms: a review, J. Environ. Manag. 226 (2018) 278–288.

[29] H.J. Gijzen, K.B. Zwart, F.J. Verhagen, G.P. Vogels, High-Rate two-phase process for the anaerobic degradation of cellulose, employing rumen microorganisms for an efficient acidogenesis, Biotechnol. Bioeng. 31 (1988) 418–425.

[30] C. Abendroth, S. Hahnke, C. Simeonov, M. Klocke, S. Casani-Miravalls, P. Ramm, C. Bürger, O. Luschnig, M. Porcar, Microbial communities involved in biogas production exhibit high resilience to heat shocks, Bioresour. Technol. 249 (2018) 1074–1079.

[31] C. Herrmann, P. Ramm, J.D. Murphy, The relationship between bioreactor design and feedstock for optimal biogas production, in: Q. Liao, J.-s. Chang, C. Herrmann, A. Xia (Eds.), Bioreactors for Microbial Biomass and Energy Conversion, Springer Singapore, Singapore, 2018, pp. 163–197.

[32] K. Beinersdorf, S. Sebök, N. Krakat, Biogasgewinnung aus stickstoffreichen Substraten: entwicklung und Optimierung von verfahrenstechnischen Lösungen zur Vermeidung von Ammoniak-Hemmungen in Biogasreaktoren, in: Biogas in der Landwirtschaft – Stand und Perspektiven. Proceedings of the KTBL/FNR conference, vols. 22–23, September 2015, pp. 195–206. Potsdam.

[33] VDLUFA - Verband deutscher landwirtschaftlicher Untersuchungs- und Forschungsanstalten, Methodenbuch Band III – Die chemische Untersuchung von Futtermitteln. 3. Auflage inklusive der 1. bis 6, Ergänzungslieferung, 1983/2006.

[34] C. Herrmann, C. Idler, M. Heiermann, Improving aerobic stability and biogas production of maize silage using silage additives, Bioresour. Technol. 197 (2015) 393–403.

[35] E. Nettmann, I. Bergmann, K. Mundt, B. Linke, M. Klocke, Archaea diversity within a commercial biogas plant utilizing herbal biomass determined by 16S rDNA and mcrA analysis, J. Appl. Microbiol. 105 (2008) 1835–1850.

[36] I. Bergmann, E. Nettmann, K. Mundt, M. Klocke, Determination of methanogenic Archaea abundance in a mesophilic biogas plant based on 16S rRNA gene sequence analysis, Can. J. Microbiol. 56 (2010) 440–444.

[37] I. Bergmann, K. Mundt, M. Sontag, I. Baumstark, E. Nettmann, M. Klocke, Influence of DNA isolation on Q-PCR-based quantification of methanogenic Archaea in biogas fermenters, Syst. Appl. Microbiol. 33 (2010) 78–84.

[38] A. Klindworth, E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, F. Glöckner, Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies, Nucleic Acids Res. 41 (2012) e1-e1.

[39] J. Hardegen, A. Latorre-Pérez, C. Vilanova, T. Günther, M. Porcar, O. Luschnig, C. Simeonov, C. Abendroth, Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass, Bioresour. Technol. 265 (2018) 275–281.

[40] E. Pruesse, C. Quast, K. Knittel, B. Fuchs, W. Ludwig, J. Peplies, F. Glockner, SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB, Nucleic Acids Res. 35 (2007) 7188–7196.

[41] P. McMurdie, S. Holmes, Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data, PLoS One 8 (2013) e61217.

[42] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (2014) 550.

[43] S. Guštin, Marinšek-Logar, Effect of pH, temperature and air flow rate on the continuous ammonia stripping of the anaerobic digestion effluent, Process Saf. Environ. Protect. 89 (1) (2011) 61–66.

[44] M. Walker, K. Iyer, S. Heaven, C.J. Banks, Ammonia removal in anaerobic digestion by biogas stripping: an evaluation of process alternatives using a first order rate model based on experimental findings, Chem. Eng. J. 178 (15) (2011) 138–145.

[45] W. Tao, K.P. Fattah, M.P. Huchzermeier, Struvite recovery from anaerobically digested dairy manure: a review of application potential and hindrances, J. Environ. Manag. 169 (2016) 46–57.

[46] Y. Ogata, T. Ishigaki, M. Nakagawa, Y. Yamada, Effect of increasing salinity on biogas production in waste landfills with leachate recirculation: a lab-scale model study, Biotechnol. Rep. 10 (2016) 111–116.

[47] E. Elbeshbishy, B.R. Dhar, G. Nakhla, H.S. Lee, A critical review on inhibition of dark biohydrogen fermentation, Renew. Sustain. Energy Rev. 79 (2017) 656–668.

[48] C. Niu, T. Hojo, W. Qiao, H. Qiang, Y.Y. Li, Characterization of methanogenesis, acidogenesis and hydrolysis in thermophilic methane fermentation of chicken manure, Chem. Eng. J. 244 (2014) 587–596.

Publication III

Systematic and Applied Microbiology 43 (2020) 126024

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: http://www.elsevier.com/locate/syapm

Shedding light on biogas: Phototrophic biofilms in anaerobic digesters hold potential for improved biogas production

Christian Abendroth[a,b,1], Adriel Latorre-Pérez[c,1], Manuel Porcar[c,d], Claudia Simeonov[a], Olaf Luschnig[e], Cristina Vilanova[c], Javier Pascual[c,*]

[a] Robert Boyle Institut e.V., Jena, Germany
[b] Technische Universität Dresden, Chair of Waste Management, Pratzschwitzer Str. 15, Pirna, Germany
[c] Darwin Bioprospecting Excellence, S.L., Paterna, Valencia, Spain
[d] Institute for Integrative Systems Biology (I2SysBio), University of Valencia–CSIC, Paterna, Valencia, Spain
[e] Bio H2 Umwelt GmbH, Jena, Germany

ARTICLE INFO

Article history:
Received 25 April 2019
Received in revised form 2 September 2019
Accepted 10 September 2019

Keywords:
Anaerobic digestion
Optimized biogas
Shotgun metagenomic sequencing
Wastewater treatment
Phototrophism
16S rRNA gene amplicon Na
Nopore sequencing

ABSTRACT

Conventional anaerobic digesters intended for the production of biogas usually operate in complete darkness. Therefore, little is known about the effect of light on their microbial communities. In the present work, 16S rRNA gene amplicon Nanopore sequencing and shotgun metagenomic sequencing were used to study the taxonomic and functional structure of the microbial community forming a biofilm on the inner wall of a laboratory-scale transparent anaerobic biodigester illuminated with natural sunlight. The biofilm was composed of microorganisms involved in the four metabolic processes needed for biogas production, and it was surprisingly rich in *Rhodopseudomonas faecalis*, a versatile bacterium able to carry out photoautotrophic metabolism when grown under anaerobic conditions. The results suggested that this bacterium, which is able to fix carbon dioxide, could be considered for use in transparent biogas fermenters in order to contribute to the production of optimized biogas with a higher CH_4:CO_2 ratio than the biogas produced in regular, opaque digesters. To the best of our knowledge, this is the first study characterising the phototrophic biofilm associated with illuminated bioreactors.

Introduction

Anaerobic digestion (AD) of organic matter is a robust technology for biogas synthesis from different types of waste [6], and numerous studies have been conducted to optimise the synthesis of biogas and evaluate potential substrates [28]. Anaerobic digesters can be fed with a wide range of substrates, such as grass biomass [2,3], sewage sludge from water treatment [24], microalgal biomass [18], and food waste [70], among others. The main goal of AD is the production of methane, a renewable energy source that can be used for heating and electricity, as well as many other operations that use combustion engines [39]. Biogas is a mixture of methane (CH_4; 55–70% of the total volume), carbon dioxide (CO_2; 30–40%) and traces of other gases, such as hydrogen sulphide (H_2S) [14,54]. Whereas methane is a flammable gas over a relatively large range of concentrations in air at standard pressure (5.4–17%), carbon dioxide is an inert gas. Therefore, increasing the CH_4:CO_2 ratio is one of the keystones for the production of high-quality biogas. The CH_4:CO_2 ratio can vary depending on the digester type, the substrate composition, and other factors such as temperature, pH, the degradation rate and substrate concentration [21]. An appropriate design for an anaerobic digester is thus central for the production of optimized biogas.

The microbial communities operating in the digester are the final key players responsible for the quality of the biogas produced. The role of different microorganisms in the four metabolic steps carried out during the AD of organic matter (hydrolysis, acidogenesis, acetogenesis, and methanogenesis) has been widely studied [74]. A diverse number of *Bacteria* are known to be involved in the hydrolysis and further acidogenesis of complex polymers, whereas the oxidation of intermediate metabolites to acetate (acetogenesis) is performed by either hydrogen- or formate-producing acetogens [61]. Lastly, methane synthesis is mainly derived from acetate and H_2/CO_2 by acetoclastic and hydrogenoclastic methanogenic *Archaea*. Therefore, an improved understanding of the microbial communities and their metabolic roles during the four stages of AD may also help to optimize biogas production in terms of quantity (yield) and quality (CH_4:CO_2 ratio of the gas produced).

[*] Corresponding author.
E-mail address: jpascual@darwinbioprospecting.com (J. Pascual).
[1] Equal contributions.

https://doi.org/10.1016/j.syapm.2019.126024
0723-2020/© 2019 Elsevier GmbH. All rights reserved.

Over the past few years, next-generation sequencing techniques, such as 16S rRNA gene amplicon sequencing and shotgun metagenomic sequencing, have been applied to study the structure and composition of microbial communities in different types of anaerobic digesters [1,2,3,23,24,43,62]. These studies have shown that each particular community is influenced by parameters such as the type of feedstock [62], temperature [5,11,19], retention time [19], salt content [17,6], viscosity [24], pH [75], or the loading rate [11,24]. Although the influence of many physicochemical parameters on microbial communities has been studied in anaerobic digesters, very little is known about the influence of light on the process [53,59,66], mainly because of the obvious fact that conventional AD systems operate in complete darkness. Interestingly, a previous study reported an increase of the relative concentration of methane when an anaerobic digester was operated under the influence of light [63]. However, the effect of light on the entire microbiome of anaerobic digesters has yet to be addressed.

Therefore, the aim of the present work was to analyse the effect of natural sunlight on the microbial community of a laboratory-scale anaerobic co-digester, in order to explore the possibility of inducing light-sensitive pathways, which might improve biogas quality due to carbon fixation. In order to reach this goal, full-length 16S rRNA gene amplicon Nanopore sequencing, shotgun metagenomic sequencing and a complete bioinformatics analysis were used to unveil the structure and composition of the microbial community growing as a red-coloured biofilm over the transparent wall of a specifically designed transparent leach-bed bioreactor.

Materials and methods

Substrate and seed sludge

Untreated grass biomass (*Graminidae*) from a pasture in Jena (Germany) was used as feedstock. Collected grass biomass was characterised by a total solids content (TS) of 30.4%, with 84.2% of the TS being volatile solids (VS). TS and VS were determined as described in Abendroth et al. [2]. One gram of fresh biomass showed a chemical oxygen demand (COD) of 260 mg O_2. Sewage sludge from an anaerobic digester of the water treatment plant in Jena (Germany) was used as seed sludge.

Digestion conditions

The experiment was carried out in an open hall during the summer of 2017. A transparent two-stage Plexiglas® leach system was built, and used to perform acidification of grass biomass in a leach-bed configuration and methanisation using an anaerobic filter, as described in Abendroth et al. [2] (Figs. S1 and S11). The anaerobic digester was placed in a sun-exposed construction hall, where it received indirect natural sunlight (ca. 10 h daylight). The acidification stage and the methane stage each had a working volume of 20 L (Fig. S1) and both stages were treated at a mesophilic temperature (37 °C). Grass biomass was retained in a strainer during acidification, and the methane stage was filled with 1.58 kg of bed packing (Hel-X, Christian Stöhr, Germany). Both stages were filled with sewage sludge at the beginning of the experiment: 8 L for the acidification stage and 11.75 L for the methane stage. For every batch cycle of acidification, 96.2 g L^{-1} of fresh grass biomass were used. During acidification, the pH was kept at 6.0 using a pH-regulation system (BL 7916, Hanna Instruments, Germany). After each cycle of acidification, the collected liquor was stored at 4 °C. Each methane stage received a daily amount of the high strength liquor produced, which corresponded to approximately 100 g COD (8.5 g COD L^{-1}). Gas production was quantified with a customized gas counter and collected in gasbags (Tecobag, Tesseraux, Germany) for further analysis.

Sample collection and metagenomic DNA isolation

After three weeks of the start-up phase, a red biofilm appeared, and a 250 μL biofilm sample was collected by scratching the inner bioreactor wall that was further mixed in a reaction tube (Fig. S1). In order to reduce the quantity of inhibiting substances, the biofilm sample was centrifuged (10 min at 20,000 g) and then washed several times with sterile phosphate-buffered saline (PBS; pH 7.2) until a clear supernatant was observed. Subsequently, metagenomic DNA was isolated using the Power Soil DNA Isolation kit (MoBio Laboratories, Carlsbad, CA, USA) following the manufacturer's instructions. DNA concentration, quality and integrity were assessed with a Nanodrop-1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and on a 0.8% (w/v) agarose gel, respectively.

Full-length 16S rRNA gene amplicon Nanopore sequencing

The bacterial full-length 16S rRNA gene was amplified with PCR using the primer pair S-D-Bact-0008-a-S-16 (5′-AGRGTTYGATYMTGGCTCAG-3′) and S-D-Bact-1492-a-A-16 (5′-TACCTTGTTAYGACTT-3′) [35]. The following reagents and concentrations were used for the first PCR reaction: 200 μM dNTPs, 200 nM of each primer, 1 U of VWR Taq DNA Polymerase (VWR®, WR International bvba/sprl, Belgium), 1 x PCR buffer supplemented with $MgCl_2$ (1.5 mM), and 10 ng of DNA template (final volume: 50 μL). PCR started with an initial denaturation step at 94 °C for 1 min, followed by 35 cycles of amplification (denaturing, 1 min at 95 °C; annealing, 1 min at 49 °C; extension, 2 min at 72 °C), and a final extension at 72 °C for 10 min. A negative control without DNA template was included. Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA) at 0.5 x concentration were used to remove primer-dimers and non-specific amplicons. DNA concentration was measured using the Qubit dsDNA HS Assay kit (Qubit 2.0 Fluorometer, Thermofisher, Waltham, USA). Then, the Ligation Sequencing Kit 1D (SQK-LSK108) was used to prepare the amplicon library to load into the MinION. The flow cell (R9.4, FLO-MIN106) was primed and then loaded as indicated in the ONT protocols.

Reads were basecalled in real time using the MinKNOW^{TM} software (version 1.13.1, standard sequencing protocol), and sequencing statistics were followed in real time using the EPI2ME debarcoding workflow. Porechop (https://github.com/rrwick/Porechop) was applied for removing the adaptors. By default, MinKNOW^{TM} software removes reads with quality values lower than 7 in the PHRED score and, thus, reads shorter than 1000 nt were discarded for subsequent analyses. The resulting sequences were analysed using the QIIME [37]. Briefly, reads were taxonomically classified through BLAST searches against the latest version of the GreenGenes database v 13.8 [15]. Rarefaction curves of the full-length 16S rRNA reads, including and excluding singletons, were obtained using the iNEXT (v. 2.0.17) R package. The full-length 16S rRNA gene amplicon Nanopore sequencing data is available through NCBI's Sequence Read Archive (SRA) database under the accession number SRR8529815.

Shotgun metagenomic sequencing

The biofilm sample was also subjected to shotgun metagenomic sequencing. Briefly, the Nextera XT Prep Kit protocol was followed for library preparation and then the Illumina MiSeq platform was used for sequencing. The parameters were adjusted to obtain pair-end sequences of 150 bp and a sequencing depth of 10 million reads. Adapters were trimmed, and quality filtering was applied with BBDuk (included in the BBTools package; Bushnell B., https://sourceforge.net/projects/bbtools/ updated January 2, 2018). Reads shorter than 50 bp and/or with a mean quality lower than Q20 (in the PHRED scale) were discarded. The quality parameters of the sequences were checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc, version 0.11.7).

The quality-checked reads were taxonomically classified via Centrifuge v. 1.0.3 [33] against a compressed database containing reference sequences from *Bacteria* and *Archaea* (updated April 15, 2018; available at https://ccb.jhu.edu/software/centrifuge/).

The shotgun sequences were assembled into contigs and scaffolds with the metaSPAdes pipeline included in SPAdes v. 3.12.0 [48]. The statistics and attributes of the assembly were explored with QUAST v. 4.6.3. Selected scaffolds were grouped into bins with MaxBin2 v. 2.2.4 [69] in order to reconstruct metagenome-assembled genomes (MAGs). CheckM v1.0.11 [50] was used for assessing the quality and completeness of each MAG. Only high-quality MAGs, with completeness values greater than 50% and contamination values lower than 10%, were considered for further analyses.

The taxonomic affiliation of each MAG was assessed with the Similar Genome Finder Service tool of the Pathosystems Resource Integration Center (PATRIC) [67]. This tool matches each MAG against a set of representative and reference genomes available in PATRIC [67] by using Mash/MinHash distances [49]. Subsequently, a phylogenomic tree was inferred for each MAG in order to determine their specific evolutionary history. The UBCG v. 3.0 pipeline (up-to-date bacterial core gene set [46]) was used to construct maximum likelihood trees based on a multiple alignment of a set of 90–92 universal and single copy gene sequences (Supplementary Table S8). Despite the fact that the UBCG pipeline is optimized for bacterial genomes, it was also used for archaeal MAGs but using 23 universal and single copy genes. In order to investigate if each MAG belonged to a known species, pairwise average nucleotide identity values (ANIb) [20] were calculated between each MAG and its closest type strain, by using the JSpeciesWS online tool [56]. Additionally, digital DNA-DNA hybridization (DDH) pairwise values were also obtained using the Genome-to-Genome Distance Calculator 2.1 (GGDC) tool [44]. Formula 2 was used for calculating the digital DDH values, as recommended for incompletely sequenced genomes [44].

Functional analysis of the microbial community

The assembled metagenomic sequences, as well as the high-quality MAGs, were annotated using the RAST toolkit (RASTtk) [8] implemented in the Genome Annotation Service in PATRIC [67]. The carbohydrate-active enzymes (CAZyme) of each MAG were determined by identifying genes containing CAZyme domains using the dbCAN2 meta server [72]. CAZyme domains were predicted by HMMER (E-Value < 1e-15, coverage > 0.35). The metabolic pathway reconstruction for each MAG was carried out by comparing the protein-coding genes against the Kyoto Encyclopedia of Genes and Genomes (KEGG) [29] and MetaCyc metabolic pathways [9]. The high-quality MAGs are publicly available at NCBI under the following genome accession numbers: SHOF00000000, SHOG00000000, SHOH00000000, SHOI00000000, SHOJ00000000, SHOK00000000, SHOL00000000 and SHOM00000000.

Results and discussion

Taxonomic diversity of the microbial community

After three weeks operating the anaerobic digester, a bright red-pigmented microbial biofilm appeared on the inner wall of the bioreactor (Supplementary Fig. S1). Since the microbiome of the main content of biodigesters has already been extensively studied [1,2,3,23,43,62], we focused on the characterisation of the microbiome of the red-pigmented biofilm that developed on the transparent wall of the bioreactor.

The taxonomic composition of the biofilm microbial community was studied via full-length 16S rRNA amplicon Nanopore sequencing. After quality filtering the raw reads, a total of 11,163 16S rRNA sequences were retrieved and taxonomically classified. The median sequence length was 1445 ± 120 nt and the mean read quality was 9.8. Similarly to Ma et al. [41], any attempt to cluster the reads into operational taxonomic units (OTUs) using the closed-reference OTU picking method available in QIIME failed (i.e. each sequence was classified as an independent OTU). For this reason, the taxonomy of each read was individually assigned and the results were then collapsed into different taxonomic levels (species, genera, and higher taxonomic ranks). The number of reads obtained was sufficient to analyse the vast majority of the microbial species (Supplementary Fig. S2A), thus enabling a comprehensive characterization of the microbial community. The saturation of the species richness was more evident when the singletons (taxa supported by only one sequence) were excluded from the dataset (Supplementary Fig. S2B).

The bacterial community was dominated by members of the phyla *Firmicutes*, *Bacteroidetes* and *Proteobacteria*, followed by *Chloroflexi*, *Spirochaetes* and the candidate phylum WS6 (Fig. 1A). Additionally, members of 43 phyla or candidate divisions were also identified (Supplementary Table S1). This profile was similar to that found by other authors in dark AD [1,2,3,23,43,62]. At the family level, *Gracilibacteraceae*, *Lachnospiraceae* and *Tissierellaceae* were the dominant *Firmicutes*, while *Porphyromonadaceae* and *Bradyrhizobiaceae* were the most abundant *Bacteroidetes* and *Alphaproteobacteria*, respectively (Fig. 1A; Supplementary Table S1). Additionally, 36 reads were classified as *Archaea*. However, since the primer pair used to amplify the 16S rRNA was optimized for *Bacteria* [35], all archaeal reads were excluded from the analysis.

Although the full-length sequence of the 16S rRNA gene was sequenced, a high number of phylotypes could only be classified to the family level (Supplementary Table S1). The high diversity of phylotypes recovered from the biofilm sample (723 phylotypes, Supplementary Table S1), might be a consequence of the high error rate of the Nanopore sequencing technology. In fact, 46.6% of microbial phylotypes were singletons and nine phyla were represented by a single read (Supplementary Table S1). Nevertheless, a study has recently demonstrated that the MinION technology has the ability to provide rRNA operon sequence data of sufficient quality for characterizing the microbiota of complex environmental samples and provides results that are reproducible, quantitative and consistent [32]. Another explanation why a high number of phylotypes were identified in the biofilm is that the seed sludge and the feedstock used were carrying a highly diverse microbial load [34], albeit that these communities might not have been metabolically active in the sampled biomass. An indication of the presence of inactive microorganisms in the community is the occurrence of obligate aerobic bacteria, such as *Arthrobacter* or *Devosia* (Supplementary Table S1). Further studies based on metagenomic RNA would be necessary to distinguish between the phylotypes that are keyplayers in the biofilm and those that are merely transported by the influent as inactive microorganisms.

In order to complement the taxonomic information of the microbial community of the red-coloured biofilm, shotgun metagenomic sequencing was also performed. A total of 8,903,087 high-quality pair-end reads with a median size of 150 nt were sequenced. The taxonomic classification of all the metagenomic reads is shown in Fig. 1B. Only 38.2% of metagenomic reads mapped against the genomic database, corroborating the taxonomic novelty of the microorganisms that formed the biofilm. Most of the reads were mapped against genomes of *Bradyrhizobiaceae*, specifically *Rhodopseudomonas palutris* (26.6% of total reads), followed by the

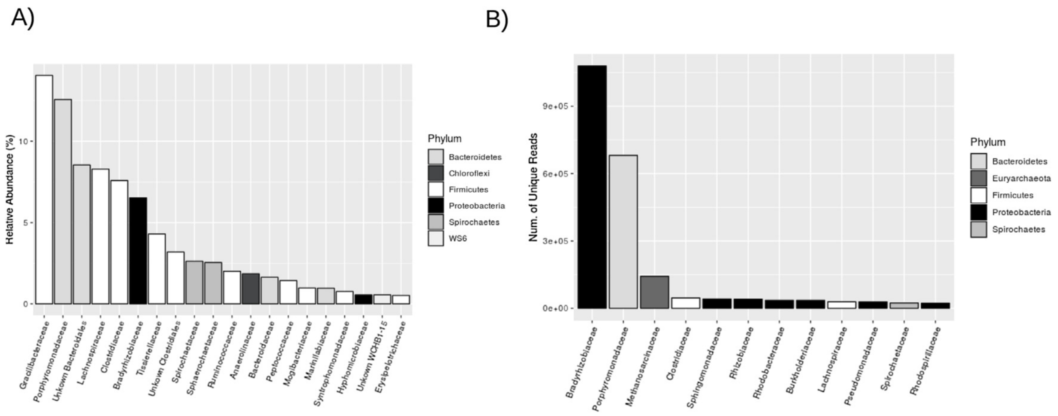


Fig. 1. (A) Taxonomic classification of the almost full-length 16S rRNA gene sequences sequenced with Nanopore technology. Reads were blasted against the GreenGenes database (v 13.8). The families with a relative abundance higher than 0.5% are shown. (B) Taxonomic classification of the shotgun metagenomic Illumina reads. Reads were mapped against a database containing reference sequences from *Bacteria* and *Archaea*. Only the 12 most abundant families are shown.

Porphyromonadaceae species *Fermentimonas caenicola* (22.9%) and the archaeal species *Methanosarcina mazei* (4.7%). Interestingly, and in sharp contrast with what has been described for regular (dark) anaerobic digesters [23], the illumination of the bioreactor triggered an enrichment of *Rhodopseudomonas* (Fig. 1B). Similarly, other authors have reported enrichment of *Rhodopseudomonas faecalis* in an illuminated anaerobic digester fed with swine sewage wastewater [68]. In contrast to the 16S rRNA sequencing results, the family *Gracilibacteraceae* was not abundant in the shotgun metagenomic data (Fig. 1B). This type of taxonomic discrepancy between both sequencing approaches has previously been discussed by other authors [64].

Functional profile of the community and definition of microbial keyplayers

A total of 6183 contigs comprised of 38,667,755 nt were assembled from the shotgun metagenomic sequencing. 40,136 coding sequences (CDS) were identified after the functional annotation of contigs with RASTtk, and 54.9% of CDS were identified as proteins with functional assignments, while the remaining CDS were annotated as hypothetical proteins. Furthermore, 556 tRNA, 53 rRNA, 583 CRISPR-repeats, 556 CRISPR-spacers and 27 CRISPR-array sequences were also reported. 38.05% of the CDS were assigned to functional subsystems (Fig. 2; Supplementary Table S2). The great majority of the CDS (36.8%) were involved in cellular metabolism, including genes engaged in the turnover of nutrients (Supplementary Table S2). Genes related to protein processing accounted for 17.3%, while those for energy and DNA processing accounted for 12.4% and 6.5%, respectively (Fig. 2).

To date, many microbial ecology studies in anaerobic digesters have been based on 16S rRNA OTUs. However, due to the great metabolic diversity of certain taxa, as well as the impossibility to classify some OTUs at lower taxonomic levels, it is difficult to predict the accurate functional roles that each microorganism plays during AD [34]. Therefore, in order to shed light on the key microorganisms and their metabolic functions in anaerobic digesters when operated under the influence of natural sunlight, assembled contigs were binned as MAGs. Eight out of the 24 MAGs obtained passed the filter of contamination and completeness (Table 1). The number of scaffolds of the eight high-quality MAGs ranged from 141 to 985 and the estimated genome size was from 1.2 Mb to 4.0 Mb (Table 1). The G + C content of four MAGs was equal to or less than 37.0% and none of them showed a value greater than 64.2% (Table 1). MAG 4 harboured chimeric ribosomal RNA operons and hence the 16S rRNA gene sequences could not be used for taxonomic purposes, while other MAGs, such as MAG 16, did not include any ribosomal RNA operon. MAG 1 was identified as a member of the species *Rhodopseudomonas faecalis* (Table 2; Supplementary Fig. S3). MAG 2 was identified as a strain of the *Fermentimonas caenicola* species (Supplementary Fig. S4), MAG 10 as *Proteiniborus ethanoligenes* (Supplementary Fig. S8), and MAG 16 as *Methanosarcina mazei* (Supplementary Fig. S10). The ANI and digital DDH values between MAGs 1, 2, 10 and 16 and the type strains of phylogenetically close species were higher than the threshold established to circumscribe prokaryotic species, namely 95% for ANI values [55] and 70% for digital DDH [44]. Therefore, both genome-related indices [10] confirmed the adscription of MAGs 1, 2, 10 and 16 to previously known species. Originally, *Rhodopseudomonas faecalis* and *Fermentimonas caenicola* were isolated from anaerobic reactors [22,71], and *Proteiniborus ethanoligenes* from a mesophilic hydrogen-producing granular sludge [47], whereas *Methanosarcina mazei* was isolated from a sewage sludge plant. Consequently, the results confirmed the prevalence of these taxa in an anaerobic environment where fermentation processes took place. Based on the number of reads obtained from the shotgun metagenomic sequencing, MAG 1 (*Rhodopseudomonas faecalis)* and MAG 2 (*Fermentimonas caenicola*) were the most abundant bacteria in the red-pigmented biofilm; while MAG 16 (*Methanosarcina mazei*) was the only archaeon identified in the community (Fig. 1). Contrarily, the other four MAGs could not be identified at the species level, but at the genus level or even at a higher taxonomic rank (Table 2). MAG 4 could represent a new species of the genus *Herbinix* (class *Clostridia*), since it was closely related to the strain *Herbinix luporum* SD1D^T (Supplementary Fig. S5), a cellulolytic bacterial strain isolated from an industrial-scale biogas plant [36]. Two MAGs were classified as members of the *Spirochaetaceae* family. Specifically, MAG 5 was related to *Sphaerochaeta globosa* Buddy^T (Supplementary Fig. S6), a strain isolated from an anoxic river sediment [57], while MAG 11 was identified as a new species of the genus *Treponema* (Supplementary Fig. S9). Finally, MAG 9 was classified as a hitherto unknown taxon of the phylum *Firmicutes*, representing a new class or order (Table 2; Supplementary Fig. S7).

Affiliation of functional CDS to the four stages of anaerobic digestion

Hydrolysis of complex polymers

A total of 108 different glycoside hydrolase families (Carbohydrate-Active enZYmes Database; [40]) were found in

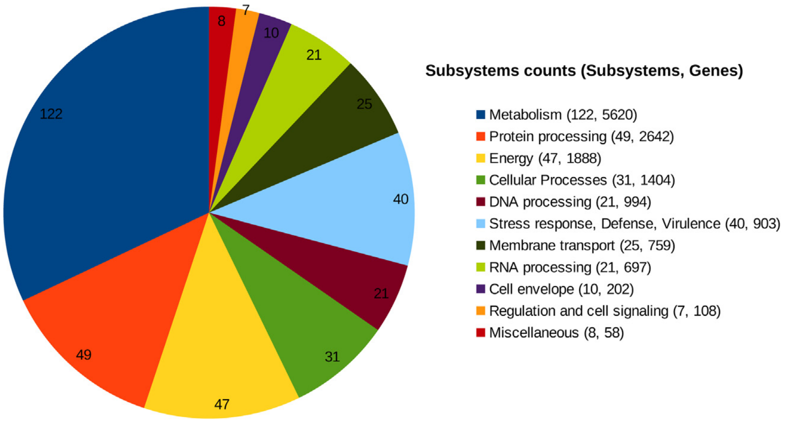


Fig. 2. Classification of annotated coding sequences (CDS) in functional subsystems. A total of 1572 CDS (in yellow) were assigned to 381 different subsystems (in green) (for interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 1
Summary statistics of the reconstructed metagenome-assembled genomes (MAGs) analyzed in this study. The completeness and contamination of each MAG were estimated with CheckM [50] and the coverage with MaxBin2 [70]. CDS, protein coding sequence; G + C, guanine-cytosine content.

MAG	Contigs	Genome length (Mb)	Coverage	Completeness (%)	Contamination (%)	G + C (%)	CDS	Proteins with functional assignments	Hypothetical proteins	Genome accession number
1	119	4.0	76.5	98.4	0.17	64.2	3797	2725	1072	SHOF00000000
2	141	2.7	31.8	100	0.55	37.0	2348	1529	819	SHOG00000000
4	319	3.2	17.8	91.2	0.67	37.0	2946	1821	1125	SHOH00000000
5	136	2.2	10.9	92.0	3.45	53.7	2270	1266	1004	SHOI00000000
9	213	1.2	6.8	96.6	4.96	36.3	1191	585	606	SHOJ00000000
10	352	1.6	6.7	53.5	5.2	31.6	1792	1027	765	SHOK00000000
11	346	2.7	5.7	73.4	3.26	53.2	2770	1297	1473	SHOL00000000
16	985	2.8	4.2	81.0	4.63	42.9	3459	2094	1365	SHOM00000000

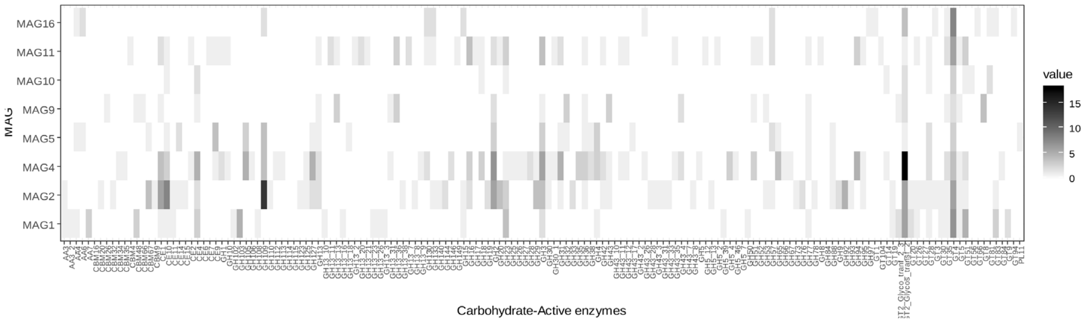


Fig. 3. Heatmap of Carbohydrate-Active Enzyme (CAZyme) families found in each of the eight metagenome-assembled genomes. Both CAZyme families and MAGs are listed alphabetically. The colour intensity corresponds to the number of protein-coding genes identified in each family. CAZyme family codes: GT, glycosyltransferases; GH, glycoside hydrolases; CE, carbohydrate esterases; PL, polysaccharide lyases; CBM, carbohydrate binding modules; AA, axillary activities (oxidative enzymes).

the eight high-quality MAGs (Fig. 3; Supplementary Table S3). The microorganism with a greater repertoire of glycoside hydrolases was MAG 4 (*Herbinix* sp.), which contained 104 glycoside hydrolases from over 49 families. MAG 2 (*Fermentimonas caenicola*) codified for 82 glycoside hydrolases from 41 different families; and MAG 11 (*Treponema* sp.) encoded 54 glycoside hydrolases belonging to 36 families. Furthermore, MAGs 4, 2 and 3 harboured the greatest repertoire of carbohydrate esterases (Fig. 3; Table S3). Representatives of glycosyl transferase families GT2_Glycos_transf_2 and GT4 were found in the eight MAGs. Moreover, MAG 4 contained a high number of GT2_Glycos_transf_2-coding genes, specifically 18 protein-coding genes (Fig. 3; Supplementary Table S3).

The nature of the substrate determines the type of bacteria involved in the hydrolysis step. Preeti Rao et al. [52], observed that digesters fed with cow manure supported more amylolytic microorganisms, whereas digesters fed with poultry waste showed higher proteolytic populations. In our case, a dominance of cellulolytic, hemicellulolytic and lignolytic bacteria, such as *Bacteroides*, *Spirochaetes* and *Clostridium*, was observed. This was expected, since the anaerobic digester used in the present study was fuelled with untreated grass biomass [51].

Acidogenesis

All the MAGs showed a potential fermentative metabolism based on a functional analysis of their genomes (Supplementary Tables S4 and S5). MAG 1 (*Rhodopseudomonas faecalis*), MAG 4 (*Herbinix* sp.), MAG 5 (*Spirochaetaceae*), MAG 9 (*Firmicutes*), MAG 10 (*Proteiniborus ethanoligenes*), MAG 11 (*Treponema* sp.) and MAG 16 (*Methanosarcina mazei*) were potentially able to carry out alcoholic fermentation with the NAD$^+$-dependent ethanol dehydrogenase

Table 2
Taxonomic affiliation and novelty of each metagenome-assembled genome (MAG). The taxonomic affiliation of each MAG was assessed with the Similar Genome Finder Service tool of the PATRIC [68] which is based on Mash distances. p-values associated with Mash distances are coded with ***0,001. The average nucleotide identity (ANI) and digital DNA-DNA hybridization (DDH) values with regard to its closest published genome and closest type strain are shown for each MAG.

MAG	Lowest common ancestor	Closest published genome				Closest type strain				Taxonomic novelty
		Taxonomic identity (GenBank assembly or RefSeq accession	Mash Distance (p-value)	Digital DDH (%)	ANI (%)	Taxonomic identity (GenBank assembly or RefSeq accession	Mash Distance (p-value)	Digital DDH (%)	ANI (%)	
1	*Bacteria;Proteobacteria*; *Alphaproteobacteria*; *Rhizobiales*; *Bradyrhizobiaceae;* *Rhodopseudomonas*	*Rhodopseudomonas faecalis* JCM 11668^T (JHAA01000000)	0.1649***	82.3	97.7	*Rhodopseudomonas faecalis* JCM 11668^T (JHAA01000000)	0.1649***	82.3	97.7	Same species
2	*Bacteria;Bacteroidetes*; *Bacteroidia*; *Bacteroidales*; *Porphyromonadaceae*; *Fermentimonas*	*Fermentimonas caenicola* ING2-E5B^T (GCA_000953535)	0.0085***	97.2	99.2	*Fermentimonas caenicola* ING2-E5B^T (GCA_000953535)	0.0085***	97.2	99.2	Same species
4	*Bacteria*; *Firmicutes*; *Clostridia*; *Clostridiales*; *Lachnospiraceae*; *Herbinix*	*Herbinix luporum* SD1D^T	0.2437***	19.1	75.1	*Herbinix luporum* SD1D^T	0.2437***	19.1	75.1	New species
5	*Bacteria*; *Spirochaetes*; *Spirochaetia*; *Spirochaetales*; *Spirochaetaceae*; *Sphaerochaeta*	*Sphaerochaeta globosa* Buddy^T (GCA_000190435.1)	0.2630***	17.7	67.6	*Sphaerochaeta globosa* Buddy^T (GCA_000190435.1)	0.2630***	17.7	67.6	New species
9	*Bacteria*; *Firmicutes*	*Aerococcus urinae* CCUG 36881^T (GCA_001543175.1)	0.2959***	28.8	63.1	*Aerococcus urinae* CCUG 3688^T (GCA_001543175.1)	0.2959***	28.8	63.1	New class or order
10	*Bacteria*; *Firmicutes*; *Clostridia*; *Clostridiales*; *Thermohalobacter*; *Proteiniborus*	*Proteiniborus ethanoligenes* DSM 21650^T (GCA_900107485.1)	0.0352***	96.5	97.2	*Proteiniborus ethanoligenes* DSM 21650^T (GCA_900107485.1)	0.0352***	96.5	97.2	Same species
11	*Bacteria*; *Spirochaetes*; *Spirochaetia*; *Spirochaetales*; *Spirochaetaceae*; *Treponema*	*Treponema primitia* ZAS-1 (GCA_000297095.1)	0.2630***	39.0	65.1	*Treponema primitia* ZAS-2^T (NZ_AEEA00000000.1)	0.2959***	20.7	65.0	New species
16	*Archaea*; *Euryarchaeota*; *Methanomicrobia*; *Methanosarcinales*; *Methanosarcinaceae*; *Methanosarcina*	*Methanosarcina mazei* Go1 (AE008384)	0.01956***	100	100	*Methanosarcina mazei* S-6^T (NZ CP009512)	0.02021***	94.5	99.3	Same species

(EC 1.1.1.1). In addition, MAG 1 (*Rhodopseudomonas faecalis*), MAG 2 (*Fermentimonas caenicola*), MAG 10 (*Proteiniborus ethanoligenes*) and MAG 11 (*Treponema* sp.) were potentially able to conduct the formation of lactate through different lactate dehydrogenases (EC 1.1.1.27 and EC 1.1.1.28). The possible fermentation of lactate to acetate plus propionate via methylmalonyl-CoA was detected in MAG 1 and MAG 2. Similarly, Heyer et al. [26] observed that lactate fermentation most likely took place in agricultural biogas plants, since large amounts of lactate are produced during the ensiling process for conservation and storage of crop material as primary or co-substrate for the anaerobic digestion process. Contrarily, the set of enzymes 3-hydroxybutyryl-CoA dehydrogenase, crotonase 3−OH-butyryl-CoA dehydratase and butyryl-CoA dehydrogenase responsible for the formation of butyryl-CoA from acetoacetyl-CoA were not detected in any MAG.

Acetogenesis

Acetogenesis pathways, such as acetogenesis by dehydrogenation or syntrophic acetogenesis, are based on the anaerobic oxidation of long- and short-chain (volatile) fatty acids. MAGs 2 (*Fermentimonas caenicola*) and 4 (*Herbinix* sp.) contained some key genes involved in the conversion of propionate to acetate (Supplementary Tables S4 and S5). Additionally, MAG 4 and MAG 11 (*Treponema* sp.) contained the enzymes necessary to perform the conversion of acetyl-CoA to acetate, namely acetate kinase (EC 2.7.2.1) and phosphate acetyltransferase (EC 2.3.1.8).

Methanogenesis

The quantity of methane produced in the anaerobic digester illuminated with natural sunlight was approximately 350 mL g^{-1} of solubilized COD. Interestingly, *Methanosarcina* (MAG 16) was the only methanogenic archaeon detected in the red-pigmented biofilm (Fig. 1B). MAG 16 harboured all the protein-coding genes in the acetoclastic pathway for methane production, except tetrahydromethanopterin S-methyltransferase (EC 2.1.1.86) and methyl-coenzyme M reductase (EC 2.4.8.1). However, although neither of these two genes were assembled into the genome of MAG 16, they were found in the whole metagenome and identified as closely related to *Methanosarcina mazei* S-6 (NZ_CP009512.1). This suggested that they were not included in MAG 16 due to a bioinformatic artefact, which would therefore support the presence of the full acetoclastic pathway. In addition, MAG 16 was the only microorganism harbouring the genes corresponding to the Wood-Ljungdahl pathway (Supplementary Tables S4 and S6). The Wood–Ljungdahl pathway coupled to methanogenesis is one of the most ancient metabolisms for energy generation and carbon fixation in *Archaea* [7].

Syntrophic acetate oxidation by anaerobic bacteria is an alternative pathway for generating methane from the anaerobic oxidation of acetate. This reaction is characterised as being energetically extremely unfavourable and the reaction takes place only when the products are subsequently utilized by the hydrogen-scavenging methanogens. Despite the fact that *Methanosarcina* has also been described as a mixotrophic methanogen [16], the key genes required for the utilization of H_2 and CO_2 for methane production were not found in MAG 16 (Supplementary Table S6). Previous research found that approximately 70% of the methane produced in the digestion of sewage sludge, which often shows low concentrations of VFAs, comes from the transformation of acetate to methane by the acetoclastic methanogens [27]. Usually, *Methanothrix* (formerly *Methanosaeta)* and *Methanosarcina* co-exist in the anaerobic digesters and their relative abundances are driven by the acetate concentration [12]. *Methanosarcina* has greater rates of acetate utilization and growth, and greater half-saturation and yield coefficients compared to *Methanothrix* [12]. Therefore, a possible cause for the dominance of *Methanosarcina mazei* over *Methanothrix* species in the illuminated anaerobic digester may have been the high concentration of acetate in the biomass. *Methanosarcina* is a very robust methanogen able to adapt to environmental changes [16], as well as being an efficient methane producer [63]. Indeed, a previous study assessing the effect of light on methane production during AD reported an enrichment of *Methanosarcina* spp. coupled with an increase in methane production [63]. This suggests, in concordance with our results, that anaerobic digesters operated under light conditions may result in an enrichment of *Methanosarcina*. Direct interspecies electron transfer (DIET) is a process that takes place during AD and is of great importance, since it may significantly accelerate methanogenesis [30,31], and it can be enhanced by adding electrically conductive particles. To date, *Methanothrix*, *Methanospirillum* and *Methanosarcina* are the only genera where DIET has been shown [42,65]. Taking into account that *Methanosarcina* is robust, efficient [30,63] and able to perform DIET, this genus is of high interest for the anaerobic digester industry. Therefore, it is very promising that anaerobic digesters operated under light conditions might enrich the genus *Methanosarcina*.

Functional novelty and implications

Unlike conventional anaerobic digesters operating in complete darkness, an enrichment of *Rhodopseudomonas faecalis* (MAG 1) took place when the anaerobic digester was illuminated with natural sunlight. *R. faecalis* is a common purple non-sulphur (PNS) bacterium able to use a wide range of metabolic pathways [68]. MAG 1 has the entire repertoire of enzymes needed to synthesize a photosystem II-type photosynthetic reaction centre (Supplementary Tables S4 and S7). Photosynthetic reaction centres are complexes composed of several proteins, pigments and other co-factors that, together, execute the primary energy conversion reactions of photosynthesis. The pigments produced by *Rhodopseudomonas palustris*, closely related to *R. faecalis*, are both bacteriochlorophyll(BChl)-a and bacteriopheophytin(BPhe)-a [45]. Therefore, the red to brownish-red colour of the biofilm developed in the bioreactor is very likely due to the massive presence of *Rhodopseudomonas faecalis* in the biofilm, although analysis with liquid chromatography-mass spectrometry would be necessary in order to confirm this hypothesis. MAG 1 is able to carry out CO_2 fixation via the Calvin-Benson cycle under anaerobic conditions (Supplementary Table S7). In this process, CO_2 and ribulose bisphosphate (5-carbon sugar) are transformed into 3-phosphoglycerate [73]. Interestingly, no other microorganisms of the community showed a phototrophic metabolism or presented the complete Calvin–Benson–Bassham (CBB) cycle. Only MAG 16 (*Methanosarcina mazei*) harboured a type III ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) (Supplementary Tables S4 and S6). Nevertheless, type III RuBisCO participates in adenosine 5′-monophosphate (AMP) metabolism, a role that is distinct from that of classical RuBisCOs of the CBB cycle [58]. MAG 1 (*Rhodopseudomonas faecalis*) also has a diazotrophic metabolism, and it is able to fix N_2 with a molybdenum-iron nitrogenase (Supplementary Tables S4 and S7).

The enrichment of *R. faecalis* linked to the illumination of anaerobic digesters reported here could be further developed and used for the production of optimized biogas with a high CH_4:CO_2 ratio. In fact, a previous study reported an increase of methane production when the anaerobic digester operated under the influence of light, although a potential increase of *Rhodopseudomonas* species was not investigated by the authors [63]. Although the biogas generated from anaerobic digestion processes is clean, carbon neutral, and environmentally friendly, raw biogas often needs to be purified prior to its use. To date, several strategies have been designed to reduce CO_2 substantially via chemical absorption [4]. However, this process is expensive. Our findings support the possibility of biologically generating optimized biogas. Since *R. faecalis* has a photoautotrophic metabolism under anaerobic conditions, it can theoretically increase the CH_4:CO_2 ratio of the produced biogas through the fixation of CO_2. The CH_4:CO_2 ratio could be further increased by the action of iron-iron (Fe-only) nitrogenases, which have been recently reported to reduce CO_2 simultaneously with nitrogen gas and protons to yield CH_4, ammonia and hydrogen gas in a single enzymatic step [73]. Even though no iron-iron nitrogenases were detected in MAG 1 (*R. faecalis*) or in the whole metagenome, other *Rhodopseudomonas* strains harbouring these enzymes could be useful for generating optimized biogas [73].

Since the enrichment of *R. faecalis* in anaerobic digesters is expected to be dependent on the amount of light that can pass through the wall of the reactor, as well as on the surface:volume ratio of the reactor, the specific design of illuminated anaerobic digesters is a critical and yet complex issue. Until recently, *Rhodopseudomonas* was of particular interest in terms of hydrogen production [13,25]. However, the recent discovery that *R. faecalis* can produce methane while fixing CO_2 [73], indicates that the genus *Rhodopseudomonas* might be useful for anaerobic digestion processes as well. Moreover, the fact that *R. palustris* [38] and *R. faecalis* [71] were originally isolated from anaerobic digesters and the fact that this genus appears to be suited for mixed culture approaches [60], indicates a high probability for the reproduction of similar biofilms, such as the one described in the present study. The results of the current study pave the way for future research aimed at optimising the development of *R. faecalis* light-dependent biofilms in order to optimize methane-rich biogas production in full-scale reactors.

Conclusions and further considerations

In the present work, a complete study was carried out of a biofilm developing on the transparent wall of a lab-scale anaerobic digester operated under sunlight conditions. The microbial community harboured members involved in the four metabolic stages needed for the anaerobic digestion of organic matter, namely breakdown of polymers into monomers, acidification, acetogenesis and methanogenesis. *Methanosarcina* was the dominant methanogen in the anaerobic digester. The key difference with regard to conventional bioreactors that operate in darkness was a very significant enrichment of *R. faecalis*, a purple non-sulphur bacterium with photoautotrophic metabolism under anaerobic conditions. The ability of this bacterium to assimilate carbon dioxide through the CBB cycle, and its compatibility with the biogas process, as well as with the rest of the microbiome, opens up the striking possibility of producing optimized biogas from biomass through specifically designed, illuminated reactors.

Acknowledgements

We are thankful for the funding received from the German Ministry of Economic Affairs and Energy (grant nos. 16KN070128, 16KN070126). We are also grateful for funding by the European Union through the BioRoboost project, H2020-NMBP-TR-IND-2018-2020/ BIOTEC-01-2018 (CSA), Project ID 210491758. Moreover, we also thank Justus Hardegen, Andreas Underberg and Anja Schmidt for technical assistance. Adriel Latorre is a recipient of a Doctorado Industrial fellowship from the Spanish Ministerio de Ciencia, Innovación y Universidades (reference DI-17-09613).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.syapm.2019.126024.

References

[1] Abendroth, C., Vilanova, C., Günther, T., Luschnig, O., Porcar, M. (2015) *Eubacteria* and *Archaea* communities in seven mesophile anaerobic digester plants. Biotechnol. Biofuels 8, 87.

[2] Abendroth, C., Hahnke, S., Simeonov, C., Klocke, M., Casani-Miravalls, M., Ramm, P., Bürger, C., Luschnig, O., Porcar, M. (2017) Microbial communities involved in biogas production exhibit high resilience to heat shocks. Bioresour. Technol. 249, 1074–1079.

[3] Abendroth, C., Simeonov, C., Peretó, J., Antúnez, O., Gavidia, R., Luschnig, O., Porcar, M. (2017) From grass to gas: microbiome dynamics of grass biomass acidification under mesophilic and thermophilic temperatures. Biotechnol. Biofuels 10, 171.

[4] Akkarawatkhoosith, N., Kaewchada, A., Jaree, A. (2019) High-throughput CO_2 capture for biogas purification using monoethanolamine in a microtube contactor. J. Taiwan Inst. Chem. Eng. 98, 113–123.

[5] Banach, A., Ciesielski, S., Bacza, T., Pieczykolan, M., Ziembińska-Buczyńska, A. (2018) Microbial community composition and methanogens' biodiversity during a temperature shift in a methane fermentation chamber. Environ. Technol. 3, 1–12.

[6] Börjesson, P., Mattiasson, B. (2008) Biogas as a resource-efficient vehicle fuel. Trends Biotechnol. 26, 7–13.

[7] Borrel, G., Adam, P.S., Gribaldo, S. (2016) Methanogenesis and the Wood–Ljungdahl pathway: an ancient, versatile, and fragile association. Genome Biol. Evol. 8, 1706–1711.

[8] Brettin, T., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Olsen, G.J., Olson, R., Overbeek, R., Parrello, B., Pusch, G.D., Shukla, M., Thomason, J.A., 3rd, Stevens, R., Vonstein, V., Wattam, A.R., Xia, F. (2015) RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci. Rep. 5, 8365.

[9] Caspi, R., Billington, R., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P.E., Ong, Q., Ong, W.K., Paley, S., Subhraveti, P., Karp, P.D. (2018) The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. 46, D633–D639.

[10] Chun, J., Rainey, F.A. (2014) Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*. Int. J. Syst. Evol. Microbiol. 64, 316–324.

[11] Ciotola, R.J., Martin, J.F., Tamkin, A., Castaño, J.M., Rosenblum, J., Bisesi, M.S., Jiyoung, L. (2014) The influence of loading rate and variable temperatures on microbial communities in anaerobic digesters. Energies 7, 785–803.

[12] Conklin, A., Stensel, H.D., Ferguson, J. (2006) Growth kinetics and competition between *Methanosarcina* and *Methanosaeta* in mesophilic anaerobic digestion. Water Environ. Res. 78, 486–496.

[13] Corneli, E., Adessi, A., Olguín, E.J., Ragaglini, G., García-López, D.A., De Philippis, R. (2017) Biotransformation of water lettuce (*Pistia stratiotes*) to biohydrogen by *Rhodopseudomonas palustris*. J. Appl. Microbiol. 123, 1438–1446.

[14] De Mes, T.Z.D., Stams, A.J.M., Reith, J.H., Zeeman, G. (2003) Methane production by anaerobic digestion of wastewater and solid wastes. Bio-methane Bio-hydrogen, 58–102.

[15] DeSantis, T., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72, 5069–5072.

[16] De Vrieze, J., Hennebel, T., Boon, N., Verstraete, W. (2012) *Methanosarcina*: the rediscovered methanogen for heavy duty biomethanation. Bioresour. Technol. 112, 1–9.

[17] De Vriezea, J., Christiaens, M.E.R., Walraedt, D., Devooght, A., Ijaz, U.Z., Boon, N. (2017) Microbial community redundancy in anaerobic digestion drives process recovery after salinity exposure. Water Res. 111, 109–117.

[18] Doloman, A., Soboh, Y., Walters, A.J., Sims, R.C., Miller, C.D. (2017) Qualitative analysis of microbial dynamics during anaerobic digestion of microalgal biomass in a UASB reactor. Int. J. Microbiol. 2017, 5291283.

[19] Gaby, J.C., Zamanzadeh, M., Horn, S.J. (2017) The effect of temperature and retention time on methane production and microbial community composition in staged anaerobic digesters fed with food waste. Biotechnol. Biofuels 10, 302.

[20] Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M. (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91.

[21] Hafner, S.D., Rennuit, C. 2017 Predicting methane and biogas production with the biogas package https://cran.r-project.org/web/packages/biogas/vignettes/predBg_function.

[22] Hahnke, S., Langer, T., Koeck, D.E., Klocke, M. (2016) Description of *Proteiniphilum saccharofermentans* sp. nov., *Petrimonas mucosa* sp. nov. and *Fermentimonas caenicola* gen. nov., sp. nov., isolated from mesophilic laboratory-scale biogas reactors, and emended description of the genus *Proteiniphilum*. Syst. Evol. Microbiol. 66, 2454.

[23] Hanreich, A., Schimpf, U., Zakrzewski, M., Schlüter, A., Benndorf, D., Heyer, R., Rapp, E., Pühler, A., Reichl, U., Klocke, M. (2013) Metagenome and metaproteome analyses of microbial communities in mesophilic biogas-producing

anaerobic batch fermentations indicate concerted plant carbohydrate degradation. Syst. Appl. Microbiol. 36, 330–338.

[24] Hardegen, J., Latorre-Pérez, A., Vilanova, C., Günther, T., Porcar, M., Luschnig, O., Simeonov, C., Abendroth, C. (2018) Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass. Bioresour. Technol. 265, 275–281.

[25] He, D., Bultel, Y., Magnin, J.P., Roux, C., Willison, J.C. (2005) Hydrogen photosynthesis by *Rhodobacter capsulatus* and its coupling to a PEM fuel cell. J. Power Sources 141, 19–23.

[26] Heyer, R., Schallert, K., Siewert, C., Kohrs, F., Greve, J., Maus, I., Klang, J., Klocke, M., Heiermann, M., Hoffmann, M., Püttker, S., Calusinska, M., Zoun, R., Saake, G., Benndorf, D., Püttker, S. (2019) Metaproteome analysis reveals that syntrophy, competition, and phage-host interaction shape microbial communities in biogas plants. Microbiome 7, 69.

[27] Jeris, J.S., McCarty, P.L. (1965) The biochemistry of methane fermentation using C^{14} tracers. J. Water Pollut. Control Fed. 37, 178–192.

[28] Jiang, Y., Banks, C., Zhang, Y., Heaven, S., Longhurst, P. (2018) Quantifying the percentage of methane formation via acetoclastic and syntrophic acetate oxidation pathways in anaerobic digesters. Waste Manage. 71, 749–756.

[29] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353–D361.

[30] Kato, S., Hashimoto, K., Watanabe, K. (2012) Methanogenesis facilitated by electric syntrophy via (semi)conductive iron-oxide minerals. Environ. Microbiol. 14, 1646–1654.

[31] Kato, S., Igarashi, K. (2018) Enhancement of methanogenesis by electric syntrophy with biogenic iron-sulfide minerals. MicrobiologyOpen 6, e00647.

[32] Kerkhof, L.J., Dillon, K.P., Häggblom, M.M., McGuinness, L.R. (2017) Profiling bacterial communities by MinION sequencing of ribosomal operons. Microbiome 5, 116.

[33] Kim, D., Song, L., Breitwieser, F., Salzberg, S. (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721–1729.

[34] Kirkegaard, R.H., McIlroy, S.J., Kristensen, J.M., Nierychlo, M., Karst, S.M., Dueholm, M.S., Albertsen, M., Nielsen, P.H. (2017) The impact of immigration on microbial community composition in full-scale anaerobic digesters. Sci. Reports 7, 9343.

[35] Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glöckner, F.O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 41, pp.e1–e1.

[36] Koeck, D.E., Ludwig, W., Wanner, G., Zverlov, V.V., Liebl, W., Schwarz, W.H. (2015) *Herbinix hemicellulosilytica* gen. nov., sp. nov., a thermophilic cellulose-degrading bacterium isolated from a thermophilic biogas reactor. Int. J. Syst. Evol. Microbiol. 65, 2365–2371.

[37] Kuczynski, J., Stombaugh, J., Walters, W., Gonzalez, A., Caporaso, J., Knight, R. (2011) Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr. Protoc. Microbiol. 27, 1E.5.

[38] Lalitha, K., Swaminathan, K.R., Vargheese, C.M., Shanthi, V.P., Padma Bai, R.P. (1994) Methanogenesis mediated by methylotrophic mixed culture. Appl. Biochem. Biotech. 49, 113–134.

[39] Lindkvist, E., Johansson, M.T., Rosenqvist, J. (2017) Methodology for analysing energy demand in biogas production plants — a comparative study of two biogas plants. Energies 10, 1822.

[40] Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B. (2014) The carbohydrate-active enzymes database (CAZy). Nucleic Acids Res. 42, D490–D495.

[41] Ma, X., Stachler, E., Bibby, K. (2017) Evaluation of Oxford nanopore MinION sequencing for 16S rRNA microbiome characterization. BioRxiv, 099960.

[42] Martins, G., Salvador, A.F., Pereira, L., Alves, M.M. (2018) Methane production and conductive materials: a critical review. Environ. Sci. Technol. 52, 10241–10253.

[43] Maus, I., Koeck, D.E., Cibis, K.G., Hahnke, S., Kim, Y.S., Langer, T., Kreubel, J., Erhard, M., Bremges, A., Off, S., Stolze, Y., Jaenicke, S., Goesmann, A., Sczyrba, A., Scherer, P., König, H., Schwarz, W.H., Zverlov, V.V., Liebl, W., Pühler, A., Schlüter, A., Klocke, M. (2016) Unraveling the microbiome of a thermophilic biogas plant by metagenome and metatranscriptome analysis complemented by characterization of bacterial and archaeal isolates. Biotechnol. Biofuels 9, 171.

[44] Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P., Göker, M. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14, 60.

[45] Mizoguchi, T., Isaji, M., Harada, J., Tamiaki, H. (2012) Isolation and pigment composition of the reaction centers from purple photosynthetic bacterium *Rhodopseudomonas palustris* species. Biochim. Biophys. Acta 1817, 395–400.

[46] Na, S.I., Kim, Y.O., Yoon, S.H., Ha, S.M., Baek, I., Chun, J. (2018) UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. J. Microbiol. 56, 280–285.

[47] Niu, L., Song, L., Dong, X. (2008) *Proteiniborus ethanoligenes* gen. nov., sp. nov., an anaerobic protein-utilizing bacterium. Int. J. Syst. Evol. Microbiol. 58, 12–16.

[48] Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P. (2017) MetaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824–834.

[49] Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132.

[50] Parks, D., Imelfort, M., Skennerton, C., Hugenholtz, P., Tyson, G. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055.

[51] Pinnell, L.J., Dunford, E., Ronan, P., Hausner, M., Neufeld, J.D. (2014) Recovering glycoside hydrolase genes from active tundra cellulolytic bacteria. Can. J. Microbiol. 60, 469–476.

[52] Preeti Rao, P., Shivaraj, D., Seenayya, G. (1993) Succession of microbial population in cow dung and poultry litter waste digesters during methanogenesis. World J. Microbiol. Biotechnol. 33, 185.

[53] Reverso, R. Patent (2017), WO2017/085080A1.

[54] Richards, B.K., Herndon, F.G., Jewell, W.J., Cummings, R.J., White, T.E. (1994) *In situ* methane enrichment in methanogenic energy crop digesters. Biomass Bioenergy 6, 275–282.

[55] Richter, M., Rosselló-Móra, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A. 106, 19126–19131.

[56] Richter, M., Rossello-Mora, R., Glockner, F.O., Peplies, J. (2015) JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics 32, 929–931.

[57] Ritalahti, K.M., Justicia-Leon, S.D., Cusick, K.D., Ramos-Hernandez, N., Rubin, M., Dornbush, J., Löffler, F.E. (2012) *Sphaerochaeta globosa* gen. nov., sp. nov. and *Sphaerochaeta pleomorpha* sp. nov., free-living, spherical spirochaetes. Int. J. Syst. Evol. Microbiol 62, 210–216.

[58] Sato, T., Atomi, H., Imanaka, T. (2007) Archaeal type III RuBisCOs function in a pathway for AMP metabolism. Science 315 (5814), 1003–1006.

[59] Sawayama, S. Patent (2000), US006106719A.

[60] Sawayama, S., Hanada, S., Kamagata, Y. (2000) Isolation and characterization of phototrophic bacteria growing in lighted upflow anaerobic sludge blanket reactor. J. Biosci, Bioeng. 89, 396–399.

[61] Stams, A.J.M., Plugge, C. (2009) Electron transfer in syntrophic communities of anaerobic bacteria and archaea. Nature 7, 568–577.

[62] Sundberg, C., Al-Soud, W.A., Larsson, M., Alm, E., Yekta, S.S., Svensson, B.H., Sørensen, S.J., Karlsson, A. (2013) 454 pyrosequencing analyses of bacterial and archaeal richness in 21 full-scale biogas digesters. FEMS Microbiol. Ecol. 85, 612–626.

[63] Tada, C., Tsukahara, K., Sawayama, S. (2006) Illumination enhances methane production from thermophilic anaerobic digestion. Appl. Microbiol. Biotechnol. 71, 363–368.

[64] Tessler, M., Neumann, J.S., Afshinnekoo, E., Pineda, M., Hersch, R., Velho, L.F.M., Segovia, B.T., Lansac-Toha, F.A., Lemke, M., DeSalle, R., Mason, C.E., Brugler, M.R. (2017) Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci. Rep. 7, 6589.

[65] Walker, D.J., Martz, E., Holmes, D.E., Zhou, Z., Nonnemann, S.S., Lovely, D.R. (2019) The archaellum of *Methanospirillum hungatei* is electrically conductive. mBio 10, e00579-19.

[66] Wang, S., Hou, X., Su, H. (2017) Exploration of the relationship between biogas production and microbial community under high salinity conditions. Sci. Rep. 7, 1149.

[67] Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E.M., Disz, T., Gabbard, J.L., Gerdes, S., Henry, C.S., Kenyon, R.W., Machi, D., Mao, C., Nordberg, E.K., Olsen, G.J., Murphy-Olson, D.E., Olson, R., Overbeek, R., Parrello, B., Pusch, G.D., Shukla, M., Vonstein, V., Warren, A., Xia, F., Yoo, H., Stevens, R.L. (2017) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res. 45, D535–D542.

[68] Wei, H., Okunishi, S., Yoshikawa, T., Kamei, Y., Maeda, H. (2016) Isolation and characterization of a purple non-sulfur photosynthetic bacterium *Rhodopseudomonas faecalis* strain a from swine sewage wastewater. Biocontrol Sci. 21, 29–36.

[69] Wu, Y., Simmons, B., Singer, S. (2015) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607.

[70] Yang, Z., Koh, S.K., Ng, W.C., Lim, R.C., Tan, H.T., Tong, Y.W., Dai, Y., Chong, C., Wang, C.H. (2016) Potential application of gasification to recycle food waste and rehabilitate acidic soil from secondary forests on degraded land in Southeast Asia. J. Environ. Manage. 172, 40–48.

[71] Zhang, D., Yang, H., Huang, Z., Zhang, W., Liu, S.J. (2002) *Rhodopseudomonas faecalis* sp. nov., a phototrophic bacterium isolated from an anaerobic reactor that digests chicken faeces. Int. J. Syst. Evol. Microbiol. 52, 2055–2060.

[72] Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P.K., Xu, Y., Yin, Y. (2018) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 46 (W1), W95–W101.

[73] Zheng, Y., Harris, D.F., Yu, Z., Fu, Y., Poudel, S., Ledbetter, R.N., Fixen, K.R., Yang, Z.Y., Boyd, E.S., Lidstrom, M.E., Seefeldt, L.C., Harwood, C.S. (2018) A pathway for biological methane production using bacterial iron-only nitrogenase. Nat. Microbiol. 3, 281.

[74] Zinder, S.H. (1984) Microbiology of anaerobic conversion of organic wastes to methane: recent developments. ASM News 50, 294–298.

[75] Zhou, J., Zhang, R., Liu, F., Yong, X., Wu, X., Zheng, T., Jiang, M., Jia, H. (2016) Biogas production and microbial community shift through neutral pH control during the anaerobic digestion of pig manure. Bioresour. Technol. 217, 44–49.

Publication IV

frontiers in Microbiology

ORIGINAL RESEARCH
published: 12 May 2020
doi: 10.3389/fmicb.2020.00867

Chemically Stressed Bacterial Communities in Anaerobic Digesters Exhibit Resilience and Ecological Flexibility

Benjamin Schwan[1]†, Christian Abendroth[1,2*]†, Adriel Latorre-Pérez[3], Manuel Porcar[3,4], Cristina Vilanova[3] and Christina Dornack[1]

[1] Institute of Waste Management and Circular Economy, Technische Universität Dresden, Pirna, Germany, [2] Robert Boyle Institut e.V., Jena, Germany, [3] Darwin Bioprospecting Excellence, S.L. Parc Cientific Universitat de València, Paterna, Spain, [4] Institute for Integrative Systems Biology, University of Valencia-CSIC, Paterna, Spain

OPEN ACCESS

Edited by:
Simona Rossetti,
Water Research Institute, Italian
National Research Council, Italy

Reviewed by:
James Scott Maki,
Marquette University, United States
Nuria Fernandez-Gonzalez,
University of Valladolid, Spain

***Correspondence:**
Christian Abendroth
christian.abendroth@tu-dresden.de

†These authors have contributed equally to this work

Specialty section:
This article was submitted to
Microbiotechnology,
a section of the journal
Frontiers in Microbiology

Received: 09 October 2019
Accepted: 14 April 2020
Published: 12 May 2020

Citation:
Schwan B, Abendroth C, Latorre-Pérez A, Porcar M, Vilanova C and Dornack C (2020) Chemically Stressed Bacterial Communities in Anaerobic Digesters Exhibit Resilience and Ecological Flexibility. Front. Microbiol. 11:867. doi: 10.3389/fmicb.2020.00867

Anaerobic digestion is a technology known for its potential in terms of methane production. During the digestion process, multiple metabolites of high value are synthesized. However, recent works have demonstrated the high robustness and resilience of the involved microbiomes; these attributes make it difficult to manipulate them in such a way that a specific metabolite is predominantly produced. Therefore, an exact understanding of the manipulability of anaerobic microbiomes may open up a treasure box for bio-based industries. In the present work, the effect of nalidixic acid, γ-aminobutyric acid (GABA), and sodium phosphate on the microbiome of digested sewage sludge from a water treatment plant fed with glucose was investigated. Despite of the induced process perturbations, high stability was observed at the phylum level. However, strong variations were observed at the genus level, especially for the genera *Trichococcus, Candidatus Caldatribacterium,* and *Phascolarctobacterium.* Ecological interactions were analyzed based on the Lotka–Volterra model for *Trichococcus*, *Rikenellaceae DMER64*, *Sedimentibacter*, *Candidatus Cloacimonas*, *Smithella, Cloacimonadaceae* W5 and *Longilinea*. These genera dynamically shifted among positive, negative or no correlation, depending on the applied stressor, which indicates a surprisingly dynamic behavior. Globally, the presented work suggests a massive resilience and stability of the methanogenic communities coupled with a surprising flexibility of the particular microbial key players involved in the process.

Keywords: anaerobic digestion, Lotka–Volterra, population modeling, anaerobic microbiomes, microbiome manipulation

BACKGROUND

In previous decades, tremendous efforts have been made to better understand the biocenosis underlying the process of anaerobic digestion. According to a recent study, approximately 300 operational taxonomic units (OTUs) represent 80% of the microorganisms involved in anaerobic digester microbiomes. If the remaining 20% are also taken into consideration, the number of

OTUs is much higher (Kirkegaard et al., 2017). Moreover, an often complex and inhomogeneous feedstock is used, which can affect microbial community structures and functions (Xu et al., 2018). To gain better access to microbial systems of such complexity, high-throughput approaches are often applied, such as 16S-rRNA gene amplicon sequencing (Abendroth et al., 2015), metagenomics (Xu et al., 2019); or metaproteomics (Hassa et al., 2018), all of which facilitate the analysis of complex microbial communities with high diversity. The continuously decreasing prices of these technologies have allowed scientists to compare many anaerobic digester plants simultaneously. For example, Sundberg et al. (2013) compared 21 full-scale anaerobic digesters, including co-digesters and sewage sludge digesters, based on 16S-rRNA amplicon sequencing at both mesophilic and thermophilic temperatures. In the study by Sundberg et al. (2013), Actinobacteria, Proteobacteria, Chloroflexi, Spirochetes and Euryarchaeota were dominant in sewage sludge digesters, while Firmicutes were especially enriched in co-digesters. Theuerl et al. (2015) indicated that even well-operating agricultural biogas plants show fluctuation in the microbial community composition due to high sensitivity to changes in the process performance.

The aforementioned studies provide good insight into microbial key players involved in the process of anaerobic digestion. However, to understand the reasons behind the observed taxonomic patterns, complex experiments are necessary, which usually involve disturbing the system to identify the changes associated with the new environment. The experiments reported include stressors like very low pH of 6.0 (Delbès et al., 2000; Hori et al., 2006; Abendroth et al., 2017), changing temperature (Shi et al., 2019), very high salt concentrations causing conductivity values up to 80 mS cm^{-1} (Ogata et al., 2016; De Vrieze et al., 2017) and varying total solids (TS) contents (Hardegen et al., 2018). For instance, to further test the hypothesis that the genus *Methanosarcina* is especially enriched in anaerobic digester sludge with low viscosity (Abendroth et al., 2015), an experiment was conducted in which sewage sludge was fed in parallel with various feedstocks with different percentages of TS (Hardegen et al., 2018). Hardegen et al. (2018) gradually increased the concentration of total volatile fatty acids (up to 10 g L^{-1} before acidosis took place); as the researchers anticipated, the approach in which a feedstock with a low percentage of TS was used resulted in higher concentrations of *Methanosarcina* than the approach with feedstocks with high concentrations of TS were fed did. In another example, Spirito et al. (2018) used antibiotics up to concentrations of 5 mg L^{-1} (monensins) to disturb the underlying microbiome. An adaptation to extremely high concentrations of monensins was possible, which was explained by the authors with a highly redundant microbiome, in which the inhibited species can be substituted by other microorganisms with similar functions.

Experiments with such harsh conditions-like those in the experiments performed by De Vrieze et al. (2017) and Spirito et al. (2018)-make it possible to study the microbial shifts caused by different stress levels; however, this provides no insight *per se* into the microbial interactions that are driving these shifts. With massive sequencing data, it would be possible to find biological correlations by, for example, pairwise comparisons or regression- and rule-based networks, enabling an approximate calculation of microbial interactions (Faust and Raes, 2012). According to Faust and Raes (2012), this would make it possible to determine whether positive, negative or neutral effects exist between different species, indicating potential ecological interactions, such as mutualism, commensalism, parasitism, amensalism or competition. Because of this, scientists are regularly trying to understand microbial interactions within anaerobic microbiomes through sequencing data. For example, Kuroda et al. (2016) analyzed the correlations between multiple OTUs within granules from an anaerobic upstream sludge blanket (UASB). In that work, many positive correlations between methanogens and syntrophic bacteria were highlighted. The existing microbial interaction between syntrophs and methanogens has been investigated since the 1980s (Baresi et al., 1978), and the work of Kuroda et al. (2016) highlighted the applicability of sequencing-based information on microbial ecology. In many more studies, based on sequencing approaches, to shed light on microbial interactions. Very often, network analysis is used to analyze the evolution of microbiomes based on 16S-rRNA gene amplicon sequencing in response to a certain environmental stress. For instance, a recently applied network analysis demonstrated that organic overloading causes microbial population shifts, which in turn affects microbial interactions (Braz et al., 2019).

Although several reports have investigated microbial interactions within anaerobic microbiomes, to date, it has not been determined whether interactions may be restricted to certain environmental conditions. For example, it is conceivable that two mutualistic bacteria shift into a state of parasitism due to changing digester conditions in which the feedstock composition changes. Using Lotka–Volterra based modeling, the presented work aims to address the question of how microorganisms in anaerobic microbiomes are ecologically adapting to externally induced fluctuations. To answer this question, four semicontinuously fed reactors were treated over 9 weeks while receiving different inhibiting substances, namely nalidixic acid, γ-aminobutyric acid (GABA) and sodium phosphate. Following this, 16S-rRNA gene amplicon sequencing and Lotka–Volterra modeling were applied to address the microbial interactions in all four reactors. Based on DNA sequencing, gLV has already been applied various times to investigate microbial interactions in the gut (Weng et al., 2017), in cheese (Mounier et al., 2008), in the coffee-machine bacteriome (Vilanova et al., 2015) and its suitability to simulate population dynamics and estimate microbial interactions based on high-throughput sequencing was recently highlighted by Kuntal et al. (2019).

MATERIALS AND METHODS

Inoculum and Substrates

As seed sludge, a digester sludge from a sewage plant in Saxonia was used. The sludge came from the digestion towers of a large sewage treatment plant in Saxony, Germany. The average solids retention time (SRT) in the digestion towers is 16.5 days. Biogas is produced under mesophilic conditions in the range of 30–35°C.

The average pH value is 7.7. The TS content varies between 3 and 5 g L^{-1} per year. The sum of the volatile fatty acids (VFA) amounts to 163 mg L^{-1} on average. At the time of sampling, this sum parameter was 169 mg L^{-1}. The ammonium content was 1157 mg L^{-1}.

The reactors were supplemented with nalidixic acid (Sigma Aldrich, Germany), GABA (Sigma Aldrich) or sodium phosphate (Sigma, Aldrich), which were applied as stressors during the last 5 weeks, as shown in **Figure 1**. To prevent starvation, glucose was used as substrate.

Reactor Performance

The anaerobic digester experiments lasted 11 weeks and were performed using custom-built continuous stirred tank reactors (CSTRs), which were used in fed-batch configuration. The reactors had a volume of 5 L, with a 3 L working volume (**Figure 1**). After 1 week without feeding, the reactors received glucose three times a week. For feeding, glucose was dissolved in 150 mL of fresh sludge from a sewage sludge digester. Since feeding events took place discontinuously and the amount of applied substrate and stressors varied during the experiment, the organic loading rate (OLR) could only be estimated. For determining the OLR, the daily flow rate of volatile solids (VS) was calculated by dividing the sum of VS per week by 7. Initially, 1 g L^{-1} of glucose was used, which is equivalent to a loading rate (OLR) of 0.43 gVS L^{-1} d^{-1}. After the third week, the reactors received three times a week 3 g of glucose per liter, which corresponds to a loading rate of 1.29 gVS L^{-1} d^{-1} (**Figure 2**), and this loading rate was retained until the end of the experiment (week 7). Before each feeding event, 150 mL of digestate was removed and used for chemical analysis. Therefore, the retention time was approximately 46.66 days.

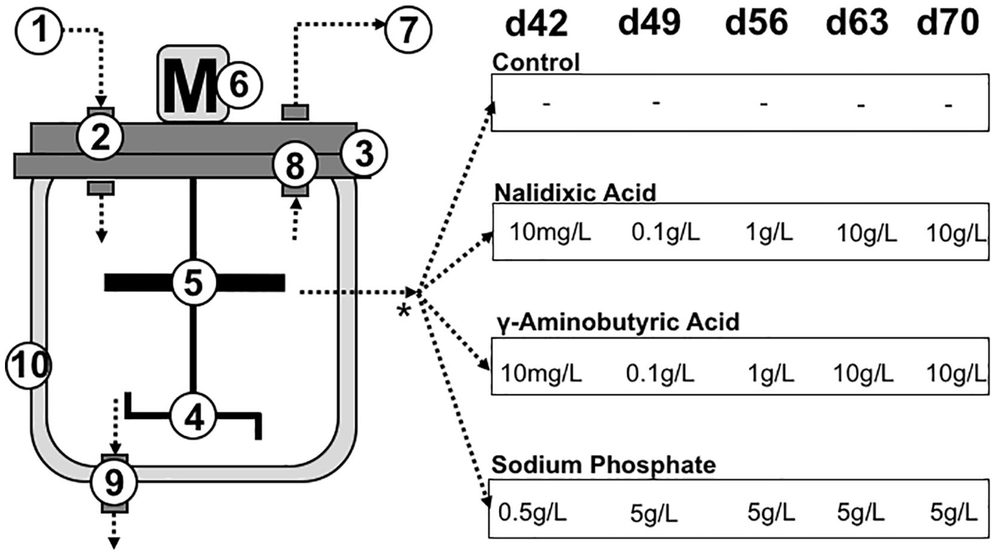
Beginning from week 7, three of the four digesters received a chemical stressor with the goal of causing disturbances in the digestion process and the underlying microbiomes (**Figure 1**). The inhibiting chemicals, which were applied to the different digesters, were fed once a week, using nalidixic acid, GABA and sodium phosphate. The fourth reactor received only the substrate (glucose) and no further supplements. The amount of stressor fed into the respective reactors was increased from 0.5 g L^{-1} to 5 g L^{-1} for sodium phosphate and from 10 mg L^{-1} to 10 g L^{-1} for nalidixic acid and GABA, as shown in **Figure 1**. Since both of them are organic substances, adding nalidixic acid and GABA increased the OLR. In weeks 7 and 8, nalidixic acid and GABA were applied in such small amounts that the OLR was only changed to the third decimal place. However, from week 9 onward, much higher amounts of stressors were applied (**Figure 1**). In week 9, the OLR was increased to 1.43 gVS L^{-1} d^{-1} and during the last 2 weeks, the OLR reached 2.72 gVS L^{-1} d^{-1}. In the case of the reactor receiving sodium phosphate, the OLR remained at 1.29 gVS L^{-1} d^{-1} throughout the experiment, since sodium phosphate is an inorganic substance.

Chemical Analysis

Biogas was analyzed simultaneously with each feeding event (three times a week, **Figure 2**). The exhaust gas measuring device "Abgasmessgerät VISIT 02 S" from Messtechnik Eheim GmbH (Germany) was used for gas analysis. This measuring device is certified according to the German legal requirements of the Federal Immission Control Act. The device is calibrated at least twice a year with test equipment according to DIN ISO 10012. The detectable gases are oxygen, methane, carbon dioxide, hydrogen sulfide and hydrogen with a volumetric flow rate of 0.8 L min^{-1}. Methane and carbon dioxide were detected with an infrared double beam sensor. Oxygen and hydrogen sulfide were detected by a carbon dioxide-compensated electrochemical sensor. The hydrogen content was determined by a palladium sensor. The detection limits for oxygen, carbon dioxide and methane are 0.1 vol%, for hydrogen and hydrogen sulfide 10 ppm with an error of ±1% of the measured value. On analyzing the gas composition, the gas was dried in a custom-built column filled with silica gel. The quantity of the dry gas was analyzed using a common gas meter (BK G6, Elster Handek GmbH Mainz, Germany). Based on the guideline VDI 4630 from the Association of German Engineers (2016), the gas volume was normalized to standard temperature (273 K) and standard pressure (1013 hPa). During the treatment (weeks 7–11), the concentration of chemical oxygen demand (COD) and total volatile fatty acids (TVFAs) were measured once a week. The COD was measured in the untreated sludge (total COD) and in the liquid phase after centrifugation (solubilized COD). The first step of solids separation was carried out via a centrifuge at 13,000 g. The second treatment step was vacuum filtration through a 0.2-μm cellulose-acetate filter (Sartorius AG, Göttingen, Germany). Finally, COD was measured with the Spectroquant COD kit (VWR, Germany) according to the manufacturer's guidelines. The spectrum of VFAs (lactic acid, formic acid, acetic acid, propionic acid, iso-butyric acid, butyric acid, and valeric acid) was determined by ion chromatography using the Metrosep Organic Acids 250/7.8 column (Model: 882 Compact IC plus, Metrohm AG, Herisau, Switzerland). The applied column is a cation exchange column, which is particularly designed for the determination of VFAs. The mobile phase contained 0.6 mmol L^{-1} of perchloric acid 10 mmol L^{-1} of lithium chloride. The detection limit is 0.25 mg L^{-1}. The amount of TVFAs was determined as the sum of all measured VFAs.

DNA Extraction and Sequencing

Before DNA extraction, samples were washed to reduce the amount of inhibiting substances (especially humic acids). For the first sample (**Figure 1**, day 0), biomass was sedimented by centrifugation for 5 min at 20,000 *g* and washed several times with sterile phosphate-buffered saline (PBS buffer). Because increasing viscosity sedimentation was impaired in the following extractions, at this point, the centrifugation time was increased to 10 min for all remaining samples. DNA extraction was performed using the DNEasy Power Soil Kit (Qiagen, Netherlands) according to the manufacturer's instructions. Extracted DNA was quantified using the Qubit dsDNA HS Assay kit (Qubit 2.0 Fluorometer, Thermo Fisher, Waltham, United States). The bacterial full-length 16S rRNA gene was amplified by polymerase chain reaction (PCR) using the following universal primers: S-D-Bact-0008-a-S-16



	d42	d49	d56	d63	d70
Control	-	-	-	-	-
Nalidixic Acid	10mg/L	0.1g/L	1g/L	10g/L	10g/L
γ-Aminobutyric Acid	10mg/L	0.1g/L	1g/L	10g/L	10g/L
Sodium Phosphate	0.5g/L	5g/L	5g/L	5g/L	5g/L

FIGURE 1 | Reactor set-up: To compare the influence of different inhibitors on microbial interactions, four experiments were performed in continuous stirred tank reactors. The left side of the figure shows the reactor design: **(1)** Substrate input, **(2)** gas tight feeding pipe, **(3)** lid, **(4)** stirrer, **(5)** upper stirrer for foam elimination, **(6)** motor for the stirrer, **(7)** gas output, **(8)** gas tight pipe, **(9)** sampling port, **(10)** double wall for heating. The experiment lasted 77 days, with chemical inhibitors applied on day 42, 49, 56, 63, and 70. Nalidixic acid, γ-aminobutyric acid (GABA) and sodium phosphate were applied as chemical inhibitors. One reactor received no additional inhibiting chemicals to function as a control.

(5′-AGRGTTYGATYMTGGCTCAG-3′) and S-D-Bact-1492-a-A-16 (5′-TACCTTGTTAYGACTT-3′) (Klindworth et al., 2012). The PCR reaction mix consisted of 200 μM dNTPs, 200 nM of each primer, 1 U of VWR Taq DNA Polymerase (VWR®, WR International bvba/sprl, Belgium), 1 x PCR buffer supplemented with MgCl2 (1.5 mM), and 1 ng of DNA template (final volume: 20 μL). The PCR amplification protocol comprised an initial denaturation step at 94°C for 1 min, followed by 35 cycles of amplification (denaturing, 1 min at 95°C; annealing, 1 min at 49°C; extension, 2 min at 72°C) and a final extension at 72°C for 10 min. A negative control (no DNA) was also included. Following the PCR reaction, DNA concentrations were measured using the Qubit dsDNA HS Assay kit (Qubit 2.0 Fluorometer, Thermo Fisher, Waltham, United States). The resulting amplicons were sequenced with Oxford Nanopore MinION, as previously described (Hardegen et al., 2018). In total, 39 samples were multiplexed in the same run using the EXP-PBC096 barcoding kit. The recommended ONT protocols were followed for priming and loading the flow cell. Raw sequences were uploaded at the National Center for Biotechnology Information[1].

Reads were basecalled with MinKNOW software (core version 3.3.2), and sequencing statistics were assessed by the EPI2ME (v2.59.1896509) 'Fastq Barcoding' protocol. Porechop[2] was applied for detection of the barcodes, demultiplexing of the samples and removal of the adaptors. Finally, reads shorter than 400 base pairs (pb) or with a mean quality below 7 (in PHRED score) were removed.

Taxonomic Analysis and Modeling

Full-length 16S rRNA sequences generated by MinION were used to obtain a taxonomic profile of each sample. Reads were classified using the Quantitative Insights Into Microbial Ecology (QIIME 1.9.1.) software (Caporaso et al., 2010). OTUs were constructed using the 'pick_otus.py' script, and uclust as the picking method (similarity threshold = 97%). Then, 'pick_rep_set.py' was run with the default parameters. Taxonomic assignment was carried out with the 'assign_taxonomy.py' script, and this consisted of BLAST searches against the latest version (v. 132) of the SILVA database. Finally, 'make_otu_table.py' was employed to obtain the final OTU table.

The QIIME results were used to perform simulations based on generalized Lotka–Volterra (gLV) models for each condition studied. The gLV model is an extension of the classic predator-prey Lotka–Volterra model, which allows the prediction of a wider range of relationships (competition, cooperation, neutralism, etc.) among the individual species —or OTUs— coexisting in the same habitat. The interaction could be directly interpreted from the algebraic sign of a coefficient incorporated to the equation (Kuntal et al., 2019). To reduce computation efforts and obtain comparable results, only the most abundant taxa detected in all the experiments were selected for the gLV simulations. Further analyses were performed using the R-software for statistical computing. Differential abundance

[1] https://www.ncbi.nlm.nih.gov/bioproject/PRJNA554976

[2] https://github.com/rrwick/Porechop

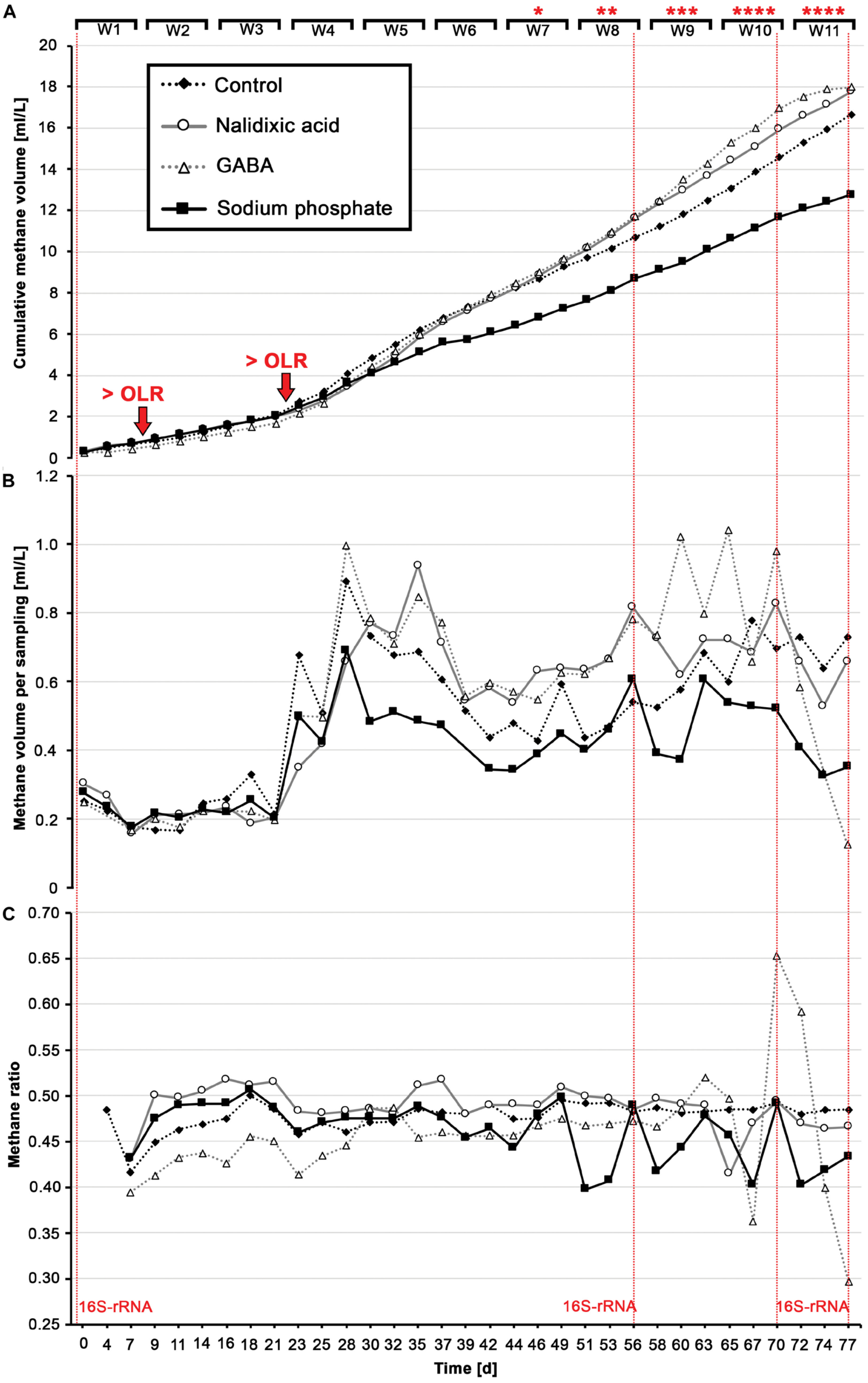


FIGURE 2 | Produced biogas: Cumulative methane **(A)**, the amount of methane per sampling day **(B)** and the ratio of methane to total biogas for each sampling day **(C)** are shown for all four digesters in response to perturbation with nalidixic acid, γ-aminobutyric acid (GABA) and sodium phosphate. The fourth reactor acted as a control, with no inhibiting substances. Organic loading rates (OLR) were increased after week 1 (0.43 g/L d^{-1}) and after week 3 (1.29 g/L d^{-1}). At days 0, 56, 70, and 77, 16S-rRNA gene samples were taken for all four reactors (highlighted with horizontal lines in red).

analyses were carried out using the DESeq2 package (Love et al., 2014; v. 1.18.1) to detect variations in the microbial composition among the different treatments and the control. The 'phyloseq_to_deseq2' function was applied to convert the phyloseq object into a DESeq2 object. Then, the DESeq2 main function was applied using the 'parametric' option for fitting the dispersion and the 'Wald test' option for calculating the significance of the resulting coefficients. The Benjamini–Hochberg method was used for adjusting the *p*-values, and only features with an adjusted *p*-value lower than 0.05 were considered significant.

RESULTS AND DISCUSSION

Methane Production Upon Addition of Microbial Stressors

The aim of the present work was to cause multiple taxonomic shifts outgoing from the same anaerobic microbiome. Extensive shifts were intended to facilitate the analysis of ecological interactions among involved microorganisms based on population dynamics analysis. Sodium phosphate was used as it is a known stressor in anaerobic digestion processes (Ogata et al., 2016). The antibiotic nalidixic acid was chosen as a stressor, as antibiotics are known to manipulate anaerobic process performance and the involved microbiomes (Mitchell et al., 2013; Mustapha et al., 2016; Bay et al., 2019; Fáberová et al., 2019). GABA was chosen, as high concentrations of butyric acid (an intermediate product from GABA degradation) is known to inhibit syntrophic metabolism in anaerobic digesters (Henson et al., 1985; Zhang et al., 2019).

The experiments started with a low OLR (0.43 gVS L^{-1} d^{-1}), with the OLR being elevated after 3 weeks (1.29 gVS L^{-1} d^{-1}; **Figure 2A**), which destabilized the digestion experiments from week 3 until week 6 (**Figure 2B**). Beginning in week 7, nalidixic acid, GABA and sodium phosphate were also added weekly, and in increasing amounts, to cause a process perturbation, and thus, multiple alterations in the underlying microbiome. Due to the addition of GABA and nalidixic acid, the OLR increased gradually to 2.72 gVS L^{-1} d^{-1} during the last 5 weeks for both cases. In the case of the reactor receiving sodium phosphate, the OLR remained at 1.29 gVS L^{-1} d^{-1} as it is an inorganic substance.

During the 11 weeks of the experiment, all reactors received a total of 78.26 g of glucose per liter, which corresponds to a theoretical methane potential of 28.96 L of methane. The control produced 16.66 L of methane per liter of working volume (**Figure 2A**). Therefore, the digestion efficiency was 57.53%. A similar methane volume would have been expected for the reactor that was supplemented with sodium phosphate, because sodium phosphate cannot be converted into methane. However, the reactor that received sodium phosphate produced only 12.75 L of methane per liter of working volume. Since the cumulative gas volume was already lower than the control, before sodium phosphate was added, a process perturbation due to sodium phosphate cannot entirely explain the lowered cumulative methane volume (**Figure 2A**). However, the fact that the ratio of methane to total biogas became highly irregular upon the addition of sodium phosphate (**Figure 2C**, weeks 7 – 11) indicates a process perturbation, which may have affected the methane productivity negatively during the last 5 weeks. This hypothesis is supported by the fact that the pH gradually decreased from 7.55 to 6.57 (**Figure 3C**).

The reactors receiving nalidixic acid and GABA, in addition to the 78.26 g of glucose, received 21.11 g L^{-1} of the respective stressor. In case of complete degradation, 0.58 L g^{-1} would be expected for nalidixic acid, and 0.49 L g^{-1} for GABA (**Supplementary Material S1**). In the case of the reactor, which received nalidixic acid, this 21.11 g L^{-1} of stressor corresponds to an additional theoretical methane potential of 12.24 L. In the case of GABA, 21.11 g L^{-1} of stressor corresponds to 10.34 L of methane. Based on the digestion efficiency of 57.53%, which was observed in the control, the reactors receiving nalidixic acid and GABA were expected to produce 7.04 L and 5.95 L more methane per liter than the control did. However, in both cases, the produced volume of methane was extremely close to the control. This suggests that the respective stressors were not entirely converted to methane. One explanation is that the respective stressors were not degradable. Another explanation is an inhibition of the underlying microbiome.

Chemical Parameters

Although the methane productivity alone did not indicate a very clear variation between the performed digestions experiments, chemical parameters did show some differences. As mentioned above, the ratio of methane to total biogas became highly irregular with the addition of sodium phosphate (**Figure 2C**, weeks 7-11). From this, one can assume a humble but continuous inhibition of the reactor receiving sodium phosphate, resulting in a pH decrease from 7.55 to 6.57 at the end of the experiment (**Figure 3C**). Comparing the result of the reactor receiving sodium phosphate to other works, it draws attention that the loading rate must usually be higher to cause acidosis. In an experiment by Goux et al. (2015), where the OLR was gradually increased, acidosis took place approximately at 4 gVS L^{-1} d^{-1}. In a recent study by Musa et al. (2018), an UASB reactor showed an even higher stability compared with that of Goux et al. (2015), as the OLR was increased until 15 gCOD L^{-1} d^{-1} before acidosis took place. In the study presented here, the loading rate in the reactor receiving sodium phosphate was based on the works from Goux et al. and Musa et al., and an OLR of 1.29 gVS L^{-1} d^{-1}, not close to a range that could cause acidosis. This supports the interpretation that the observed process disturbance was caused by high concentrations of sodium phosphate.

In contrast to the reactor receiving sodium phosphate, a very sudden and heavy shock was observed in the reactor receiving GABA as stressor, which resulted in a strong increase in solubilized COD and TVFAs beginning in week 9 (**Figures 3B,D** and **Supplementary Figure S1**). In addition, as expected, this aforementioned COD and TVFA shock coincided with strong irregularities in methane productivity, which was almost fully disrupted by the end of the experiment (**Figures 2A,B**;

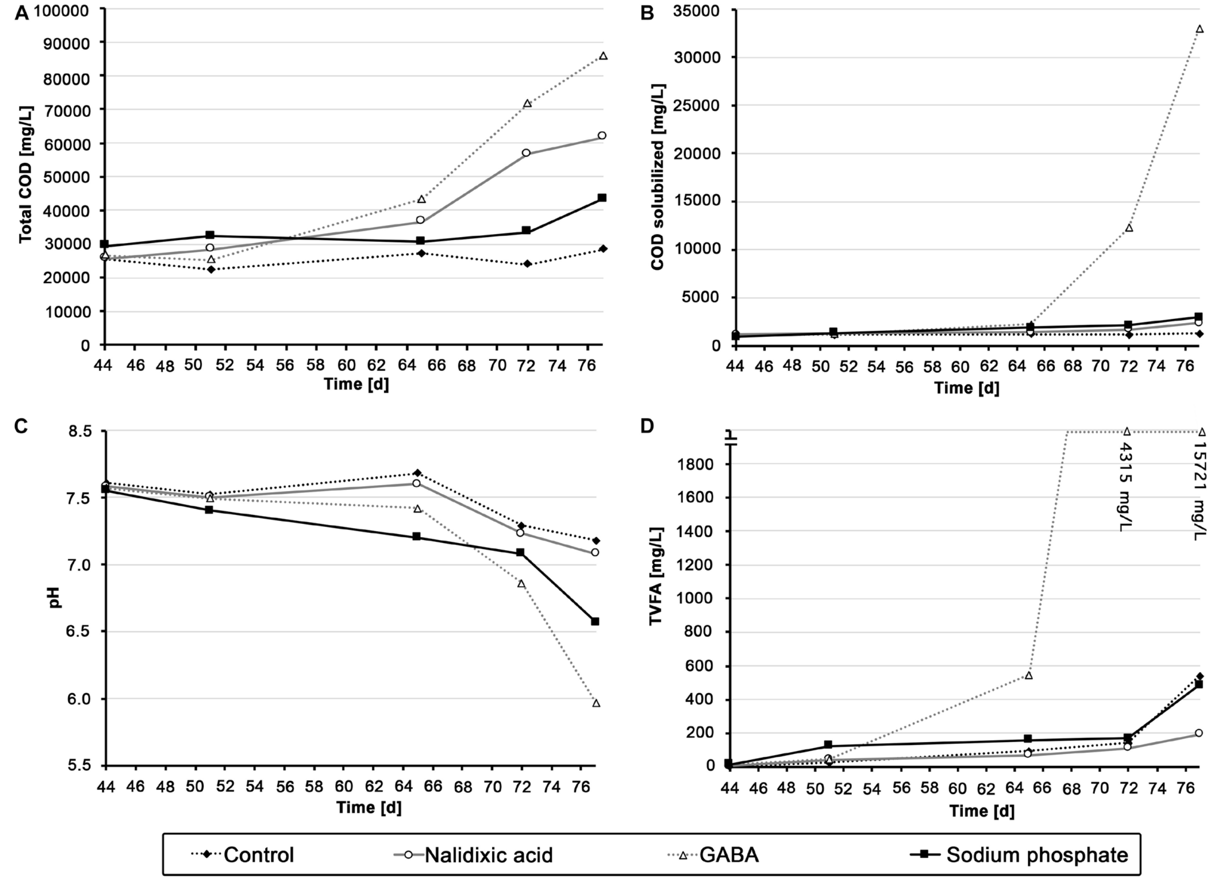


FIGURE 3 | Chemical analysis of the different digesters: Total chemical oxygen demand **(A)**, solubilized chemical oxygen demand **(B)**, pH **(C)** and amount of total volatile fatty acids **(D)** are shown for all the digester experiments, which were perturbed with nalidixic acid, γ-aminobutyric acid (GABA) and sodium phosphate.

day 77) and showed a strongly reduced ratio of methane (**Figure 3C**, day 77).

Compared with the acidosis events in the aforementioned works from Goux et al. (2015) and Musa et al. (2018), it appears that the OLR in the present study (max. 2.72 gVS L^{-1} d^{-1}) was still too small to cause acidosis. As discontinuous fed-batch reactors were used in the present work, one could argue that shock loads may have destabilized the process. However, in an experiment by Nachaiyasit and Stuckey (1997), shock loads with OLR as high as 18 gCOD L^{-1} d^{-1} were applied over a duration of 20 days without causing acidosis. Therefore, it appears unlikely that a substrate overload caused acidosis in the present experiment. A potential explanation could be the aforementioned release of butyric acid, which is a known inhibitor of anaerobic digestion processes (van den Heuvel et al., 1988).

The chemical parameters for the reactor treated with nalidixic acid were particularly unexpected. As explained in the previous section, the methane yield was lower than anticipated, indicating an uncomplete degradation and/or inhibitory effect in the process. Due to the low methane yield, one would expect an increase in TVFAs or COD in the liquid fraction. However, TVFAs and solubilized COD remained at a low level, with a concentration of less than 600 mg L^{-1} (**Figures 3B,D**). However, at the end of the experiment, a strong increase in the total COD up to 61.60 g L^{-1} was observed. A potential explanation for these findings is an impaired degradation due to adsorption. This hypothesis is supported by the fact that the antibiotics ampicillin, norfloxacin, ciprofloxacin, ofloxacin, tetracycline, roxithromycin, and trimethoprim are mainly removed from sewage systems due to adsorption (Li and Zhang, 2010).

Taxonomic Profiles After Treatment

As the basis for population modeling based on the Lotka–Volterra equations, high-throughput sequencing of 16S-rRNA gene amplicons was applied. To create a general overview of the produced data, Bray–Curtis dissimilarities were calculated and analyzed based on a principal component analysis for ordination (**Figure 4**). The control was extremely different from the rest of the time points. At the beginning of the time period, in which supplementation with the respective chemical stressors started (day 56), all the samples clustered close to each other. However, at day 70, the underlying microbiomes had already clearly diverged.

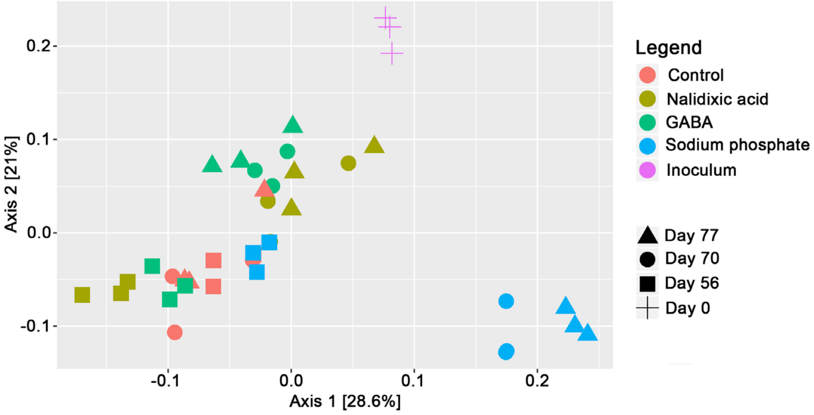


FIGURE 4 | Principal component analysis of 16S-rRNA gene amplicon sequences after calculation of Bray-Curtis dissimilarities at the genus level.

For days 70 and 77, the samples from the reactor receiving sodium phosphate clustered far away from the reactors receiving nalidixic acid and GABA. Interestingly, and despite showing clear differences in the underlying chemical parameters (**Figure 3**), the reactors receiving nalidixic acid and GABA clustered together. The respective taxonomic profiles for all reactors are shown in **Figure 5**.

The dominant phyla observed during the experiment were Bacteroidetes (38.82% ± 5.08%) and Firmicutes (19.87% ± 6.68%), which is in line with other studies (Klocke et al., 2007; Sundberg et al., 2013; Abendroth et al., 2015, 2017). Phyla that were observed in minor ratios, were Patescibacteria (9.13% ± 2.73%), Chloroflexi (8.10% ± 1.62%), Proteobacteria (6.56% ± 1.76%), Cloacimonetes (5.36% ± 1.91%), Verrucomicrobia (3.18% ± 1.14) and Spirochaetes (2.41% ± 1.45%). These minor phyla are also typical of digested sewage sludge (Abendroth et al., 2015). The taxonomic patterns were surprisingly similar in all the experiments, despite of the process perturbations due to the addition of nalidixic acid, GABA and sodium phosphate. Such stability at the phylum level has been indicated in other studies. For example, in the work from Calusinska et al. (2018) 20 mesophilic full-scale bioreactors were monitored over a time period of 1 year, and a surprisingly stable core microbiome was revealed. In addition, with harsher conditions, the underlying microbiome shows robustness. For example, the effect of thermoshocks on high-strength liquor from an acidifying pre-treatment stage for an anaerobic digester sludge was investigated, and the frequencies of phyla remained stable despite the harsh heat shocks applied (Abendroth et al., 2018).

Despite of the high robustness of anaerobic digester microbiomes at the phylum level, a shift was detected for Bacteroidetes with addition of nalidixic acid at day 56 (**Figure 5A**). As the antibiotic nalidixic acid affects gram-negative bacteria, the inhibition of Bacteroidetes was expected. However, more Gram-negative groups should also have been affected. Moreover, the primordial ratio of Bacteroidetes was already re-established at day 70, indicating a rapid adaptation by the involved Gram-negative bacteria. To obtain a deeper understanding of the respective taxonomic shifts, a differential analysis was applied, in which differences among perturbated reactors and the control experiment were analyzed (**Supplementary Tables S1–S3**). Although the difference for Bacteroidetes at day 56 appeared to be clear compared with days 70 and 77, a differential abundance analysis indicated no significant differences, when comparing the results from day 56 to the control experiment.

In the subsequent discussion, only significant changes with $p < 0.05$ were considered. Compared with the control, it appeared that nalidixic acid caused significant increases in the ratio of Firmicutes, Tenericutes, Cloacimonetes, and Lentispheara. In contrast, Patescibacteria and Nitrospirae showed a significant decrease. Particularly Tenericutes and Nitrospirae seem to have been strongly affected by nalidixic acid, as they were affected at more than one time point. Despite of their statistical significance, it must be highlighted that the respective shifts were extremely small (**Figure 5A**). An explanation for this robustness may be a high antibiotic resistance of microbiomes from digested sewage sludge, which has already been highlighted by multiple authors (e.g., Amador et al., 2015; Naquin et al., 2015; Karkman et al., 2018; Yin et al., 2019).

Although the performed principal component analysis indicated a high similarity for the microbiomes that were treated with nalidixic acid and GABA (**Figure 4**), they showed some differences in relation to the control. Atribacteria and Fibrobacteres were reduced in the reactor receiving GABA but not in the reactor receiving nalidixic acid. An increase was observed for the phyla Epsilonbacteraeota and Spirochaetes, which was not observed in the reactor receiving nalidixic acid neither. Interestingly, the phyla Tenericutes and Nitrospirae were also affected by GABA, as was the case with nalidixic acid and with sodium phosphate. This similar shift behavior indicates a high robustness for Tenericutes, as well as a high sensitivity for Nitrospirae. Nitrospirae are known to occur regularly in wastewater treatment plants (Zhang et al., 2017); however, to our knowledge, there are no reports that link Nitrospirae with perturbated conditions in anaerobic digesters. At any rate, the described sensitivity is supported by Daims (2014) work, which highlighted the difficulties in cultivating Nitrospirae, especially

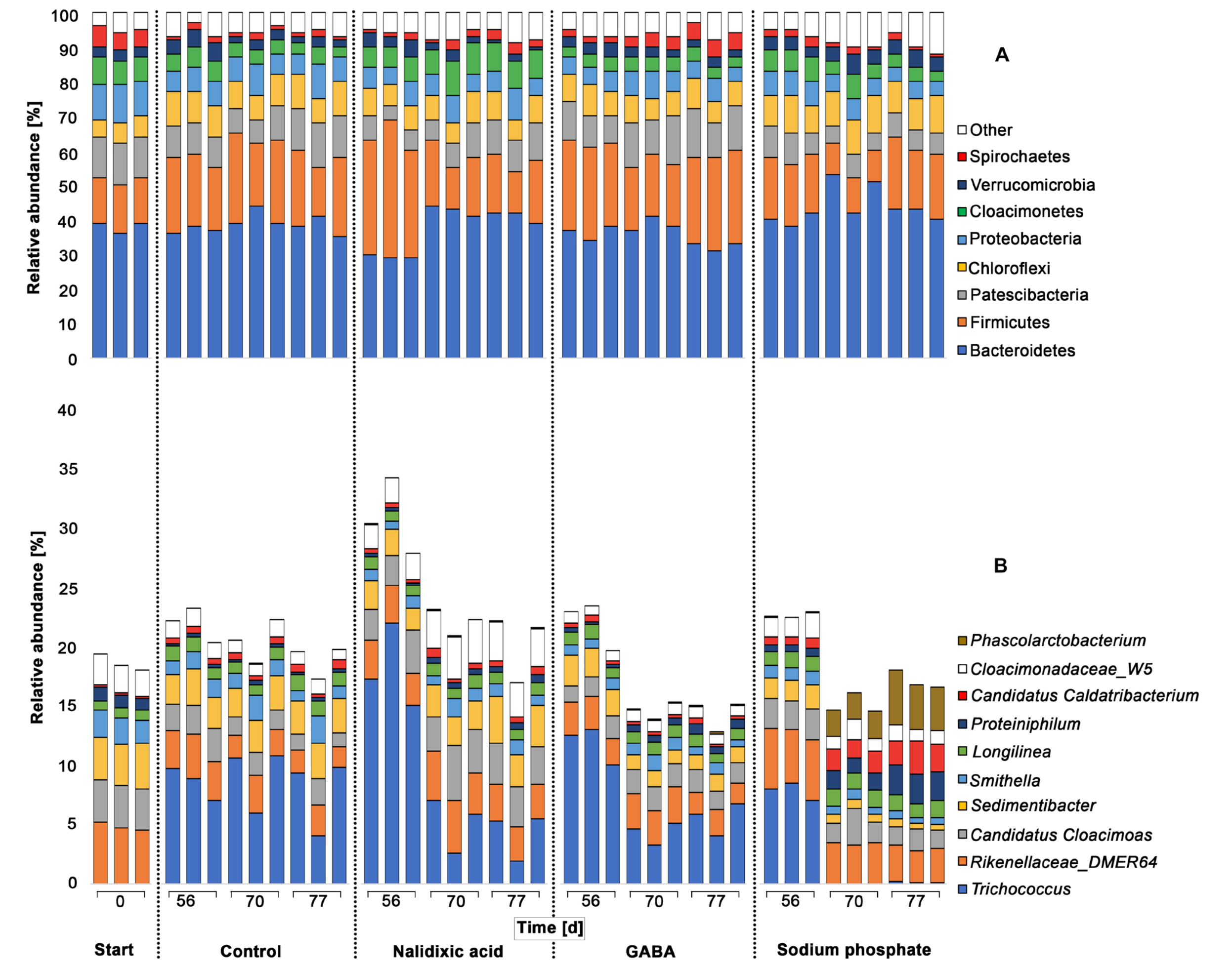


FIGURE 5 | Taxonomic profiles of chemically stressed digester microbiomes: Taxonomic profiles are shown for all experiments (perturbation with nalidixic acid, γ-aminobutyric acid [GABA], sodium phosphate and the control). Results are shown for the 8 most abundant phyla **(A)** and 10 most abundant genera **(B)**. For determining the most abundant phyla and genera, they were sorted after summing up their relative abundancies in all samples. The 16S-rRNA gene amplicons were analyzed for the seed sludge (start) and at three time points during treatment (days 56, 70, and 77). For each timepoint 3 sludge samples were taken and analyzed.

the genus Nitrospira. The observed increase for Tenericutes due to the application of all tested stressors is of particular interest, as it is in concordance with a recent work by Braz et al. (2018), where the increase in the abundance of Tenericutes was descrobed as a consequence of an OLR shock.

Other phyla that were significantly impaired due to the application of sodium phosphate were Aegiribacteria, Firmicutes, Proteobacteria, Patescibacteria, and Fibrobacteres. Moreover, there was a significant increase in the ratio of Verrucomicrobia, Synergistetes, Lentisphearae and Atribacteria. Like the reactor receiving nalidixic acid, reactors receiving GABA and sodium phosphate showed only small taxonomic shifts (**Figure 5**), which again highlights the robustness of the underlying microbiome.

To compare the differences in relative abundancies at the genus level among perturbated reactors and the control experiment (**Figure 5B**), differential abundance analyses were applied here (**Supplementary Tables S4–S6**). The most abundant genera, for which significant changes with $p < 0.05$ were observed, were *Trichococcus*, *Sedimentibacter*, *Phascolarctobacterium*, *Cadidatus Caldatribacterium,* and *Proteiniphilum.*

With the addition of nalidixic acid, *Trichococcus* showed a ratio 18.36% ± 2.93% at day 56, which was significantly higher, by 8.80%, than the control. However, no significant differences were detectable at day 70 between the control and the nalidixic acid-receiving reactor anymore, suggesting a fast adaptation. A similar observation was made by Mitchell et al. (2013), where ampicillin with concentrations between 280 and 350 mg L^{-1} inhibited the process only during the early stages. In concordance with this observation, it has recently been described that sewage sludge from wastewater treatment often contains considerable amounts of antibiotic resistance genes (Mengli et al., 2019).

With the addition of sodium phosphate, *Trichococcus* showed a significantly lower ratio than in the control at day 70 Interestingly, *Trichococcus* was not detected in the initial sample (anaerobic digested sludge from a waste water treatment plant). The overall increase of *Trichococcus* at day 56 cannot be explained by the addition of nalidixic acid, GABA or sodium phosphate, as *Trichococcus* was enriched in the control as well.

Like *Trichococcus*, *Sedimentibacter* significantly decreased with the addition of sodium phosphate. At day 56, *Sedimentibacter* showed a ratio of 1.81% ± 0.17% and decreased to a ratio of 0.63% ± 0.09% on days 70 and 77. There were several genera that were significantly enriched upon addition of sodium phosphate in comparison with the control samples, namely, *Phascolarctobacterium*, *Candidatus Caldatribacterium,* and *Proteiniphilum*. At day 56, these three genera showed ratios of 0.02% ± 0.01%, 0.77% ± 0.12 and 0.59% ± 0.09%, respectively. During the last two sampling time points the ratio of these three genera increased to 3.19% ± 1.07%, 2.07% ± 0.46% and 1.96 ± 0.59%.

It should be stressed that the high sensitivity of *Trichococcus and Sedimentibacter,* as well as the increase in relative abundance of *Candidatus Caldatribacterium, Phascolarctobacterium,* and *Proteiniphilum,* is likely linked to phosphate but not conductivity. The highest observed conductivity values for the reactor receiving sodium phosphate was 12.06 mS cm^{-1}, but process disturbance due to high conductivity values are usually observed at values higher than 35 mS cm^{-1} (Ogata et al., 2016). By contrast, an inhibitory effect due to high phosphate levels has already been reported at a concentration of 70 mM (Paulo et al., 2005). According to Paulo et al. (2005) the phosphate concentration in the present study reached a level that already inhibited the underlying biocenosis process; in total, approximately 20.5 g L^{-1} was added, corresponding to 125 mM. Other authors have described inhibiting effects due to elevated phosphorous levels too. For example, Sharma and Singh (2001) described phosphate as detrimental for anaerobic sludge granulation during the treatment of distillery effluents. Mancipe-Jiménez et al. (2017) described inhibitory effects due to a sudden increase of phosphorus in the influent during anaerobic liquid waste treatment.

From a total of 2995 OTUs, 25 changed their relative abundance on day 56 significantly. On day 70, the number increased to 80 significant changes, which was elevated again on day 77 to 119 significant changes. This number might appear small, but it has to be considered that 2960 OTUs had a relative abundance of less than 1% in the total pool of sequences. To reach a better impression of the severity of the induced stresses at the community level, all significant changes (**Supplementary Tables S4–S6**) were compared in Venn diagrams. These showed that, with increasing concentrations of stressors, the number of significant taxonomic shifts also increased (**Figure 6**). From the eight genera that were affected similarly in all three reactors, five showed a significant decrease and three showed a significant increase. The five decreasing genera were *Gracilibacter*, *Geobacter*, *Syntrophobacter*, and two uncultured bacteria. One of these two uncultured bacteria could only be classified on class level (Thermodesulfovibrionia) and the other one on family level (Gracilibacteraceae). The three increasing genera were *Fermentimonas*, *Proteiniphilum* and an uncultured bacterium belonging to the family Acidaminococcaceae.

Comparing the shown taxonomic profiles to the existing literature, it is immediately apparent that the number of works addressing acidosis events on a bacterial level is limited. Many works address acidosis events based only on chemical parameters. Authors of more recent works also address the methanogenic community (e.g., Steinberg and Regan, 2011; Lerm et al., 2012; Tale et al., 2015), but bacterial communities remain underrepresented in most of the works. Among the few works addressing the bacterial community and in relation to the results presented here, an article from Goux et al. (2015) is of particular interest; as in the present study, Goux et al. (2015) observed only small variations at the phylum level. Moreover, the phyla Bacteroidetes, Firmicutes, Chloroflexi, Proteobacteria, Cloacimonetes, Verrucomicrobia and Spirochaetes were also abundant, and on lower taxonomic levels, Goux et al. (2015) described a more intense shift behavior as well. Another work addressing the bacterial community during organic overloading is that of Braz et al. (2019). One of their findings was an increased abundance of fatty acid fermenters and a disturbance of syntrophic bacteria. These two findings are in concordance with the finding presented here of decreased ratios for the genera *Geobacter* and *Syntrophobacter*, which are known syntrophic bacteria (Meher and Ranade, 1993; Liu et al., 2018). The aforementioned increase in *Fermentimonas* and *Proteiniphilum* is also in concordance with the described increase of fatty acid fermenters in the work from Braz et al. (2019). Both *Fermentimonas* and *Proteiniphilum* are known to produce VFAs from a wide range of substrates (Hahnke et al., 2016).

In respect to the observed taxonomic profiles and the detected changes it has to be highlighted that the repeated input of 150 ml of digested sewage sludge during each feeding event might have influenced the results. Invasion of microbial communities is a problem, which has recently been highlighted by Kinnunen et al. (2016). However, the used setting reproduces the normal conditions in the industry and, additionally there are multiple reasons for which it is likely that this had a minor impact on the presented results: The sludge that was used as fed was the same, which was used originally as inoculum. Therefore, the fed did not introduce new kinds of organisms into the system. Moreover, a comparative analysis was performed, in which all the reactors shared the same feeding conditions and, thus, the same "input" microbiota. Therefore, the comparisons are not influenced by this factor. This is supported by a PCA (**Figure 4**), which shows that the microbiomes diverged and that they were in the end very different from the control.

Generalized Lotka–Volterra Modeling

To investigate the effect of the different perturbations on the interactions between microorganisms, a gLV was applied. The possibility for fast and robust assessment of microbial interactions directly from microbial time series was recently emphasized by Faust et al. (2018). This model can be used not only to predict the predator-prey interactions in the shape of Lotka–Volterra equations but also to detect a wider range of

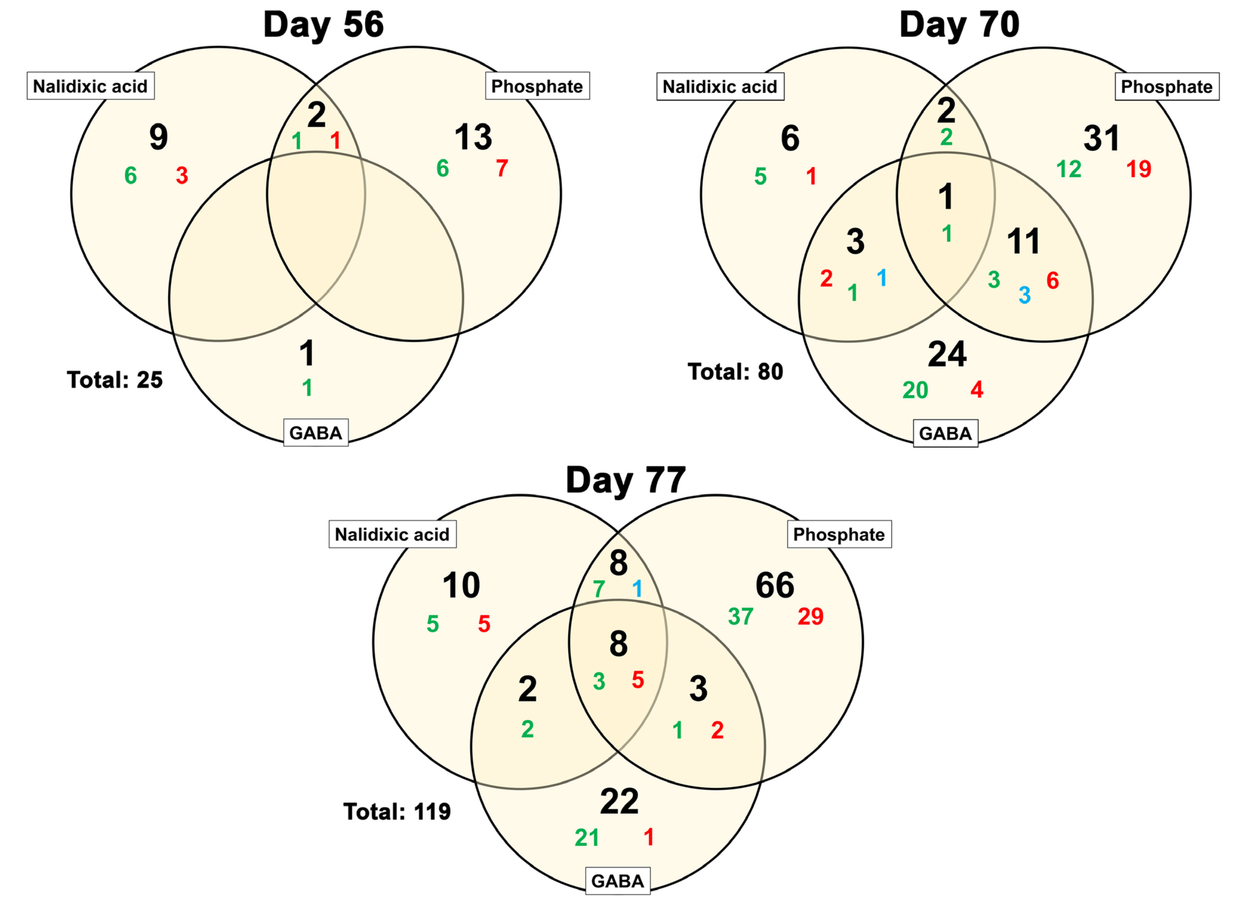


FIGURE 6 | Venn diagrams for genera exhibiting a dynamic behavior in comparison with the control in all reactors: Significant changes in frequency are shown for all the days where DNA samples were analyzed (day 56, 70, and 77). Genera showing an increase in relative abundance are highlighted in green. Genera exhibiting a decrease are highlighted in red. In some cases, the relative abundance of a genus was significantly increased by one stressor but significantly decreased with another stressor. Such cases are highlighted in blue. The sum of all changes (blue, red, and green) is given in black. The total number of OTUs, which were significantly affected in all the reactors is shown to the left of each diagram. Genera from each reactor were compared with the control and only results with a significance of $p < 0.05$ were considered.

relationships, including competition, cooperation and neutralism (Kuntal et al., 2019). Based on DNA sequencing, gLV has already been applied various times to investigate microbial interactions in the gut (Weng et al., 2017), in cheese (Mounier et al., 2008), in the coffee-machine bacteriome (Vilanova et al., 2015) or in bacteria grown on pine-tree resin-based medium (Dorado-Morales et al., 2015).

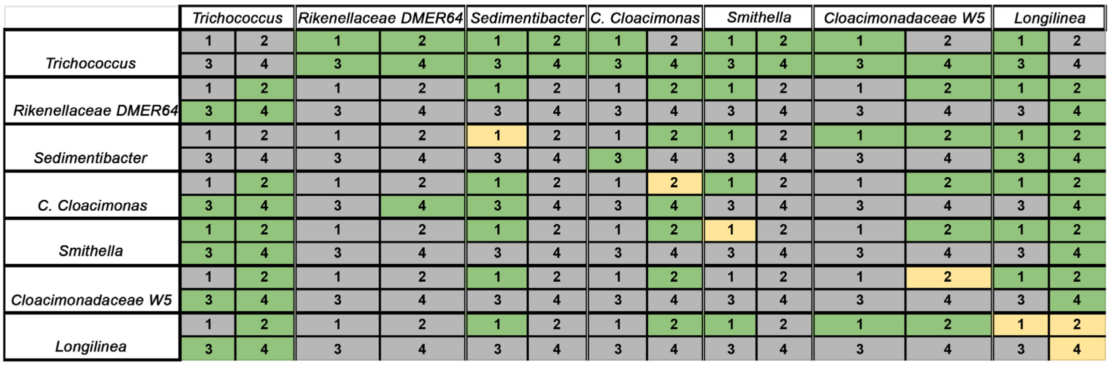
Recently, a graphical user interface (GUI) based interactive platform was published by Kuntal et al. (2019); this is available online[3], and it automates the estimation of the respective gLV parameters, based on the following equation:

$$\frac{dx_i}{dt} = x_i(r_i + \sum_{j=1}^{n} \propto_{ij} x_j). \qquad (1)$$

Here, $\frac{dx_i}{dt}$ corresponds to the rate of growth of species x_i, r_i represents the intrinsic growth rate and $\propto_{ij}$ is the 'interaction coefficient'. gLV predictions are based on the algebraic sign of the interaction coefficient. If this coefficient is positive, a beneficial effect is assumed, while prejudicial effects are derived from negative values of the parameter. Finally, if the interaction coefficient is equal to zero, no interaction is assumed between the two taxa.

In the present study, the most abundant bacteria were selected for each condition according to their average relative abundance. Only those OTUs present among the top-10 abundant bacteria in all groups were kept for further Lotka–Volterra modeling (7 OTUs). Applying the gLV on the here presented set of taxonomic data (**Figure 7**), more positive interactions among the studied taxa were observed in the control experiment (24) than in the rest of the conditions (23 with nalidixic acid, 15 with GABA and 18 with sodium phosphate). In contrast, there were more

[3] http://web.rniapps.net/webglv



	Trichococcus		*Rikenellaceae DMER64*		*Sedimentibacter*		*C. Cloacimonas*		*Smithella*		*Cloacimonadaceae W5*		*Longilinea*	
Trichococcus	▨ 1	▨ 2	🟩 1	🟩 2	🟩 1	🟩 2	🟩 1	▨ 2	🟩 1	🟩 2	🟩 1	▨ 2	🟩 1	▨ 2
	▨ 3	▨ 4	🟩 3	🟩 4	🟩 3	🟩 4	🟩 3	🟩 4	🟩 3	🟩 4	🟩 3	🟩 4	🟩 3	▨ 4
Rikenellaceae DMER64	▨ 1	🟩 2	▨ 1	▨ 2	🟩 1	▨ 2	▨ 1	🟩 2	🟩 1	▨ 2	▨ 1	🟩 2	🟩 1	🟩 2
	🟩 3	🟩 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	🟩 4
Sedimentibacter	▨ 1	▨ 2	▨ 1	▨ 2	🟨 1	▨ 2	▨ 1	🟩 2	🟩 1	▨ 2	🟩 1	🟩 2	🟩 1	🟩 2
	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	🟩 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	🟩 3	🟩 4
C. Cloacimonas	▨ 1	🟩 2	▨ 1	▨ 2	🟩 1	▨ 2	▨ 1	🟨 2	🟩 1	▨ 2	▨ 1	🟩 2	🟩 1	🟩 2
	🟩 3	🟩 4	▨ 3	🟩 4	🟩 3	▨ 4	▨ 3	🟩 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	🟩 4
Smithella	🟩 1	🟩 2	▨ 1	▨ 2	🟩 1	▨ 2	▨ 1	🟩 2	🟨 1	▨ 2	▨ 1	🟩 2	🟩 1	🟩 2
	🟩 3	🟩 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	🟩 4
Cloacimonadaceae W5	▨ 1	🟩 2	▨ 1	▨ 2	🟩 1	▨ 2	▨ 1	🟩 2	▨ 1	▨ 2	▨ 1	🟨 2	🟩 1	🟩 2
	🟩 3	🟩 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	🟩 4
Longilinea	▨ 1	🟩 2	▨ 1	▨ 2	🟩 1	▨ 2	▨ 1	🟩 2	🟩 1	▨ 2	🟩 1	🟩 2	🟨 1	🟨 2
	🟩 3	🟩 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	▨ 4	▨ 3	🟨 4

FIGURE 7 | Ecological interactions among the most abundant bacteria in all samples, as deduced from generalized Lotka–Volterra model. Gray: negative interaction; Green: positive interaction; Yellow: no interaction. The numbers 1 – 4 indicate the reactors with the respective stressors: 1: control; 2: perturbation with nalidixic acid; 3: γ-aminobutyric acid (GABA; FG); 4: sodium phosphate (FP).

negative interactions detected in the reactors with nalidixic acid (23), GABA (34) and sodium phosphate (31) than in the control (22). These results suggest that the perturbations introduced in the system tend to create a more competitive environment, in which microorganisms are more likely to interact negatively with each other.

Apart from the total number of microbial interactions (positive, negative, or neutral), it is important to determine which types of pairwise interaction are observed among the taxa in the different set conditions (**Figure 7**). Interactions involving *Trichococcus* spp. or *DMER64*, in general, were stable in the four conditions. In other words, *Trichococcus* spp. and *DMER64* tend to behave the same way (positively, negatively, or neutrally) with the rest of the studied taxa in all the conditions. However, the pairwise relationships involving other taxa were less homogeneous (i.e., *Cloacimonadaceae W5* negatively interacted with *Trichococcus* in the control experiment, but positive interactions between these two taxa were detected in the rest of the conditions).

Of all the alternative perturbations, the treatment with nalidixic acid proved to be the one with the deepest effects in the interaction patterns compared with the control, whereas the treatments with GABA and sodium phosphate tended to reproduce the same microbial interactions observed in the control (shared interactions of the control with: GABA = 26; sodium phosphate = 27; nalidixic acid = 20). Indeed, the treatment with nalidixic acid displayed a higher number of interactions that were not found in the rest of the conditions (unique interactions in the treatments with nalidixic acid = 11; GABA = 5; sodium phosphate = 0).

Together, our results suggest that antibiotic treatment affects the community interactions present in the anaerobic digesters in a deeper way. Interestingly, all the applied digester conditions resulted in changes in the interaction patterns of the studied microbial taxa. This is of interest in terms of a work from Scherlach and Hertweck (2018), which highlights that microbe–microbe interactions can shape the specific "microenvironment" due to the secretion of chemical mediators. In the context mentioned above, therefore, it would be a promising approach to combine the Lotka–Volterra model (based on 16S-rRNA gene amplicon sequencing) with transcriptomics and metabolomics in future works.

Although the Lotka–Volterra model does not guarantee causality, the high number of genera for which the described correlational behavior was observed suggests that this reflects a biological relationship.

It should be highlighted that the present work was not focused on methanogenic archaea, but rather, it concentrated on bacteria. In the past, the stress responses of methanogenic archaea were extensively investigated using stressors, such as ammonium (Dai et al., 2016), light (Olson et al., 1991), pH and VFAs (Staley et al., 2011). The common view of such works is that, when comparing them to involved bacteria, methanogenic archaea show high sensitivity. Although methanogenic archaea are the most important microorganisms in methane production, since they are performing the final step of anaerobic digestion (methanogenesis), bacteria are key players. Bacteria are responsible for the hydrolysis of complex polymers and the conversion of resulting monomers into hydrogen, acetate, and carbon dioxide, which are the main substrates for methanogenic archaea (Robles et al., 2018). This degradation process involves three phases (hydrolysis, acidogenesis, and acetogenesis). Especially during acidogenesis, various metabolic intermediates are formed; these are of high value for the bio-based industry (Wainaina et al., 2019). The possibility of producing such metabolites during anaerobic digestion also raises the question of how the robustness of the involved bacteria might be overcome in order to manipulate the spectrum of yielded metabolites. In this vein, a recent review article from Strous and Sharp (2018), which explained the importance of 'designer microbiomes for environmental, energy and health biotechnology,' can be highlighted.

Results from applying the Lotka–Volterra model for the first time on anaerobic digestion show that microbiomes of anaerobic digesters are not only robust and redundant, but also surprisingly flexible in terms of microbial interactivity. This flexibility indicates that the manipulation of anaerobic microbiomes at the level of microbial interactivity is an ambitious goal that may be achieved more easily with constant digester conditions to prevent the alteration of microbial interaction patterns.

CONCLUSION

Emanating from the same microbiome and using different stressors (nalidixic acid, GABA and sodium phosphate), multiple taxonomic shifts were caused for subsequent analysis of populational dynamics. Although the aim of the present work was not to characterize the respective stressors in detail, it can be concluded that sodium phosphate has a particularly strong effect on the bacterial biocenosis, and in contrast, taxonomic profiles were surprisingly stable after addition of nalidixic acid and GABA (in spite of a clear acidosis for the latter case). Taxonomic profiles on phylum level were surprisingly robust. At the genus level, important taxonomic variations were observed especially for the genera *Trichococcus, Candidatus Caldatribacterium, Phascolarctobacterium*, *Proteiniphilum, Gracilibacter*, *Geobacter*, *Syntrophobacter,* and *Fermentimonas*. Therefore, these genera may be promising targets for the surveillance of anaerobic digester microbiomes.

Main objective in the present study was to trigger —and thus shed light— on microbial interactions, based on the gLV model. Except for sodium phosphate, the addition of the respective stressors did not alter taxonomic profiles drastically, indicating a high robustness for the bacterial biocenosis in digested sewage sludge. Interestingly, potential ecological interactions among the key players were strongly affected by all treatments, and in some cases, two pairs of genera showed negative, positive or no correlation, depending on the treatment. Although the presented work suggests a massive resilience and stability of the underlying bacterial biocenosis in respect to the relative abundance of involved bacteria, a highly flexible behavior was observed in terms of microbial interactivity."

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the https://www.ncbi.nlm.nih.gov/bioproject/PRJNA554976.

AUTHOR CONTRIBUTIONS

BS and CA performed anaerobic digestions experiments. AL-P, CV, and CA performed the taxonomic analyses. BS, CA, AL-P, MP, CV, and CD were writing the manuscript.

FUNDING

This research was funded by the German Ministry of Economic Affairs and Energy (grant numbers 16KN070128 and 16KN070126). Funding by the European Union through the BioRoboost project, H2020-NMBP-TR-IND-2018-2020/BIOTEC-01-2018 (CSA), Project ID 210491758 is acknowledged.

ACKNOWLEDGMENTS

We are grateful for funding of the work by the German Ministry of Economic Affairs and Energy (grant numbers 16KN070128 and 16KN070126). Moreover, we thank the Spanish Ministry of Science, Innovation and Universities for funding the Ph.D. of Adriel Latorre-Pérrez (Doctorado Industrial Fellowship, reference DI-17-09613). Finally, we are grateful for open access funding by the publication fund of the TU Dresden and for funding by the European Union through the BioRoboost project, H2020-NMBP-TR-IND-2018-2020/BIOTEC-01-2018 (CSA), Project ID 210491758.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.00867/full#supplementary-material

FIGURE S1 | Analysis of total volatile fatty acids (TVFAs): The concentrations of formic acid, acetic acid, lactic acid, propionic acid, iso-butyric acid, butyric acid and valeric acid are shown for the control **(A)**, and for the reactions including nalidixic acid **(B)**, GABA **(C)**, and sodium phosphate **(D)**.

TABLE S1 | Differential abundance analysis at the phylum level to compare the control and the reactor receiving nalidixic acid: The log2FoldChange of the normalized abundance was calculated using the DESeq2-package (Love et al., 2014). *p*-values of the respective changes were adjusted using the Benjamini-Hochberg method.

TABLE S2 | Differential abundance analysis at the phylum level to compare the control and the reactor receiving γ-aminobutyric acid (GABA): The log2FoldChange of the normalized abundance was calculated using the DESeq2-package (Love et al., 2014). *p*-values of the respective changes were adjusted using the Benjamini–Hochberg method.

TABLE S3 | Differential abundance analysis at the phylum level to compare the control and the reactor receiving sodium phosphate: The log2FoldChange of the normalized abundance was calculated using the DESeq2-package (Love et al., 2014). *p*-values of the respective changes were adjusted using the Benjamini–Hochberg method.

TABLE S4 | Differential abundance analysis at the genus level to compare the control and the reactor receiving nalidixic acid. The log2FoldChange of the normalized abundance was calculated using the DESeq2-package (Love et al., 2014). The *p*-values of the respective changes were adjusted using the Benjamini–Hochberg method.

TABLE S5 | Differential abundance analysis at the genus level to compare the control and the reactor receiving γ-aminobutyric acid (GABA). The log2FoldChange of the normalized abundance was calculated using the DESeq2-package (Love et al., 2014). The *p*-values of the respective changes were adjusted using the Benjamini–Hochberg method.

TABLE S6 | Differential abundance analysis at the genus level to compare the control and the reactor receiving sodium phosphate. The log2FoldChange of the normalized abundance was calculated using the DESeq2-package (Love et al., 2014). The *p*-values of the respective changes were adjusted using the Benjamini–Hochberg method.

MATERIAL S1 | Calculations.

REFERENCES

Abendroth, C., Hahnke, S., Simeonov, C., Klocke, M., Casani-Miravalls, S., Ramm, P., et al. (2018). Microbial communities involved in biogas production exhibit high resilience to heat shocks. *Bioresour. Technol.* 249, 1074–1079. doi: 10.1016/j.biortech.2017.10.093

Abendroth, C., Simeonov, C., Peretó, J., Antúnez, O., Gavidia, R., Luschnig, O., et al. (2017). From grass to gas: microbiome dynamics of grass biomass acidification under mesophilic and thermophilic temperatures. *Biotechno Biofuels.* 10:171. doi: 10.1186/s13068-017-0859-0

Abendroth, C., Vilanova, C., Günther, T., Luschnig, O., and Porcar, M. (2015). Eubacteria and Archaea communities in seven mesophile anaerobic digester plants. *Biotechnol. Biofuels* 8:87. doi: 10.1186/s13068-015-0271-6

Amador, P. P., Fernandes, R. M., Prudêncio, M. C., Barreto, M. P., and Duarte, I. M. (2015). Antibiotic resistance in wastewater: occurrence and fate of *Enterobacteriaceae* producers of class A and class C β-lactamases. *J. Environ. Sci. Health A Tox Hazard. Subst. .Environ Eng.* 50, 26–39. doi: 10.1080/10934529.2015.964602

Association of German Engineers (2016). *Fermentation of Organic Materials – Characterization of the Substrate, Sampling, Collection of Material Data, Fermentation Tests*. Düsseldorf: Verlag des Vereins Deutscher Ingenieure.

Baresi, L., Mah, R. A., Ward, D. M., and Kaplan, I. R. (1978). Methanogenesis from acetate: enrichment studies. *Appl. Environ. Microbiol.* 36, 186–197.

Bay, Y., Xu, R., Wang, Q. P., Zhang, Y. R., and Yang, Z. H. (2019). Sludge anaerobic digestion with high concentrations of tetracyclines and sulfonamides: dynamics of microbial communities and change of antibiotic resistance genes. *Bioresour. Technol.* 276, 51–59. doi: 10.1016/j.biortech.2018.12.066

Braz, G. H. R., Fernandez-Gonzales, N., Lema, J. M., and Carballa, M. (2018). The time response of anaerobic digestion microbiome during an organic loading rate shock. *Appl. Microbiol. Biotechnol.* 102, 10285–10297. doi: 10.1007/s00253-018-9383-9

Braz, G. H. R., Fernandez-Gonzalez, N., Lema, J. M., and Carballa, M. (2019). Organic overloading affects the microbial interactions during anaerobic digestion in sewage sludge reactors. *Chemosphere* 222, 323–332. doi: 10.1016/j.chemosphere.2019.01.124

Calusinska, M., Goux, X., Fosséпré, M., Muller, E. E. L., Wilmes, P., and Delfosse, P. (2018). A year of monitoring 20 mesophilic full-scale bioreactors reveals the existence of stable but different core microbiomes in bio-waste and wastewater anaerobic digestion systems. *Biotechnol. Biofuels* 11:196. doi: 10.1186/s13068-018-1195-8

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Dai, X., Yan, H., Li, N., He, J., Ding, Y., Dai, L., et al. (2016). Metabolic adaptation of microbial communities to ammonium stress in a high solid anaerobic digester with dewatered sludge. *Sci. Rep.* 6:28193. doi: 10.1038/srep28193

Daims, H. (2014). "The family nitrospiraceae," in *The Prokaryotes*, eds E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson (Berlin: Springer), doi: 10.1007/978-3-642-38954-2_126

De Vrieze, J., Christiaens, M. E. R., Walraedt, D., Devooght, A., Ijazb, U. Z., and Boon, N. (2017). Microbial community redundancy in anaerobic digestion drives process recovery after salinity exposure. *Water Res.* 111, 109–117. doi: 10.1016/j.watres.2016.12.042

Delbès, C., Moletta, R., and Godon, J. J. (2000). Monitoring of activity dynamics of an anaerobic digester bacterial community using 16S rRNA polymerase chain reaction–single-strand conformation polymorphism analysis. *Environ. Microbiol.* 2, 506–515. doi: 10.1046/j.1462-2920.2000.00132.x

Dorado-Morales, P., Vilanova, C., Garay, C. P., Martí, J. M., and Porcar, M. (2015). Unveiling bacterial interactions through multidimensional scaling and dynamics modeling. *Sci. Rep.* 5:18396. doi: 10.1038/srep18396

Fáberová, M., Ivanová, L., Szabová, P., Štolcová, M., and Bodík, I. (2019). The influence of selected pharmaceuticals on biogas production from laboratory and real anaerobic sludge. *Environ. Sci. Pollut. Res. Int.* 26, 31846–31855. doi: 10.1007/s11356-019-06314-4

Faust, K., Bauchinger, F., Laroche, B., de Buyl, S., Lahti, L., Washburne, A. D., et al. (2018). Signatures of ecological processes in microbial community time series. *Microbiome* 6:120. doi: 10.1186/s40168-018-0496-2

Faust, K., and Raes, J. (2012). Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550. doi: 10.1038/nrmicro2832

Goux, X., Calusinska, M., Lemaigre, S., Marynowska, M., Klocke, M., Udelhoven, T., et al. (2015). Microbial community dynamics in replicate anaerobic digesters exposed sequentially to increasing organic loading rate, acidosis, and process recovery. *Biotechnol. Biofuels* 8:122. doi: 10.1186/s13068-015-0309-9

Hahnke, S., Langer, T., Koeck, D. E., and Klocke, M. (2016). Description of *Proteiniphilum saccharofermentans* sp. nov., *Petrimonas mucosa* sp. nov. and *Fermentimonas caenicola* gen. nov., sp. nov., isolated from mesophilic laboratory-scale biogas reactors, and emended description of the genus *Proteiniphilum*. *Int. J. Syst. Evol. Microbiol.* 66, 1466–1475. doi: 10.1099/ijsem.0.000902

Hardegen, J., Latorre-Pérez, A., Vilanova, C., Günther, T., Porcar, M., Luschnig, O., et al. (2018). Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass. *Bioresour. Technol.* 265, 275–281. doi: 10.1016/j.biortech.2018.06.005

Hassa, J., Maus, I., Off, S., Pühler, A., Scherer, P., Klocke, M., et al. (2018). Metagenome, metatranscriptome, and metaproteome approaches unraveled compositions and functional relationships of microbial communities residing in biogas plants. *Appl. Microbiol. Biotechnol.* 102, 5045–5063. doi: 10.1007/s00253-018-8976-7

Henson, J. M., Bordeaux, F. M., Rivard, C. J., and Smith, P. H. (1985). Quantitative influences of butyrate or propionate on thermophilic production of methane from biomass. *Appl. Environ. Microbiol.* 51, 288–292.

Hori, T., Haruta, S., Ueno, Y., Ishii, M., and Igarashi, Y. (2006). Dynamic transition of a methanogenic population in response to the concentration of volatile fatty acids in a thermophilic anaerobic digester. *Appl. Environ. Microbiol.* 72, 1623–1630. doi: 10.1128/AEM.72.2.1623-1630.2006

Karkman, A., Do, T. T., Walsh, F., and Virta, M. P. J. (2018). Antibiotic-resistance genes in waste water. *Trends Microbiol.* 26, 220–228.

Kinnunen, M., Dechesne, A., Proctor, C., Hammes, F., Johnson, D., Quintela-Baluja, M., et al. (2016). A conceptual framework for invasion in microbial communities. *ISME J.* 10, 2773–2775. doi: 10.1038/ismej.2016.75

Kirkegaard, R. H., McIlroy, S. J., Kristensen, J. M., Nierychlo, M., Karst, S. M., Dueholm, M. S., et al. (2017). The impact of immigration on microbial community composition in full-scale anaerobic digesters. *Sci. Rep.* 7:9343. doi: 10.1038/s41598-017-09303-0

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2012). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next- generation sequencing-based diversity studies. *Nucleic Acids Res.* 41, e1. doi: 10.1093/nar/gks808

Klocke, M., Mähnert, P., Mundt, K., Souidi, K., and Linke, B. (2007). Microbial community analysis of a biogas-producing completely stirred tank reactor fed continuously with fodder beet silage as mono-substrate. *Syst. Appl Microbiol.* 30, 139–151. doi: 10.1016/j.syapm.2006.03.007

Kuntal, B., Gadgil, C., and Mandel, S. S. (2019). Web-gLV: a web based platform for lotka-volterra based modeling and simulation of microbial populations. *Front. Microbiol.* 10:288. doi: 10.3389/fmicb.2019.00288

Kuroda, K., Nobu, M. K., Mei, R., Narihiro, T., Bocher, B. T. W., Yamaguchi, T., et al. (2016). A single-granule-level approach reveals ecological heterogeneity in an upflow anaerobic sludge blanket reactor. *PLoS ONE* 11:e0167788. doi: 10.1371/journal.pone.0167788

Lerm, S., Kleyböcker, A., Miethling-Graff, R., Alawi, M., Kasina, M., Liebrich, M., et al. (2012). Archaeal community composition affects the function of anaerobic co-digesters in response to organic overload. *Waste Manage.* 32, 389–399. doi: 10.1016/j.wasman.2011.11.013

Li, B., and Zhang, T. (2010). Biodegradation and adsorption of antibiotics in the activated sludge process. *Environ. Sci. Technol.* 2010:9. doi: 10.1021/es903490h

Liu, X., Zhuo, S., Rensing, C., and Zhou, S. (2018). Syntrophic growth with direct interspecies electron transfer between pili-free *Geobacter* species. *ISME* 12, 2142–2151. doi: 10.1038/s41396-018-0193-y

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome. Biol.* 15:550.

Mancipe-Jiménez, D. C., Costa, C., and Márquez, M. C. (2017). Methanogenesis inhibition by phosphorus in anaerobic liquid waste treatment. *De Gruyter Liq Waste Recycl.* 2, 1–8.

Meher, K. K., and Ranade, D. R. (1993). Isolation of propionate degrading bacterium in co-culture with a methanogen from a cattle dung biogas plant. *J. Biosci.* 18, 271–277.

Mengli, W., Ruying, L., and Zhao, Q. (2019). Distribution and removal of antibiotic resistance genes during anaerobic sludge digestion with alkaline, thermal hydrolysis and ultrasonic pretreatments. *Front. Environ. Sci. Eng.* 13:43. doi: 10.1007/s11783-019-1127-2

Mitchell, S. M., Ullman, J. L., Teel, A. L., Watts, R. J., and Frear, C. (2013). The effects of the antibiotics ampicillin, florfenicol, sulfamethazine, and tylosin on biogas production and their degradation efficiency during anaerobic digestion. *Bioresour. Technol.* 149, 244–252. doi: 10.1016/j.biortech.2013.09.048

Mounier, J., Monnet, C., Vallaeys, T., Arditi, R., Sarthou, A. S., Hélias, A., et al. (2008). Microbial interactions within a cheese microbial community. *Appl. Environ. Microbiol.* 74, 172–181. doi: 10.3389/fmicb.2019.01901

Musa, M. A., Idrus, S., Hasfalina, C. M., and Daud, N. N. N. (2018). Effect of organic loading rate on anaerobic digestion performance of mesophilic (UASB) reactor using cattle slaughterhouse wastewater as substrate. *Int. J. Environ. Res. Public Health* 15:2220. doi: 10.3390/ijerph15102220

Mustapha, N. A., Sakai, K., Shirai, Y., and Maeda, T. (2016). Impact of different antibiotics on methane production using waste-activated sludge: mechanisms and microbial community dynamics. *Appl. Microbiol. Biotechnol.* 100, 9355–9364. doi: 10.1007/s00253-016-7767-2

Nachaiyasit, S., and Stuckey, D. C. (1997). The effect of shock loads on the performance of an anaerobic baffled reactor (ABR). 1. Step changes in feed concentration at constant retention time. *Water Res.* 31, 2737–2746.

Naquin, A., Shrestha, A., Sherpa, M., Nathaniel, R., and Boopathy, R. (2015). Presence of antibiotic resistance genes in a sewage treatment plant in Thibodaux, Louisiana, USA. *Bioresour. Technol.* 188, 79–83. doi: 10.1016/j.biortech.2015.01.052

Ogata, Y., Ishigaki, T., Nakagawa, M., and Yamada, Y. (2016). Effect of increasing salinity on biogas production in waste landfills with leachate recirculation: a lab-scale model study. *Biotechnol. Rep.* 10, 111–116. doi: 10.1016/j.btre.2016.04.004

Olson, K. D., McMahon, C. W., and Wolfe, R. S. (1991). Light sensitivity of methanogenic archaebacteria. *Appl. Environ. Microbiol.* 57, 2683–2686.

Paulo, P. L., dos Santos, A. B., Ide, C. N., and Lettinga, G. (2005). Phosphate inhibition on thermophilic acetoclastic methanogens: a warning. *Water Sci. Technol.* 52, 331–336.

Robles, G., Nair, R. B., Kleinsteuber, S., Nikolausz, M., and Horváth, I. S. (2018). "Biogas production: microbiological aspects," in *Biogas: Fundamentals, Process and Operation; Biogas, Biofuel and Biorefinery Technologies 6*, eds M. Tabatabaei and H. Ghanavati (Berlin: Springer), doi: 10.1007/978-3-319-77335-3_7

Scherlach, K., and Hertweck, C. (2018). Mediators of mutualistic microbe-microbe interactions. *Nat. Prod. Rep.* 35, 303–308. doi: 10.1039/c7np00035a

Sharma, J., and Singh, R. (2001). Effect of nutrients supplementation on anaerobic sludge development and activity for treating distillery effluent. *Bioresour. Technol.* 79, 203–206. doi: 10.1016/s0960-8524(00)00131-0

Shi, X., Zhao, J., Chen, L., Zuo, J., Yang, Y., Zhang, Q., et al. (2019). Genomic dynamics of full-scale temperature-phased anaerobic digestion treating waste activated sludge: focusing on temperature differentiation. *Waste Manage.* 87, 621–628. doi: 10.1016/j.wasman.2019.02.041

Spirito, C. M., Daly, S. E., Werner, J. J., and Angenent, L. T. (2018). Redundancy in anaerobic digestion microbiomes during disturbances by the antibiotic monensin. *Appl. Environ. Microbiol.* 84, e2692–e2617. doi: 10.1128/AEM.02692-17

Staley, B. F., de los Reyes, F. L. III, and Barlaz, M. A. (2011). Effect of spatial differences in microbial activity, pH, and substrate levels on methanogenesis initiation in refuse. *Appl. Environ. Microbiol.* 77, 2381–2391. doi: 10.1128/AEM.02349-10

Steinberg, L. M., and Regan, J. M. (2011). Response of lab-scale methanogenic reactors inoculated from different sources to organic loading rate shocks. *Bioresour. Technol.* 102, 8790–8798. doi: 10.1016/j.biortech.2011.07.017

Strous, M., and Sharp, C. (2018). Designer microbiomes for environmental, energy and health biotechnology. *Curr. Opin. Microbiol.* 43, 117–123. doi: 10.1016/j.mib.2017.12.007

Sundberg, C., Al-Soud, W. A., Larsson, M., Alm, E., Yekta, S. S., Svensson, B. H., et al. (2013). 454 pyrosequencing analyses of bacterial and archaeal richness in 21 full-scale biogas digesters. *FEMS Microbiol. Ecol.* 85, 612–626. doi: 10.1111/1574-6941.12148

Tale, V. P., Maki, J. S., and Zitomer, D. H. (2015). Bioaugmentation of overloaded anaerobic digesters restores function and archaeal community. *Water Res.* 70, 138–147. doi: 10.1016/j.watres.2014.11.037

Theuerl, S., Kohrs, F., Benndorf, D., Maus, I., Wibberg, D., Schlüter, A., et al. (2015). Community shifts in a well-operating agricultural biogas plant: how process variations are handled by the microbiome. *Appl. Microbiol. Biotechnol.* 99, 7791–7802. doi: 10.1007/s00253-015-6627-9

van den Heuvel, J. C., Beeftink, H. H., and Verschuren, P. G. (1988). Inhibition of the acidogenic dissimilation of glucose in anaerobic continuous cultures by free butyric acid. *Environ. Microbiol.* 29, 89–94.

Vilanova, C., Iglesias, A., and Porcar, M. (2015). The coffee-machine bacteriome: biodiversity and colonisation of the wasted coffee tray leach. *Sci Rep.* 5:17163. doi: 10.1038/srep17163

Wainaina, S., Lukitawesa, Kumar Awasthi, M., and Taherzadeh, M. J. (2019). Bioengineering of anaerobic digestion for volatile fatty acids, hydrogen or methane production: a critical review. *Bioengineered* 10, 437–458. doi: 10.1080/21655979.2019.1673937

Weng, F. C., Shaw, G. T., Weng, C. Y., Yang, Y. J., and Wang, D. (2017). Inferring microbial interactions in the gut of the hong kong whipping frog (*Polypedates megacephalus*) and a validation using probiotics. *Front. Microbiol.* 30:25. doi: 10.3389/fmicb.2017.00525

Xu, R., Yang, Z. H., Zheng, Y., Wang, Q. P., Bai, Y., Liu, J. B., et al. (2019). Metagenomic analysis reveals the effects of long-term antibiotic pressure on sludge anaerobic digestion and antimicrobial resistance risk. *Bioresour. Technol.* 282, 179–188. doi: 10.1016/j.biortech.2019.02.120

Xu, R., Zhang, K., Liu, P., Khan, A., Xiong, J., Tian, F., et al. (2018). A critical review on the interaction of substrate nutrient balance and microbial community structure and function in anaerobic co-digestion. *Bioresour. Technol.* 247, 1119–1127. doi: 10.1016/j.biortech.2017.09.095

Yin, X., Deng, Y., Ma, L., Wang, Y., Chan, L. Y. L., and Zhang, T. (2019). Exploration of the antibiotic resistome in a wastewater treatment plant by a nine-year longitudinal metagenomic study. *Environ. Int.* 133:105270. doi: 10.1016/j.envint.2019.105270

Zhang, B., Xiangyang, X., and Liang, Z. (2017). Structure and function of the microbial consortia of activated sludge in typical municipal wastewater treatment plants in winter. *Sci. Rep.* 7:17930. doi: 10.1038/s41598-017-17743-x

Zhang, M., Ma, Y., Ji, D., Li, X., Zhang, J., and Zang, L. (2019). Synergetic promotion of direct interspecies electron transfer for syntrophic metabolism of propionate and butyrate with graphite felt in anaerobic digestion. *Bioresour. Technol.* 287:121373. doi: 10.1016/j.biortech.2019.121373

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 13 December 2021
doi: 10.3389/fmicb.2021.768240

A Round Trip to the Desert: *In situ* Nanopore Sequencing Informs Targeted Bioprospecting

Adriel Latorre-Pérez[1], Helena Gimeno-Valero[1], Kristie Tanner[1], Javier Pascual[1], Cristina Vilanova[1]* and Manuel Porcar[1,2]

[1] Darwin Bioprospecting Excellence S.L., Paterna, Spain, [2] Institute for Integrative Systems Biology I2SysBio (University of València-CSIC), Paterna, Spain

Bioprospecting expeditions are often performed in remote locations, in order to access previously unexplored samples. Nevertheless, the actual potential of those samples is only assessed once scientists are back in the laboratory, where a time-consuming screening must take place. This work evaluates the suitability of using Nanopore sequencing during a journey to the Tabernas Desert (Spain) for forecasting the potential of specific samples in terms of bacterial diversity and prevalence of radiation- and desiccation-resistant taxa, which were the target of the bioprospecting activities. Samples collected during the first day were analyzed through 16S rRNA gene sequencing using a mobile laboratory. Results enabled the identification of locations showing the greatest and the least potential, and a second, informed sampling was performed focusing on those sites. After finishing the expedition, a culture collection of 166 strains belonging to 50 different genera was established. Overall, Nanopore and culturing data correlated well, since samples holding a greater potential at the microbiome level also yielded a more interesting set of microbial isolates, whereas samples showing less biodiversity resulted in a reduced (and redundant) set of culturable bacteria. Thus, we anticipate that portable sequencers hold potential as key, easy-to-use tools for *in situ*-informed bioprospecting strategies.

Keywords: Nanopore sequencing, bioprospecting, *in situ* sequencing, 16S rRNA gene sequencing, microbiome analysis, Tabernas Desert

OPEN ACCESS

Edited by:
Angel Valverde,
Spanish National Research Council (CSIC), Spain

Reviewed by:
Jeremy Dodsworth,
California State University, San Bernardino, United States
Juan Antonio Ugalde,
Universidad del Desarrollo, Chile

****Correspondence:***
Cristina Vilanova
cristina@darwinbioprospecting.com

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *31 August 2021*
Accepted: *15 November 2021*
Published: *13 December 2021*

Citation:
Latorre-Pérez A, Gimeno-Valero H, Tanner K, Pascual J, Vilanova C and Porcar M (2021) A Round Trip to the Desert: In situ Nanopore Sequencing Informs Targeted Bioprospecting. Front. Microbiol. 12:768240. doi: 10.3389/fmicb.2021.768240

INTRODUCTION

Scaling laws have predicted that the Earth is home to 1 trillion (10^{12}) microbial species (Locey and Lennon, 2016). A large fraction of this biodiversity still remains to be explored and very likely harbors novel molecules, enzymes and/or biological activities with potential applications in industrial processes, drug development, cosmetics or environment-related issues (i.e., bioremediation). The search for these novel products from biological sources and, in particular, from microorganisms, is known as microbial bioprospecting. Extreme environments, such as the deep sea or hyper-arid deserts, are of special interest for bioprospecting studies, as they tend to be sources of undiscovered biodiversity (Bull and Goodfellow, 2019).

The characteristics (i.e., nutrient and oxygen availability, humidity, irradiation, pH, etc.) of a given environment shape the composition of its microbiota, often leading to the existence of temporal and spatial variations in the microbial community composition (Lauber et al., 2009; DiGiulio et al., 2015). Spatial changes have also been observed at microscale: for example, in gradients of soil depths as recently demonstrated with the SoilBox system (Bhattacharjee et al., 2020).

In this context, sequencing technologies can be used for elucidating whole microbial profiles from samples, thus enabling to unveil changes in microbiome composition which are usually not detected with culture-based approaches. Illumina sequencing platforms—such as the MiSeq System—are the current standard for microbiome sequencing. Nevertheless, this technology is time-consuming and usually requires shipping the samples to a centralized sequencing facility. Therefore, *in situ* third-generation sequencing (TGS) strategies emerge as a promising alternative to this traditional approach (Latorre-Pérez et al., 2021).

Among TGS technologies, the Oxford Nanopore Technologies (ONT) MinION system is especially relevant for *in situ* sequencing as it is the smallest sequencing device currently available, it is inexpensive in comparison to other TGS devices, and the obtention of long reads can be assessed in real-time (Latorre-Pérez et al., 2021). Thus, sequencing data can be directly analyzed through bioinformatic pipelines that can be run on servers, laptops, or even mobile phones (Mitsuhashi et al., 2017; Palatnick et al., 2021).

Nanopore sequencing has previously been used in range of real-time applications, such as pathogen detection and surveillance (Quick et al., 2016; Charalampous et al., 2019; Chan et al., 2020); forensic identification (Tytgat et al., 2020; Vasiljevic et al., 2021); or industrial process monitoring (Hardegen et al., 2018; McHugh et al., 2021). Among all the potential uses of MinION, *in situ* sequencing is especially interesting for those situations where no alternative analyses are feasible due to a lack of equipment (i.e., second-generation sequencing platforms, qPCR instruments.). This is the case for most bioprospecting expeditions, which are usually carried out far away from microbiology laboratories. Previous works have demonstrated that both sample preparation and microbiome sequencing can be achieved using a reduced, mobile laboratory. Indeed, Nanopore sequencing has been successfully applied in extremely remote locations such as the Antarctic Dry Valleys (Johnson et al., 2017), the Canadian High Arctic (Goordial et al., 2017), the largest European ice cap (Vatnajökull, Iceland) (Gowers et al., 2019) or the International Space Station (Castro-Wallace et al., 2017; Burton et al., 2020). Beyond the undoubtedly scientific interest of analyzing microbial samples up to hundreds of kilometers away from the nearest laboratory, microbial bioprospecting could further benefit from *in situ* sequencing, as it would allow for a more directed and evidence-based sampling procedure focused on those sampling locations that prove to be enriched with the microbial taxa and/or biological activities of interest.

To test this hypothesis, we planned a two-night expedition to the Tabernas Desert (Almería, Spain). This dryland has recently been reported to harbor a previously unexplored high bacterial biodiversity, significantly enriched in radiation- and desiccation-resistant microorganisms, which were the target of our study (Molina-Menor et al., 2021). A minimum setup of both laboratory and bioinformatic tools was designed and used for analyzing biocrust and soil samples *via* 16S rRNA gene sequencing throughout the expedition. The obtained taxonomic profiles were used to identify sample types enriched in taxa that have been described to be radiation resistant, allowing us to collect additional samples before ending the journey. Overall, this work demonstrates the feasibility of using portable, nanopore-based sequencing devices to study microbial communities without the need of returning to the lab, which could potentially inform decision-making during sampling.

RESULTS

Sampling Expedition Roadmap

Based on previous sampling experiences in the Tabernas Desert and sequencing tests performed in the laboratory, a detailed roadmap for the expedition and the experimental procedures was designed (**Figure 1**). The total duration of the expedition was less than 60 h, including traveling (∼25% of hands-on time) and two nights. The rest of hands-on time was spent on library preparation (∼28%), sequencing and basecalling (∼26%), sampling and setup (13%), and data analysis (8%). The first set of sequencing data was generated approximately 24 h after sample collection.

Microbiome Sequencing and Bioinformatic Analysis

Twelve biocrust and two bulk soil samples ("control" samples) were collected and analyzed through full-length 16S rRNA gene sequencing using the ONT MinION platform. A total of 1,657,804 raw reads were generated. After length and quality filtering, an average of 101,972 ± 20,949 sequences per sample were obtained (min: 50,051; max: 128,282; median $Q = 10.3$). Reads were subsequently analyzed by using a custom pipeline (Spaghetti), which was inspired by previous works (Cuscó et al., 2018; Santos et al., 2020; Urban et al., 2021). Spaghetti relied on minimap2 (Li, 2018) alignments against the SILVA v. 138 database (Quast et al., 2013), and taxonomic assignments were obtained in ∼2 h. Other alignment tools were tested as alternatives to minimap2, but they were discarded for different reasons: BLAST took ∼26 h to finish a ∼1M reads analysis, while LAST exceeded the available laptop's RAM (16 Gb).

Taxonomic and Diversity Analysis

Spaghetti data analysis and visualization pipeline generated several plots designed to provide a rapid overview of the taxonomy and the diversity of the samples (**Supplementary File 1**). At the phylum level, biocrust samples were dominated by *Cyanobacteria* (∼34.5% of average relative abundance), *Bacteroidota* (∼22.7%), *Proteobacteria* (∼19.2%), *Acidobacteriota* (∼6.0%) and *Actinobacteriota* (∼4.7%), while soil samples were mainly characterized by *Actinobacteriota* (∼24.8%), *Acidobacteriota* (∼18.6%), *Proteobacteria* (∼14.2%). *Planctomycetota* (∼14.2%) and *Gemmatimonadota* (∼7.3%) (**Supplementary Figure 1** and **Supplementary Table 1**).

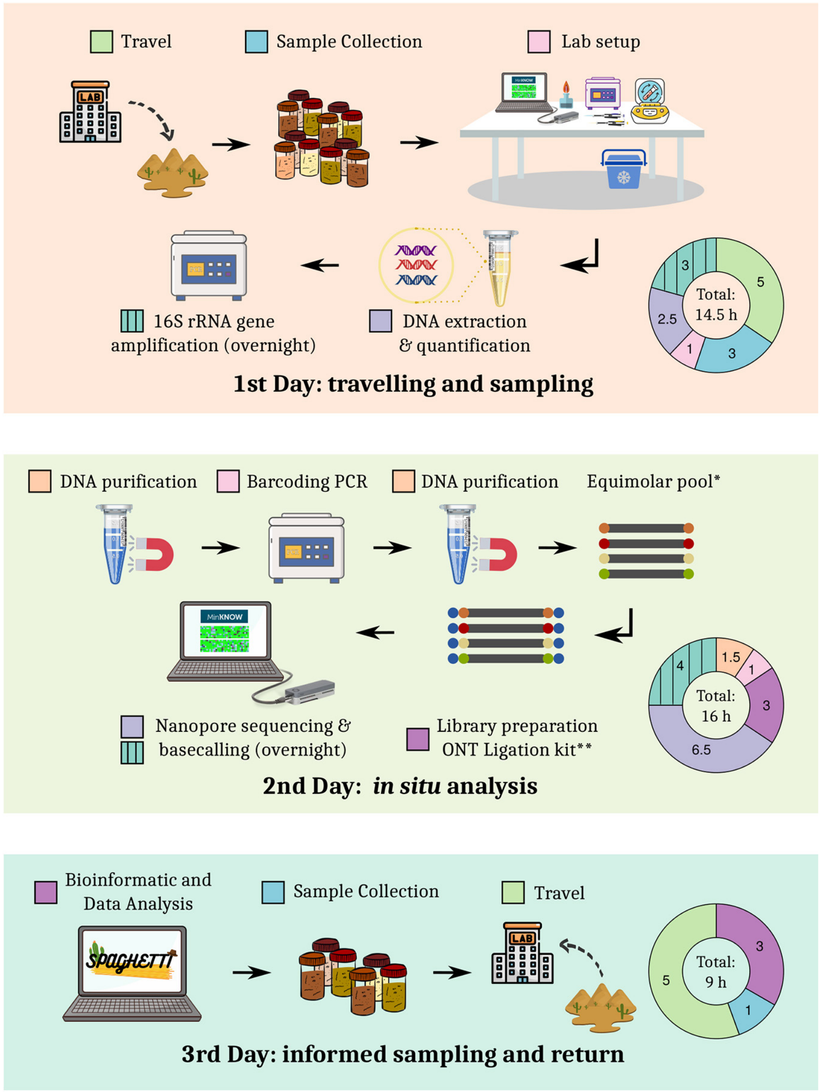

Travel
Sample Collection
Lab setup
LAB
16S rRNA gene amplification (overnight)
DNA extraction & quantification
Total: 14.5 h
3
5
2.5
1
3
1st Day: travelling and sampling
DNA purification
Barcoding PCR
DNA purification
Equimolar pool*
MinKNOW
Nanopore sequencing & basecalling (overnight)
Library preparation ONT Ligation kit**
Total: 16 h
4
1.5
1
3
6.5
2nd Day: *in situ* analysis
Bioinformatic and Data Analysis
Sample Collection
Travel
SPAGHETTI
LAB
Total: 9 h
3
5
1
3rd Day: informed sampling and return

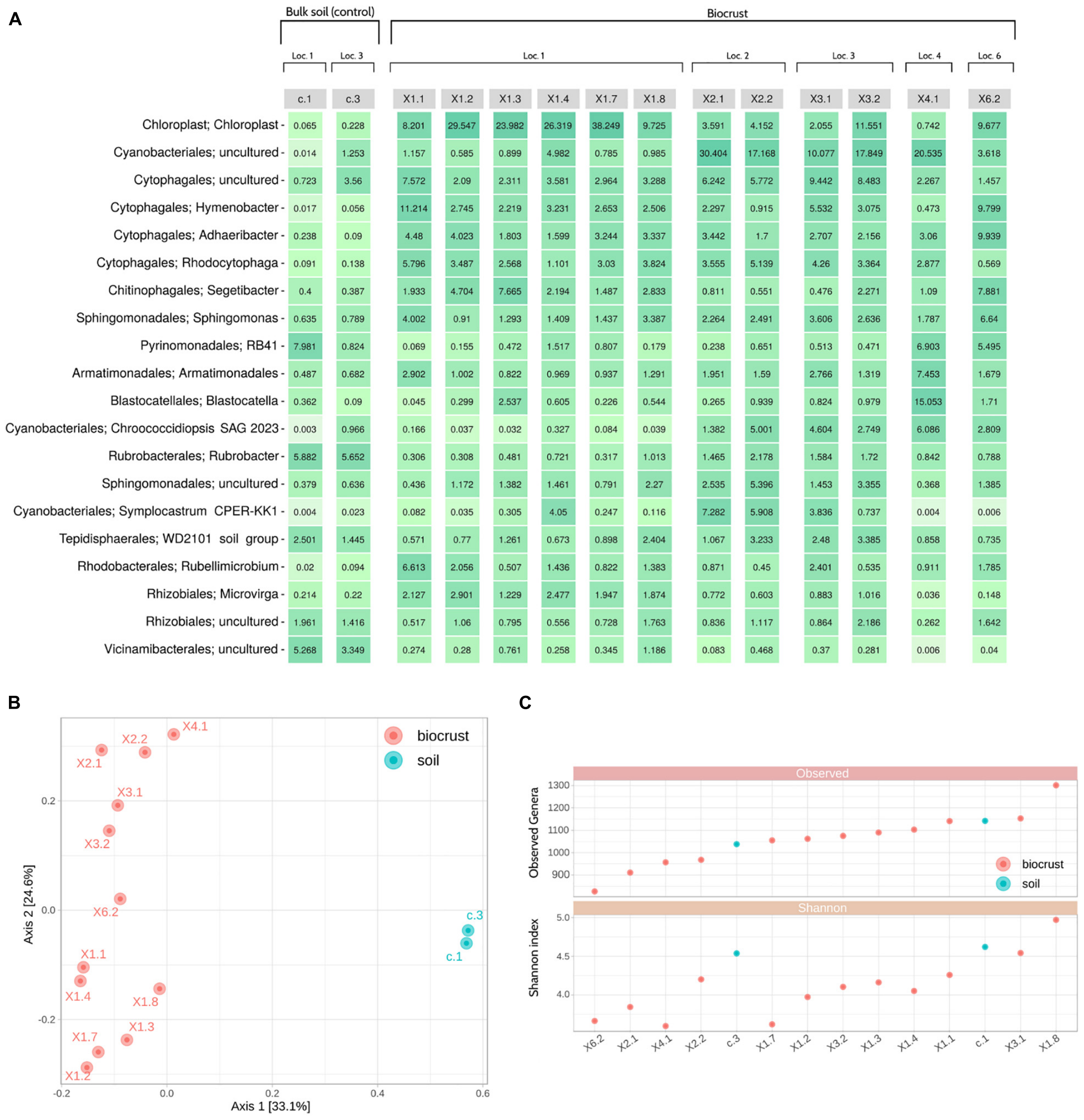
FIGURE 1 | Detailed roadmap of the sampling expedition.

As expected, a higher variability in the microbiome composition was detected at the genus level, with an uncultured *Cyanobacteriales* (∼4.7% of average relative abundance), *Hymenobacter* (∼3.9%), an uncultured *Chroococcidiopsaceae* (∼3.8%), an uncultured *Spirosomaceae* (∼3.7%) and *Adhaeribacter* (∼3.5%) being the most dominant taxa for biocrust samples. Moreover, a considerable amount of reads (∼14.0%) were assigned to chloroplasts in these samples. On the other hand, soil samples were mainly characterized by *Rubrobacter* (∼5.8%), *Vicinamibacteraceae* (∼4.8%), an uncultured *Pirellulaceae* (∼4.7%), *Pyrinomonadaceae* RB41 (∼4.4%), an uncultured *Vicinamibacterales* (∼4.1%), and a low presence of reads assigned to chloroplasts (∼0.15%) (**Figure 2A** and **Supplementary Table 2**).

Beta diversity analyses showed that biocrust and soil samples were clearly distinguishable at the microbiome level. Moreover, samples tend to cluster based on their sampling location (X1, X2, X3, X4 or X6), instead of other characteristics (i.e., color and shape of the biocrust) (**Figure 2B**). Alpha diversity indices were used to identify the most and least rich and diverse samples, which were X1.8/X1.3/c.1 and X6.2/X2.1/X4.1, respectively (**Figure 2C**).

Radiation- and Desiccation-Resistant Bacteria Detection

Once the general taxonomic and diversity profiles were obtained, special attention was paid to 29 bacterial genera that had proven



A

	Bulk soil (control)		Biocrust											
	Loc. 1	Loc. 3	Loc. 1						Loc. 2		Loc. 3		Loc. 4	Loc. 6
	c.1	c.3	X1.1	X1.2	X1.3	X1.4	X1.7	X1.8	X2.1	X2.2	X3.1	X3.2	X4.1	X6.2
Chloroplast; Chloroplast	0.065	0.228	8.201	29.547	23.982	26.319	38.249	9.725	3.591	4.152	2.055	11.551	0.742	9.677
Cyanobacteriales; uncultured	0.014	1.253	1.157	0.585	0.899	4.982	0.785	0.985	30.404	17.168	10.077	17.849	20.535	3.618
Cytophagales; uncultured	0.723	3.56	7.572	2.09	2.311	3.581	2.964	3.288	6.242	5.772	9.442	8.483	2.267	1.457
Cytophagales; Hymenobacter	0.017	0.056	11.214	2.745	2.219	3.231	2.653	2.506	2.297	0.915	5.532	3.075	0.473	9.799
Cytophagales; Adhaeribacter	0.238	0.09	4.48	4.023	1.803	1.599	3.244	3.337	3.442	1.7	2.707	2.156	3.06	9.939
Cytophagales; Rhodocytophaga	0.091	0.138	5.796	3.487	2.568	1.101	3.03	3.824	3.555	5.139	4.26	3.364	2.877	0.569
Chitinophagales; Segetibacter	0.4	0.387	1.933	4.704	7.665	2.194	1.487	2.833	0.811	0.551	0.476	2.271	1.09	7.881
Sphingomonadales; Sphingomonas	0.635	0.789	4.002	0.91	1.293	1.409	1.437	3.387	2.264	2.491	3.606	2.636	1.787	6.64
Pyrinomonadales; RB41	7.981	0.824	0.069	0.155	0.472	1.517	0.807	0.179	0.238	0.651	0.513	0.471	6.903	5.495
Armatimonadales; Armatimonadales	0.487	0.682	2.902	1.002	0.822	0.969	0.937	1.291	1.951	1.59	2.766	1.319	7.453	1.679
Blastocatellales; Blastocatella	0.362	0.09	0.045	0.299	2.537	0.605	0.226	0.544	0.265	0.939	0.824	0.979	15.053	1.71
Cyanobacteriales; Chroococcidiopsis SAG 2023	0.003	0.966	0.166	0.037	0.032	0.327	0.084	0.039	1.382	5.001	4.604	2.749	6.086	2.809
Rubrobacterales; Rubrobacter	5.882	5.652	0.306	0.308	0.481	0.721	0.317	1.013	1.465	2.178	1.584	1.72	0.842	0.788
Sphingomonadales; uncultured	0.379	0.636	0.436	1.172	1.382	1.461	0.791	2.27	2.535	5.396	1.453	3.355	0.368	1.385
Cyanobacteriales; Symplocastrum CPER-KK1	0.004	0.023	0.082	0.035	0.305	4.05	0.247	0.116	7.282	5.908	3.836	0.737	0.004	0.006
Tepidisphaerales; WD2101 soil group	2.501	1.445	0.571	0.77	1.261	0.673	0.898	2.404	1.067	3.233	2.48	3.385	0.858	0.735
Rhodobacterales; Rubellimicrobium	0.02	0.094	6.613	2.056	0.507	1.436	0.822	1.383	0.871	0.45	2.401	0.535	0.911	1.785
Rhizobiales; Microvirga	0.214	0.22	2.127	2.901	1.229	2.477	1.947	1.874	0.772	0.603	0.883	1.016	0.036	0.148
Rhizobiales; uncultured	1.961	1.416	0.517	1.06	0.795	0.556	0.728	1.763	0.836	1.117	0.864	2.186	0.262	1.642
Vicinamibacterales; uncultured	5.268	3.349	0.274	0.28	0.761	0.258	0.345	1.186	0.083	0.468	0.37	0.281	0.006	0.04

FIGURE 2 | Nanopore sequencing results. **(A)** Heatmap showing the top 20 genera detected in the samples and their relative abundances. **(B)** Principal Coordinates Analysis (PCoA) using the Bray-Curtis dissimilarity metric. **(C)** Alpha diversity analysis: Observed genera (richness) (top); Shannon index (bottom). Samples are ordered by richness. Loc, Location.

to be radiation- and/or desiccation-resistant according to the literature (Montero-Calasanz et al., 2013; Yu et al., 2015; Deng et al., 2016; Etemadifar et al., 2016; Paulino-Lima et al., 2016; Golinska et al., 2020; Tanner, 2020). The objective of this analysis was to identify those samples which maximized the richness and abundance of those radiation- and/or desiccation-resistant taxa, since they should hold a greater potential for isolating and discovering microbial strains and substances of biotechnological interest.

Overall, the number of radiation- and desiccation-resistant genera detected in the samples by Nanopore sequencing was high, ranging from 23 (X2.1 and X2.2) to 29 (X1.1 and X1.8) (**Figure 3A**). Although some of the taxa were present in low abundance (< 0.01%), the selected bacteria accounted for 11.5% of the relative abundance of the samples, in average (**Figure 3B**). Biocrust profiles were dominated by *Hymenobacter* (∼3.9% of the total relative abundance), *Sphingomonas* (∼2.7%), *Rubellimicrobium* (∼1.7%), *Microvirga* (∼1.3%) and *Rubrobacter* (∼1%). The two bulk soil samples were mainly characterized by the presence of *Rubrobacter* (∼5.8%), *Arhtrobacter* (∼0.9%) and *Sphingomonas* (∼0.7%) (**Figure 3** and **Supplementary Table 2**).

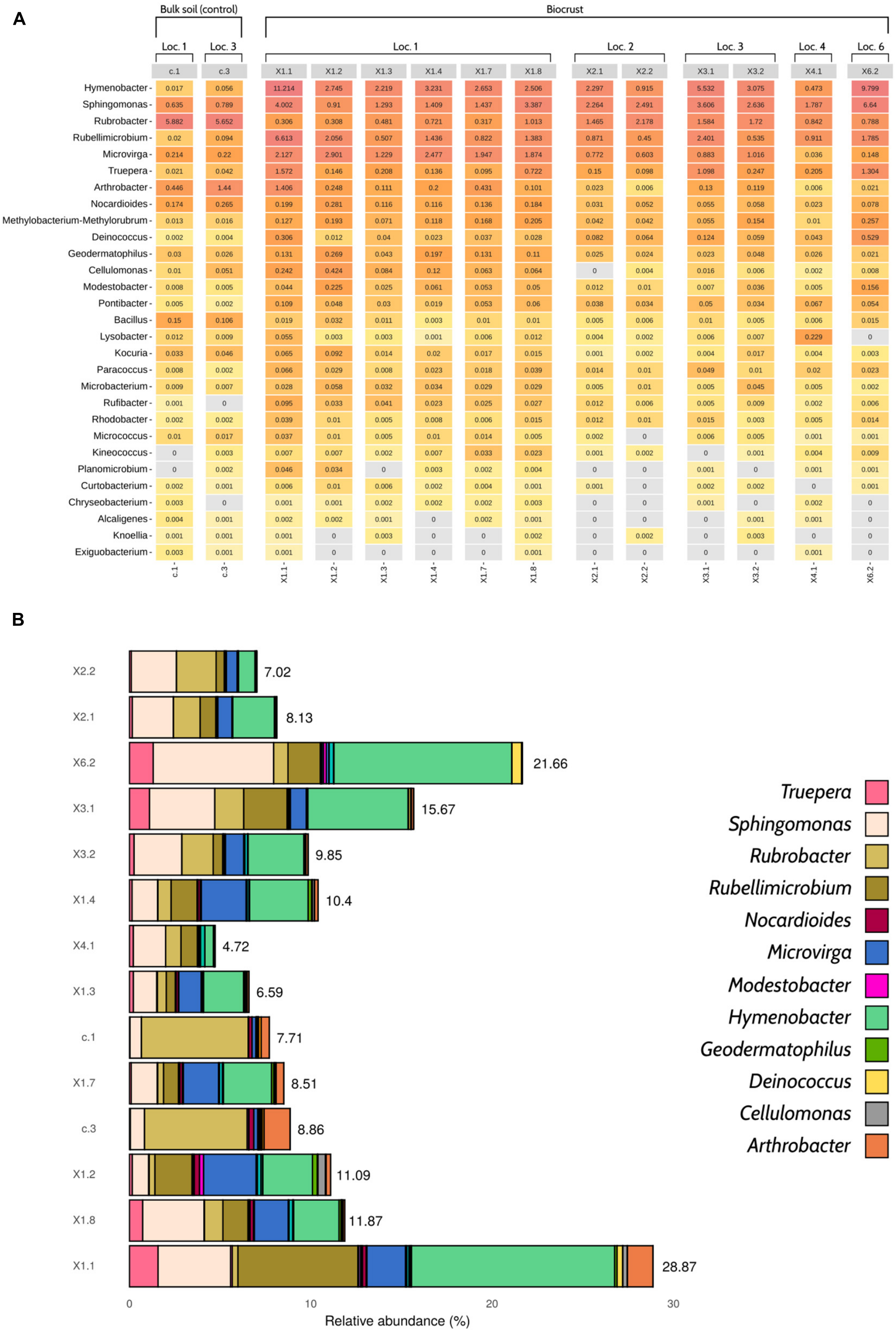


FIGURE 3 | Profile of desiccation- and radiation-resistant bacteria according to Nanopore sequencing data. **(A)** Heatmap showing the 29 genera of interest and their relative abundances (%). **(B)** Barplot displaying the cumulative relative abundances of the selected taxa (*n* = 29). Only 12 genera have been colored in order to improve visualization, as the abundance of some taxa was so low that they cannot be properly distinguished in the figure. An interactive version of this figure including the 29 genera of interest can be found in **Supplementary Figure 4**. The relative abundance of desiccation- and radiation-resistant genera was calculated considering the whole microbial community, not only the taxa of interest. Loc, Location.

After analyzing all the results provided by the pipeline, additional samples were collected. This time, bioprospecting activities focused on obtaining biological replicates of three selected samples: (a) biocrust X1.1, with the highest number of radiation- and desiccation-resistant genera (29); (b) bicrust X2.1, with the lowest number of radiation- and desiccation-resistant genera (23); and (c) bulk soil c.1, taken as a control for comparisons between biocrust and bulk soil samples.

GPS positions of the original samples were traced back and samples were identified based on the pictures that were taken on the first sampling day. Finally, two additional replicates were collected for each type of sample.

Microbial Collection Establishment and Identification

Back in the laboratory after the expedition, all the collected samples (n = 20) were cultured under three different conditions: (1) Tryptic Soy Agar (TSA) medium, (2) SSE/HD medium (SSE/HD), and (3) SSE/HD medium + uninterrupted artificial light (SSE/HD + light). A total of 166 strains comprising 50 different genera were isolated and identified through Sanger sequencing of the partial 16S rRNA gene. The bacterial colonies displayed differences in morphology and appearance, with white, yellow, pink, red, orange and brown being the most predominant colors (**Supplementary Table 3**). Initially, samples cultured on SSE/HD + light did not display any microbial growth after 4 weeks of incubation. For that reason, plates were removed from the artificial light, and a few days later, different bacterial colonies started to grow.

The genus *Arthrobacter* was the most represented in the microbial collection, with up to 37 isolates belonging to this taxonomic group (**Figure 4**). A total of 15 strains, which were mainly isolated from soil samples, were classified as *Streptomyces.* Other predominant genera in the collection were *Pseudoarthrobacter* (9 isolates), *Kocuria* (6), *Bacillus* (6), *Skermanella* (5), *Blastococcus* (5), and *Belnapia* (5). At the sample level, biocrusts collected from Location 1 (X1) presented the highest number of bacterial isolates. Specifically, samples X1.1B (21 isolates/15 unique genera), X1.2 (17/13), X1.3 (16/12), and X1.1 (16/10) showed the highest diversity of cultured bacteria. On the other hand, samples X3.2 (0/0), X6.2 (0/0), X4.1 (2/2), and X2.1A (2/2) presented the lowest diversity of isolates (**Figure 4**).

The taxonomic profiles obtained by Nanopore sequencing were compared to the results from the molecular identification of the isolated strains. Overall, Nanopore sequencing and culture-based data correlated well. In fact, only 14 out of the 166 isolated strains belonged to genera that were not detected in the original sample by *in situ* microbiome sequencing (**Figure 4** and **Supplementary Table 3**). Interestingly, three of the isolated genera (*Mycolicibacterium, Lentzea,* and *Sinorhizobium*) were not detected in any sample of the dataset. After revising the database used for assigning the taxonomy of the reads (see section "Materials and Methods"), a mislabeling of those taxa at the genus level was detected. Specifically, *Mycolicibacterium* was labeled as *Mycobacterium, Lentzea as Lechevalieria* and *Sinorhizobium* as *Ensifer*. These three genera were indeed detected by Nanopore sequencing in all the samples where the strains were isolated from **Supplementary Table 2**. Finally, it is worth highlighting that some of the most abundant radiation-resistant bacteria detected by *in situ* 16S rRNA sequencing (i.e., *Hymenobacter, Rubrobacter, Rubellimicrobium, Microvirga, Truepera. . .*) were not cultured from any sample. Indeed, only 50 out of the 441 genera (11.3%) with an average relative abundance > 0.01% according to Nanopore sequencing were represented in the microbial culture collection.

Among the selected samples, X1.1 yielded the highest number of total cultured strains (48), the highest number of different cultured genera -as deduced by partial 16S rRNA gene Sanger sequencing- (26), and the highest number of cultured genera classified as radiation- or desiccation-resistant according to literature (8) (**Figure 5**). In contrast, and as expected considering the results from *in situ* sequencing (**Figure 3B**), X2.1 samples displayed the lowest diversity of cultured bacteria and radiation- and desiccation-resistant genera. Moreover, almost all the genera isolated from X2.1 samples were also isolated from X1.1 (**Figure 5B**, **Supplementary Figure 2**, and **Supplementary Tables 4**, **5**), thus confirming the hypothesis that this sample was less valuable from the bioprospecting point of view. A different profile of bacteria was isolated from C1 bulk soil samples (**Supplementary Figure 2**), with only one radiation-resistant genus -*Sphingomonas*- cultured exclusively from this type of sample (**Figure 5B**). Interestingly, the relative abundance of *Sphingomonas* was higher in all the biocrust samples than in bulk soil (**Figure 3A**), although this genus was isolated only from samples C1B and X1.8.

Focusing on the culture conditions, 124 strains were isolated from TSA (38 different genera), 24 from SSE/HD (17 different genera), and 18 from SSE/HD + light (13 different genera) (**Figure 6** and **Supplementary Table 6**). Nevertheless, strains isolated from SSE/HD + light presented a significantly lower similarity to their closest type strain than strains isolated from TSA (FDR adjusted p-value < 0.05; Mann–Whitney U test), based on partial 16S rRNA gene sequencing. In fact, ∼89% of the strains isolated from SSE/HD + light showed a similarity lower than 98.7% to their closest neighbor, a common threshold for defining new species (Chun et al., 2018), compared to ∼46 and 66% displayed by TSA- and SSE/HD-isolated strains, respectively (**Figure 6B**).

DISCUSSION

Bioprospecting is often a unidirectional process, with scientists leaving their research institute for several days or weeks to collect samples that are only screened upon arrival at the laboratory. This is usually a blind task, since the screening results are obtained once the expedition is over. As sampling sites are generally remote and far from the researcher's laboratory, returning to the locations where bioprospecting occurred is not always viable, thus preventing further exploitation of the samples that showed a greater potential based on the screening. This work is a proof of concept of the use of portable Nanopore sequencing as a tool for guiding and informing bioprospecting activities during

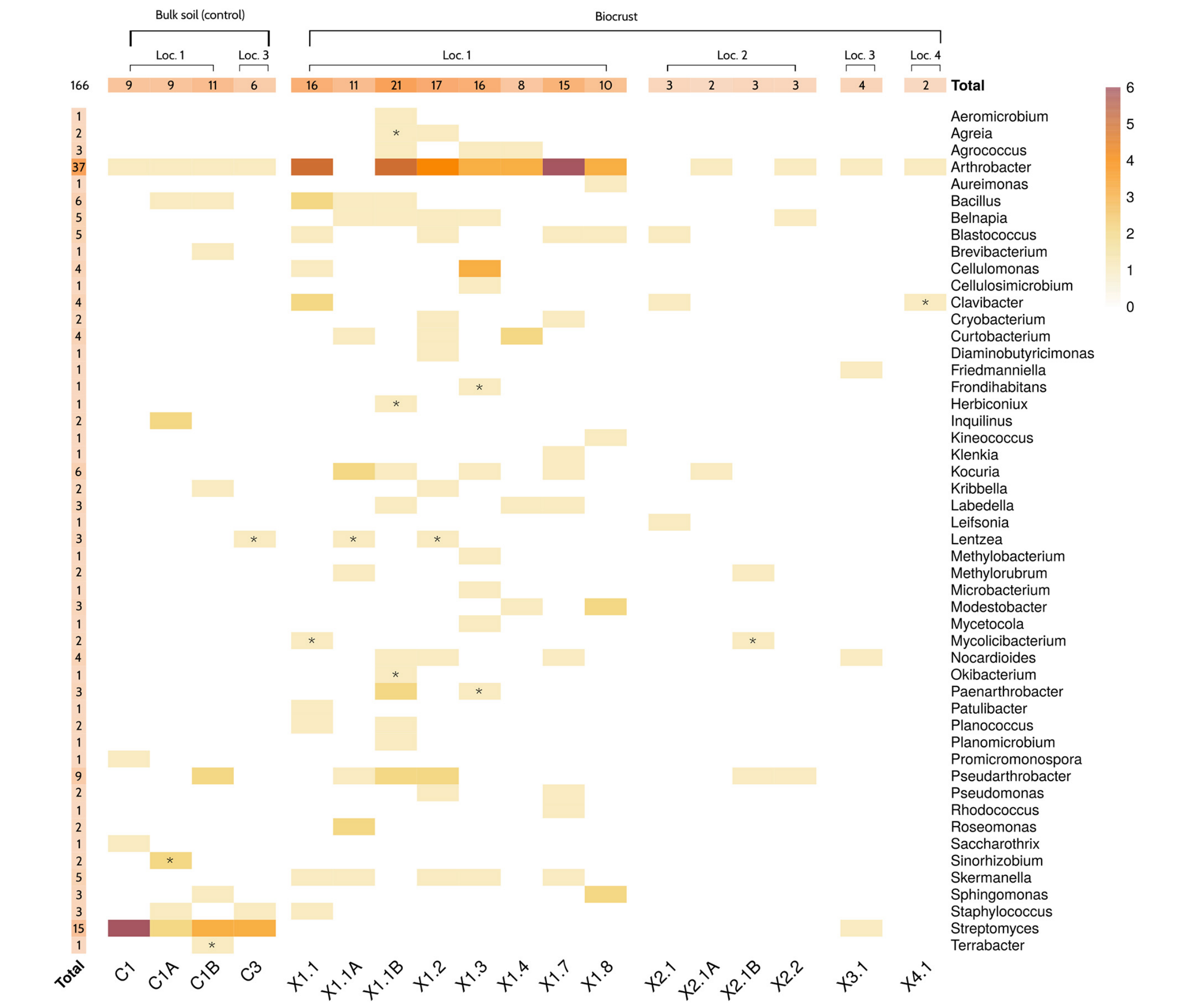


FIGURE 4 | Culture collection description. Heatmap showing the number of strains isolated from each sample. Genus-level taxonomy of the strains was obtained by partial 16S rRNA gene sequencing of isolates. Letters "A" and "B" indicate the samples that were collected on the third day, after analyzing the original samples by Nanopore sequencing. Symbol "*" highlights those genera that were not originally detected in that sample by *in situ* sequencing. Only genera with a relative abundance higher than 0.001% were considered as detected. Samples 3.2 and 6.2 are not shown, since no bacterial strain was isolated from them. Loc, Location.

a sampling expedition, in our case to the only European desert, the Tabernas Desert (Almería, Spain).

ONT sequencing is a well-established technique for studying microbial communities (Ciuffreda et al., 2021), and portable sequencing (i.e., MinION) has indeed been applied to characterize microbiomes in some of the most remote places of the universe that are accessible to human beings (Castro-Wallace et al., 2017; Goordial et al., 2017; Johnson et al., 2017; Gowers et al., 2019; Burton et al., 2020). Although some authors have demonstrated the utility of *in situ* sequencing to assess the animal biodiversity in the rainforest (Menegon et al., 2017; Pomerantz et al., 2018), the present work is, to the best of our knowledge, the first confirmation that this technology can be applied during a microbial bioprospecting expedition to improve the bioprospecting strategy itself.

Our results demonstrate that DNA analyses can be integrated into the sampling roadmap, while keeping the duration of the journey under 72 h (**Figure 1**). The obtained sequencing yield was substantially higher than the output reported in other on-site studies (Latorre-Pérez et al., 2021), and it was comparable to the yield of runs performed on fully equipped laboratories (Nygaard et al., 2020; Urban et al., 2021). It must be noted that instead of directly sequencing in the field, we decided to set up a mobile laboratory 15 km away from the sampling location in an apartment with internet and electricity access. This allowed us to apply the same protocols that we routinely use in the

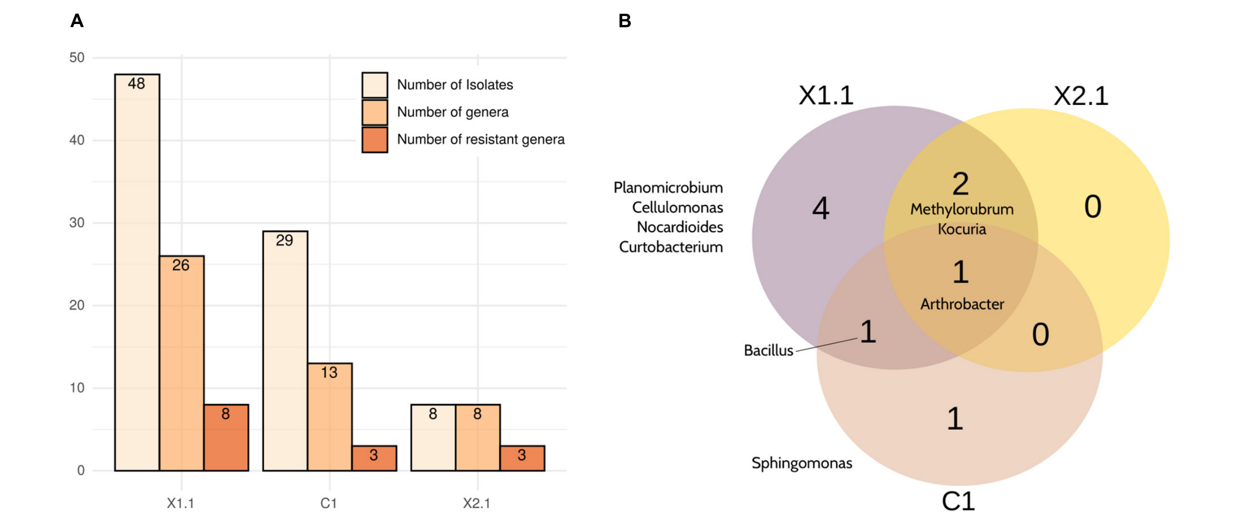


FIGURE 5 | Summary of the strains isolated from the selected samples (X1.1, X2.1, and C1). **(A)** Number of different strains, genera and desiccation-/ radiation-resistant genera isolated from each sample. **(B)** Venn diagram comparing the desiccation- and radiation-resistant genera isolated from any replicate of each selected sample. Note that all the identifications were obtained by partial 16S rRNA gene Sanger sequencing.

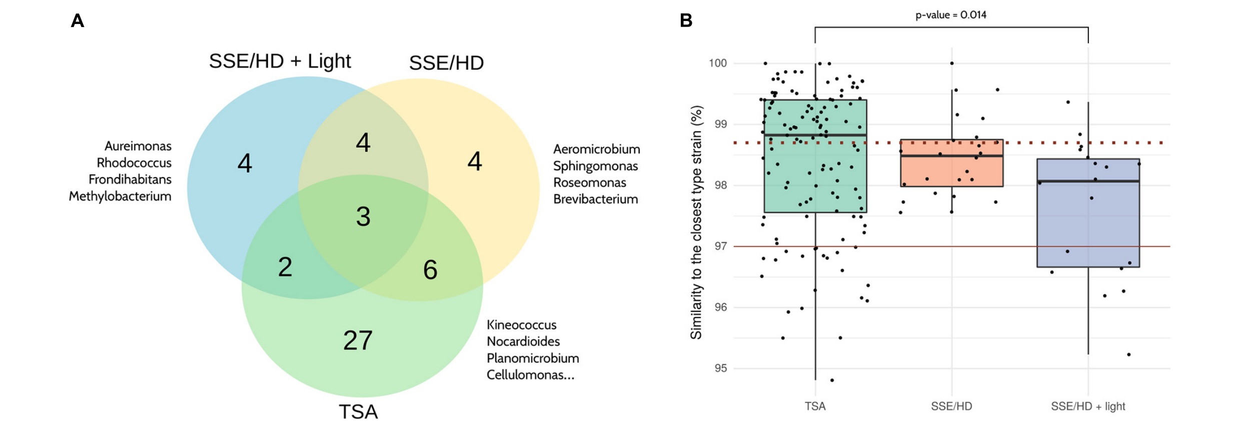


FIGURE 6 | Comparison of the different culture conditions. **(A)** Venn diagram showing the bacterial genera isolated from each culture condition. The complete list of genera isolated from each condition can be found in **Supplementary Table 6**. **(B)** Percentage of similarity shared by each strain and its closest phylogenetic neighbor according to partial 16S rRNA gene sequencing. The dotted and the solid red lines are drawn on 98.7 and 97% of similarity, respectively. The Mann–Whitney *U* test was applied for comparing between groups, and *p*-values were corrected using the Benjamini-Hochberg method. Only significant results are highlighted. Note that all the identifications were obtained by partial 16S rRNA gene Sanger sequencing.

laboratory with little modifications, thus reducing the risk of failure during the expedition. Nevertheless, simplified protocols (i.e., Field Sequencing Kit; ONT, Oxford, United Kingdom, Cat. No.: SQK-LRK001) involving shorter preparation time and less equipment could be employed, even with the lack of electricity or internet, as has been previously demonstrated (Edwards et al., 2019; Gowers et al., 2019). Indeed, Spaghetti does not require an internet connection, so this pipeline could be also used for on-site analyses.

Different sample types (i.e., biocrust and bulk soil) were clearly distinguishable according to microbial profiles (**Figure 2**). As expected, *Cyanobacteria* was more abundant in biocrust samples, since these microorganisms are a crucial part of biological soil crusts, which often also harbor other organisms such as lichens, microalgae, microfungi or mosses (Williams et al., 2016; Machado-de-Lima et al., 2019). This would explain the higher presence of sequences assigned to chloroplasts in this type of samples. Overall, phylum-level taxonomy was concordant with the microbial profiles expected for soil samples, with *Bacteroidota, Proteobacteria, Acidobacteriota, Actinobacteriota, Planctomycetota, Verrucomicrobiota,* and *Gemmatimonadota* dominating the microbiomes (Buckley et al., 2006; Spain et al., 2009; Bergmann et al., 2011; DeBruyn et al., 2011; Zhang et al., 2019; Kalam et al., 2020; Larsbrink and McKee, 2020). At the genus level, differences and similarities between samples were resolved. In consequence, *in situ* Nanopore sequencing could be especially helpful for choosing those samples that maximize the microbial diversity -according to beta diversity or any other metric-, preventing the selection of samples with poor diversity or little variation for further screening, thus saving time and resources.

Taxonomic information could also be used for identifying those samples that contain the microorganisms of interest. As a proof of concept, we focused on genera that were previously described to be desiccation- and/or radiation-resistant, and which thus hold potential for biotechnological applications (Gabani and Singh, 2013; Molina-Menor et al., 2021). The prevalence of these taxa in the samples collected from the Tabernas Desert was high (**Figure 3**). This was expected, since most of these bacteria are often found in or isolated from other arid soils and biocrusts (Holmes et al., 2000; Rainey et al., 2005; Zhang et al., 2007; Abed et al., 2010; Amin et al., 2016; Hu et al., 2016; Wübbeler et al., 2017; Liang et al., 2019). Nevertheless, *in situ* sequencing in combination with our analysis pipeline led to the categorization and identification of samples that showed a greater diversity and abundance of the genera of interest. Thanks to such information, those samples -technically, biological replicates of the samples- could be further collected and thoroughly analyzed back in the laboratory.

It is well known that detecting a certain taxon by high-throughput sequencing does not necessarily mean that this taxon can be successfully isolated from the sample. In our case, culture-based and Nanopore sequencing data correlated well (**Figure 4**), although an important fraction of the genera detected with the sequencing approach was not represented in the microbial culture collection. This could be expected given that a significant number of prokaryotic taxa are virtually "unculturable." In any case, the sample that held the greatest potential at the microbiome level -according to Nanopore data- (X1.1) also resulted in the most interestingly complex set of culturable bacteria (**Figure 5**), despite using a relatively simple culturing approach. On the other hand, some of the most dominant bacteria according to sequencing data could not be isolated from any sample (i.e., *Hymenobacter* or *Rubrobacter*) very likely due to culturing biases. Although this limitation is inherent to bioprospecting strategies that rely on obtaining microbial cultures, knowing the presence of a certain taxonomic group in the sample would allow for the use of microorganism-specific culture conditions or enrichment methods, thus increasing the chances of success.

In general, the profile of bacteria isolated from the Tabernas Desert was similar to the one previously described (Molina-Menor et al., 2021). *Arthrobacter* was the predominant genus, which is consistent with the observations of da Rocha et al. (2015). Other bacteria, such as *Belnapia, Kocuria* or *Skermanella* were also recurrent in biocrust samples. Nevertheless, up to 29 genera isolated in this study were not recovered by Molina-Menor et al. (2021).

Interestingly, some of the isolated bacteria may represent new species according to partial 16S rRNA gene sequencing, showing the great, yet to be discovered, ecological and biotechnological potential hidden in the Tabernas Desert. Although full 16S rRNA gene sequences and genomes should be retrieved for circumscribing new taxa (Chun et al., 2018), bacteria isolated from SSE/HD + light displayed a lower similarity to any other previously described type strain (**Figure 6**). These results were indeed obtained by serendipity, as bacterial growth was only detected after removing the culture plates from artificial light (∼4 weeks after plating), which was not the original idea.

Despite the promising results obtained in this proof of concept, we have identified some limitations of *in situ* Nanopore sequencing. The first one is the taxonomic resolution of 16S rRNA gene sequencing. Although long-read platforms have the ability to sequence the full-length 16S rRNA gene, the intrinsic error associated to ONT sequencing hampers species-level identification. This error also hinders the direct comparison between Nanopore-based microbiome sequencing and the 16S rRNA gene sequences obtained from the isolates by Sanger sequencing, as it would be difficult to discern if a particular fraction of Nanopore reads actually comes from a specific strain in the collection or from a phylogenetically related strain (or even species) that may or may not have been isolated. For that reason, we decided to perform the analyses at the genus level and to compare the taxonomic profiles instead of comparing the sequences. Nanopore-based, 16S rRNA gene sequencing has proved to be robust for microbiome characterization at this taxonomic level, showing a performance similar to Illumina sequencing (Cuscó et al., 2018; Heikema et al., 2020; Matsuo et al., 2020; Nygaard et al., 2020; Winand et al., 2020). However, as the final objective of bioprospecting is to actually isolate the bacterial strains, it must be noted that phenotype can greatly vary among members of the same genus or even species, so genus-resolved taxonomy could be insufficient in some cases. Recent studies have shown that species-level resolution is feasible thanks to advances in software (Curry et al., 2021; Rodríguez-Pérez et al., 2021), while other works demonstrated that improved taxonomic resolution could be achieved by using longer amplicons (16S-ITS-23S) (Benítez-Páez and Sanz, 2017; Cuscó et al., 2018). Moreover, Nanopore sequencing errors are also decreasing due to improvements in basecallers and chemistries, which have allowed to reach up to 99.3% of modal accuracy on raw reads (accessed on July 17, 2021).[1] If accuracy continues to increase at this rate, it is reasonable to think that species-level identifications, and even strain-level resolution in some cases, may be achieved in the near future. Nevertheless, high-accuracy basecalling models are based on complex machine learning methods that require longer execution time, so improvements on the speed of these models are still required for being used in real-time applications (Xu et al., 2020).

It must be highlighted that this study was focused on the detection and isolation of potential radiation- and desiccation-resistant bacteria according to their taxonomic affiliation and according to the previous bibliography describing this type of

[1] https://nanoporetech.com/accuracy

features in particular genera. Our approach is thus a proof of concept that a wide taxonomic group can be identified in the samples by using Nanopore sequencing, but 16S rRNA gene itself would not be an accurate predictor of the actual ability of the isolates to resist radiation or desiccation (Steen et al., 2019). If the purpose of the bioprospecting expedition is to detect specific functional activities, shotgun metagenomic data would be needed to resolve the taxonomy at the strain level (Dilthey et al., 2019) and to ascertain the functional potential of the different members of the microbial community according to their gene content. In this regard, it has to be noted that ONT sequencers tend to incorporate indel errors on the reads that complicate the functional prediction (Watson and Warr, 2019; Latorre-Pérez et al., 2020), and this is therefore a current limitation of the informed bioprospecting strategy we are describing in this work.

Finally, sequencing strategies show the microbiome composition based on relative abundances, which may mislead the results interpretation. For instance, if a taxon is detected in Sample 1 and in Sample 2 at 10 and 1% of relative abundance, respectively, that does not imply that Sample 1 has a higher absolute abundance of the target bacteria, since the total microbial load of the samples has not been measured. This should be taken into account when selecting the samples of interest for further exploitation.

Notwithstanding the limitations, our results clearly show that Nanopore sequencing is a powerful tool for deciphering the microbial composition of different samples during a bioprospecting expedition, and that it can contribute to optimize the sampling strategy *in situ*. With microorganisms colonizing almost any known biotope (Archer et al., 2019; Sielaff et al., 2019; Tanner et al., 2020), an instrument able to resolve microbial communities inhabiting different niches is a valuable resource that can be used for targeting sample collection. Therefore, it can be envisaged a close future in microbial ecology, in which bioprospecting journeys will start with a preliminary sampling step, coupled to nanopore-based *in situ* analysis, which will enable a second, more targeted sampling (of specific plant species, soil depths, geological substrates, salt concentration, humidity level, etc.) in a very short time lapse. This strategy will both ease further work in the lab and increase the chances of identification of the target microbial taxa and/or biomolecule of interest.

MATERIALS AND METHODS

Sample Collection

Sampling was carried out in November 2020 at the Tabernas Desert Natural Park (Almeria, Spain), under the permission of the competent authorities. Biocrust and bulk soil samples were collected in two different days. Biocrust samples were gathered using a laboratory spatula that was sterilized with ethanol 96% immediately before collecting each sample. Bulk soil (∼5 cm deep) was directly introduced into sterile falcon tubes. On the first day, fourteen different samples were taken, and then analyzed through *in situ* microbiome sequencing. Based on the results, six additional samples were gathered in the second sampling day. These samples were, indeed, biological replicates of the least and most promising samples based on sequencing data. Metadata (geolocation, type of sample, appearance and pictures) was collected and associated to each sample (**Supplementary Figure 3**).

Laboratory Setup

Requirements for DNA extraction, PCR amplification, library preparation and sequencing were evaluated, and a minimum laboratory setup was designed accordingly (**Supplementary Table 7**). The necessary equipment fitted in the trunk of a compact car, and it was transferred to an apartment in Viator (Almería, Spain), 15 km away from the Tabernas Desert, where the mobile laboratory was established and all the experimental and data analysis procedures were carried out. The apartment was equipped with electricity, internet connection, a fridge and a freezer.

DNA Extraction and 16S rRNA Gene Amplification

Approximately 0.25 g of the samples were used to perform DNA extraction with the DNEasy Power Soil Kit (QIAGEN, Germany, Cat. No.: 12888) according to the manufacturer's instructions, with an additional incubation step at 65°C after the addition of the C1 solution. DNA was resuspended in 30 μL of sterile Mili-Q water. Qubit x1 dsDNA High-Sensitivity Assay kit (Qubit 2.0 Fluorometer, Thermo Fisher, Waltham, United States, Cat. No.: Q33230) was used for DNA quantification. PCR amplification of the full-length bacterial 16S rRNA gene (V1-V9; ∼1.45 kbp) was carried out by using the S-D-Bact-0008-a-S-16 (5′-AGR GTT YGA TYM TGG CTC AG-3′) and S-D-Bact-1492-a-A-16 (5′-TAC CTT GTT AYG ACT T-3′) primers (Klindworth et al., 2013), which were tailed with the ONT Universal Tags: 5′-TTT CTG TTG GTG CTG ATA TTG C-3′ for forward primer, and 5′-ACT TGC CTG TCG CTC TAT CTT C-3′ for reverse primer. The PCR reaction mix for each sample consisted of 22 μL of H_2O, 25 μL of NZYTaq II 2x Green Master Mix (NZYTech, Lisboa, Portugal, Cat. No.: MB358), 1 μL of both forward and reverse primers and 1 μL of template DNA. For the negative control, 1 μL of Mili-Q water was used instead. The following conditions were used for PCR: initial denaturation (94°C; 1 min); amplification (35 cycles) comprising denaturation (95 °C; 1 min), annealing (49 °C; 1 min) and extension (72°C; 2 min); final extension (72 °C; 10 min). The resulting amplicons were purified with the NucleoMag kit for PCR clean up with magnetic beads (Macherey-Nagel, Germany, Cat. No.: 744100.4). Magnetic beads were used at 0.5 x concentration, and manufacturer's instructions were followed.

Barcodes were added by employing the PCR Barcoding Expansion Pack 1-96 (ONT, Oxford, United Kingdom, Cat. No.: EXP-PBC096). PCR mix consisted of 22 μL of H_2O, 25 μL of NZYTaq II 2x Green Master Mix, 1 μL of the specific barcode and 2 μL of the purified DNA. The following conditions were used for PCR: initial denaturation (95°C; 3 min); amplification (15 cycles) comprising denaturation (95°C; 15 s), annealing (62°C; 15 s) and extension (72°C; 90 s); final extension (72°C; 5 min). Amplicons were purified with the NucleoMag kit and quantified with the

Qubit x1 dsDNA High-Sensitivity Assay kit. Finally, an equimolar pool of amplicons was prepared for library construction.

Library Preparation and Nanopore Sequencing

The Ligation Sequencing Kit (ONT, Oxford, United Kingdom, Cat. No.: SQK-LSK109) was used to prepare the sequencing library. Briefly, the NEBNext FFPE DNA Repair Mix (New England Biolabs, Ipswich, United States, Cat. No.: M6630) was used for DNA repair and end-prep. Then, a purification with the NucleoMag kit was carried out. Finally, adapter ligation and clean-up was performed by following the ONT SQK-LSK109 protocol.

A R9.4.1 MinION flow cell (ONT, Oxford, United Kingdom, Cat. No.: FLO-MIN106D) was primed and loaded as indicated by the manufacturer. Sequencing was performed during ~6.5 h. Reads were basecalled with MinKNOW software (v. 20.06.5; core v. 4.0.5) using Guppy's (v. 4.0.9) fast basecalling model, and sequences with $Q < 7$ (default threshold implemented in MinKNOW) were discarded.

Bioinformatic and Statistical Analysis

Reads were analyzed with Spaghetti, a custom pipeline for automatic bioinformatic analysis of Nanopore sequencing data and semi-automatic exploratory analysis and data visualization. Briefly, Spaghetti bioinformatic pipeline consists of the following steps:

1. Porechop (v. 0.2.4)[2] is run with default parameters for removing sequencing adapters from reads.
2. Nanofilt (v. 2.7.1) (De Coster et al., 2018) is used to filter reads shorter than 1,200 bp or longer than 1,800 bp.
3. Quality check is carried out with NanoStat (v. 1.4.0) using default parameters (De Coster et al., 2018).
4. Chimeras are detected and removed by using yacrd (v. 0.6.2) with -c and -n parameters set to 4 and 0.4, respectively, as suggested by the authors for Nanopore data (Marijon et al., 2020).
5. Filtered reads are mapped against the SILVA database (v. 138) (Quast et al., 2013), as formatted and provided by Qiime2,[3] by using minimap2 (v. 2.17-r9419) (Li, 2018) with "-x map-ont" and "–secondary = no" options. In order to reduce minimap2's memory usage, -K option was set to 10M, as previously suggested (Gamaarachchi et al., 2019).
6. Alignments are subsequently filtered with in-house python scripts (included in the pipeline), and taxonomy and abundance tables are obtained.

A detailed explanation of the pipeline and the specific commands that were used can be found on Spaghetti's GitHub repository.[4]

Spaghetti data visualization and analysis module was mainly based on the phyloseq R package (v. 1.30.0) (McMurdie and Holmes, 2013). For alpha diversity tests, all the samples were rarefied to the lowest library size (50,051 reads/sample) to mitigate uneven sequencing depth. For beta diversity, Principal Coordinates Analysis (PCoA) were created using the Bray-Curtis dissimilarity metric and relative abundances. Heatmaps were produced with ampvis2 (v. 2.6.5) (Andersen et al., 2018). Custom figures were created using ggplot2 (v. 3.3.1). Plotly (v. 4.9.2.1) was used for producing interactive plots. Venn diagrams were obtained using an online tool.[5]

All the analyses were run on a MSI GF63 Thin 9SC-047XES laptop (CPU: Intel Corei7-9750H, 6 core, 12 threads; RAM: 16GB; SSD: 512 Gb; Graphics Card: GeForce GTX 1650).

Isolation of Bacterial Strains

Upon arrival at the laboratory, the samples were homogenized by mixing 1 g of the sample with 1 mL of sterile Phosphate Buffered Saline (PBS) and serial dilutions up to 10^{-7} were performed. Then, 50 μL of the 10^{-2} to 10^{-7} dilutions were spread on Petri dishes containing either TSA medium (composition in g/L: 15.0 tryptone, 5.0 soya peptone, 5.0 sodium chloride, 15.0 agar) or SSE/HD 1:10 medium (composition detailed on the DSMZ media database, medium number 1,426). In the case of the SSE/HD 1:10 medium, duplicates of each dilution were cultured, with one of the replicates being incubated under uninterrupted artificial light and the other replicate being incubated, together with the TSA plates, under natural light. All plates were incubated in oxygenic conditions and at room temperature.

Individual colonies were selected based on their color and morphology from the TSA and SSE/HD 1:10 plates incubated under natural light after 6, 11, 18, 30 and 35 days of incubation (**Supplementary Table 3**). These colonies were re-streaked on fresh culture medium to isolate them in pure culture. Most of the isolates were obtained from TSA medium and from the more concentrated dilutions (10^{-2} and 10^{-4}). Regarding the samples cultured on SSE/HD 1:10 under uninterrupted artificial light, these were removed from the artificial light after 4 weeks of incubation as they did not display any microbial growth. A few days after removal, different bacterial colonies started to grow. These colonies were re-streaked on fresh culture medium and isolated in pure culture. All pure strains were cryo-preserved in glycerol (20% glycerol in an over-night culture of the strain) at −80°C for further uses.

Molecular Identification of Isolates

A loopful of each isolate, grown on solid medium, was resuspended in 100 μL of sterile Milli-Q water and subjected to a rapid DNA extraction that consisted of three cycles of boiling and freeze-thawing. Then, a PCR was performed to amplify the 16S rRNA gene using the following universal primers: 8F (5′-AGAGTTTGATCCTGGCTCAG-3′) (Edwards et al., 1989) and 1492R (5′-GGTTACCTTGTTACGACTT-3′) (Stackebrandt and Liesack, 1993). The following conditions were used for PCR: initial denaturation (95°C; 5 min); amplification (24 cycles) comprising denaturation (94°C; 15 s), annealing (48°C; 15 s) and extension (72°C; 90 s); final extension (72°C; 5 min).

[2] https://github.com/rrwick/Porechop

[3] https://docs.qiime2.org/2020.8/data-resources/

[4] https://github.com/adlape95/Spaghetti

[5] http://bioinformatics.psb.ugent.be/webtools/Venn/

Amplicons were visualized by electrophoresis in a 1% agarose gel stained with GoldView DNA Safe Stain (UVAT Nerium Scientific, Valencia, Spain) (100 V, 30 min). Amplicons were precipitated overnight at −20°C in a mixture of isopropanol 1:1 (vol:vol) and potassium acetate 1:10 (vol:vol) (3M, pH 5). The next day, DNA was pelleted by centrifugations for 10 min at 12,000 rpm, then washed with 70% ethanol and resuspended in 15 μL of sterile Milli-Q water. Amplicons were tagged using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Carlsbad, CA, United States) and sent for Sanger sequencing of the partial 16S rRNA gene at the SCSIE (Serveis Centrals de Suport a la Investigació Experimental) of the University of Valencia (Spain), using the same universal primers as previously mentioned (8F and 1492R).

All resulting sequences were edited with UGENE v.33 (Okonechnikov et al., 2012) to remove low quality base calls, and taxonomic identification was performed using the BLASTn tool and the 16S ribosomal RNA sequences (Bacteria and Archaea) database (NCBI). Finally, clones were dereplicated using the BLASTn tool to compare each partial 16S rRNA sequence to the rest of strains belonging to the collection of microorganisms established in this project. Any strain displaying > 99.9% similarity to another strain in the collection and isolated from the same sample was considered to be a replicate and therefore discarded from the collection. This was performed to avoid an overestimation of the culturable diversity, as bacterial clones of the same species are not relevant for the microbial collection. The comparison between results from Nanopore sequencing and microbial culture collection was based on taxonomic information. Nanopore and Sanger 16S rRNA gene sequences were taxonomically classified independently, as described above. Then, the genus-level profiles were evaluated to find those taxa that had been identified by both approaches.

DATA AVAILABILITY STATEMENT

Nanopore raw sequences have been deposited in the NCBI (BioProject ID: PRJNA749463). Spaghetti is available on GitHub (https://github.com/adlape95/Spaghetti).

AUTHOR CONTRIBUTIONS

AL-P and HG-V performed the in-field experimental and bioinformatic work. AL-P and JP performed the data analysis, while HG-V and KT established, and characterized the microbial culture collection. AL-P prepared the figures. AL-P, MP, and CV designed the experiment and the expedition. All the authors wrote and revised the manuscript.

FUNDING

Financial support from the Spanish Government grant SETH (ref: RTI2018-095584-B-C41-42-43-44) is acknowledged. AL-P is a recipient of a Doctorado Industrial fellowship from the Spanish Ministerio de Ciencia, Innovación y Universidades (reference DI-17-09613).

ACKNOWLEDGMENTS

We are grateful to the Junta de Andalucía, and especially to the Consejería de Agricultura, Ganadería, Pesca y Desarrollo Sostenible (Delegación Territorial en Almería) for granting the permission for collecting samples in the Paraje Natural Desierto de Tabernas. We also want to thank all their colleagues from Darwin Bioprospecting Excellence who contributed to the development of MinION sequencing protocols and field applications. The name of the analysis pipeline ("Spaghetti") is a tribute to the cinematographic legacy of the Tabernas Desert, which was the setting chosen by Sergio Leone in the 60s to shoot several of his most famous films (e.g., Dollars Trilogy film series), thus establishing the Spaghetti Western genre.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.768240/full#supplementary-material

REFERENCES

Abed, R. M. M., Al Kharusi, S., Schramm, A., and Robinson, M. D. (2010). Bacterial diversity, pigments and nitrogen fixation of biological desert crusts from the Sultanate of Oman. *FEMS Microbiol. Ecol.* 72, 418–428. doi: 10.1111/j.1574-6941.2010.00854.x

Amin, A., Ahmed, I., Habib, N., Abbas, S., Hasan, F., Xiao, M., et al. (2016). *Microvirga pakistanensis* sp. nov., a novel bacterium isolated from desert soil of Cholistan, Pakistan. *Arch. Microbiol.* 198, 933–939. doi: 10.1007/s00203-016-1251-3

Andersen, K. S., Kirkegaard, R. H., Karst, S. M., and Albertsen, M. (2018). ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *BioRxiv* [Preprint]. doi: 10.1101/299537

Archer, S. D. J., Lee, K. C., Caruso, T., Maki, T., Lee, C. K., Cary, S. C., et al. (2019). Airborne microbial transport limitation to isolated Antarctic soil habitats. *Nat. Microbiol.* 4, 925–932. doi: 10.1038/s41564-019-0370-4

Benítez-Páez, A., and Sanz, Y. (2017). Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinIONTM portable nanopore sequencer. *Gigascience* 6:gix043. doi: 10.1093/gigascience/gix043

Bergmann, G. T., Bates, S. T., Eilers, K. G., Lauber, C. L., Caporaso, J. G., Walters, W. A., et al. (2011). The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol. Biochem.* 43, 1450–1455. doi: 10.1016/j.soilbio.2011.03.012

Bhattacharjee, A., Velickovic, D., Wietsma, T. W., Bell, S. L., Jansson, J. K., Hofmockel, K. S., et al. (2020). Visualizing microbial community dynamics *via* a controllable soil environment. *mSystems* 5:e00645-19. doi: 10.1128/msystems.00645-19

Buckley, D. H., Huangyutitham, V., Nelson, T. A., Rumberger, A., and Thies, J. E. (2006). Diversity of Planctomycetes in soil in relation to soil history and environmental heterogeneity. *Appl. Environ. Microbiol.* 72, 4522–4531. doi: 10.1128/AEM.00149-06

Bull, A. T., and Goodfellow, M. (2019). Dark, rare and inspirational microbial matter in the extremobiosphere: 16 000 m of bioprospecting campaigns. *Microbiology* 165, 1252–1264. doi: 10.1099/mic.0.000822

Burton, A. S., Stahl, S. E., John, K. K., Jain, M., Juul, S., Turner, D. J., et al. (2020). Off earth identification of bacterial populations using 16S rDNA nanopore sequencing. *Genes* 11, 1–10. doi: 10.3390/genes11010076

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B. R., et al. (2017). Nanopore DNA sequencing and genome assembly on the international space station. *Sci. Rep.* 7, 1–12. doi: 10.1038/s41598-017-18364-0

Chan, W. S., Au, C. H., Lam, H. Y., Wang, C. L. N., Ho, D. N. Y., Lam, Y. M., et al. (2020). Evaluation on the use of Nanopore sequencing for direct characterization of coronaviruses from respiratory specimens, and a study on emerging missense mutations in partial RdRP gene of SARS-CoV-2. *Virol. J.* 17, 1–13. doi: 10.1186/s12985-020-01454-3

Charalampous, T., Kay, G. L., Richardson, H., Aydin, A., Baldan, R., Jeanes, C., et al. (2019). Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat. Biotechnol.* 37, 783–792. doi: 10.1038/s41587-019-0156-5

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., da Costa, M. S., et al. (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* 68, 461–466. doi: 10.1099/ijsem.0.002516

Ciuffreda, L., Rodríguez-Pérez, H., and Flores, C. (2021). Nanopore sequencing and its application to the study of microbial communities. *Comput. Struct. Biotechnol. J.* 19, 1497–1511. doi: 10.1016/j.csbj.2021.02.020

Curry, K. D., Wang, Q., Nute, M. G., Tyshaieva, A., Reeves, E., Soriano, S., et al. (2021). Emu: species-level microbial community profiling for full-length nanopore 16S reads. *bioRxiv* [Prepirnt]. doi: 10.1101/2021.05.02.442339

Cuscó, A., Catozzi, C., Viñes, J., Sanchez, A., and Francino, O. (2018). Microbiota profiling with long amplicons using nanopore sequencing: full-length 16s rRNA gene and whole rrn operon [version 1; referees: 2 approved, 3 approved with reservations]. *F1000Research* 7:1755. doi: 10.12688/f1000research.16817.1

da Rocha, U. N., Cadillo-Quiroz, H., Karaoz, U., Rajeev, L., Klitgord, N., Dunn, S., et al. (2015). Isolation of a significant fraction of non-phototroph diversity from a desert biological soil crust. *Front. Microbiol.* 6:277. doi: 10.3389/fmicb.2015.00277

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34, 2666–2669. doi: 10.1093/bioinformatics/bty149

DeBruyn, J. M., Nixon, L. T., Fawaz, M. N., Johnson, A. M., and Radosevich, M. (2011). Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil. *Appl. Environ. Microbiol.* 77, 6295–6300. doi: 10.1128/AEM.05005-11

Deng, W., Yang, Y., Gao, P., Chen, H., Wen, W., and Sun, Q. (2016). Radiation-resistant *Micrococcus luteus* SC1204 and its proteomics change upon gamma irradiation. *Curr. Microbiol.* 72, 767–775. doi: 10.1007/s00284-016-1015-y

DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., et al. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci. U.S.A.* 112, 11060–11065. doi: 10.1073/pnas.1502875112

Dilthey, A. T., Jain, C., Koren, S., and Phillippy, A. M. (2019). Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat. Commun.* 10:3066. doi: 10.1038/s41467-019-10934-2

Edwards, A., Debbonaire, A. R., Nicholls, S. M., Rassner, S. M., Sattler, B., Cook, J. M., et al. (2019). In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. *bioRxiv* [Preprint]. doi: 10.1101/073965

Edwards, U., Rogall, T., Blöcker, H., Emde, M., and Böttger, E. C. (1989). Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res.* 17, 7843–7853. doi: 10.1093/nar/17.19.7843

Etemadifar, Z., Gholami, M., and Derikvand, P. (2016). UV-resistant bacteria with multiple-stress tolerance isolated from desert areas in Iran. *Geomicrobiol. J.* 33, 1–7. doi: 10.1080/01490451.2015.1063025

Gabani, P., and Singh, O. V. (2013). Radiation-resistant extremophiles and their potential in biotechnology and therapeutics. *Appl. Microbiol. Biotechnol.* 97, 993–1004. doi: 10.1007/s00253-012-4642-7

Gamaarachchi, H., Parameswaran, S., and Smith, M. A. (2019). Featherweight long read alignment using partitioned reference indexes. *Sci. Rep.* 9:4318. doi: 10.1038/s41598-019-40739-8

Golinska, P., Montero-Calasanz, M. C., Świecimska, M., Yaramis, A., Igual, J. M., Bull, A. T., et al. (2020). *Modestobacter excelsi* sp. nov., a novel actinobacterium isolated from a high altitude Atacama Desert soil. *Syst. Appl. Microbiol.* 43:126051. doi: 10.1016/j.syapm.2019.126051

Goordial, J., Altshuler, I., Hindson, K., Chan-Yam, K., Marcolefas, E., and Whyte, L. G. (2017). *in situ* field sequencing and life detection in remote (79°26'N) Canadian high arctic permafrost ice wedge microbial communities. *Front. Microbiol.* 8:2594. doi: 10.3389/fmicb.2017.02594

Gowers, G.-O. F., Vince, O., Charles, J.-H., Klarenberg, I., Ellis, T., and Edwards, A. (2019). Entirely off-grid and solar-powered DNA sequencing of microbial communities during an ice cap traverse expedition. *Genes* 10:902. doi: 10.3390/genes10110902

Hardegen, J., Latorre-Pérez, A., Vilanova, C., Günther, T., Porcar, M., Luschnig, O., et al. (2018). Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass. *Bioresour. Technol.* 265, 275–281. doi: 10.1016/j.biortech.2018.06.005

Heikema, A. P., Horst-Kreft, D., Boers, S. A., Jansen, R., Hiltemann, S. D., de Koning, W., et al. (2020). Comparison of illumina versus nanopore 16s rRNA gene sequencing of the human nasal microbiota. *Genes* 11:1105. doi: 10.3390/genes11091105

Holmes, A. J., Bowyer, J., Holley, M. P., O'Donoghue, M., Montgomery, M., and Gillings, M. R. (2000). Diverse, yet-to-be-cultured members of the Rubrobacter subdivision of the Actinobacteria are widespread in Australian arid soils. *FEMS Microbiol. Ecol.* 33, 111–120. doi: 10.1016/S0168-6496(00)00051-9

Hu, Q. W., Chu, X., Xiao, M., Li, C. T., Yan, Z. F., Hozzein, W. N., et al. (2016). *Arthrobacter deserti* sp. Nov., isolated from a desert soil sample. *Int. J. Syst. Evol. Microbiol.* 66, 2035–2040. doi: 10.1099/ijsem.0.000986

Johnson, S. S., Zaikova, E., Goerlitz, D. S., Bai, Y., and Tighe, S. W. (2017). Real-time DNA sequencing in the antarctic dry valleys using the Oxford nanopore sequencer. *J. Biomol. Tech.* 28, 2–7. doi: 10.7171/jbt.17-2801-2809

Kalam, S., Basu, A., Ahmad, I., Sayyed, R. Z., El-Enshasy, H. A., Dailin, D. J., et al. (2020). Recent understanding of soil acidobacteria and their ecological significance: a critical review. *Front. Microbiol.* 11:580024. doi: 10.3389/fmicb.2020.580024

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41, e1. doi: 10.1093/nar/gks808

Larsbrink, J., and McKee, L. S. (2020). Bacteroidetes bacteria in the soil: glycan acquisition, enzyme secretion, and gliding motility. *Adv. Appl. Microbiol.* 110, 63–98. doi: 10.1016/bs.aambs.2019.11.001

Latorre-Pérez, A., Pascual, J., Porcar, M., and Vilanova, C. (2021). A lab in the field: applications of real-time, *in situ* metagenomic sequencing. *Biol. Methods Protoc.* 5:baa016. doi: 10.1093/biomethods/bpaa016

Latorre-Pérez, A., Villalba-Bermell, P., Pascual, J., and Vilanova, C. (2020). Assembly methods for nanopore-based metagenomic sequencing: a comparative study. *Sci. Rep.* 10:5125. doi: 10.1038/s41598-020-70491-3

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* 75, 5111–5120. doi: 10.1128/AEM.00335-09

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Liang, Y., Tang, K., Wang, Y., Yuan, B., Tan, F., Feng, F., et al. (2019). *Hymenobacter crusticola* sp. nov., isolated from biological soil crust. *Int. J. Syst. Evol. Microbiol.* 69, 547–551. doi: 10.1099/ijsem.0.003196

Locey, K. J., and Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U.S.A.* 113, 5970–5975. doi: 10.1073/pnas.1521291113

Machado-de-Lima, N. M., Fernandes, V. M. C., Roush, D., Velasco Ayuso, S., Rigonato, J., Garcia-Pichel, F., et al. (2019). The compositionally distinct *Cyanobacterial biocrusts* from *Brazilian savanna* and their environmental drivers of community diversity. *Front. Microbiol.* 10:2798. doi: 10.3389/fmicb.2019.02798

Marijon, P., Chikhi, R., and Varré, J.-S. (2020). yacrd and fpa: upstream tools for long-read genome assembly. *Bioinformatics* 36, 3894–3896. doi: 10.1093/bioinformatics/btaa262

Matsuo, Y., Komiya, S., Yasumizu, Y., Yasuoka, Y., Mizushima, K., Takagi, T., et al. (2020). Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinIONTM nanopore sequencing confers species-level resolution. *BMC Microbiol.* 21:35. doi: 10.1186/s12866-021-02094-5

McHugh, A. J., Yap, M., Crispie, F., Feehily, C., Hill, C., and Cotter, P. D. (2021). Microbiome-based environmental monitoring of a dairy processing facility highlights the challenges associated with low microbial-load samples. *NPJ Sci. Food* 5, 1–13. doi: 10.1038/s41538-021-00087-2

McMurdie, P. J., and Holmes, S. (2013). Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e0061217. doi: 10.1371/journal.pone.0061217

Menegon, M., Cantaloni, C., Rodriguez-Prieto, A., Centomo, C., Abdelfattah, A., Rossato, M., et al. (2017). On site DNA barcoding by nanopore sequencing. *PLoS One* 12:e0184741. doi: 10.1371/journal.pone.0184741

Mitsuhashi, S., Kryukov, K., Nakagawa, S., Takeuchi, J. S., Shiraishi, Y., Asano, K., et al. (2017). A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Sci. Rep.* 7, 1–9. doi: 10.1038/s41598-017-05772-5

Molina-Menor, E., Gimeno-Valero, H., Pascual, J., Peretó, J., and Porcar, M. (2021). High culturable bacterial diversity from a european desert: the tabernas desert. *Front. Microbiol.* 11:583120. doi: 10.3389/fmicb.2020.583120

Montero-Calasanz, M. C., Göker, M., Broughton, W. J., Cattaneo, A., Favet, J., Pötter, G., et al. (2013). *Geodermatophilus tzadiensis* sp. nov., a UV radiation-resistant bacterium isolated from sand of the Saharan desert. *Syst. Appl. Microbiol.* 36, 177–182. doi: 10.1016/j.syapm.2012.12.005

Nygaard, A. B., Tunsjø, H. S., Meisal, R., and Charnock, C. (2020). A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. *Sci. Rep.* 10:3209. doi: 10.1038/s41598-020-59771-0

Okonechnikov, K., Golosova, O., Fursov, M., Varlamov, A., Vaskin, Y., Efremov, I., et al. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi: 10.1093/bioinformatics/bts091

Palatnick, A., Zhou, B., Ghedin, E., and Schatz, M. C. (2021). IGenomics: comprehensive DNA sequence analysis on your Smartphone. *Gigascience* 9:giaa138. doi: 10.1093/gigascience/giaa138

Paulino-Lima, I. G., Fujishima, K., Navarrete, J. U., Galante, D., Rodrigues, F., Azua-Bustos, A., et al. (2016). Extremely high UV-C radiation resistant microorganisms from desert environments with different manganese concentrations. *J. Photochem. Photobiol. B Biol.* 163, 327–336. doi: 10.1016/j.jphotobiol.2016.08.017

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L. A., et al. (2018). Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* 7:giy033. doi: 10.1093/gigascience/giy033

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, 590–596. doi: 10.1093/nar/gks1219

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232. doi: 10.1038/nature16996

Rainey, F. A., Ray, K., Ferreira, M., Gatz, B. Z., Nobre, M. F., Bagaley, D., et al. (2005). Extensive diversity of ionizing-radiation-resistant bacteria recovered from Sonoran Desert soil and description of nine new species of the genus *Deinococcus* obtained from a single soil sample. *Appl. Environ. Microbiol.* 71, 5225–5235. doi: 10.1128/AEM.71.9.5225-5235.2005

Rodríguez-Pérez, H., Ciuffreda, L., and Flores, C. (2021). NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics* 37, 1600–1601. doi: 10.1093/bioinformatics/btaa900

Santos, A., van Aerle, R., Barrientos, L., and Martinez-Urtaza, J. (2020). Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Comput. Struct. Biotechnol. J.* 18, 296–305. doi: 10.1016/j.csbj.2020.01.005

Sielaff, A. C., Urbaniak, C., Babu, G., Mohan, M., Stepanov, V. G., Tran, Q., et al. (2019). Characterization of the total and viable bacterial and fungal communities associated with the International Space Station surfaces. *Microbiome* 7:50. doi: 10.1186/s40168-019-0666-x

Spain, A. M., Krumholz, L. R., and Elshahed, M. S. (2009). Abundance, composition, diversity and novelty of soil *Proteobacteria*. *ISME J.* 3, 992–1000. doi: 10.1038/ismej.2009.43

Stackebrandt, E., and Liesack, W. (1993). "Nucleic acids and classification," in *Handbook of New Bacterial Systematics*, eds M. Goodfellow, and A. G. O'Donnell (London: Academic Press), 152–189.

Steen, A. D., Crits-Christoph, A., Carini, P., DeAngelis, K. M., Fierer, N., Lloyd, K. G., et al. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J.* 13, 3126–3130. doi: 10.1038/s41396-019-0484-y

Tanner, K. (2020). *Life Under the Sun: Microbial Ecology and Applications of The Solar Panel Microbiota*. Valencia: University of Valencia.

Tanner, K., Molina-Menor, E., Latorre-Pérez, A., Vidal-Verdú, À, Vilanova, C., Peretó, J., et al. (2020). Extremophilic microbial communities on photovoltaic panel surfaces: a two-year study. *Microb. Biotechnol.* 13, 1819–1830. doi: 10.1111/1751-7915.13620

Tytgat, O., Gansemans, Y., Weymaere, J., Rubben, K., Deforce, D., and Van Nieuwerburgh, F. (2020). Nanopore sequencing of a forensic STR multiplex reveals loci suitable for single-contributor STR profiling. *Genes* 11:381. doi: 10.3390/genes11040381

Urban, L., Holzer, A., Baronas, J. J., Hall, M. B., Braeuninger-Weimer, P., Scherm, M. J., et al. (2021). Freshwater monitoring by nanopore sequencing. *eLife* 10:e61504. doi: 10.7554/eLife.61504

Vasiljevic, N., Lim, M., Humble, E., Seah, A., Kratzer, A., Morf, N. V., et al. (2021). Developmental validation of Oxford Nanopore Technology MinION sequence data and the NGSpeciesID bioinformatic pipeline for forensic genetic species identification. *Forensic Sci. Int. Genet.* 53:102493. doi: 10.1016/j.fsigen.2021.102493

Watson, M., and Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* 37, 124–126. doi: 10.1038/s41587-018-0004-z

Williams, L., Loewen-Schneider, K., Maier, S., and Büdel, B. (2016). Cyanobacterial diversity of western European biological soil crusts along a latitudinal gradient. *FEMS Microbiol. Ecol.* 92:fiw157. doi: 10.1093/femsec/fiw157

Winand, R., Bogaerts, B., Hoffman, S., Lefevre, L., Delvoye, M., Van Braekel, J., et al. (2020). Targeting the 16s rRNA gene for bacterial identification in complex mixed samples: comparative evaluation of second (illumina) and third (oxford nanopore technologies) generation sequencing technologies. *Int. J. Mol. Sci.* 21:298. doi: 10.3390/ijms21010298

Wübbeler, J. H., Oppermann-Sanio, F. B., Ockenfels, A., Röttig, A., Osthaar-Ebker, A., Verbarg, S., et al. (2017). *Sphingomonas jeddahensis* sp. nov., isolated from Saudi Arabian desert soil. *Int. J. Syst. Evol. Microbiol.* 67, 4057–4063. doi: 10.1099/ijsem.0.002249

Xu, Z., Mai, Y., Liu, D., He, W., Lin, X., Xu, C., et al. (2020). Fast-Bonito: a faster basecaller for nanopore sequencing. *bioRxiv* [Preprint]. doi: 10.1101/2020.10.08.318535

Yu, L. Z. H., Luo, X. S., Liu, M., and Huang, Q. (2015). Diversity of ionizing radiation-resistant bacteria obtained from the Taklimakan Desert. *J. Basic Microbiol.* 55, 135–140. doi: 10.1002/jobm.201300390

Zhang, B., Wu, X., Tai, X., Sun, L., Wu, M., Zhang, W., et al. (2019). Variation in actinobacterial community composition and potential function in different soil ecosystems belonging to the arid heihe river basin of Northwest China. *Front. Microbiol.* 10:2209. doi: 10.3389/fmicb.2019.02209

Zhang, Q., Liu, C., Tang, Y., Zhou, G., Shen, P., Fang, C., et al. (2007). *Hymenobacter xinjiangensis* sp. nov., a radiation-resistant bacterium isolated from the desert of Xinjiang, China. *Int. J. Syst. Evol. Microbiol.* 57, 1752–1756. doi: 10.1099/ijs.0.65033-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch

OPEN

Assembly methods for nanopore-based metagenomic sequencing: a comparative study

Adriel Latorre-Pérez[1,2], Pascual Villalba-Bermell[1,2], Javier Pascual[1] & Cristina Vilanova[1]✉

Metagenomic sequencing has allowed for the recovery of previously unexplored microbial genomes. Whereas short-read sequencing platforms often result in highly fragmented metagenomes, nanopore-based sequencers could lead to more contiguous assemblies due to their potential to generate long reads. Nevertheless, there is a lack of updated and systematic studies evaluating the performance of different assembly tools on nanopore data. In this study, we have benchmarked the ability of different assemblers to reconstruct two different commercially-available mock communities that have been sequenced using Oxford Nanopore Technologies platforms. Among the tested tools, only metaFlye, Raven, and Canu performed well in all the datasets. These tools retrieved highly contiguous genomes (or even complete genomes) directly from the metagenomic data. Despite the intrinsic high error of nanopore sequencing, final assemblies reached high accuracy (~ 99.5 to 99.8% of consensus accuracy). Polishing strategies demonstrated to be necessary for reducing the number of indels, and this had an impact on the prediction of biosynthetic gene clusters. Correction with high quality short reads did not always result in higher quality draft assemblies. Overall, nanopore metagenomic sequencing data-adapted to MinION's current output-proved sufficient for assembling and characterizing low-complexity microbial communities.

Background

Metagenomic sequencing has revolutionized the way we study and characterize microbial communities. This culture-independent technique based on shotgun sequencing has been applied in a broad range of biological fields, ranging from microbial ecology[1] to evolution[2], or even clinical microbiology[3]. In recent years, metagenomics has also become a powerful tool for recovering individual genomes directly from complex microbiomes[2,4,5], leading to the identification and description of new- and mostly unculturable-taxa with meaningful implications[6].

Illumina has been the most widely used platform for metagenomic studies. Illumina reads are characterized by their short length (75–300 bp) and high accuracy (~ 0.1% of basecalling errors)[7]. When performing de novo assemblies, Illumina sequences often result in highly fragmented genomes, even when sequencing pure cultures[8,9]. This is a consequence of the inability to correctly assemble genomic regions containing repetitive elements that are longer than the read length[9]. This fragmentation problem is magnified when handling metagenomic sequences due to the existence of intergenomic repeats that are shared by more than one taxon present in the microbial community[10]. It has to be noted that microbial communities often contain related species or sub-species in different-and unknown- abundances, resulting in extensive intergenomic overlaps that can hinder the assembly process[11,12].

Third generation sequencing platforms have recently emerged as a solution to resolve ambiguous repetitive regions and to improve genome contiguity. Despite the considerable error associated to these technologies (~ 5 to 15% of basecalling errors)[13,14], their ability to produce long reads (up to 10–12 kb of mean read length)[7,15] has allowed them to generate genomes with a high degree of completeness[16,17]. Currently, the most widely-used third generation technologies are Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), both based on single molecule sequencing, and therefore, PCR-free. PacBio was the first long-read technology established in the market[18]. However, PacBio instruments require particular operation conditions and large capital investments[19]. On the other hand, ONT platforms are becoming increasingly popular among researchers, especially in the case of MinION sequencers. Although the cost of GridION and PromethION devices is also notable

[1]Darwin Bioprospecting Excellence S.L., Paterna, Spain. [2]These authors contributed equally: Adriel Latorre-Pérez and Pascual Villalba-Bermell. ✉email: cristina@darwinbioprospecting.com

Species	Abbreviations	Estimated size (Mbp)	Composition even (CS) (%)	Composition log (CSII) (%)
Bacillus subtilis	*B. subtilis*	4.134	12.00	0.89
Cryptococcus neoformans	*C. neoformans*	18.599	2.00	0.00089
Enterococcus faecalis	*En. faecalis*	2.965	12.00	0.00089
Escherichia coli	*E. coli*	5.140	12.00	0.089
Lactobacillus fermentum	*L. fermentum*	2.012	12.00	0.0089
Listeria monocytogenes	*Li. monocytogenes*	3.008	12.00	89.1
Pseudomonas aeruginosa	*P. aeruginosa*	6.592	12.00	8.9
Saccharomyces cerevisiae	*S. cerevisiae*	11.864	2.00	0.89
Salmonella enterica	*Sa. enterica*	4.781	12.00	0.089
Staphylococcus aureus	*St. aureus*	2.838	12.00	0.000089

Table 1. Description of the microorganisms comprising the ZymoBIOMICS mock communities and their theoretical composition.

(~ 50,000$ to 170,000$), MinION is a cost-effective (~ 1,000$), portable sequencing platform, which enables real-time analysis pipelines[20]. This platform has been broadly applied over the last few years, due to its suitability for in-field and clinical studies[21,22], but also for sequencing complete prokaryotic and eukaryotic genomes[17,23–25], and characterizing microbial communities[26,27].

Benchmarking is the usual way to evaluate genomic methodologies (i.e. DNA extraction, library preparations, etc.) and bioinformatic tools. In the metagenomic context, benchmarking studies are frequently based on mock communities. A mock community is an artificial microbial community in which the abundance of all the microorganisms is known[28]. Mock communities can be generated in silico[29] or experimentally, as a mixture of defined DNA proportions. For de novo assemblies, a great effort has been made in order to benchmark all the available tools and methodologies suitable for studying microbial ecosystems via Illumina shotgun sequencing[12,30,31]. Nevertheless, due to the highly dynamic development of new software applicable to ONT platforms, we found that the few evaluation studies that have been focused to date on nanopore-based metagenomic assembly did not cover the current spectrum of available assemblers[32–34].

In the present study, we have used the data generated by Nicholls et al.[15] to comprehensively assess the current state-of-art of *de novo* assembly tools suitable for nanopore-based, metagenomic sequencing. Original data was generated through metagenomic sequencing of two microbial communites (ZymoBIOMICS Microbial Community Standards CS and CSII) with both GridION and PromethION platforms. Overall, this work demonstrates the suitability of using nanopore sequencing exclusively for assembling low-complexity microbial communities, and paves the way towards the standardization of bioinformatic pipelines for long-read sequencing data.

Methods

Dataset description. Benchmarking datasets were extracted from Nicholls et al.[15] (PRJEB29504), and consisted of high-coverage sequencing of two individual mock communities (ZymoBIOMICS Microbial Community Standards CS Even ZRC190633 and CSII Log ZRC190842) with both GridION and PromethION platforms. The mock communities contained the same species (eight bacteria; two yeasts), but differed in the expected proportion for each microorganism. CS mock community has a homogeneous distribution of microorganisms (12% for each bacteria and 2% for the yeasts), while the species present in CSII are distributed on a logarithmic scale, with relative abundances ranging from 89.1 to 0.000089% (Table 1). Following the nomenclature from Nicholls et al.[15], we have used the terms "Even" when referring to CS mock community, and "Log" when referring to CSII.

Nicholls et al.[15] yielded ~ 14 Gbp of data on a single GridION flowcell (48 h of sequencing) and ~ 152 Gbp on the PromethION platform (64 h of sequencing). In order to reduce the computational effort, we performed an initial subsampling of this data. In particular, GridION and PromethION datasets were subsampled at two different sequencing depths (3 Gbp and 6 Gbp) to recreate MinION runs with different outputs, and the yield matched the output described in recent shotgun sequencing experiments based on MinION[9,34–39]. Subsampling was performed by selecting the top lines of the FASTQ files. Nevertheless, the most promising tools were further tested on the original GridION data to check their computational demands and general performance. All the datasets were trimmed with porechop (https://github.com/rrwick/Porechop; v. 0.2.4) in order to remove adapters from read ends and split sequences with internal adapters.

De novo assembly. As first proposed by Lindgreen et al.[40], the tools selected for the present benchmarking were required to meet the following criteria:

- The tool should be freely available.
- The tool should have a suitable user guide, both for installation and usage.
- The tool should have been extensively used or show potential to become widely used.

In our study, a total of three widely used metagenomic short-read assemblers and ten long-read tools (or different versions of the same tool) were taken into consideration. Nevertheless, it was not possible to install and/or run all the software due to different reasons (Supplementary Table S1). The commands used for running each assembler are provided in Supplementary Table S2. It is worth highlighting that tools were run with default parameters when no metagenomic configuration was explicitly recommended in the user guide.

Reference genomes. All the species included in the mock community had an available reference genome sequenced with a combination of Illumina and nanopore reads (available at https://doi.org/10.5281/zenodo.3935737). These assemblies provided by Zymo Research Corporation (Irvine, CA, USA) consisted of eight complete genomes for the bacterial strains, and two draft genomes for the yeasts.

Nicholls et al.[15] sequenced and assembled each genome again from pure cultures using Illumina reads only. In the present work, however, ZymoBIOMICS genomes were used as a reference for carrying out the comparative analyses, due to their higher level of completeness. Although these reference genomes cannot be considered as "gold standards", Goldstein et al.[9] demonstrated that the error profile obtained through hybrid assembly (ONT + Illumina MiSeq) was similar to the one obtained with MiSeq-only assembly, but the former resulted in higher contiguity. Reference genomes were gathered in a single multi-FASTA file to create a single-reference metagenome.

Evaluation of the assemblies. All the assemblers were run on the same desktop computer (CPU: AMD RYZEN 7 1700X 3.4GHZ; Cores: 8; Threads: 16; RAM: Corsair Vengeance 64 GB; SSD: Samsung 860 EVO Basic SSD 500 GB) working under Ubuntu 18.04 operative system. The time required by each tool to perform the assembly was measured with the built-in bash version of the "time" command.

Completeness and contiguity of de novo assemblies were first evaluated via QUAST (v. 5.0.2)[41]. MetaQUAST (v. 5.0.2)[42] was used for obtaining assembly statistics based on the alignment of the generated contigs against the reference genomes. Only contigs longer than 500 bp and with > X10 coverage were selected for calculating the general statistics. MetaQUAST failed to run with some draft metagenomes and, for that reason, minimap2 (v. 2.15)[43] was used instead to align the assemblies to the reference metagenome. Then, the percentage of metagenome covered by the draft assemblies was calculated using the 'pileup.sh' script from BBTools suite (http://sourceforge.net/projects/bbmap/).

The resulting assemblies were further evaluated in order to determine their error profile. Due to the lack of a standard methodology, the presence of SNPs and indels was analyzed using two different strategies. The first one consisted of the alignment of the contigs against the reference metagenome via minimap2. BAM files were then analysed using bcftools (https://samtools.github.io/bcftools/; v. 1.9) and the in-house script 'indels_and_snps.py' (https://doi.org/10.5281/zenodo.3935763) was applied to quantify the variants. The second strategy was based on the use of MuMmer4 (https://sourceforge.net/projects/mummer/files/; v. 3.23). This tool was employed to align the draft assemblies to the reference metagenome. Then, the script 'count_SNPS_indels.pl' from Goldstein et al.[9] was used to calculate the final number of SNPs and indels. In both strategies, the number of variants was normalized to the total assembly size of each metagenome.

Biosynthetic gene clusters (BGCs) are usually composed of repetitive genetic structures that are hard to assemble with short reads, and with long-read technologies being therefore more suitable to overcome this issue. However, BGCs are also very sensitive to frameshift errors, which have been reported to frequently occur in nanopore data[9]. For that reason, AntiSMASH web service (v. 5.0)[44] was used to compare the performance on BGC prediction BGCs number and profile among the different assembly tools.

Assembly polishing. Draft assemblies (Even GridION 6 Gbp dataset) were further polished with Racon [45] and Medaka (https://nanoporetech.github.io/medaka/), using the commands specified in Supplementary Table S2. As the Medaka model for the specific version of Guppy (v2.2.2 GPU basecaller) originally used for basecalling the data was not available, we used the Medaka default model (r941_min_high_g351) for polishing. ONT or Illumina reads were used for iteratively running 4 rounds of Racon. Polishing was carried out using the same ONT input reads as those used for assembling each dataset, whereas Illumina reads (MiSeq plataform) were retrieved from the shotgun metagenomic sequencing data available for the Even mock community (ERR2984773)[15]. Only the draft assemblies corrected with ONT reads were further polished with Medaka, again using the original ONT sequences as input. Indels and SNPs were evaluated after each polishing step using the MumMer-based strategy, as detailed above.

Results

Subsampling. In the present study, the data released by Nicholls et al.[15] (ultra-deep sequencing of two different mock communities using GridION and PromethION platforms) was used in order to study the suitability of nanopore sequencing to characterize low complex microbial communities. The mock communities were composed by the same ten microorganisms, but in different proportions (Table 1). With the aim of reducing the computational resources needed for the first screening of the selected assemblers, the GridION and PromethION datasets were subsampled to obtain an output comparable with recent genomic or metagenomic studies based on MinION (approximately 3 Gbp and 6 Gbp)[9,34–39]. In general, mean read length remained the same in the subsampled datasets in comparison to the original sequencing data[15]. However, read quality was higher in the subsampled dataset. This fact suggested a bias towards higher qualities at the start of the run, since subsampling was carried out by selecting the top reads of the original files (Table 2). In fact, the bottom reads which are acquired later in the sequencing run displayed the same quality than the whole dataset.

Metagenome assembly. From the selected tools, we were able to correctly install and run nine out of the ten long-read assemblers, and two out of the three short-read assemblers (Supplementary Table S1). In total,

Dataset name	Original dataset				New dataset				
	Gbp	Number of reads	Mean read length	Mean read quality	Gbp	Number of reads	Mean read length	Mean read quality	SRA accession number
Even GridION	14.007	3,491,078.0	4,012.3	8.4	3.042	747,682.0	4,069.5	8.9	SRX6817349
Log GridION	16.032	3,667,007.0	4,372.0	8.0	3.053	685,926.0	4,451.0	8.7	SRX6817351
Even PromethION	146.291	36,527,376.0	4,005.0	7.3	2.979	748,367.0	3,981.0	8.2	SRX6817353
Log PromethION	148.028	35,118,078.0	4,215.2	7.6	2.990	711,524.0	4,203.3	8.3	SRX6817355
Even GridION	14.007	3,491,078.0	4,012.3	8.4	6.092	1,495,377.0	4,073.9	8.8	SRX6817350
Log GridION	16.032	3,667,007.0	4,372.0	8.0	6.094	1,371,820.0	4,442.4	8.5	SRX6817352
Even PromethION	146.291	36,527,376.0	4,005.0	7.3	5.970	1,496,919.0	3,988.8	8.2	SRX6817354
Log PromethION	148.028	35,118,078.0	4,215.2	7.6	5.956	1,422,918.0	4,185.8	8.2	SRX6817356

Table 2. Sequencing statistics for the original and the subsampled datasets.

74 assemblies were generated, 40 for the Even mock community and 34 for the Log community. Six assemblies could not be completed because miniasm and Pomoxis failed to run with the 6 Gbp Log datasets, whereas Unicycler failed to run with the 3 Gbp Log datasets. The total size of each draft assembly and the fraction of metagenome recovered from the reference genomes were evaluated for the Even datasets in order to obtain a first view of the general tool performance.

Overall, long-read assemblers resulted in a total assembly size closer to the theoretical size, and also recovered a larger metagenome fraction, with some exceptions (Fig. 1). Nevertheless, large differences were detected for both metrics among the assemblers. All the assemblers were far from recovering the totality of the metagenome, both in the 3 Gbp and the 6 Gbp datasets (Fig. 1A). It must be noted that metaQUAST and minimap2 results were consistent for the long-read assemblers, but not for the short-read assemblers, where minimap2 metric was significantly higher (Fig. 1B). MetaFlye (both versions) yielded the best assemblies in terms of total metagenome size and metagenome recovery except for the minimap2 metric, followed by Pomoxis, Canu and Raven (previously known as Ra). Interestingly, assembly pipelines based on the miniasm algorithm (Pomoxis, Unycicler, and miniasm itself) presented huge variations in their performance. Unicycler and miniasm performed relatively well for the 3 Gbp dataset, but when using 6 Gb, the final assembly did not improve significantly in the case of miniasm, and the general performance was highly reduced for Unicycler. This is in contrast to Pomoxis, which produced the second most complete assemblies with both dataset sizes. Although based on miniasm, it is worth highlighting that Unicycler's pipeline is designed for single isolate assembly, so reduced performance was expected for metagenomic studies. Finally, Redbean (previously known as wtdbg2) and Shasta resulted in poor assembly performance in comparison to the other long-read tools.

MetaQUAST was used for further evaluating the degree of completeness of each individual draft genome (Fig. 2). As expected, yeast genomes were generally less recovered than bacterial ones, due to their lower abundance (2%) and higher size, explaining the low metagenome fraction generally recovered by all the assemblers (Fig. 1). In fact, the maximum average recovery fraction for the bacterial genomes was 99.92% (Supplementary Fig. S1). Minia and Megahit were not able to recover any single genome with high completeness (> 95% of genome coverage) in any dataset. For the 3 Gbp dataset, metaFlye (both versions) and Unicycler recovered the eight bacterial genomes with a high completeness level (> 98.6%), while Pomoxis achieved lower recovery fractions for two genomes (~ 96.9 to 97.4%). Raven and Canu resulted in reduced recovery percentages, but still retrieved all the prokaryotic genomes with a mean covered fraction greater than 85% and 87%, respectively. Redbean and Shasta achieved particularly low fractions of genome recovery.

For the 6 Gbp dataset, Unicycler performance decreased substantially as noted in Fig. 1, while Canu, Pomoxis, Raven and metaFlye achieved similar or better results. In general, metaFlye displayed the best performance on both dataset sizes in terms of genome recovery, closely followed by Pomoxis. This trend was also observed when analyzing the proportion of yeast genomes recovered by each tool. In this context, it is important to highlight that metaFlye's ability to recover eukaryotic genomes was reduced when using metaFlye v2.7. This is due to the lower number of missassemblies retrieved by this metaFlye version, indicating that the reduced fraction of genome recovery is compensated with more reliable assemblies (Supplementary Fig. S2).

These results were confirmed when analyzing the Log mock community (Supplementary Fig. S3). Canu, metaFlye, Raven and Pomoxis were able to recover *Listeria monocytogenes* and *Pseudomonas aeruginosa* genomes (89.1% and 8.9% of total genomic DNA in the Log mock community, respectively) with a level of completeness higher than 99%. These assemblers also recovered a significant fraction of *Bacillus subtilis* (0.89% of total genomic DNA in the Log mock community). In fact, Raven was able to reconstruct > 99% of its genome using the 6 Gbp datasets, whereas metaFlye recovered ~ 98%. In this case, both tools outperformed Canu. Nevertheless, Raven did not recover a significant fraction of *Saccharomyces cerevisiae*, whereas Canu and metaFlye did (> 8%). Pomoxis worked correctly when using the 3 Gbp datasets, but failed to run with both 6 Gbp files. The other tools based on the miniasm algorithm also failed to run the 3 Gbp (Unicycler) and/or 6 Gbp datasets (miniasm). In all cases, the error was related to memory usage and accession (segmentation violation), and could not be solved. Nevertheless, using a computer with more RAM would help to easily overcome this problem. Shasta, RedBean, Minia and Megahit performed poorly in comparison to the other tools (Supplementary Fig. S3). It has to be noted that Shasta and RedBean were not originally designed to work with metagenomic data, which could result in problems to handle uneven coverages.

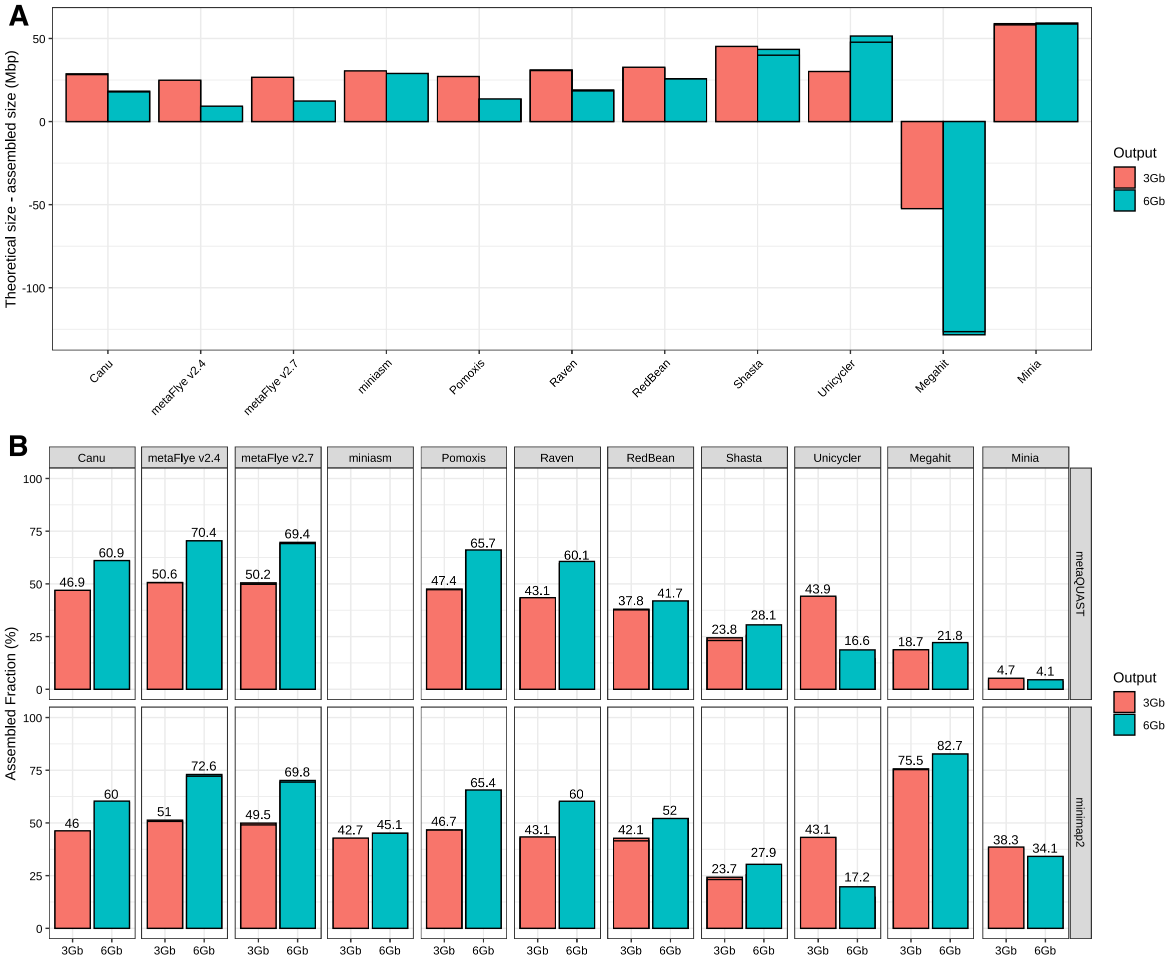


Figure 1. Evaluation of metagenome assembly size corresponding to each tested tool for the subsampled Even datasets. (**A**) Total assembled size of draft assemblies with respect to the total size of the reference metagenome; (**B**) fraction of the reference metagenome covered by the draft assembly, calculated by two different methods: metaQUAST (top panel) and minimap2 + BBTools (bottom panel).

Regarding the time consumed by each tool, Shasta was the fastest assembler (Fig. 3A). This tool was able to assemble the 6 Gbp datasets in only 285 s, approximately. RedBean and miniasm were the second and third most fast software, followed by Raven (1.5–1.9 times faster than metaFlye v2.7). MetaFlye was 1.4–1.7 times faster than Pomoxis, and 3.8–5.5 times faster than Canu, which proved to be the slowest tool. These trends were also found in the Log mock community (Supplementary Fig. S4), where Canu spent up to 22 h reconstructing a draft metagenome assembly from the 6 Gbp datasets. In this case, Raven was faster than metaFlye v2.7 for the 3 Gbp datasets, but not for the 6 Gbp ones.

General metagenome statistics (N50, L50, and number of contigs) were evaluated using QUAST (Fig. 3; Supplementary Table S3). It has to be stressed that the comparisons based on these metrics are difficult to analyze due to the large variation in the general performance among the different assemblers. For instance, Shasta resulted in the highest N50 and the lowest L50 values for the 6 Gbp dataset, but this tool was able to cover less than 35% of the metagenome. In fact, the total assembly size for Shasta was approximately 18–21 Mbp, in comparison to the 49–53 Mbp assembled by metaFlye.

As expected, short-read assemblers did not perform well with nanopore data, resulting in thousands (Minia), or even hundreds of thousands of contigs (Megahit). Interestingly, long-read assemblers resulted in more fragmented draft genomes when using the 6 Gbp datasets. Except for Shasta, the other long-read assemblers also reduced their N50 and increased their L50 and number of contigs score when using 6 Gbp. Goldstein et al.[9] demonstrated that Canu assemblies improved with higher coverage when assembling bacterial isolates. This fact suggests that the loss of contiguity detected may be a direct consequence of a higher recovery rate of yeast genomes, which might be more fragmented. Indeed, assembly statistics of the Canu draft assemblies remained almost the same for the bacterial species when using 3 or 6 Gbp (Supplementary Table S4). Finally, metaFlye and Raven resulted in a more contiguous assembly with higher N50 and lower L50 in comparison to the other best

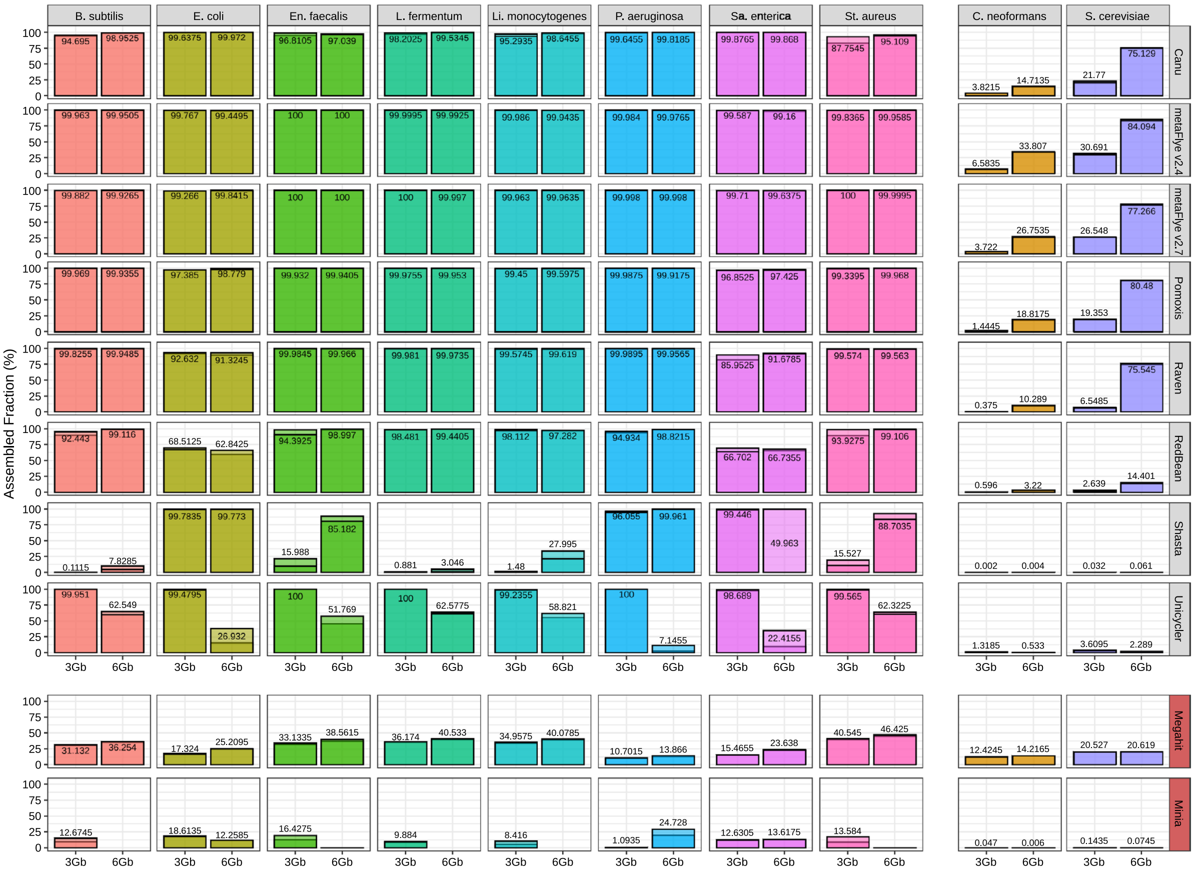


Figure 2. Fraction of the genome covered by the draft assemblies obtained using each tool, and for each individual microorganism (subsampled Even datasets). Miniasm assemblies are not shown, since it was not possible to evaluate them with metaQUAST.

performing tools (Canu and Pomoxis), for both 3 and 6 Gbp datasets (Fig. 3; Supplementary Table S3). Remarkably, metaFlye v2.7 yielded slightly better results than metaFlye v2.4 (Fig. 3B–D), and required less time (Fig. 3A).

ONT hardware, protocols and software are in constant development, leading to large improvements in short periods of time. Recently, an optimized DNA extraction and purification methodology has allowed to reach an average yield of ~ 15.9 Gbp per flowcell[46]. For that reason, we decided to run the most promising assemblers directly on GridION's original data (Even mock community; 14 Gbp). RedBean was included because of its computational efficiency, which is a key factor for the analysis of deeply sequenced microbiomes. Results were similar to those obtained for the 3 and 6 Gbp (Fig. 4). Canu recovered the highest proportion of bacterial genomes, closely followed by metaFlye. Raven, once again, displayed problems when reconstructing the whole *Escherichia coli* and *Salmonella enterica* genomes, an issue also detected for RedBean in a more notable way. MetaFlye and Raven achieved a better recovery ratio than Canu for the yeast genomes. Overall, metaFlye genomes were more complete but less contiguous than the Raven draft assemblies, which presented a lower number of contigs for all the species with the exception of *E. coli* and *S. enterica* (Fig. 4B). This trend was also observed for the Log datasets (Supplementary Fig. S4). Remarkably, Raven was able to assemble two bacterial genomes in only one contig (*Lactobacillus fermentum* and *P. aeruginosa*), and retrieved four additional genomes in only 2–3 contigs. Finally, it was not possible to run Pomoxis on this dataset because of the unsolvable error previously described.

Assembly accuracy. Sequencing errors are the biggest drawback of third generation sequencing platforms. These errors can reach the final assemblies, resulting in lower quality draft genomes. In order to evaluate how the different assemblers handle the specific error profile of ONT platforms, we analysed the total number of SNPs and indels present in each draft metagenome. As described in Methods, two different and complementary strategies were used to quantify these types of errors: (1) minimap2 + bcftools, and (2) MuMmer (Fig. 5). Both strategies relied on the alignment of the draft assemblies to the reference metagenome, composed by a mix of all the complete genomes of each strain present in the mock community.

Results were not fully consistent between the two methodologies, especially for the indels estimation, but they still showed similar trends. All the long-read assemblers retrieved draft metagenomes with an average similarity higher than ~ 98.9%, with the exception of miniasm, which resulted in an approximate accuracy of only 96%.

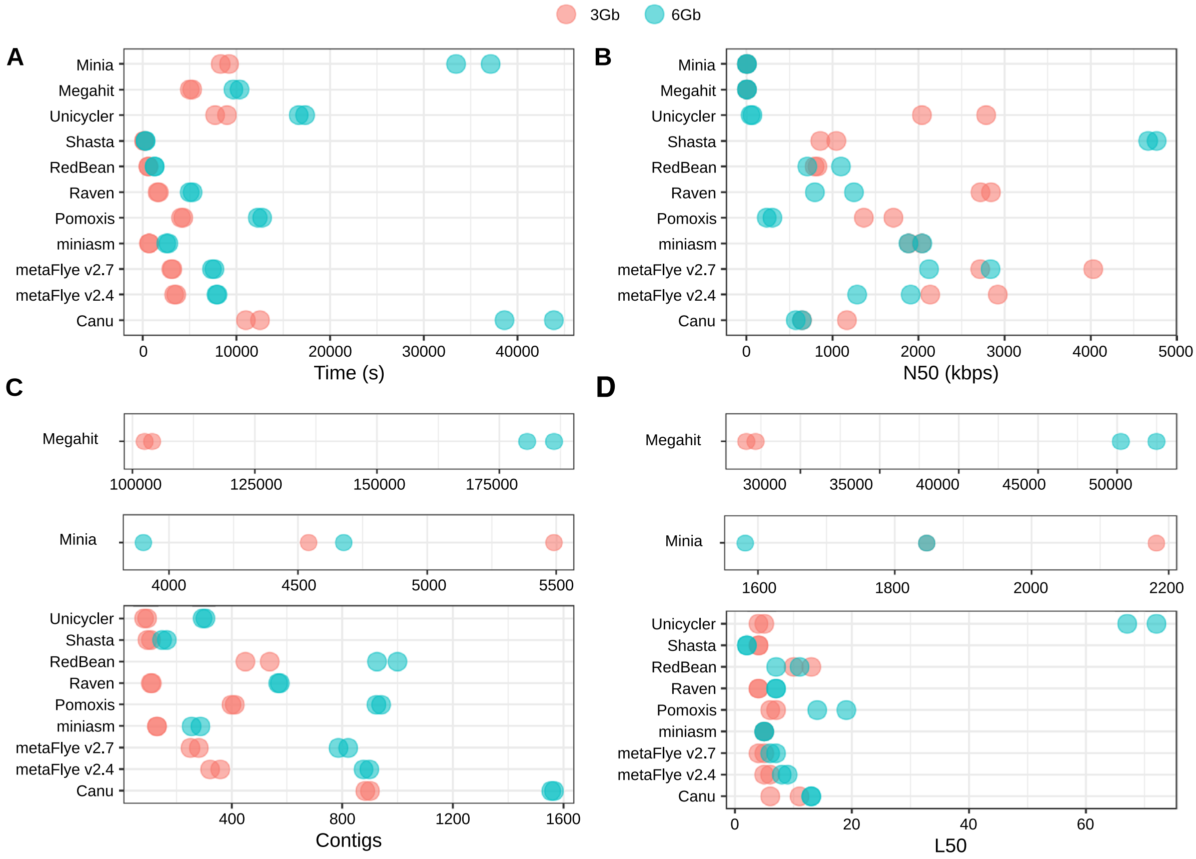


Figure 3. General assembly performance of each tool for the subsampled Even datasets. (**A**) Run time; (**B**) N50; (**C**) number of contigs; (**D**) L50.

This low accuracy could explain the inability of metaQUAST to evaluate miniasm assemblies. It has to be noted that the other pipelines based on miniasm, Pomoxis and Unicycler, incorporated several rounds of polishing via Racon[45], which substantially reduced the number of SNPs and indels in the final draft assembly (see below).

Canu displayed a higher percentage of similarity for both methodologies and datasets, followed by Unicycler for the 3 Gbp dataset, and Shasta for the 6 Gbp one. Pomoxis, metaFlye, and Raven presented similarities over 99.5%. In the case of the indel profile, Unicycler and metaFlye v2.7 clearly outperformed Canu. Raven and Pomoxis also achieved a better indel ratio than Canu, except for the 6 Gbp dataset and the bcftools metric. Redbean, miniasm, and Shasta results were inconsistent between the two methodologies tested (Fig. 5).

Biosynthetic gene cluster prediction. Gene prediction is highly affected by genome contiguity, completeness and accuracy. BGCs are especially influenced by these factors, since they are usually found in repetitive regions which are often poorly assembled. AntiSMASH was used to assess the number of clusters found in the draft assemblies retrieved by each tool in comparison to the reference metagenome with the aim of evaluating BGC prediction on nanopore-based metagenomic assemblies (Fig. 6). As expected, none of the tools recovered the entire BCG profile, since metagenomes were not completely reconstructed (Fig. 1). Using the entire GridION dataset (14 Gbp) did not improve the number of BCGs recovered (Supplementary Table S5). Overall, when considering the total number of BGCs predicted and the similarity of the obtained profile compared to the reference profile, Raven displayed the best performance for both 3 Gbp datasets, whereas metaFlye v2.7 displayed the best performance for the 6 Gbp datasets. Pomoxis also achieved good predictions, outperforming Canu. All the predicted profiles presented an enrichment in lasso peptides (ribosomally-synthesized short peptides), which were not present in the reference profile. To further study this phenomenon, lasso peptides predicted by the different tools were searched using BLAST against the BGCs predicted in the reference metagenome. No hits were found, suggesting that these results might be prediction artifacts mainly caused by indels, which are probably introducing frameshift errors, and artificially increasing the number of short peptides being predicted (i.e. lasso peptides). In fact, metaFlye v2.7, which had a significantly lower indel ratio, retrieved fewer lasso peptides than metaFlye 2.4 (Fig. 5). We also corrected Pomoxis assemblies with Medaka, leading to a lower indel ratio (see the following section). Lasso peptides were not detected in Pomoxis + Medaka assemblies, highlighting the importance of indel correction for functional prediction (Supplementary Fig. S5).

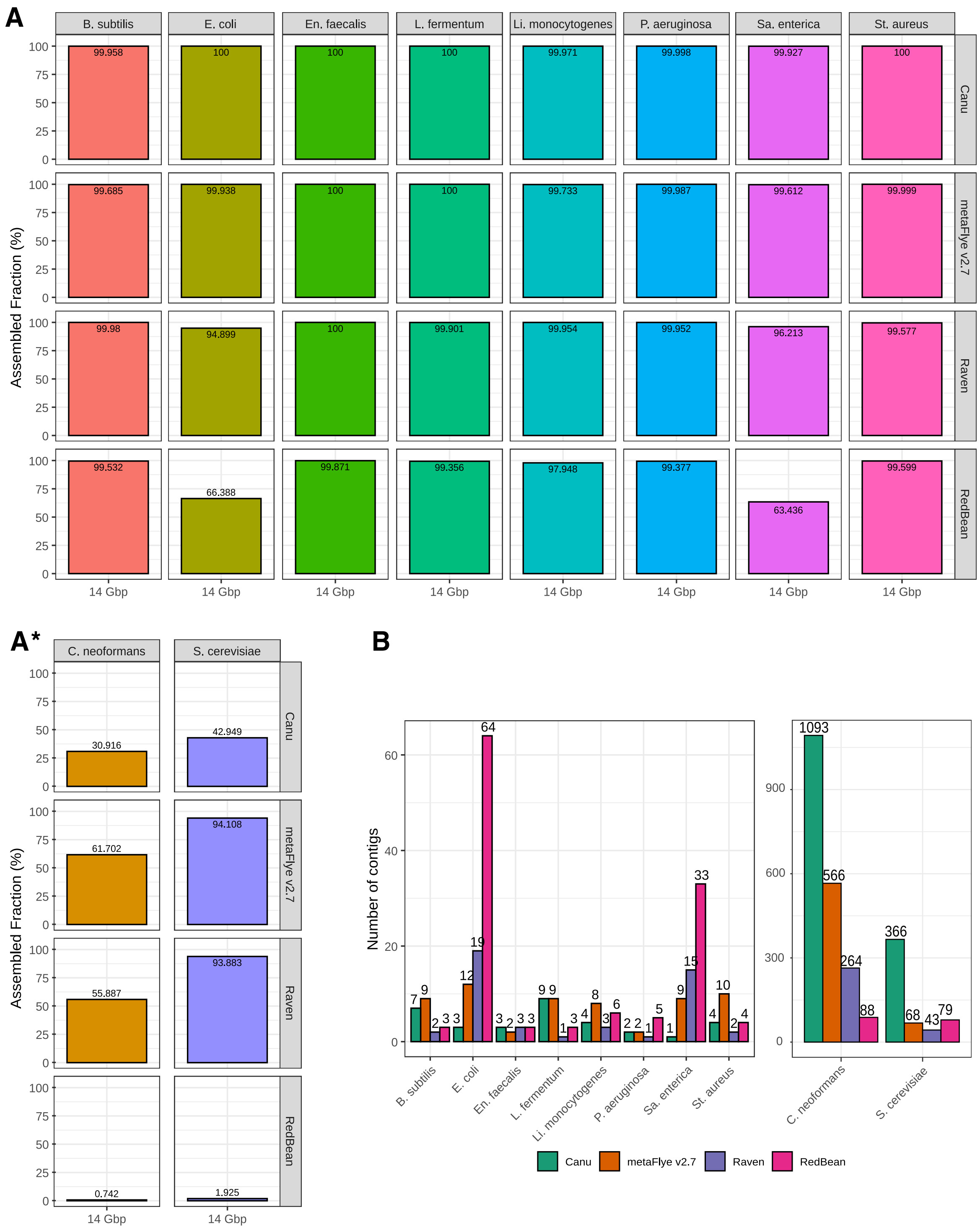


Figure 4. Even GridION (14 Gbp) assembly evaluation for the best performing tools. (**A**) Fraction of the genome covered by draft assemblies; (**B**) number of contigs for each microorganism.

Polishing evaluation. Polishing is the process of correcting assemblies in order to generate improved consensus sequences. Input for polishing nanopore-based assemblies can be raw ONT reads (i.e. Racon or Medaka)[45], raw electric signal (i.e. Nanopolish) (https://github.com/jts/nanopolish), or even high quality short

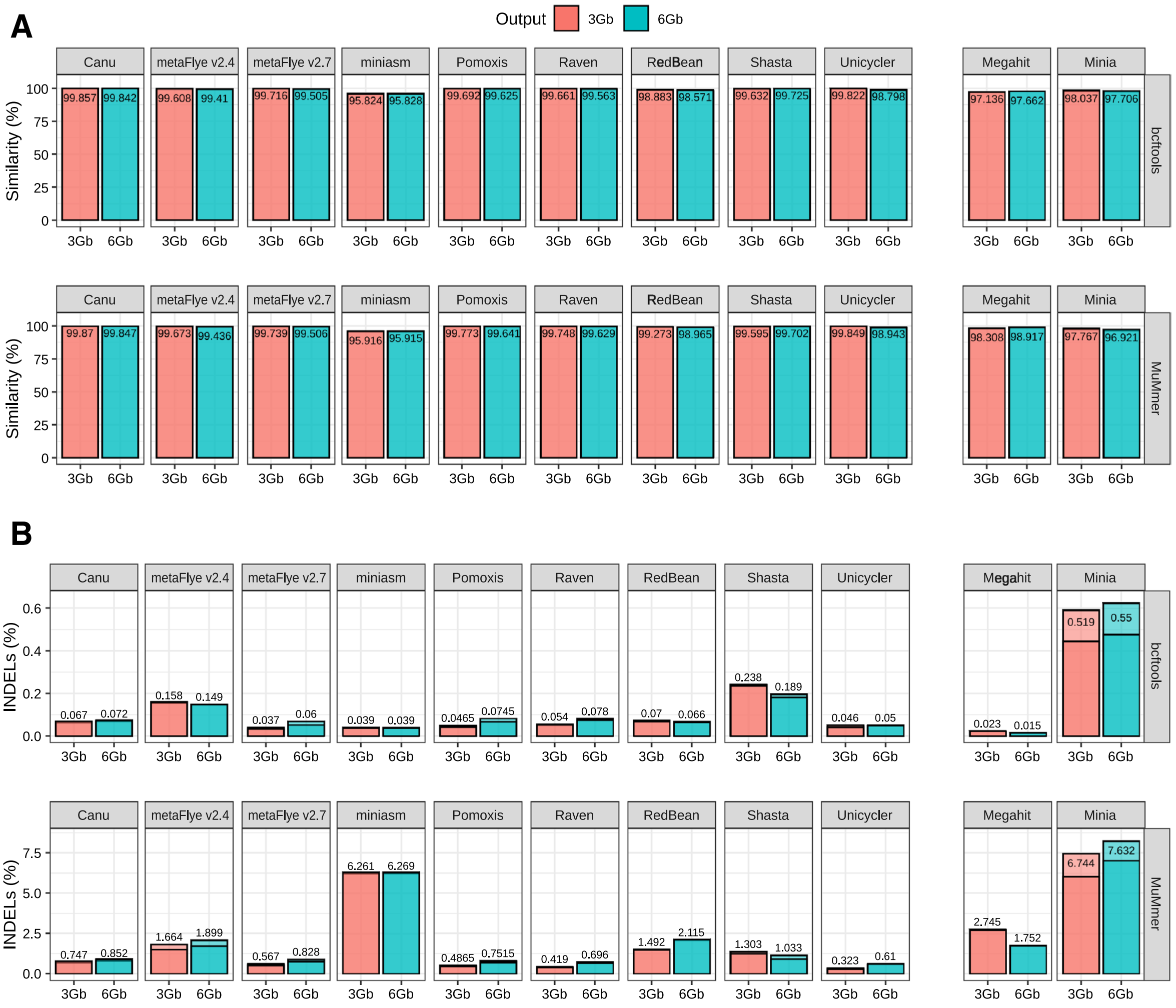


Figure 5. Assembly accuracy for the draft assemblies (subsampled Even datasets). (**A**) Percentage of similarity calculated as the total number of matches normalized by the metagenome size; (**B**) percentage of indels calculated as the total number of indels normalized by the metagenome size. In both cases, two different strategies were used: (top panel) alignment with minimap and evaluation with bcftools + 'indels_and_snps.py' in-house script; (bottom panel) alignment with MuMMer and evaluation with 'count_SNPS_indels.pl' script from Goldstein et al.[9].

reads (i.e. Racon)[45]. The state-of-art polishing workflow for nanopore sequencing consists of correcting the draft assemblies through several rounds of Racon (typically 2–4), followed by a single Medaka step.

Some of the tested tools automatically incorporated Racon (Raven, Pomoxis and Unicylcer) in their pipelines, whereas the others included different algorithms for correcting the reads before (Canu) or after (metaFlye and ReadBean) the assembly process. Thus, we wanted to assess how various steps of polishing could affect the SNP and indel ratio of the different assemblers. Results were highly heterogenous (Fig. 7; Supplementary Table S6). Pomoxis and Raven drastically improved their accuracy after several rounds of polishing with the original ONT reads (Supplementary Table S6). In fact, accuracy with no polishing steps was close to 96%, as reported for miniasm (Fig. 5). Higher similarity percentages were observed after one round of Racon (1R) for Raven, and four rounds of Racon + one round of Medaka (4R + m) for Pomoxis. Redbean and metaFlye -which were run again without using their built-in polishers- also improved their accuracy after 1R or 4R + m, respectively. Canu presented a lower percentage of SNPs when no polishing steps were added to the pipeline (Supplementary Table S6). Nevertheless, all the tools drastically improved their indel ratio after 4R + m. The percentage of improvement varied between 41% (Canu) and 91% (Raven and Pomoxis) (Fig. 7A). It has to be highlighted that the lowest number of SNPs and indels was achieved by Canu, which is the only tool that carries out error correction before assembling the reads.

The error profiles were evaluated again to further assess whether polishing draft assemblies with high quality short reads led to improved assemblies. Albeit yielding heterogeneous results, all the tools achieved better indel ratios after four rounds of Racon correction with Illumina reads (Supplementary Table S6). In this case, all the assemblers improved their accuracy (% of similarity) after one (Canu and metaFlye) or four (Pomoxis, Raven and

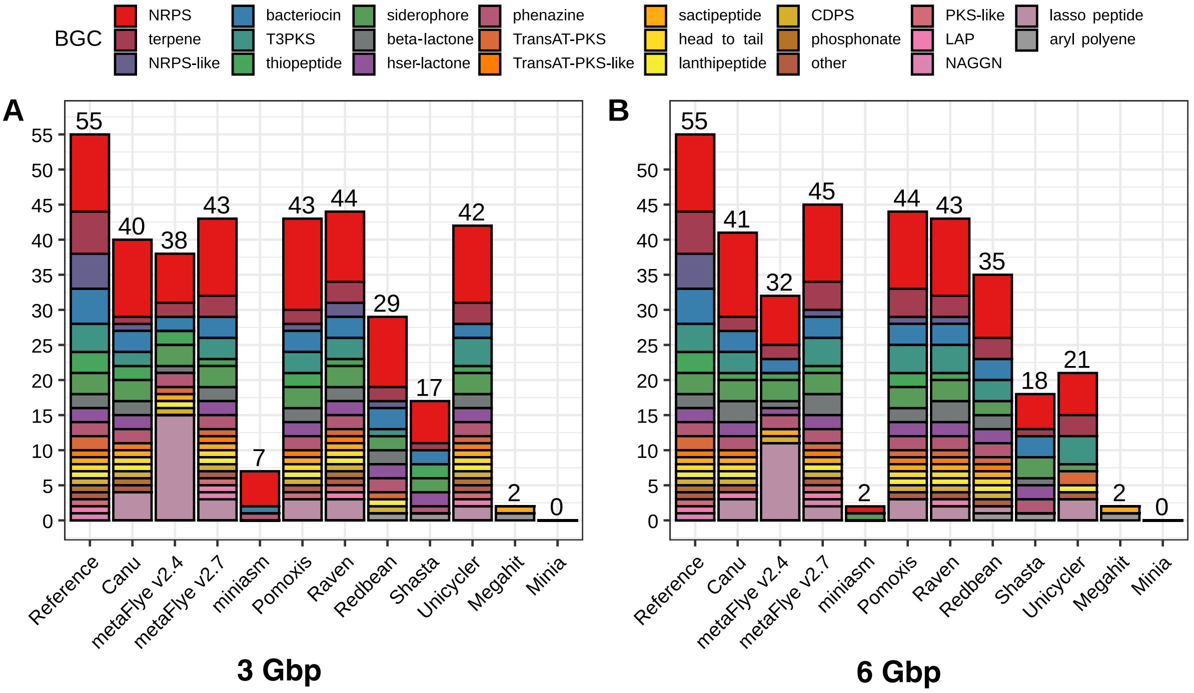


Figure 6. Number of biosynthetic gene clusters (BGCs) predicted by antiSMASH for each draft assembly in the Even GridION datasets. (**A**) BGCs predicted using the 3 Gbp dataset; (**B**) BGCs predicted using the 6 Gbp dataset.

RedBean) Racon rounds. When comparing the highest scores obtained with Illumina-based correction to the highest scores achieved after ONT-based polishing (Fig. 7), the percentage of similarity was higher for metaFlye and Canu assemblies corrected with Illumina reads, and lower for Pomoxis, Raven and RedBean, where ONT polishing outperformed Illumina's. A similar trend was observed for the indel ratio. This time, Illumina correction clearly enhanced the indel correction for metaFlye and Canu. In fact, Canu + Illumina correction retrieved the lowest indel ratio. Pomoxis, Raven and RedBean achieved a better indel correction with ONT reads.

Discussion

Assembling shotgun sequencing data is often a key factor for characterizing the functional and taxonomic diversity of microbial communities. In the recent years, nanopore-based sequencers (Oxford Nanopore Technologies; ONT) are rapidly growing in popularity due to four basic reasons: (1) low cost, (2) long-read generation, (3) portability, and (4) real-time analysis. Several bioinformatic tools have been developed to handle nanopore sequences during the assembly process. Nevertheless, there is a lack of systematic, up-to-date, independent studies comparing the performance of the currently available tools. This work is aimed at filling this gap using the data previously published by Nicholls et al.[15], which consisted of the ultra-deep sequencing of two different mock communities (Table 1) on GridION and PromethION platforms (ONT). These platforms follow the same sequencing principles as MinION, but they have a significantly higher output. For that reason, the datasets were subsampled in order to adapt their output to the current yield offered by MinION (3–6 Gbp)[9,34–39], then extending the study to higher yields comparable with other recent works[46].

Despite the relatively low complexity of the mock communities analyzed in this evaluation study, our results show that there is a huge variation in assembly results depending on the software chosen to perform the analysis. Minia and Megahit poorly reconstructed the microbial genomes (Figs. 1, 2) and produced highly fragmented draft assemblies (Fig. 3). This output was expected, since these assemblers are highly optimized to work on short reads, which are very different to the data generated by ONT sequencers.

Long-read assemblers (Canu, metaFlye, Unicycler, miniasm, Raven, Shasta and Readbean) also presented significant differences in the general assembly performance. This was expected too, since some of the tools were not specifically designed for assembling metagenomes (Supplementary Table S1). Overall, only metaFlye, Raven, and Canu worked well on all the tested datasets. They were able to recover the eight bacterial genomes from the Even dataset with a high degree of completeness, and also reconstructed a significant fraction of the yeast genomes. Draft assemblies were highly contiguous when using these three tools, as they were able to reconstruct bacterial genomes in only 1–19 contigs (Fig. 4B). Unicycler and, especially, Pomoxis, also performed well for some datasets and metrics, but failed to run in some cases (Supplementary Table S1). Both tools are pipelines based on miniasm that include further polishing steps by Racon. Miniasm alone was also unable to assemble the Log 6 Gbp dataset, indicating a lack of consistency of the algorithm for different microbial community structures.

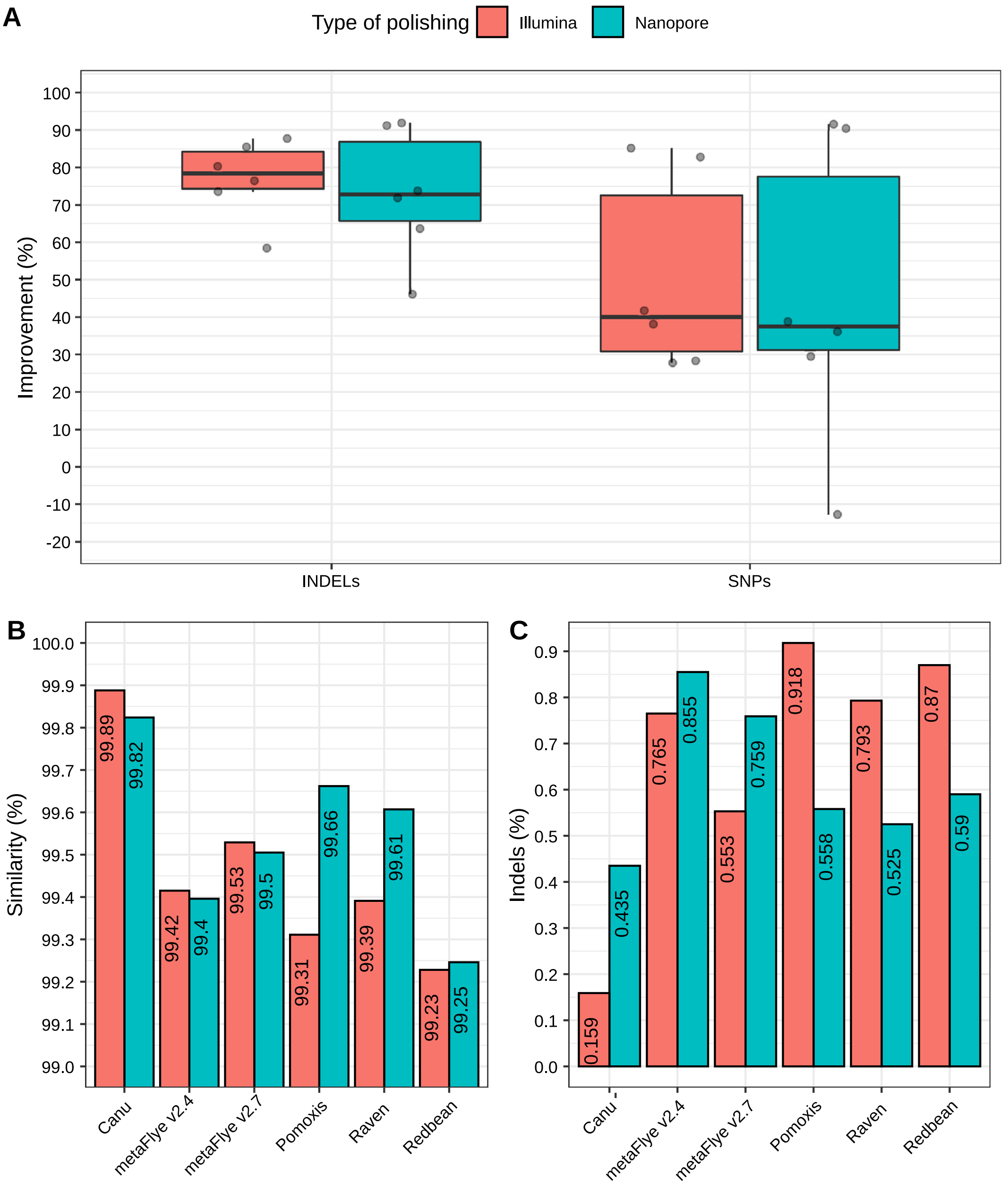


Figure 7. Polishing evaluation. (**A**) Percentage of improvement within the whole metagenome, taking as a reference the number of errors prior to polishing; (**B**) highest similarity percentage achieved by each tool; (**C**) best indel ratio achieved by each tool. Note that a different number of polishing rounds may be needed for achieving the highest similarity and the lowest indel ratio depending on the tool.

Finally, Shasta and RedBean (wtdgb2) retrieved incomplete assemblies and they did not provide any additional advantage other than computational efficiency.

Our results are in accordance with previous studies. MetaFlye has proved to outperform other tools in terms of metagenome recovery when using different mock communities[32,33], although it must be noted that these previous studies did not include all the tools selected in the present benchmark. Canu also performed well in other studies[33], and has been proposed for increasing the contiguity of metagenome assembled genomes recovered from real samples[46]. Nevertheless, its high computational cost limits the use of Canu for bigger datasets (Fig. 3, Supplementary Fig. S4)[33,46]. RedBean displayed a reduced performance in comparison to other long-read assemblers[32,33,46,47]. To the best of our knowledge, no other metagenome assembly benchmark has included Pomoxis, Shasta, or Raven. Wick and Holt[47] evaluated different tools for single isolate assembly (not metagenomic assembly), and reported that Shasta was more likely to produce incomplete draft assemblies, while Raven was reliable for chromosome assembly, as also seen in our work. Although Pomoxis was not included in this last benchmark, another miniasm + Racon strategy was used. This strategy, that was reported to perform robustly among different genomic datasets, is equivalent to one of the pipelines used in the present study (here referred to as Unicycler). This observed robustness is in contrast to our results, supporting the idea that the intrinsic differential coverage of metagenomic datasets could be the cause of the inconsistency detected for miniasm in this benchmark.

Although sequencing errors are one of the main drawbacks of third generation sequencing platforms, the best performing tools (metaFlye v2.7, Canu, Raven and Pomoxis) achieved > 99.5% of accuracy in the final assemblies. Indels may be especially problematic, since they can introduce frameshift errors, which hinder functional prediction. After analyzing the different BGCs profiles, metaFlye and Raven demonstrated to reach better results and they outperformed Canu. This is in accordance with the indel ratio calculated for each tool (Fig. 5B). It has to be highlighted that these results were obtained by using ONT configurations explicitly recommended in the manual of each tool. The use of other tools (i.e. polishers) led to assemblies with enhanced quality. The lowest number of SNPs and indels were achieve after different rounds of polishing for some assemblers (Supplementary Table S6). As a consequence, the number of polishing rounds is variable and must be carefully chosen by the user. Correction with Illumina reads is a useful strategy for reducing the number of indels and SNPs produced by metaFlye and Canu, as also reported in Moss et al.[46]. The combination of Canu with polishing tools resulted in the best accuracy, especially when using Illumina reads for the correction.

Finally, time is a crucial parameter when choosing a bioinformatic tool, even more if considering MinION's ability to generate real-time data. In this sense, metaFlye v2.7 was up to 6.7 times faster than Canu, which was the slowest tool tested on this benchmark. Raven was even faster than metaFlye, and tended to generate fewer contigs (Fig. 3, Supplementary Fig. S4; Supplementary Table S3).

Taken together, our results show that nanopore data (accommodated to current MinION's output) can lead to highly contiguous and accurate assemblies when using the proper tools, with no need of complementary sequencing with Illumina. From all the tested software, metaFlye v2.7 resulted the best in terms of metagenome recovery fraction and total metagenome assembled size. Raven achieved slightly lower genome fractions than metaFlye, but was faster and generally retrieved a lower number of contigs. Canu was the most accurate tool and introduced fewer indels when combined with polishing tools, but its assembly process also demonstrated to be time consuming. Pomoxis and other miniasm-based pipelines are also promising, but their inconsistency problems should be addressed. This work may help software developers to design new bioinformatic tools optimized for nanopore-based shotgun metagenomic sequencing, although further research is still needed in order to benchmark the different assemblers on more complex microbial communities.

Conclusions

Shotgun metagenomic sequencing based on short reads usually results in highly fragmented metagenomes, thus complicating downstream analyses such as the recovery of individual genomes or the prediction of complex and repetitive gene structures (i.e. biosynthetic gene clusters, CRISPR-CAS systems, etc.). This work demonstrates that, despite the high error intrinsic to third-generation sequencing platforms, nanopore data alone can overcome these limitations and retrieve extremely contiguous genomes directly from simple microbial communities. However, there is a huge variation in assembly performance depending on the chosen software. In general terms, metaFlye could be defined as the best suited assembler for nanopore metagenomic data. This tool leads to the highest metagenome recovery ratio and performs robustly among the tested datasets. Raven also performed well and required less time to perform the analyses. Canu is more suitable when lower error rates are required, but draft assemblies should be polished in order to reduce the number of indels. Polishing with short reads does not necessarily improve the quality of the draft assemblies but, in combination with Canu, it can lead to the most accurate metagenome reconstruction. Overall, this work demonstrates the suitability of using nanopore sequencing alone for assembling low-complexity microbial communities, and paves the way towards the standardization of bioinformatic pipelines for long-read sequencing data.

Data availability

Raw data was deposited in the NCBI database under the BioProject number PRJNA564477 (https://www.ncbi.nlm.nih.gov/bioproject/564477). Raw datasets from Nicholls et al.[15] can be downloaded from the ENA (https://www.ebi.ac.uk/ena/data/view/PRJEB29504). All the code used in this study is publicly available at doi: https://doi.org/10.5281/zenodo.3935763. It includes the bash scripts designed for the automatic execution of the different bioinformatic analysis, the R code and CSV tables for figure construction, and other in-house and third-party scripts needed to reproduce the analyses.

Received: 10 June 2020; Accepted: 22 July 2020
Published online: 12 August 2020

References

1. Hiraoka, S., Yang, C. & Iwasaki, W. Metagenomics and bioinformatics in microbial ecology: Current status and beyond. *Microbes Environ.* **31**(3), 204–212 (2016).
2. Hug, L. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**(5), 16048. https://doi.org/10.1038/nmicrobiol.2016.48 (2016).
3. Nutman, A. & Marchaim, D. How to: Molecular investigation of a hospital outbreak. *Clin. Microbiol. Infect.* **25**(6), 688–695 (2019).
4. Tully, B., Graham, E. & Heidelberg, J. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5**, 170203. https://doi.org/10.1038/sdata.2017.203 (2018).
5. Nayfach, S. *et al.* New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510. https://doi.org/10.1038/s41586-019-1058-x (2019).

6. Fettweis, J. M. *et al.* The vaginal microbiome and preterm birth. *Nat. Med.* **25**, 1012–1021. https://doi.org/10.1038/s41591-019-0450-2 (2019).
7. Goodwin, S., McPherson, J. & McCombie, W. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351. https://doi.org/10.1038/nrg.2016.49 (2016).
8. Wick, R., Judd, L., Gorrie, C. & Holt, K. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**(6), e1005595 (2017).
9. Goldstein, S., Beka, L., Graf, J. & Klassen, J. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genom.* **20**, 1 (2019).
10. Olson, N. *et al.* Metagenomic assembly through the lens of validation: Recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinform.* **20**(4), 1140–1150 (2017).
11. Ayling, M., Clark, M. & Leggett, R. New approaches for metagenome assembly with short reads. *Brief. Bioinform.* **21**(2), 584–594 (2019).
12. Sczyrba, A. *et al.* Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071. https://doi.org/10.1038/nmeth.4458 (2017).
13. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 1–11 (2018).
14. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**, 100 (2017).
15. Nicholls, S., Quick, J., Tang, S. & Loman, N. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience* **8**, 5 (2019).
16. Jayakumar, V. & Sakakibara, Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief. Bioinform.* **20**(3), 866–876 (2017).
17. Loman, N., Quick, J. & Simpson, J. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735. https://doi.org/10.1038/nmeth.3444 (2015).
18. Koren, S. *et al.* Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**(9), R101 (2013).
19. González-Escalona, N., Allard, M., Brown, E., Sharma, S. & Hoffmann, M. Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. *PLoS One* **14**, e0220494 (2019).
20. Lu, H., Giordano, F. & Ning, Z. Oxford nanopore MinION sequencing and genome assembly. *Genom. Proteom. Bioinforma.* **14**, 265–279 (2016).
21. Pomerantz, A. *et al.* Real-time DNA barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building. *GigaScience* **7**, 20 (2018).
22. Orsini, P. *et al.* Design and MinION testing of a nanopore targeted gene sequencing panel for chronic lymphocytic leukemia. *Sci. Rep.* **8**, 11798. https://doi.org/10.1038/s41598-018-30330-y (2018).
23. Wick, R., Judd, L., Gorrie, C. & Holt, K. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial. Genom.* **3**, 20 (2017).
24. Deschamps, S. *et al.* A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat. Commun.* **9**, 4844. https://doi.org/10.1038/s41467-018-07271-1 (2018).
25. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345. https://doi.org/10.1038/nbt.4060 (2018).
26. Hardegen, J. *et al.* Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass. *Biores. Technol.* **265**, 275–281 (2018).
27. Benítez-Páez, A. & Sanz, Y. Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION^{TM} portable nanopore sequencer. *GigaScience* **6**, 20 (2017).
28. Bokulich, N. *et al.* mockrobiota: A public resource for microbiome bioinformatics benchmarking. *mSystems* **1**, 20 (2016).
29. Fritz, A. *et al.* CAMISIM: Simulating metagenomes and microbial communities. *Microbiome* **7**, 20 (2019).
30. Vollmers, J., Wiegand, S. & Kaster, A. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective—not only size matters!. *PLoS One* **12**, e0169662 (2017).
31. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
32. Hu, Y., Fang, L., Nicholson, C. and Wang, K. Implications of error-prone long-read whole-genome shotgun sequencing on characterizing reference microbiomes. Prepint available at https://www.biorxiv.org/content/10.1101/2020.03.05.978866v1.full (2020).
33. Kolmogorov, M., Rayko, M., Yuan, J., Polevikov, E. & Pevzner, P. metaFlye: Scalable long-read metagenome assembly using repeat graphs. Prepint available at https://www.biorxiv.org/content/10.1101/637637v1 (2019).
34. Sevim, V. *et al.* Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Sci. Data* **6**, 285. https://doi.org/10.1038/s41597-019-0287-z (2019).
35. Dhar, R. *et al.* De novo assembly of the Indian blue peacock (*Pavo cristatus*) genome using Oxford Nanopore technology and Illumina sequencing. *GigaScience* **8**, 5 (2019).
36. Parajuli, P., Deimel, L. & Verma, N. Genome analysis of *Shigella flexneri* serotype 3b strain SFL1520 reveals significant horizontal gene acquisitions including a multidrug resistance cassette. *Genome Biol. Evol.* **11**, 776–785 (2019).
37. Leidenfrost, R. M. *et al.* Benchmarking the MinION: Evaluating long reads for microbial profiling. *Sci. Rep.* **10**, 5125. https://doi.org/10.1038/s41598-020-61989-x (2020).
38. Hamner, S. *et al.* Metagenomic profiling of microbial pathogens in the Little Bighorn River, Montana. *Int. J. Environ. Res. Public Health* **16**, 1097 (2019).
39. Che, Y. *et al.* Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome* **7**, 44. https://doi.org/10.1186/s40168-019-0663-0 (2019).
40. Lindgreen, S., Adair, K. & Gardner, P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* **6**, 19233. https://doi.org/10.1038/srep19233 (2016).
41. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
42. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2015).
43. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
44. Blin, K. *et al.* AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
45. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
46. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0422-6 (2020).

47. Wick, R. & Holt, K. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research* **8**, 2138 (2020).

Acknowledgements

We are grateful to Dr. Manuel Porcar, who kindly revised the manuscript and gave some good advice for improving our work. We are indebted to Kristie Tanner for language correction. AL is a recipient of a Doctorado Industrial fellowship from the Spanish Ministerio de Ciencia, Innovación y Universidades (reference DI-17-09613).

Author contributions

A.L. and P.V. performed the data analyses. A.L., J.P., and C.V. designed the research and discussed the results. All the authors wrote the manuscript, and have read and approved the final manuscript version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-70491-3.

Correspondence and requests for materials should be addressed to C.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Publication VII

BIOLOGY
Methods & Protocols

Biology Methods and Protocols, 2020, 1–11

doi: 10.1093/biomethods/bpaa016
Advance Access Publication Date: 20 August 2020
Review

REVIEW

A lab in the field: applications of real-time, *in situ* metagenomic sequencing

Adriel Latorre-Pérez[1], Javier Pascual[1], Manuel Porcar[1,2] and Cristina Vilanova[1,*]

[1]Darwin Bioprospecting Excellence SL, Valencia, Spain and [2]Institute for Integrative Systems Biology, I2SysBio, University of Valencia-CSIC, Valencia, Spain

*Correspondence address. Darwin Bioprospecting Excellence SL, Valencia, Spain; E-mail: cristina@darwinbioprospecting.com

Abstract

High-throughput metagenomic sequencing is considered one of the main technologies fostering the development of microbial ecology. Widely used second-generation sequencers have enabled the analysis of extremely diverse microbial communities, the discovery of novel gene functions, and the comprehension of the metabolic interconnections established among microbial consortia. However, the high cost of the sequencers and the complexity of library preparation and sequencing protocols still hamper the application of metagenomic sequencing in a vast range of real-life applications. In this context, the emergence of portable, third-generation sequencers is becoming a popular alternative for the rapid analysis of microbial communities in particular scenarios, due to their low cost, simplicity of operation, and rapid yield of results. This review discusses the main applications of real-time, *in situ* metagenomic sequencing developed to date, highlighting the relevance of this technology in current challenges (such as the management of global pathogen outbreaks) and in the next future of industry and clinical diagnosis.

Keywords: third-generation sequencing; *in situ* metagenomics; microbial ecology

Introduction

For many years, culture-dependent approaches were the only tools available for the study of microorganisms, although the vast majority of microbial species (>99%) cannot be cultivated [1]. This limitation lasted until the development of molecular techniques, such as the automation of Sanger sequencing [2], molecular markers [3], cloning [4], or fluorescence *in situ* hybridization [5], among many others. However, these molecular techniques presented other weaknesses, like the inability to access low-abundance microorganisms, generating a bias towards the most abundant taxa.

The term metagenomics was proposed in the 1990s [6] to define the set of genomes that could be found in a given environment. The fundamental aim of metagenomics is the study of microorganisms in the context of their community by means of sequencing genomic fragments from the entire microbiome simultaneously. Nevertheless, this goal can be partially accomplished by sequencing marker genes, even though this approach should not be considered as true metagenomics [7]. In marker-gene studies, generic, relatively universal primers are used to amplify a fragment of a given gene through polymerase chain reaction (PCR) (e.g. 16S rRNA for bacteria/archaea, 18S/ITS for fungi) from all genomes present in a given sample, and the resulting pool of amplicons is sequenced. Next, the sequences are clustered into operational taxonomic units (OTUs), each OTU is taxonomically identified, and compared across samples. Traditionally, OTUs were constructed by grouping sequences according to a defined similarity threshold (typically 97%).

Received: 18 May 2020; **Revised:** 7 August 2020; **Accepted:** 18 August 2020

However, OTUs are being replaced by amplicon sequence variants, which group sequences that are completely identical [8]. While fast and inexpensive, this method does not give any information on the hundreds of thousands of functional genes encoded by other parts of the (meta)genomes as these remain unsequenced. Whole-genome sequencing (WGS)—or shotgun—metagenomics can offer an alternative and complementary method since it is based on the application of sequencing techniques to the entirety of the genomic material in the microbiome of an environmental sample. Sequencing the genomes of all microorganisms can provide information about the diversity of functional genes, and allow the assignment of each metabolic function to specific taxa, to identify novel genes or proteins so far unknown, and to assemble genomes in order to study evolutionary relationships.

The number of metagenomic studies has dramatically increased in the last years, mainly due to the emergence of high-throughput sequencing technologies and the development of bioinformatic tools that facilitate the assembly of data and the assignment of sequences through a process called binning [9]. The binning process consists of grouping assembled sequences (contigs) into discrete units (bins), which ideally represent draft genomes of individual microorganisms [10]. Overall, both high-throughput sequencing and bioinformatics have proven powerful tools that have generated, at a relatively low cost, a huge amount of genetic information [11].

High-throughput sequencing technologies can be divided into second- and third-generation ones. Two of the most widely used second-generation sequencing (SGS) technologies are Illumina and Ion Torrent. Albeit both techniques are based on sequence-by-synthesis, they have methodological differences. In Illumina sequencers, short DNA fragments are attached to a glass slide or micro-well and amplified to produce clusters. Fluorescence-labelled nucleotides are then washed across the flowcell and are incorporated to the complementary DNA sequence of the clustered fragment. Then, fluorescence from the incorporated nucleotides is detected, revealing the DNA sequence. On the other hand, Ion Torrent is based on the use of semiconductor materials that detect the release of H^+ protons while the DNA molecule is synthesized [12, 13].

Third-generation sequencing (TGS), also known as long-read sequencing, is based on single-molecule sequencing, which speeds up the sequencing process. This technology is currently under active development and includes platforms such as Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT). PacBio is based on single-molecule, real-time sequencing technology. An engineered DNA polymerase is attached to a single strand of DNA, and these are placed into micro-wells called Zero Mode Waveguides (ZMWs) [14]. During polymerization, the incorporated phospholinked nucleotides carry a fluorescent tag (different for each nucleotide) on their terminal phosphate. The tag is excited and emits light which is captured by a sensitive detector (through a powerful optical system). Eventually, the fluorescent label is cleaved off and the polymerization complex is then ready to extend the strand [15]. On the other hand, in ONT, a single-strand of DNA passes through a protein nanopore, resulting in changes in the electric current that can be measured. The DNA polymer complex consists of double-stranded DNA and an enzyme that unwinds the double-strand and passes the single-stranded DNA through the nanopore. As the DNA bases pass through the pore, there is a detectable disruption in the electric current, and this allows the identification of the bases on the DNA strand [16, 17].

Three substantial improvements have been made in TGS technologies with regard to SGS:

1. Increase in read length. While the SGS technologies produce many millions of short reads (150–400 bp), TGS typically produce much longer reads (6–20 kb)—without theoretical length limit for ONT—albeit far fewer reads per run (typically hundreds of thousands). Short reads produced by SGS lead to highly fragmented assemblies when it comes to *de novo* assembly of larger genomes because of difficulties in resolving repetitive sequences in the genome.
2. Reduction of sequencing time (from days to hours or even minutes for real-time applications). While major SGS platforms use sequencing by synthesis technologies, TGS technologies directly target single DNA molecules, and in the case of ONT platforms, reads are available for analysis as soon as they have passed through the sequencer.
3. Reduction or elimination of sequencing biases introduced by PCR amplification [18]. Despite this improvement, TGS technologies have high systematic error rates (~5–15%) unlike SGS technologies (<1%) [19]. Nevertheless, the accuracy of TGS can improve up to 99.9% in consensus sequences thanks to recent software developments [20].

In 2014, ONT released the MinION sequencing system which, unlike the bulk sequencing installations needed for the other technologies, is a palm-sized device producing long reads in real-time. When launched, the MinION read length was ~6–8 kb [21, 22]; however, lab protocols enabling the obtention of longer sequences (>100 kb) have been reported [23]. MinION is the smallest sequencing device currently available (10 × 3 x 2 cm and weighing just 90 g). It is inexpensive (less than €1000) in comparison with PacBio (more than €100 000), allowing laboratories with few economic resources to be able to access this technology. It can be directly plugged into a standard USB3 port on a computer with a simple configuration. Specifically, a computer with a solid-state drive, >8 GB of RAM, and >128 GB of hard disk space can be used for sequencing. The sequencer periodically outputs a group of reads in the form of raw current signals, which are then base-called on a laptop or on an ultra-portable ONT's MinIT. Furthermore, sequence analyses (such as sequence alignment and genome polishing) can be performed on a mobile phone [24]. Therefore, the ultra-portability, affordability, and speed in data production make the MinION technology suitable for real-time sequencing in a variety of environments, such as Ebola surveillance in West Africa during the last outbreak [25], microbial communities inspection in the Arctic [26], DNA sequencing on the International Space Station (ISS) [27], and even the recently emerging pandemic coronavirus SARS-CoV-2 [28, 29]. This review describes a range of applications in which having portable, low-cost, fast, and robust technologies allowing an *in situ* analysis of samples is key to address important challenges.

Portable sequencing in natural environments

Exploring the microbial diversity of natural environments via DNA sequencing techniques has become a routine in the last decade. Long-scale studies like the Earth Microbiome Project have led to the massive characterization of microbial populations inhabiting different environments on our planet [30]. Moreover, metagenomic sequencing has proved to be very useful for a wide range of applications such as recovering new genomes from unculturable organisms, mining microbial enzymes with potential applications in the industry, or discovering new biosynthetic gene clusters [31–33]. These studies have typically relied on next-generation sequencing platforms like Illumina, which usually requires shipping samples to a centralized sequencing facility. Nevertheless, biodiversity assessment studies are usually carried out in remote locations with limited access to DNA sequencing services, forcing scientists to design-intensive sampling expeditions and returning to their home institutions to perform the sequencing and the data analysis.

ONT sequencers have emerged as an alternative to these traditional approaches, allowing the creation of mobile, in-field laboratories. Figure 1 depicts a general workflow for the metagenomic analysis of samples using adapted protocols and a MinION device. Pomerantz *et al.* [34] and Menegon *et al.* [35] designed portable laboratories that included thermocyclers and centrifuges powered by external batteries, and a MinION device connected to a laptop to perform *in situ* DNA sequencing. Both works were not focused on metagenomic applications, but on evaluating the taxonomic identity of different animal specimens (reptiles and amphibians) via targeted sequencing of the 16S rRNA gene or other mitochondrial genes. However, the applied methodologies and lab configurations could be easily adapted to perform metataxonomic approaches relying on the amplification and massive sequencing of marker genes.

The feasibility of MinION-based metagenomic sequencing protocols has been specially tested in extremely cold environments. Edwards *et al.* [36] reported for the first time the use of mobile laboratories for the *in situ* characterization of the microbiota of a High Arctic glacier. They were able to adapt the widely used PowerSoil DNA Isolation kit (MoBio, Inc.) for its in-field use, and to perform the data analysis either online and offline. The report included new results from *in situ* metagenomics and 16S rRNA sequencing of different glaciers samples, and a benchmarking of the performance of in-field sequencing protocols by using mock communities as well as real samples. In the latter case, they compared the resulting taxonomic profiles with the microbial composition assessed by SGS platforms, describing strongly positive Pearson correlations at the phylum level. Goordial *et al.* [26] were also able to perform *in situ* MinION sequencing in the McGill Arctic Research Station. In this case, a permafrost sample was analysed using two different library preparation kits on the same extracted DNA. A similar percentage of Bacteria and Archaea was detected using both kits, but differences in the relative abundance of viruses and eukaryotic organisms were noted. The taxonomic profile of the same permafrost sample was also obtained by means of 16S rRNA Illumina sequencing. Notably, similar taxonomic groups were identified in all the cases at the phylum level, although relative abundances varied among the different methodologies. In a parallel work, Johnson *et al.* [37] used portable field techniques to isolate DNA from desiccated microbial mats collected in the Antarctic Dry Valleys, construct metagenomic libraries, and sequence the samples outdoors (Taylor Valley; Temperature = −1°C) and in the McMurdo Station (Room Temperature , RT). Longer reads were achieved by sequencing at RT, but average and median read length did not depend on ambient temperature. The study also reported that cold temperatures (4°C) reduced the quality of the generated sequences, even when working with high-quality DNA (Lambda Phage). Finally, Gowers *et al.* [38] designed and transported a miniaturized lab across Europe's largest ice cap (Vatnajökull, Iceland) by ski and sledge. They adapted DNA extraction and sequencing protocols to be performed in a tent during the expedition, using solar energy and external batteries to power the hardware. Offline basecalling was achieved *in situ* by using Guppy (Oxford Nanopore, Oxford, UK), but the metagenomic data analysis could not be carried out due to code errors while running the local version of Kaiju [39].

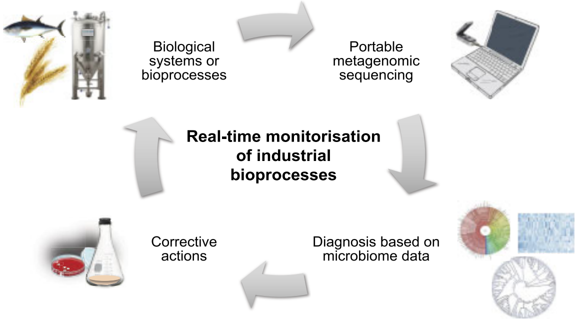

Biological systems or bioprocesses
Portable metagenomic sequencing
Real-time monitorisation of industrial bioprocesses
Corrective actions
Diagnosis based on microbiome data

Figure 1: Schematic representation of an *in situ* metagenomics workflow for the analysis of environmental and clinical samples.

In addition to cold environments, ONT sequencers have been also applied for sequencing a biofilm sample at a depth of 100 m within a Welsh coal mine [40]. This work presented the 'MetageNomad', a suite of off-the-shelf tools for metagenomic sequencing in remote areas using battery-powered equipment. The authors were able to perform the data analysis *in situ* by using Centrifuge [41] and a local database for characterizing the microbial composition of the sample.

Interestingly, MinION devices have allowed DNA sequencing off the Earth. A first study from Castro-Wallace *et al.* [27] compared the performance of nanopore sequencing in the ISS with experiments carried out on Ground Control, obtaining similar results. As a proof-of-concept, the authors used equimolar mixtures of genomic DNA from lambda bacteriophage, *Escherichia coli* (strain K12, MG1655) and *Mus musculus* (female BALB/c mouse) for the metagenomic sequencing. Data analysis could not be carried out at the ISS because of the lack of a laptop with the necessary tools installed, but it was demonstrated on the ground that sequencing analysis and microbial identification are completely feasible aboard the ISS. Recently, Burton *et al.* [42] have reported that the preparation and sequencing of 16S rRNA libraries are also achievable at the ISS. Specifically, the ZymoBIOMICS Microbial Community DNA Standard (Zymo Research) was used as the input DNA. Again, the results were comparable to the microbial profiles obtained on Earth. Remarkably, Carr *et al.* [43] determined that ONT sequencers performed consistently in reduced gravity environments, which would allow the use of nanopore sequencing in space expeditions to Mars or icy moons.

Although the viability of nanopore sequencing has been widely demonstrated even under extremely harsh conditions, the vast majority of the studies resulted in reduced yield compared to current MinION's metagenomic output (Table 1), which could reach up to 27 Gbp using a single flowcell [48] . This highlights the need to optimize in-field protocols in order to maximize the use of sequencing resources and reduce the price per sample, which is a key factor in some applications. Recently, a work from Urban *et al.* [44] studied the microbial communities present in the surface water of Cam River (Cambridge). All the protocols were carried out in the lab, and the authors were able to achieve up to ∼5.5 M 16S rRNA full-length sequences with exclusive barcode assignments in a single MinION run. Other groups have used MinION devices for characterizing river water [45], seawater [46], and marine sediments [47] through metagenomic sequencing. Even though these experiments were not implemented in the field, they demonstrated the possibility of obtaining higher sequencing yields (Table 1). The described outputs are compatible with more ambitious metagenomic analyses, such as the *de novo* recovery of single genomes directly from complex environmental samples. For that reason, the adaptation of sequencing protocols to field conditions is still to be further optimized.

Supporting microbiome-driven industrial processes

Microbiology has been present in the industry for centuries. In fact, human beings already used microorganisms for their own benefits long before they even knew that microscopic life existed. Nowadays, most of the microbiome-driven industrial processes are still not completely understood. Metagenomic sequencing has been widely applied in order to shed light on the microbial and metabolic transitions occurring on these industrial transformations. Some examples include the investigation of the link between microorganisms and their key roles or prevalence in microbial-based food products [49, 50]; the interaction of plants and root-associated bacteria for enhancing plant mineral nutrition [51]; or the description of the adverse effects of industrial subproducts used as soil fertilizers [52].

ONT portable sequencers are not only a valuable tool for characterizing industrial microbiomes, but for detecting and monitoring crucial microorganisms in real time (Fig. 2). Hardegen *et al.* [53] used full-length 16S rRNA sequencing for analysing changes in the archaeal community present in anaerobic digesters operating under different conditions. Higher proportions of *Methanosarcina* spp. were detected in the reactors achieving elevated biogas production. Although the sequencing was not carried out *in situ*, the suitability of MinION for monitoring and evaluating an industrial process through a microbial marker was demonstrated. Bacteriomes involved in the biogas production have been also studied through nanopore sequencing [54, 55], producing results which could be coupled with the Lotka–Volterra model for analysing the microbial interactions occurring in the reactor [56].

Water quality and wastewater management is another area of great interest for microbial monitoring. In fact, it has been proposed that sewage could serve for tracking infectious agents excreted in urine or faeces, such as SARS-CoV-2 [57]. In this particular context, the *in situ* and real-time assessment of pathogenic microorganisms by means of MinION sequencing would be especially advantageous. Hu *et al.* [58] reported correlations between *E. coli* culturing counts and the proportion of nanopore reads mapping a comprehensive human gut microbiota gene dataset, highlighting the potential use of this molecular technique as an indicator of faecal contamination. ONT metagenomic sequencing results were similar to those obtained with Illumina 16S rRNA sequencing, but a reduced time was achieved using MinION. Nanopore sequencing could be also employed for evaluating antibiotic resistance genes (ARGs) and antimicrobial-resistant pathogens present in wastewater treatment plants [59]. In this case, both Illumina and nanopore shotgun sequencing revealed comparable abundances of major ARG types. The agreement between the two platforms has been also described for the analysis of different water sources in Nepal through 16S rRNA sequencing [60]. Although long-reads allowed the classification of 59.41% of the reads down to the species level—no Illumina reads were classified at this level—a significant number of false-positives arose. These results were consistent with observations from [61], which showed that the bacterial identification at the genus level was reliable. Species-level missclassifications could be partially addressed by employing different—and optimized—bioinformatic approaches for the taxonomic classification [45, 62], by sequencing the complete 16S-ITS-23S region of the ribosomal operon [63, 64], or by coupling MinION sequencing with complementary quantitative PCR assays [60].

Agro-food industry would also benefit from real-time sequencing. For instance, nanopore metagenomic sequencing could be useful for the quick detection of plant pathogens infecting crops. Hu *et al.* [65] were able to identify the fungal species causing diseases on wheat plants, which were previously infected with known microbes. Co-occurrences between fungal and bacterial genera were also detected. Viral infectious diseases could be *in situ* monitored by using this technology, allowing rapid and improved response to outbreaks [66]. Other successful applications of ONT in the food industry included the characterization of the microbiome of a salmon ectoparasite (*Caligus rogercresseyi*), revealing its potential role as a reservoir for fish pathogens [67]; and the determination of the fish species present in complex mixtures, which would help to prevent—and rapidly detect—food fraud [68].

Overall, nanopore results generally agreed with those obtained by Illumina sequencing when available, thus validating the use of this technology for the vast majority of applications. Despite the huge potential shown, the suitability of MinION sequencing in an industrial context has yet to be ascertained, since all the discussed works were not carried out under field conditions. In fact, there are some critical points to be addressed before this technique could become a standard in the industry: (i) sequencing cost should be reduced; (ii) rapid and reliable *in situ* DNA extraction and library preparation protocols should be designed and validated; (iii) minimal sequencing yields should be determined for each specific application; (iv) fast and real-time pipelines should be created and tested; and (v) level of expertise for managing the data and the samples should be notably reduced.

Real-time analysis of clinical samples

Microbial infections are an increasingly relevant problem in intensive care units worldwide. Especially, the emergence of multi-drug resistance microorganisms is one of the main threats our society is facing from a clinical point of view [69]. Current diagnostics for pathogen identification in hospitals is still mainly dependent on culture- and molecular-based approaches, which have several limitations regarding specificity, bias, sensitivity, and time to diagnosis. The revolution of high-throughput sequencing and the decreasing costs associated to SGS has strongly empowered clinical diagnostics and other aspects of medical care [70]. In the particular case of clinical infections, high-throughput metagenomic sequencing allowed for the first time the precise strain-level identification of multiple pathogenic agents in single, all-inclusive diagnostic tests [71]. However, the limitations of SGS regarding cost and time to results (as described in previous sections) hamper its application when a fast analysis is needed. For instance, in the case of sepsis, patients are usually treated with broad-spectrum antibiotics until the first results of culture-based analysis (including determination of antibiotic susceptibility) are obtained 36–48 h later. When available, SGS approaches can speed up the process to ∼24 h, but result is expensive, labour intensive, and

Table 1: Summary of procedures and sequencing yield obtained under in-field and regular lab conditions

References	Samples	Library type	Lab type	Equipment	Yield (no of reads)
Edwards *et al.* [36]	Cryoconite and mock communities	Metagenomic	In-field	Vortex, Microcentrifuge, Fluorometer, PCR cycler (optional), Laptop, and Miscellaneous power sources	3514–52 000
Edwards *et al.* [36]	Cryoconite and mock communities	16S rRNA	In-field	Vortex, Microcentrifuge, Fluorometer, MiniPCR, Laptop, and Miscellaneous power sources	20,000–220 051
Menegon *et al.* [35]	Animal tissue	16S rRNA and other marker genes	In-field	Microcentrifuge, Fluorometer, Thermocycler, Laptop, Portable Refrigerator, and 12 V portable batteries	5039
Pomerantz *et al.* ([34]	Animal tissue	16S rRNA and other marker genes	In-field	Benchtop centrifuge, MiniPCR, Laptop, and External batteries	16 663
Goordial *et al.* [26]	Permafrost	Metagenomic	Research station	Vortex, Microcentrifuge, Fluorometer, Magnetic rack, and Computer	6348–9530
Johnson *et al.* [37]	Microbial mats	Metagenomic	In-field	Laptop and Insulating materials	573–6026
Gowers *et al.* [38]	Soil	Metagenomic	In-field	USB Vortex, Hand-powered Centrifuge, Fluorometer, Laptop, Solar panels, and External batteries	19 839–133 538
Edwards *et al.* [40]	Sediment	Metagenomic	In-field	TerraLyser (Zymo Research, Inc.), 12 V Microcentrifuge, Fluorometer, MiniPCR, two Laptops, and External batteries	1184
Castro-Wallace *et al.* [27]	Mock DNA	Metagenomic	ISS	Microcentrifuge, Fluorometer, MiniPCR, and Computer	14 903–60 864 (libraries prepared in a regular lab on Earth)
Burton *et al.* [42]	Mock DNA and pure cultures	16S rRNA	ISS	MiniPCR, Computer, Refrigerator, and Freezer	>15 000
Urban *et al.* [44]	River water	16S rRNA	Regular lab	Fully equipped	737 164–5 491 510
Hamner *et al.* [45]	River water	Metagenomic	Regular lab	Fully equipped	397 884–1 261 165
Liem *et al.* [46]	Seawater	Metagenomic	Regular lab	Fully equipped	225 200–1 316 823
Cáceres *et al.* [47]	Marine sediments	Metagenomic	Regular lab	Fully equipped	1 500 000

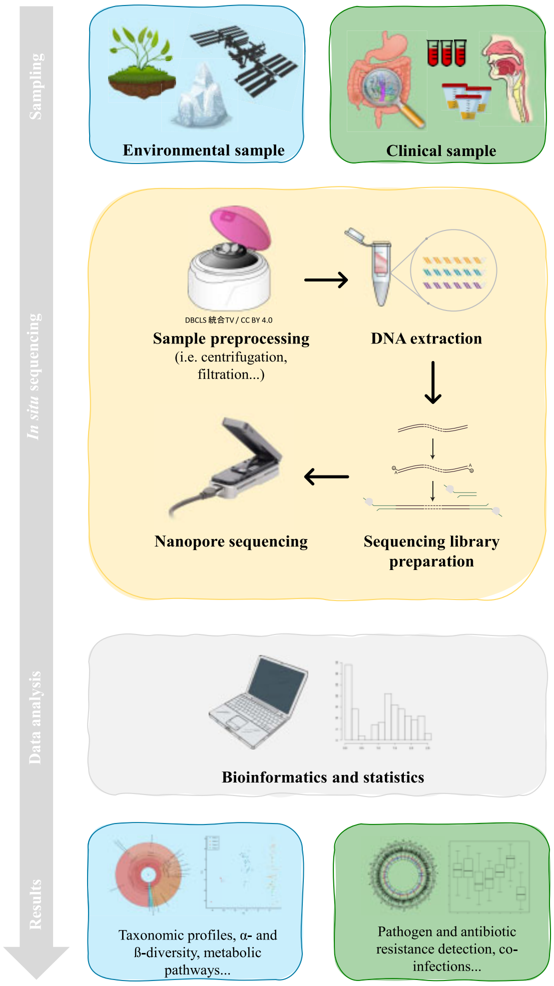

Sampling
In situ sequencing
Data analysis
Results
Environmental sample
Clinical sample
DBCLS 統合TV / CC BY 4.0
Sample preprocessing
(i.e. centrifugation, filtration...)
DNA extraction
Nanopore sequencing
Sequencing library preparation
Bioinformatics and statistics
Taxonomic profiles, α- and ß-diversity, metabolic pathways...
Pathogen and antibiotic resistance detection, co-infections...

Figure 2: Real-time, *in situ* sequencing as a monitoring tool for industrial bioprocesses. Relevant systems (digesters, crops, farmed animals, etc.) are sampled and analysed through metagenomic sequencing with MinION. Sequencing and bioinformatic analysis result in the rapid diagnosis of problems, for which corrective actions (antimicrobial treatments, bioaugmentation, change in control process parameters, etc.) can be early implemented.

informatically challenging for most hospitals and healthcare centres [72]. In this context, MinION sequencing (Fig. 1) paves the way towards a diagnostic alternative in a clinically critical timeframe, which could reduce the morbidity and mortality associated to major microbial infections.

The first reports on MinION sequencing in clinical diagnosis were focused on the detection of single pathogens during outbreaks. Flagship examples of such applications are the fast (<24 h) detection of Ebola virus during the 2015 outbreak in West Africa [16, 73], or the fast (<6 h) phylogenomic analysis of *Salmonella* strains during a hospital outbreak [74]. Other significant efforts have focused on the fast identification of single clinical isolates [75], including the analysis of ARGs in a timeframe of <6 h [76, 77]. However, a range of use cases in the clinical field requires the use of metagenomic sequencing to unveil the identity of viral or microbial communities rather than single isolates. In the case of viruses, the seminal work of Greninger *et al.* [78] reported the detection of several viral pathogens in human blood in <6 h since the obtention of the samples, by using cDNA conversion and random amplification prior to sequencing. Despite the notable error rate observed in the sequences, all viruses (chikungunya virus, Ebola virus, and hepatitis C virus) were correctly identified and most of their genomes were recovered with high accuracy (97–99%). A similar approach was reported for the rapid identification of mosquito-borne arbovirus [79], and other viruses causing co-infections, including dengue, from human serum samples [80].

On the other hand, an extensive number of reports have been focused on the analysis of infections caused by bacterial communities (Table 2), using different approaches which resulted in different analysis times. Even though a range of PCR-free protocols have been developed for MinION sequencing, one of the main problems associated to the analysis of microbial communities in clinical samples is the overwhelming concentration of host DNA, which hampers the detection of bacterial sequences during the first hours of the sequencing runs [89, 90]. Several strategies have been applied to partially overcome this limitation. On the one hand, PCR-based approaches targeting the 16S rRNA gene proved the most rapid methods to identify pathogenic agents from human samples. Particular examples of this are the metagenomic analysis in empyema patients with pleural effusion [83] and the metagenomic analysis of patients with acute respiratory distress syndrome [84], both studies reporting the obtention of the first results in only 2 h after the collection of samples. On the other hand, the use of human cell-free samples allows the application of WGS protocols for the analysis of the communities, yielding not only taxonomic information but also the identification of putative antimicrobial resistance genes, which are of outstanding relevance for the selection of effective treatments. Pendleton *et al.* analysed in 2017 [86] lavage fluids from patients with pneumonia and managed to identify the bacterial pathogens in the lungs in <9 h using a WGS strategy. Similar approaches performed on urine samples [87] and resected valves from patients with endocarditis [85] yielded a diagnosis in 4 h. For the analysis of bacterial sepsis, recent reports describe the application of MinION metagenomic sequencing on cell-free samples (<6 h from samples to results) [81] and on faecal samples from preterm infants (obtaining results in <5 h) [82]. The depletion of human DNA prior to metagenomic sequencing proved also a useful alternative to reduce total analysis time [88].

In the current SARS-CoV-2 outbreak, MinION sequencing is proposed not only as a rapid tool for WGS, but also as a metagenomics-based approach for the rapid diagnosis of polymicrobial/viral infections associated to coronavirus disease COVID19. This is especially relevant to optimize the treatment of patients suffering severe symptoms of the disease.

Finally, other advantages of MinION sequencing besides the reduction of analysis are also to be highlighted. Given the low price of the devices and consumables (in comparison to SGS equipment), MinION has enabled the metagenomic analysis of clinical samples in areas with limited resources [25, 91]. Also, from a technical point of view, the generation of long reads increases the resolution of the taxonomic analysis of the samples, reaching in most cases a species-level identification of the most abundant members of the communities [92, 93].

The ‘read until’ strategy: towards cost-effective *in situ* metagenomics

Metagenomic applications are often limited by the nature of the samples to be analysed. For instance, the characterization of prokaryotes or viruses present in a sample dominated by host

Table 2: Summary of procedures and analysis times (from sample to results) reported for MinION-based metagenomic analyses of clinical samples

References	Clinical application	Sample type	Approach	Total analysis time, h
Grumaz *et al.* [81]	Bacteremia in septic patients	Blood cell-free samples	Whole-genome amplification + ligation sequencing	5–6
Leggett *et al.* [82]	Rapid diagnosis of preterm infants with suspected sepsis	Faeces	Different approaches tested	<5
Mitsuhashi *et al.* [83]	Unveiling microbial communities in empyema patients	Pleural effusions	16S rRNA amplification + rapid sequencing	2
Greninger *et al.* [78]	Identification of viral pathogens in clinical samples	Blood samples	Amplified cDNA + ligation sequencing	<6
Tanaka *et al.* [84]	Metagenomic analysis in patients with acute respiratory distress syndrome (ARDS)	Airway secretions	16S rRNA amplification + rapid sequencing	2
Cheng *et al.* [85]	Metagenomic analysis in culture-negative infective endocarditis cases	Resected valves	Ligation sequencing	4
Pendleton *et al.* [86]	Identification of bacterial pathogens in the lungs of patients with pneumonia	Lavage fluid	Ligation sequencing	9
Batovska *et al.* [79]	Metagenomics of mosquito-borne arbovirus	Mosquitoes	cDNA conversion + ligation sequencing	<10
Schmidt *et al.* [87]	Identification of pathogens and AMR in urine infections	Urine	Ligation sequencing and rapid sequencing	4
Charalampous *et al.* [88]	Diagnosis of bacterial lower respiratory infections	Sputa and endotracheal secretions	Human DNA depletion + Rapid PCR sequencing	6
Kafetzopoulou *et al.* [80]	Metagenomic analysis of viral infections and co-infections	Plasma and serum	Ligation sequencing and rapid sequencing	Not reported
Sanderson *et al.* [89]	Metagenomic sequencing from infected orthopaedic devices	Sonication fluid from explanted prostheses	Different approaches tested	4
Gong *et al.* [90]	Metagenomic analysis of liver abscess	Abscess aspirates	Ligation sequencing	Not reported

DNA via direct shotgun sequencing could be really challenging, and would require high sequencing depth, thus increasing the cost of the analysis [94, 95]. Although it is possible to enrich samples in particular fractions (i.e. differential centrifugation and filtration) or DNA fragments (i.e. PCR amplification and DNA hybridization) [96, 97], several factors should be taken into account when considering a fast, *in situ* application. Mainly, it would be especially difficult to adapt enrichment protocols to field conditions, and they could cause substantial losses of genetic material, add extra time to sample preparation, and result in a significant bias.

In this context, targeted or selective real-time sequencing—also known as 'Read Until'—is a new approach for focusing the sequencing process to specific DNA fragments of interest. Read Until is based on the ability of programming nanopore sequencers to reject individual DNA molecules while they are being read [98], releasing the individual nanopore to sequence another DNA fragment. ONT sequencing speed is estimated to be 450 bp/s [98–100], and it is relatively common to achieve sequences longer than 100 kbp [24, 101]. Theoretically, to discard a read for being read after a few seconds of translocation through the nanopore would prevent wasting sequencing capacity, which could be saved for sequencing targeted DNA fragments [99]. In a metagenomic context, the Read Until strategy could be used to deplete sequencing of undesirable DNA (i.e. host DNA) or for enriching specific genes/genomes. This depletion/enrichment procedures would not require any experimental steps, thus facilitating their use under field conditions.

Selective sequencing was first demonstrated by Loose *et al.* [102]. Later, Edwards *et al.* [103] showed the ability of Read Until strategies to enrich *E. coli* genomic sequences over human DNA. However, the actual revolution in targeted ONT sequencing is taking place in the recent months, with three different approaches being simultaneously released (Table 3). The first one, named BOSS-RUNS, introduced the dynamic selection of DNA regions of interest [100]. This method consists of focusing sequencing efforts on areas that have achieved low coverage during the run, thus leading to the compensation of sequencing bias. With this methodology, De Maio *et al.* [100] were able to effectively enrich multiple loci of interest within a bacterial genome, enabling up to 5-fold coverage improvement. In the field of metagenomics, BOSS-RUNS could be applied for improving the characterization of samples by ensuring the deep sequencing of clade-specific genetic markers [104]. On the other hand, Kovaka *et al.* [99] recently developed UNCALLED, a tool able to directly map ONT raw signals in order to detect wanted/unwanted sequences. They used this approach for sequencing a mock community (ZymoBIOMICS high molecular weight) containing seven bacteria and one yeast. The objective was to map the generated signals to a database containing the references for the bacterial genomes (29 Mbp), rejecting DNA strands when a match was detected. Bacterial sequencing depletion resulted

Table 3: Summary of Read-Until strategies developed for ONT sequencing

References	Method's name	Based on	Mapping algorithm/tool	Main objectives	Comments
Loose *et al.* [102]	–	Raw signal	Dynamic time warping	Enriching target regions of lambda virus genome, and obtaining uniform coverage for 11 different amplicons	Poor scalability
Edwards *et al.* [103]	RUBRIC	Sequence	LAST	Enriching *E. coli* genome (1%) in a sample dominated by human host DNA (99%)	Modest enrichment (15%)
De Maio *et al.* [100]	BOSS-RUNS	Sequence	NA	Enriching multi-locus regions in a bacterial genome, and retrieving uniform coverage in shotgun sequencing	Tested on simulated data. Limited scalability
Kovaka *et al.* [99]	UNCALLED	Raw signal	UNCALLED	Enriching a yeast genome by depleting bacterial genomes	Limited to mapping to non-repetitive references smaller than ∼100 Mbp. Implemented on CPU. Prior knowledge of the sample needed
Payne *et al.* [98]	–	Sequence	Minimap	Enriching a yeast genome, by dynamically adjusting genome coverage for every species in the mock community	No prior knowledge of the sample needed. Implemented only on GPU. Scalable to Gbp references.

in up to 4.46-fold of yeast genome enrichment. Finally, another strategy based on DNA sequences comparison has been proposed by Payne *et al.* [98]. In this work, the same ZymoBIOMICS mock community was used, but the enrichment of the yeast genome was achieved in a different way. Briefly, sequencing started with default parameters, but when a pre-defined coverage was reached for a specific microorganism, its genome sequence was given to the Read Until application in order to reject DNA strands coming from this microorganism. Interestingly, the pipeline was adapted to incorporate a metagenome classifier (Centrifuge) [41], allowing the use of this strategy without prior knowledge of the sample.

Overall, selective sequencing has proved useful for different metagenomic applications. Nevertheless, an associated reduced total yield per flowcell has been reported [98, 99]. This could be explained by two main reasons: (i) rejecting DNA strands increase the time that a nanopore is not reading a molecule and (ii) voltage changes needed for rejecting the fragments may produce pore blockages [98]. Nuclease flush could potentially help to overcome this situation, although current throughputs are enough for enriching DNA sequences and reducing the time needed to reach the desired coverage [98, 99], which is a key factor in many *in situ* applications.

Concluding remarks

In this work, we have reviewed the state-of-the-art, current research, and applications of real-time, *in situ* metagenomics. The spectacular development of metagenomic technologies in the last years as well as the number and importance of current and new challenges—including biomedical hazards—that could be addressed with portable metagenomic sequencing, reveals the importance of further developing this technology to match a variety of niches that we can, already, forecast. For example, we can envisage a close future in which microbial ecologists will be equipped with small, MinION-like devices that will allow to both extract DNA, carry out a fast sequencing, and yield the results in a very short time. The understandability of the results and the minimization of the—visible—bioinformatic background will be very important to allow non-specialized staff to use such portable devices. The recent COVID-19 outbreak as well as the surveillance of Ebola, Zika, and many other emergent diseases will need an army of—not necessarily specialized—detectors, for which easy-to-run, easy-to-understand platforms will be needed. Alternatively, raw sequencing data will have to be transmitted through secure Internet-based applications to centralized points, in which specialist staff will further process and finally analyse the information. Such portable, easy-to-use, cheap devices will be used in quality control of all sorts of foods and ingredients; in the identification of crop pathogens on an individual plant basis; in forensic investigations; in the assessment of the energetic potential of different substrates or batches for biogas production; or for the identification of the best soils for specific crops, as deduced by the soil microbial (either taxonomic or functional) profile. In order to meet all these possibilities (which we have ambitiously described in future and not conditional tense), the combination of five traits will have to take place. The *in situ*, portable platform of the future will (have to) be: inexpensive, robust, fast, easy to use, and connectable. A platform with these features will have a game-changing effect on the way we perform—and understand—microbial ecology.

Acknowledgements

The authors want to thank all their colleagues from Darwin Bioprospecting Excellence who contributed to the development of MinION sequencing protocols and field applications. Adriel Latorre is a recipient of a Doctorado Industrial fellowship from the Ministerio de Ciencia, Innovación y Universidades (reference DI-17-09613).

Conflict of interest statement. None declared.

References

1. Rappé MS, Giovannoni SJ. The uncultured microbial majority. *Annu Rev Microbiol* 2003;**57**:1–11

2. Sanger F, Air GM, Barrell BG *et al*. Nucleotide sequence of bacteriophage φX174 DNA. *Nature* 1977;**265**:687–95
3. Pace NR, Stahl DA, Lane DJ *et al*. The analysis of natural microbial populations by ribosomal RNA sequences. *Advances in Microbial Ecology*. Boston, MA: Springer, 1986, 1–55
4. Pace NR. Analyzing natural microbial populations by rRNA sequences. *ASM News* 1985;**51**:4–12
5. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Mol Biol Rev* 1995;**59**: 143–69
6. Handelsman J, Rondon MR, Brady SF *et al*. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 1998;**5**:R245–9
7. Quince C, Walker A, Simpson J *et al*. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017;**35**:833–44
8. Callahan B, McMurdie P, Holmes S. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 2017;**11**:2639–43
9. 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
10. Wu Y, Tang Y, Tringe S *et al*. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2014;**2**:26
11. Hiraoka S, Yang CC, Iwasaki W. Metagenomics and bioinformatics in microbial ecology: current status and beyond. *Microbes Environ* 2016;**31**:204–12
12. Rothberg JM, Hinz W, Rearick TM *et al*. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011;**475**:348–52
13. Merriman BRD, Team IT, Rothberg JM. Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis* 2012; **33**:3397–417
14. Levene MJ, Korlach J, Turner SW *et al*. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 2003;**299**:682–6
15. Buermans HPJ, Den Dunnen JT. Next generation sequencing technology: advances and applications. *Biochim Biophys Acta* 2014;**1842**:1932–41
16. Kasianowicz J, Brandin E, Branton D *et al*. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci* 1996;**93**:13770–3
17. Howorka S, Cheley S, Bayley H. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nat Biotechnol* 2001;**19**:636–9
18. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet* 2010;**19**:R227–40
19. Ip CL, Loose M, Tyson JR *et al*. MinION analysis and reference consortium: phase 1 data release and analysis. *F1000 Res* 2015;**4**:1075
20. Zhang Y, Liu C, Leung H *et al*. CONNET: accurate genome consensus in assembling nanopore sequencing data via deep learning. *iScience* 2020;**23**:101128
21. Jain M, Fiddes IT, Miga KH *et al*. Improved data analysis for the MinION nanopore sequencer. *Nat Methods* 2015;**12**: 351–6
22. Loman NJ, Watson M. Successful test launch for nanopore sequencing. *Nat Methods* 2015;**12**:303–4
23. Tyson JR, O'Neil NJ, Jain M *et al*. MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res* 2018;**28**:266–74.
24. Samarakoon H, Punchihewa S, Senanayake A *et al*. F5N: nanopore sequence analysis toolkit for android smartphones. *bioRxiv* 2020.
25. Quick J, Loman NJ, Duraffour S *et al*. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 2016;**530**: 228–32
26. Goordial J, Altshuler I, Hindson K *et al*. In situ field sequencing and life detection in remote (79 26′ N) Canadian high arctic permafrost ice wedge microbial communities. *Front Microbiol* 2017;**8**:2594
27. Castro-Wallace SL, Chiu CY, John KK *et al*. Nanopore DNA sequencing and genome assembly on the International Space Station. *Sci Rep* 2017;**7**:1–12
28. Harcourt J, Tamin A, Lu X *et al*. Isolation and characterization of SARS-CoV-2 from the first US COVID-19 patient. *bioRxiv* 2020.
29. Moore SC, Penrice-Randal R, Alruwaili M *et al*. Amplicon based MinION sequencing of SARS-CoV-2 and metagenomic characterisation of nasopharyngeal swabs from patients with COVID-19. *medRxiv* 2020.
30. Thompson L, Sanders J, McDonald D *et al*. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017;**551**:457–63
31. Alneberg J, Karlsson C, Divne A *et al*. Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* 2018; **6**:173
32. Alma'abadi A, Gojobori T, Mineta K. Marine mtagenome as A resource for novel enzymes. *Genom Proteom Bioinformat* 2015;**13**:290–5
33. Cuadrat R, Ionescu D, Dávila A *et al*. Recovering genomics clusters of secondary metabolites from lakes using genome-resolved metagenomics. *Front Microbiol* 2018;**9**:251
34. Pomerantz A, Peñafiel N, Arteaga A *et al*. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *GigaScience* 2018;**7**:giy033
35. Menegon M, Cantaloni C, Rodriguez-Prieto A *et al*. On site DNA barcoding by nanopore sequencing. *PLoS One* 2017;**12**: e0184741
36. Edwards A, Debbonaire A, Nicholls S *et al*. In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. *bioRxiv* 2016
37. Johnson S, Zaikova E, Goerlitz D *et al*. Real-time DNA sequencing in the Antarctic dry valleys using the Oxford nanopore sequencer. *J Biomol Tech* 2017;**28**:2–7
38. Gowers G, Vince O, Charles J *et al*. Entirely off-grid and solar-powered DNA sequencing of microbial communities during an ice cap traverse expedition. *Genes* 2019;**10**:902
39. Menzel P, Ng K, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* 2016;**7**: 11257
40. Edwards A, Soares A, Rassner S *et al*. Deep sequencing: intraterrestrial metagenomics illustrates the potential of off-grid Nanopore DNA sequencing. *bioRxiv* 2017.
41. Kim D, Song L, Breitwieser F *et al*. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016;**26**:1721–9
42. Burton A, Stahl S, John K *et al*. Off earth identification of bacterial populations using 16S rDNA nanopore sequencing. *Genes* 2020;**11**:76
43. Carr C, Bryan N, Saboda K *et al*. Nanopore sequencing at mars, Europa and microgravity conditions. *bioRxiv* 2020.

44. Urban L, Holzer A, Baronas J *et al*. Freshwater monitoring by nanopore sequencing. *bioRxiv* 2020.
45. Hamner S, Brown B, Hasan N *et al*. Metagenomic profiling of microbial pathogens in the little bighorn river. *Int J Environ Res Public Health* 2019;**16**:1097
46. Liem M, Regensburg-Tuïnk A, Henkel C *et al*. Microbial diversity characterization of seawater in a pilot study using Oxford Nanopore Technologies long-read sequencing. *bioRxiv* 2020.
47. Cáceres E, Lewis W, Homa F *et al*. Near-complete Lokiarchaeota genomes from complex environmental samples using long and short read metagenomic analyses. *bioRxiv* 2019.
48. Moss E, Maghini D, Bhatt A. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* 2020;**38**:701–7
49. Jonnala BY, McSweeney P, Sheehan J *et al*. Sequencing of the cheese microbiome and its relevance to industry. *Front Microbiol* 2018;**9**:1020
50. Morgan H, Du Toit M, Setati M. The grapevine and wine microbiome: insights from high-throughput Amplicon sequencing. *Front Microbiol* 2017;**8**:820
51. Jacoby R, Peukert M, Succurro A *et al*. The role of soil microorganisms in plant mineral nutrition—current knowledge and future directions. *Front Plant Sci* 2017;**8**:1617
52. Cassman N, Lourenço K, do Carmo J *et al*. Genome-resolved metagenomics of sugarcane vinasse bacteria. *Biotechnol Biofuels* 2018;**11**:270
53. Hardegen J, Latorre-Pérez A, Vilanova C *et al*. Methanogenic community shifts during the transition from sewage mono-digestion to co-digestion of grass biomass. *Bioresour Technol* 2018;**265**:275–81
54. Ramm P, Abendroth C, Latorre-Pérez A *et al*. Ammonia removal during leach-bed acidification leads to optimized organic acid production from chicken manure. *Renew Energy* 2020;**146**:1021–30
55. Abendroth C, Latorre-Pérez A, Porcar M *et al*. Shedding light on biogas: phototrophic biofilms in anaerobic digesters hold potential for improved biogas production. *Syst Appl Microbiol* 2020;**43**:126024
56. Schwan B, Abendroth C, Latorre-Pérez A *et al*. Chemically stressed bacterial communities in anaerobic digesters exhibit resilience and ecological flexibility. *Front Microbiol* 2020; **11**:867
57. Mallapaty S. How sewage could reveal true scale of coronavirus outbreak. *Nature* 2020;**580**:176–7
58. Hu Y, Ndegwa N, Alneberg J *et al*. Stationary and portable sequencing-based approaches for tracing wastewater contamination in urban stormwater systems. *Sci Rep* 2018;**8**: 11907
59. Che Y, Xia Y, Liu L *et al*. Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome* 2019;**7**:44
60. Acharya K, Khanal S, Pantha K *et al*. A comparative assessment of conventional and molecular methods, including MinION nanopore sequencing, for surveying water quality. *Sci Rep* 2019;**9**:15726
61. Winand R, Bogaerts B, Hoffman S *et al*. Targeting the 16S rRNA gene for bacterial identification in complex mixed samples: comparative evaluation of second (Illumina) and third (Oxford Nanopore Technologies) generation sequencing technologies. *Int J Mol Sci* 2019;**21**:298
62. Santos A, van Aerle R, Barrientos L *et al*. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Comput Struct Biotechnol J* 2020;**18**:296–305
63. Cuscó A, Catozzi C, Viñes J *et al*. Microbiota profiling with long amplicons using nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the Rrn operon. *F1000Res* 2019;**7**:1755
64. Benítez-Páez A, Sanz Y. Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION^{TM} portable nanopore sequencer. *GigaScience* 2017;**6**:1–12
65. Hu Y, Green G, Milgate A *et al*. Pathogen detection and microbiome analysis of infected wheat using a portable DNA Sequencer. *Phytobiomes J* 2019;**3**:92–101
66. Boykin L, Ghalab A, Rossitto De Marchi B *et al*. Real time portable genome sequencing for global food security. *F1000Res* 2018;**7**:1101
67. Gonçalves A, Collipal-Matamal R, Valenzuela-Muñoz V *et al*. Nanopore sequencing of microbial communities reveals the potential role of sea lice as a reservoir for fish pathogens. *Sci Rep* 2020;**10**:2895
68. Voorhuijzen-Harink M, Hagelaar R, van Dijk J *et al*. Toward on-site food authentication using nanopore sequencing. *Food Chem X* 2019;**2**:100035
69. Rossolini GM, Arena F, Pecile P *et al*. Update on the antibiotic resistance crisis. *Curr Opin Pharmacol* 2014;**18**:56–60
70. Koboldt DC, Steinberg KM, Larson DE *et al*. The next-generation sequencing revolution and its impact on genomics. *Cell* 2013;**155**:27–38
71. Wilson MR, Naccache SN, Samayoa E *et al*. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med* 2014;**370**:2408–17
72. Dickson RP, Erb-Downward JR, Prescott HC *et al*. Analysis of culture-dependent versus culture-independent techniques for identification of bacteria in clinically obtained bronchoalveolar lavage fluid. *J Clin Microbiol* 2014;**52**:3605–13
73. Hoenen T, Groseth A, Rosenke K *et al*. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerg Infect Dis* 2016;**22**:331–4
74. Quick J, Ashton P, Calus S *et al*. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol* 2015;**16**:114
75. Ashikawa S, Tarumoto N, Imai K *et al*. Rapid identification of pathogens from positive blood culture bottles with the MinION nanopore sequencer. *J Med Microbiol* 2018;**67**:1589–95
76. Lemon J, Khil P, Frank K *et al*. Rapid nanopore sequencing of plasmids and resistance gene detection in clinical isolates. *J Clin Microbiol* 2017;**55**:3530–43
77. Phelan J, O'Sullivan D, Machado D *et al*. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med* 2019;**11**:41
78. Greninger AL, Naccache SN, Federman S *et al*. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med* 2015;**7**:99
79. Batovska J, Lynch SE, Rodoni BC *et al*. Metagenomic arbovirus detection using MinION nanopore sequencing. *J Virol Methods* 2017;**249**:79–84
80. Kafetzopoulou LE, Efthymiadis K, Lewandowski K *et al*. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Eurosurveillance* 2018;**23**:1800228
81. Grumaz C, Hoffmann A, Vainshtein Y *et al*. Rapid next-generation sequencing–based diagnostics of bacteremia in septic patients. *J Mol Diagn* 2020;**22**:405–18

82. Leggett RM, Alcon-Giner C, Heavens D *et al.* Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. *Nat Microbiol* 2020;**5**:430–42
83. Mitsuhashi S, Kryukov K, Nakagawa S *et al.* A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Sci Rep* 2017;**7**:5657
84. Tanaka H, Matsuo Y, Nakagawa S *et al.* Real-time diagnostic analysis of MinION^{TM}-based metagenomic sequencing in clinical microbiology evaluation: a case report. *JA Clin Rep* 2019;**5**:24
85. Cheng J, Hu H, Kang Y *et al.* Identification of pathogens in culture-negative infective endocarditis cases by metagenomic analysis. *Ann Clin Microbiol Antimicrob* 2018;**17**:43
86. Pendleton KM, Erb-Downward JR, Bao Y *et al.* Rapid pathogen identification in bacterial pneumonia using real-time metagenomics. *Am J Respir Crit Care Med* 2017;**196**: 1610–2
87. Schmidt K, Mwaigwisya S, Crossman LC *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J Antimicrob Chemother* 2017;**72**:104–14
88. Charalampous T, Kay GL, Richardson H *et al.* Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat Biotechnol* 2019;**37**: 783–92
89. Sanderson ND, Street TL, Foster D *et al.* Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. *BMC Genom* 2018;**19**:714
90. Gong L, Huang YT, Wong CH *et al.* Culture-independent analysis of liver abscess using nanopore sequencing. *PLoS One* 2018;**13**:e0190853
91. Nakagawa S, Inoue S, Kryukov K *et al.* Rapid sequencing-based diagnosis of infectious bacterial species from meningitis patients in Zambia. *Clin Transl Immunol* 2019;**8**:e1087
92. Ibironke O, McGuinness L, Lu S *et al.* Species-level evaluation of the human respiratory microbiome. *GigaScience* 2020;**9**: giaa038
93. D'Andreano S, Cuscó A, Francino O. Rapid and real-time identification of fungi up to the species level with long amplicon Nanopore sequencing from clinical samples. *bioRxiv* 2020.
94. Nelson M, Pope C, Marsh R *et al.* Human and extracellular DNA depletion for metagenomic analysis of complex clinical infection samples yields optimized viable microbiome profiles. *Cell Rep* 2019;**26**:2227–40.e5
95. Feigelman R, Kahlert C, Baty F *et al.* Sputum DNA sequencing in cystic fibrosis: non-invasive access to the lung microbiome and to pathogen details. *Microbiome* 2017;**5**:20
96. O'Flaherty B, Li Y, Tao Y *et al.* Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing. *Genome Res* 2018;**28**:869–77
97. Lewandowski K, Xu Y, Pullan S *et al.* Metagenomic nanopore sequencing of influenza virus direct from clinical respiratory samples. *J Clin Microbiol* 2019;**58**:e00963–19
98. Payne A, Holmes N, Clarke T *et al.* Nanopore adaptive sequencing for mixed samples. Whole exome capture and targeted panels. *bioRxiv* 2020.
99. Kovaka S, Fan Y, Ni B *et al.* Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *bioRxiv* 2020.
100. De Maio N, Manser C, Munro R *et al.* BOSS-RUNS: a flexible and practical dynamic read sampling framework for nanopore sequencing. *bioRxiv* 2020.
101. Jain M, Olsen H, Turner D *et al.* Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* 2018;**36**: 321–3
102. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. *Nat Methods* 2016;**13**:751–4
103. Edwards H, Krishnakumar R, Sinha A *et al.* Real-time selective sequencing with RUBRIC: read until with basecall and reference-informed criteria. *Sci Rep* 2019;**9**:11475
104. Truong D, Franzosa E, Tickle T *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015;**12**:902–3

Appendix D

Other publications

scientific reports

OPEN

The Spanish gut microbiome reveals links between microorganisms and Mediterranean diet

Adriel Latorre-Pérez[1]✉, Marta Hernández[2]✉, Jose Ramón Iglesias[2], Javier Morán[3], Javier Pascual[1], Manuel Porcar[1,4], Cristina Vilanova[1] & Luis Collado[5]

Despite the increasing evidence of links between human gut and health, the number of gut microbiomes that have been studied to date at a country level are surprisingly low. Mediterranean countries, including some of the most long-lived and healthy countries in the world, have not been considered so far in those studies at a large scale. The main objective of this work is to characterize the gut microbiome of a healthy adult population of a Mediterranean, paradigmatically healthy country: Spain. Stool samples from 530 healthy volunteers were collected, total metagenomic DNA extracted, and the microbial profiles determined through 16S rRNA metataxonomic sequencing. Our results confirm the associations between several microbial markers and different variables, including sex, age, BMI and diet choices, and bring new insights into the relationship between microbiome and diet in the Spanish population. Remarkably, some of the associations found, such as the decrease of *Faecalibacterium* with age or the link of *Flavonifractor* with less healthy dietary habits, have been barely noticed in other large-scale cohorts. On the other hand, a range of links between microorganisms, diet, and lifestyle coincide with those reported in other populations, thus increasing the robustness of such associations and confirming the importance of these microbial markers across different countries. Overall, this study describes the Spanish "normal" microbiome, providing a solid baseline for future studies investigating the effects of gut microbiome composition and deviations in the adherence to the Mediterranean diet.

Human gut is one of the most diverse ecosystems on Earth. As a result of millions of years of co-evolution, gut microorganisms perform essential activities for human health and nutrition, from the digestion of vegetal fiber[1] to the regulation of complex signalling pathways acting beyond our gut[2,3]. Since the development of metagenomic sequencing techniques, the human gut microbiome has been a recurrent object of study. In 2007, The Human Microbiome Project (HMP) was launched with two main objectives: (1) understanding the dimension of the microbial communities associated to the human body, regarding variability among individuals; and (2) shedding light on the interplay between gut microbiota and a range of diseases[4,5]. To date, nearly 6.000 gut microbiome samples (out of more than 31.000 corresponding to different body sites) have been analyzed in the framework of the HMP. These datasets originate from individuals of different sex, age, culture, geographic location, and health status, which implies multiple potential connections between the composition of the gut microbiome and a range of health issues and diseases[6,7]. In order to shed light on those correlations, the microbiome profiles of different cohorts (usually, healthy vs diseased) are often compared, and differences involving single microorganisms[8,9], microbial consortia[10], or dynamic behaviors of the community are identified[11].

The definition of a "normal" (or "healthy") microbiome is crucial to understand how this microbiome is altered as a consequence of any factor. However, the systematic analysis of the microbiome of healthy individuals has only been addressed by a few studies, and the very concept of "normal" microbiome is still controversial[12,13]. Since a range of factors associated to climate, geography, and culture are known to influence the gut microbiome[14–16],

[1]Darwin Bioprospecting Excellence S.L., Paterna, Spain. [2]Instituto Central Lechera Asturiana para la Nutrición Personalizada (ICLANP), Siero, Spain. [3]Instituto de Innovación Alimentaria, Universidad Católica de Murcia, Murcia, Spain. [4]Institute for Integrative Systems Biology (I2SysBio), University of València-CSIC, Paterna, Spain. [5]Department of Medicine, Complutense University of Madrid, Madrid, Spain. ✉email: alatorre@darwinbioprospecting.com; marta.hernandez@capsa.es

Heitkamp *et al. Biotechnol Biofuels* (2021) 14:185
https://doi.org/10.1186/s13068-021-02034-5

Biotechnology for Biofuels

RESEARCH Open Access

Monitoring of seven industrial anaerobic digesters supplied with biochar

Kerstin Heitkamp[1†], Adriel Latorre-Pérez[2†], Sven Nefigmann[3], Helena Gimeno-Valero[2], Cristina Vilanova[2], Efri Jahmad[5] and Christian Abendroth[4,5*]

Abstract

Background: Recent research articles indicate that direct interspecies electron transfer (DIET) is an alternative metabolic route for methanogenic archaea that improves microbial methane productivity. It has been shown that multiple conductive materials such as biochar can be supplemented to anaerobic digesters to increase the rate of DIET. However, the industrial applicability, as well as the impact of such supplements on taxonomic profiles, has not been sufficiently assessed to date.

Results: Seven industrial biogas plants were upgraded with a shock charge of 1.8 kg biochar per ton of reactor content and then 1.8 kg per ton were added to the substrate for one year. A joint analysis for all seven systems showed a decreasing trend for the concentration of acetic acid ($p < 0.0001$), propionic acid ($p < 0.0001$) and butyric acid ($p = 0.0022$), which was significant in all cases. Quantification of the cofactor F420 using fluorescence microscopy showed a reduction in methanogenic archaea by up to a power of ten. Methanogenic archaea could grow within the biochar, even if the number of cells was 4 times less than in the surrounding sludge. 16S-rRNA gene amplicon sequencing showed a higher microbial diversity in the biochar particles than in the sludge, as well as an accumulation of secondary fermenters and halotolerant bacteria. Taxonomic profiles indicate microbial electroactivity, and show the frequent occurrence of *Methanoculleus*, which has not been described in this context before.

Conclusions: Our results shed light on the interplay between biochar particles and microbial communities in anaerobic digesters. Both the microbial diversity and the absolute frequency of the microorganisms involved were significantly changed between sludge samples and biochar particles. This is particularly important against the background of microbial process monitoring. In addition, it could be shown that biochar is suitable for reducing the content of inhibitory, volatile acids on an industrial scale.

Keywords: Anaerobic digestion, Biochar, DIET, Microbial communities

Background

Anaerobic digestion is a methane-yielding process carried out by a microbial biocenosis composed of bacteria and methanogenic archaea. Firstly, substrate is hydrolyzed by bacteria. Further degradation by acetogenic bacteria leads to the formation of mainly organic acids, alcohols, hydrogen, and carbon dioxide. Eventually, the aforementioned metabolites are transformed into acetate, hydrogen, and carbon dioxide during acetogenesis. Metabolites produced by acetogenic bacteria are transformed by methanogenic archaea into methane [1]. Methanogenesis is usually divided into three major pathways: acetoclastic, hydrogenotrophic and methylotrophic methanogenesis [2]. In all three pathways, acetate, formate, hydrogen and several methyl compounds (mono-, di- and trimethylamines) serve as electron carriers for

*Correspondence: christian.abendroth@tu-dresden.de
†Heitkamp Kerstin and Adriel Latorre-Pérez contributed equally to this work
[4] Institute of Waste Management and Circular Economy, Technische Universität Dresden, Pirna, Germany
Full list of author information is available at the end of the article

microbial biotechnology

Open Access

Extremophilic microbial communities on photovoltaic panel surfaces: a two-year study

Kristie Tanner,[1,2,†] Esther Molina-Menor,[2,†] Adriel Latorre-Pérez,[1] Àngela Vidal-Verdú,[2] Cristina Vilanova,[1] Juli Peretó[1,2,3] and Manuel Porcar[1,2]*

[1] *Darwin Bioprospecting Excellence S.L., Calle Catedrático Agustín Escardino 9, Paterna, 46980, Spain.*
[2] *Institute for Integrative Systems Biology I2SysBio, University of Valencia – CSIC, Catedrático José Beltrán 2, Paterna, 46980, Spain.*
[3] *Department of Biochemistry and Molecular Biology, University of Valencia, Dr. Moliner 50, Burjassot, 46100, Spain.*

Summary

Solar panel surfaces can be colonized by microorganisms adapted to desiccation, temperature fluctuations and solar radiation. Although the taxonomic and functional composition of these communities has been studied, the microbial colonization process remains unclear. In the present work, we have monitored this microbial colonization process during 24 months by performing weekly measurements of the photovoltaic efficiency, carrying out 16S rRNA gene high-throughput sequencing, and studying the effect of antimicrobial compounds on the composition of the microbial biocenosis. This is the first time a long-term study of the colonization process of solar panels has been performed, and our results reveal that species richness and biodiversity exhibit seasonal fluctuations and that there is a trend towards an increase or decrease of specialist (solar panel-adapted) and generalist taxa, respectively. On the former, extremophilic bacterial genera *Deinococcus*, *Hymenobacter* and *Roseomonas* and fungal *Neocatenulostroma*, *Symmetrospora* and *Sporobolomyces* tended to dominate the biocenosis; whereas *Lactobacillus* sp or *Stemphyllium* exhibited a decreasing trend. This profile was deeply altered by washing the panels with chemical agents (Virkon), but this did not lead to an increase of the solar panels efficiency. Our results show that solar panels are extreme environments that force the selection of a particular microbial community.

Introduction

Extreme environments are characterized by their strong selective pressures, which can include physical (i.e., temperature or radiation), geochemical (i.e., desiccation or salinity) and/or biological stresses (i.e., limited nutrient availability) (Lynn and Rocco, 2001). The microorganisms that inhabit these environments, known as extremophiles or extremotolerants, are selected due a variety of mechanisms, such as biofilm formation (Flemming *et al.*, 2016; Blanco *et al.*, 2019); the production of extremolytes and extremozymes (Gabani and Singh, 2013); or highly efficient DNA repair systems (Singh and Gabani, 2011). Microorganisms inhabiting extreme environments evolve faster than those inhabiting 'benign' environments, mainly due to the high mutation rates associated to stressful environmental conditions (Li *et al.*, 2014), and this could lead to these microorganisms being rich sources of new specialized metabolites (Sayed *et al.*, 2019).

A diversity of physical, geochemical and biological extremes (solar radiation, temperature fluctuations, desiccation and limited nutrient availability) concur on solar panel surfaces. A study performed on subaerial solar panel biofilms in São Paulo revealed that dust, pollen and other debris covering the solar panel surfaces accumulated in time and included abundant fungi and pigmented bacterial genera, and this was associated with a

Received 20 March, 2020; revised 2 June, 2020; accepted 9 June, 2020.

*For correspondence. E-mail manuel.porcar@uv.es; Tel. +34 628821813.

[†] These authors contributed equally to this work.

Microbial Biotechnology (2020) **0**(0), 1–12
doi:10.1111/1751-7915.13620

Funding information
Financial support from the Spanish Government (grant Helios. reference: BIO2015-66960-C3-1-R co-financed by FEDER funds and Ministerio de Economía y Competitividad) and from the European CSA on biological standardization BIOROBOOST (EU grant number 820699) are acknowledged. KT is a recipient of a Doctorado Industrial fellowship from the Ministerio de Ciencia, Innovación y Universidades (Spain), with reference DI-16-08976. AL is a recipient of a Doctorado Industrial fellowship from the Ministerio de Ciencia, Innovación y Universidades (Spain), with reference DI-17-09613. EMM is funded with a Formación de Profesorado Universitario (FPU) grant from the Spanish Government (Ministerio de Ciencia, Innovación y Universidades), with reference FPU17/04184. AVV is funded with a Formación de Profesorado Universitario (FPU) grant from the Spanish Government (Ministerio de Ciencia, Innovación y Universidades), with reference FPU18/02578. CV is a recipient of a Torres Quevedo fellowship from the Ministerio de Ciencia, Innovación y Universidades (Spain), with reference PTQ-16-08227.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 15 April 2020
doi: 10.3389/fmicb.2020.00645

Out of the Abyss: Genome and Metagenome Mining Reveals Unexpected Environmental Distribution of Abyssomicins

*Alba Iglesias[1], Adriel Latorre-Pérez[2], James E. M. Stach[1,3], Manuel Porcar[2,4] and Javier Pascual[2]**

[1] School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom, [2] Darwin Bioprospecting Excellence S.L., Paterna, Spain, [3] Centre for Synthetic Biology and the Bioeconomy, Newcastle University, Newcastle upon Tyne, United Kingdom, [4] Institute for Integrative Systems Biology (I^{2}SysBio), University of Valencia-CSIC, Paterna, Spain

OPEN ACCESS

Edited by:
John R. Battista,
Louisiana State University,
United States

Reviewed by:
Sergey B. Zotchev,
University of Vienna, Austria
Francisco (Paco) Barona-Gomez,
Center for Research and Advanced Studies of the National Polytechnic Institute, Mexico

****Correspondence:***
Javier Pascual
jpascual@darwinbioprospecting.com;
javier.pascual.fjpm@gmail.com

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *24 October 2019*
Accepted: *20 March 2020*
Published: *15 April 2020*

Citation:
Iglesias A, Latorre-Pérez A, Stach JEM, Porcar M and Pascual J (2020) Out of the Abyss: Genome and Metagenome Mining Reveals Unexpected Environmental Distribution of Abyssomicins. Front. Microbiol. 11:645. doi: 10.3389/fmicb.2020.00645

Natural products have traditionally been discovered through the screening of culturable microbial isolates from diverse environments. The sequencing revolution allowed the identification of dozens of biosynthetic gene clusters (BGCs) within single bacterial genomes, either from cultured or uncultured strains. However, we are still far from fully exploiting the microbial reservoir, as most of the species are non-model organisms with complex regulatory systems that can be recalcitrant to engineering approaches. Genomic and metagenomic data produced by laboratories worldwide covering the range of natural and artificial environments on Earth, are an invaluable source of raw information from which natural product biosynthesis can be accessed. In the present work, we describe the environmental distribution and evolution of the abyssomicin BGC through the analysis of publicly available genomic and metagenomic data. Our results demonstrate that the selection of a pathway-specific enzyme to direct genome mining is an excellent strategy; we identified 74 new Diels–Alderase homologs and unveiled a surprising prevalence of the abyssomicin BGC within terrestrial habitats, mainly soil and plant-associated. We also identified five complete and 12 partial new abyssomicin BGCs and 23 new potential abyssomicin BGCs. Our results strongly support the potential of genome and metagenome mining as a key preliminary tool to inform bioprospecting strategies aimed at the identification of new bioactive compounds such as -but not restricted to- abyssomicins.

Keywords: abyssomicins, genome mining, metagenome mining, bioprospecting, biosynthetic gene cluster distribution and evolution

INTRODUCTION

Natural products are the main source of pharmaceutically interesting biomolecules. In particular, the search of microbial specialized metabolites has yielded a broad range of chemical structures with bioactivities, from antibiotics or antimycotics to immunosuppressants and anticancer compounds. Among those, compounds featuring tetronate moieties are attractive due to their versatile biological

OPEN ACCESS

EMBO reports

Words, images and gender

Lessons from a survey on the public perception of synthetic biology and related disciplines

Manuel Porcar[1,2], Adriel Latorre-Pérez[2], Esther Molina-Menor[1] & Martí Domínguez[3]

The fast development of new research fields, such as genetic engineering or synthetic biology, is often met with public concerns or even resistance, the fate of genetically modified crops being a prime example. There are many factors at play that determine how laypeople perceive new technologies and a better understanding of these can help to inform debate. Foremost, however, it is necessary to obtain reliable information on public opinion of emerging technologies that have the potential to affect their lives. To this end, we conducted a survey to gauge public opinion on genetic engineering and biotechnology as part of a special exhibition at the CosmoCaixa Museum in Barcelona, Spain. The large sample size of 38,113 respondents allowed us to assess the effect of age, gender or education on the perception of three related terms: "biotechnology", "genetic engineering" and "synthetic biology". In addition, by randomly associating these terms with the image of either a male or a female scientist, we looked at the effect of gender on people's perception of these technologies. In short, two conclusions can be reached: the terms "biotechnology" and "genetic engineering" were preferred to "synthetic biology". Second, terms associated with an image of a female scientist were better rated compared to the same terms associated with a male researcher. These results show an interesting gender dimension of public perception of new technologies.

Public perception of biotechnology

Synthetic biology, genetic engineering and biotechnology are interrelated terms with blurred boundaries. Biotechnology uses living organisms, cells or cellular components to synthesize products for agriculture, medicine, industry and research and has been used for centuries, albeit unconsciously. Genetic engineering is one of the subdisciplines of biotechnology: it involves the manipulation of an organism's DNA sequence by addition, deletion or modification in order to expand the product range of biotechnology. While both generally are based on using organisms, genes or metabolic pathways from nature, synthetic biology aims to design novel artificial systems. Synthetic biology can thus be seen as both an extension of genetic engineering, as well as a new view on biotechnology by using engineering principles such as standardization, modularity or orthogonality [1].

> "There are many factors at play that determine how laypeople perceive new technologies and a better understanding of these can help to inform debate."

In the public eye, however, biotechnology, genetic engineering and synthetic biology are often reduced to genetically modified organisms (GMOs). This, combined with a critical perception of GMOs, has fuelled a generally negative attitude of biotechnology. The last Eurobarometer survey (2010) on GM food showed that only 5% of Europeans completely support it, 18% "tend to agree", but as much as 61% totally disagree, that is, are against GM food. Moreover, 83% of Europeans had not heard about synthetic biology before. The main concerns were the possible risks rather than potential benefits from these technologies (http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/index#p = 1&instruments = SPECIAL&search = 341). Indeed, genetic engineering is perceived with a higher degree of concern compared to other scientific fields [2].

> "... Generation T (2011–present), also known as Generation Alpha, is growing up with an iPad or a smartphone in their hand in front of a screen."

In relation to perceptions of gender, a number of recent studies have shown biases of how men and women are evaluated and perceived at work [3,4]. A randomized double-blind study of professors in biology, chemistry and physics showed that identical academic profiles were more positively evaluated when they belonged to a male student than a female student. The result of such biases is that women in academia have to work harder than their male peers to obtain the same recognition [5] and that males are often seen as more capable than women [6]. Just to highlight one common example of gender stereotyping, when using neutral or non-gender-specific language, people tend to assume that a specialist in question is a man [7].

The exhibition and the survey

The survey was carried out in the CosmoCaixa museum, a flagship science museum in Barcelona that is sponsored by La Caixa

[1] Institute for Integrative Systems Biology (I2SysBio), Universitat de València-CSIC, Valencia, Spain. E-mail: manuel.porcar@uv.es
[2] Darwin Bioprospecting Excellence SL, Valencia, Spain
[3] Language Theory and Communication Sciences Department (UV), Universitat de València, Valencia, Spain

DOI 10.15252/embr.201948401 | EMBO Reports (2019) 20: e48401 | Published online 26 June 2019

frontiers in Microbiology

ORIGINAL RESEARCH
published: 29 October 2021
doi: 10.3389/fmicb.2021.714110

Beyond Archaea: The Table Salt Bacteriome

Leila Satari[1], Alba Guillén[1], Adriel Latorre-Pérez[2] and Manuel Porcar[1,2]*

[1] Institute for Integrative Systems Biology (I2SysBio), Universitat de València-CSIC, Paterna, Spain, [2] Darwin Bioprospecting Excellence S.L., Parc Científic Universitat de València, Paterna, Spain

Commercial table salt is a condiment with food preservative properties by decreasing water activity and increasing osmotic pressure. Salt is also a source of halophilic bacteria and archaea. In the present research, the diversity of halotolerant and halophilic microorganisms was studied in six commercial table salts by culture-dependent and culture-independent techniques. Three table salts were obtained from marine origins: Atlantic Ocean, Mediterranean (Ibiza Island), and Odiel marshes (supermarket marine salt). Other salts supplemented with mineral and nutritional ingredients were also used: Himalayan pink, Hawaiian black, and one with dried vegetables known as Viking salt. The results of 16S rRNA gene sequencing reveal that the salts from marine origins display a similar archaeal taxonomy, but with significant variations among genera. Archaeal taxa *Halorubrum*, *Halobacterium*, *Hallobellus*, *Natronomonas*, *Haloplanus*, *Halonotius*, *Halomarina*, and *Haloarcula* were prevalent in those three marine salts. Furthermore, the most abundant archaeal genera present in all salts were *Natronomonas*, *Halolamina*, *Halonotius*, *Halapricum*, *Halobacterium*, *Haloarcula*, and uncultured *Halobacterales*. *Sulfitobacter* sp. was the most frequent bacteria, represented almost in all salts. Other genera such as *Bacillus*, *Enterococcus*, and *Flavobacterium* were the most frequent taxa in the Viking, Himalayan pink, and black salts, respectively. Interestingly, the genus *Salinibacter* was detected only in marine-originated salts. A collection of 76 halotolerant and halophilic bacterial and haloarchaeal species was set by culturing on different media with a broad range of salinity and nutrient composition. Comparing the results of 16S rRNA gene metataxonomic and culturomics revealed that culturable bacteria *Acinetobacter*, *Aquibacillus*, *Bacillus*, *Brevundimonas*, *Fictibacillus*, *Gracilibacillus*, *Halobacillus*, *Micrococcus*, *Oceanobacillus*, *Salibacterium*, *Salinibacter*, *Terribacillus*, *Thalassobacillus*, and also Archaea *Haloarcula*, *Halobacterium*, and *Halorubrum* were identified at least in one sample by both methods. Our results show that salts from marine origins are dominated by Archaea, whereas salts from other sources or salt supplemented with ingredients are dominated by bacteria.

Keywords: table salt microbiome, halotolerant bacteria, halophilic bacteria, haloarchaea, 16S rRNA gene sequencing analysis

OPEN ACCESS

Edited by:
Joaquin Bautista-Gallego,
University of Extremadura, Spain

Reviewed by:
Heng-Lin Cui,
Jiangsu University, China
Hua Xiang,
Institute of Microbiology, Chinese Academy of Sciences (CAS), China

***Correspondence:**
Manuel Porcar
manuel.porcar@uv.es

Specialty section:
This article was submitted to Food Microbiology, a section of the journal Frontiers in Microbiology

Received: 24 May 2021
Accepted: 27 September 2021
Published: 29 October 2021

Citation:
Satari L, Guillén A, Latorre-Pérez A and Porcar M (2021) Beyond Archaea: The Table Salt Bacteriome. Front. Microbiol. 12:714110. doi: 10.3389/fmicb.2021.714110

Received: 6 February 2018 | Revised: 23 March 2018 | Accepted: 7 April 2018

DOI: 10.1002/mbo3.648

ORIGINAL ARTICLE

WILEY MicrobiologyOpen Open Access

Thermoelectric heat exchange and growth regulation in a continuous yeast culture

Adriel Latorre-Pérez[1] | Cristina Vilanova[1] | José J. Alcaina[2] | Manuel Porcar[1,2]

[1]Darwin Bioprospecting Excellence SL, Paterna, Spain

[2]Biotechnology and Synthetic Biology Laboratory, I2SysBio (Institute for Integrative Systems Biology), University of Valencia-CSIC, Paterna, Spain

Correspondence
Manuel Porcar, Biotechnology and Synthetic Biology Laboratory, I2SysBio (Institute for Integrative Systems Biology), University of Valencia-CSIC, Paterna, Spain.
Email: manuel.porcar@uv.es

Funding information
University of Valencia, Grant/Award Number: CPI-13-128

Abstract

We have designed a thermoelectric heat exchanger (TEHE) for microbial fermentations that is able to produce electric power from a microbial continuous culture using the intrinsic heat generated by microbial growth. While the TEHE was connected, the system proved able to stably self-maintain both the temperature and the optical density of the culture. This paves the way toward a more sustainable operation of microbial fermentations, in which energy could be saved by converting part of the metabolic heat into usable electric power.

KEYWORDS

continuous culture, heat exchange, Peltier–Seebeck effect, power production, temperature regulation

1 | INTRODUCTION

A range of parameters such as temperature, pH, or substrate concentration need to be stable in order to sustain a suitable microbial growth and/or a stable biosynthesis of a bioproduct (Walker, 2000). Temperature strongly affects a range of fundamental cellular processes (Goldberg, 2003; Haas, 2010), and thus keeping a microbial culture in a suitable range of temperatures is of high importance in terms of strain performance (Amillastre, Aceves-Lara, Uribelarrea, Alfenore, & Guillouet, 2012). Large-scale growth of most microorganisms is accompanied by the production of heat (Brettel, Lamprecht, & Schaarschmidt, 1981), which, when large culture volumes are set, often results in an undesirable increase in the temperature of the batch culture that has to be alleviated through refrigeration (von Stockar & van der Wielen, 1997; Türker, 2004).

In a previous work, we described the first microbial thermoelectric cell (MTC), a system designed for batch cultures that allows the partial conversion of microbial metabolic heat into electricity. MTC is based on the Seebeck effect, a thermoelectric property that allows direct conversion of temperature differences to electricity voltage. Taking into account that microbial growth is mainly exothermic, theoretically it is possible to produce an electrical current with the generated metabolic heat by using a thermoelectric cell (Rodríguez-Barreiro, Abendroth, Vilanova, Moya, & Porcar, 2013). Nevertheless, a range of industrial fermentations are carried out in continuous culture, where stable cellular densities can be maintained during long periods thanks to the supply of fresh medium, which is introduced at a rate that is equal to the volume of product that is removed from the fermenter. In this work, we aimed at designing, constructing, and characterizing a continuous culture system in which temperature is automatically controlled and electric power is constantly obtained during all the fermentation process. To do that, we envisaged, constructed, and set in place a thermoelectric heat exchanger (hereafter called TEHE), a device also based on the Seebeck effect, which facilitates a fine control of temperature and fresh medium input—and thus microbial growth—while electric power is produced.

2 | MATERIALS AND METHODS

2.1 | Experimental set-up

A medium-scale continuous culture of budding yeast *Saccharomyces cerevisiae* strain D170 (kindly provided by Prof. Emilia Matallana, IATA, Valencia, Spain) in YPD medium supplemented with 18%

Received: 30 September 2021 | Accepted: 6 December 2021

DOI: 10.1002/mbo3.1259

ORIGINAL ARTICLE

MicrobiologyOpen Open Access WILEY

Living in a bottle: Bacteria from sediment-associated Mediterranean waste and potential growth on polyethylene terephthalate

Àngela Vidal-Verdú[1] | Adriel Latorre-Pérez[2] | Esther Molina-Menor[1] | Joaquin Baixeras[3] | Juli Peretó[1,2,4] | Manuel Porcar[1,2]

[1]Institute for Integrative Systems Biology (I2SysBio), University of Valencia-CSIC, Paterna, Spain

[2]Darwin Bioprospecting Excellence S.L., Paterna, Spain

[3]Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia, Paterna, Spain

[4]Department of Biochemistry and Molecular Biology, University of Valencia, Burjassot, Spain

Correspondence

Manuel Porcar, Institute for Integrative Systems Biology (I2SysBio), University of Valencia-CSIC, Catedrático José Beltrán, 2, Paterna 46980, Spain.
Email: manuel.porcar@uv.es

Funding information

Formación del Profesorado Universitario, Grant/Award Numbers: FPU17/04184, FPU18/02578; Ministerio de Ciencia e Innovación, Grant/Award Numbers: DI-17-09613, RTI2018-095584-B-C41-42-43-44; H2020 Environment, Grant/Award Numbers: 210491758, 101000470; Agencia Estatal de la Innovación AEI, Grant/Award Number: PCI2019-111845-2; Agència Valenciana de la Innovación AVI, Grant/Award Number: INNEST/2021/334

Abstract

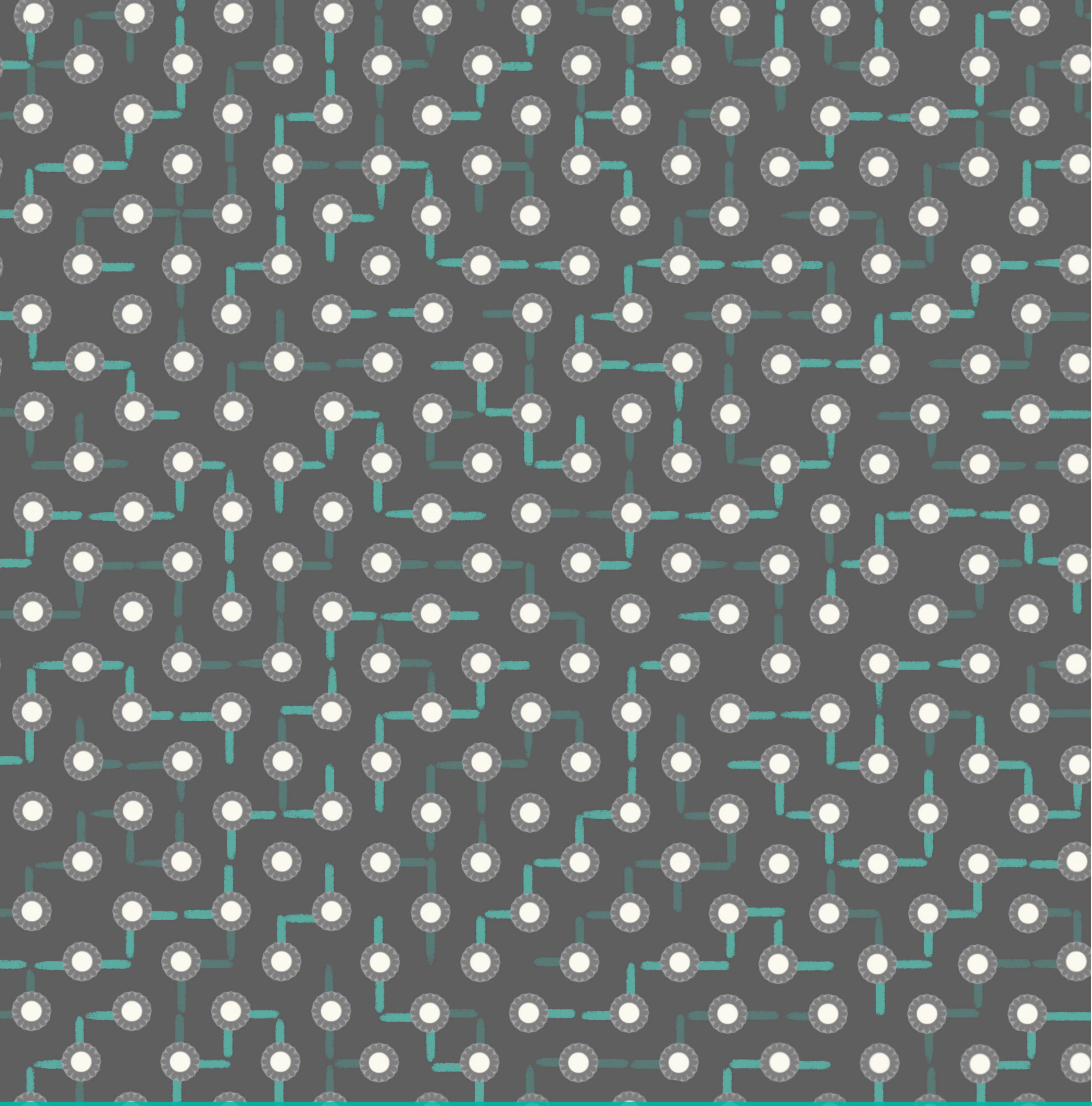
Ocean pollution is a worldwide environmental challenge that could be partially tackled through microbial applications. To shed light on the diversity and applications of the bacterial communities that inhabit the sediments trapped in artificial containers, we analyzed residues (polyethylene terephthalate [PET] bottles and aluminum cans) collected from the Mediterranean Sea by scanning electron microscopy and next generation sequencing. Moreover, we set a collection of culturable bacteria from the plastisphere that were screened for their ability to use PET as a carbon source. Our results reveal that *Proteobacteria* are the predominant phylum in all the samples and that *Rhodobacteraceae*, *Woeseia*, *Actinomarinales*, or *Vibrio* are also abundant in these residues. Moreover, we identified marine isolates with enhanced growth in the presence of PET: *Aquimarina intermedia*, *Citricoccus* spp., and *Micrococcus* spp. Our results suggest that the marine environment is a source of biotechnologically promising bacterial isolates that may use PET or PET additives as carbon sources.

KEYWORDS

bioprospecting, bioremediation, marine sediments, marine waste, plastic-degrading microorganisms, polyethylene terephthalate

1 | INTRODUCTION

Plastic production and, subsequently, plastic waste have increased exponentially through the last decades (Worm et al., 2017). The poor management of these residues, and their resistance to natural degradation (in some cases it comprises from hundreds to thousands of years) (Barnes et al., 2009), has resulted in a major, worldwide problem of plastic accumulation in all ecosystems on Earth. Even though the amount of recycled plastic has doubled from 2006 to 2018, the amount of post-consumer waste plastic that is sent to landfills in Europe was still 25% in 2018 (PlasticsEurope, 2020).



Universitat de València

Codirectores: Dra. Cristina Vilanova Serrador y Dr. Manuel Porcar Miralles

Valencia, Enero 2022