UNIVERSITA’ DEGLI STUDI DI VERONA

DEPARTMENT OF

Biotechnology

GRADUATE SCHOOL OF

Natural Sciences and Engineering

DOCTORAL PROGRAM IN

Biotechnology

XXXIII cycle

TITLE OF THE DOCTORAL THESIS

Development of novel bioinformatic pipelines for MinION-based DNA barcoding

S.S.D. BIO/18

Coordinator: Prof. Matteo Ballottari

Tutor: Prof. Massimo Delledonne

Doctoral Student: Simone Maestri

Development of novel bioinformatic pipelines for MinION-based DNA barcoding
Simone Maestri

Tesi di Dottorato
Verona, 29/04/2021

Abstract

DNA-barcoding is the process of taxonomic identification based on the sequence of a marker gene. When complex samples are analysed, we refer in particular to meta-barcoding. Barcoding has traditionally been performed with Sanger sequencing platform. The emergence of second-generation sequencing platforms, mainly represented by Illumina, enabled the high-throughput sequencing of hundreds of samples, and allowed the characterization of complex samples through meta-barcoding experiments. However, fragments sequenced with the Illumina platform are shorter than 600 bp, and this greatly limits taxonomic resolution of closely related species. Moreover, both these platforms suffer of long turnaround time, since they require shipping the samples to a sequencing facility, and complex regulations may hamper the export of material out of the country of origin. More recently, Oxford Nanopore Technologies provided the MinION, a portable and cheap third-generation sequencer, which has the potential of overcoming issues of currently available platforms, thanks to the production of long sequencing reads. However, MinION reads suffer of high error rate, therefore suitable analysis pipelines are needed to overcome this issue.

In this thesis I describe the development of bioinformatic pipelines for MinION-based DNA barcoding. Starting from the analysis of single samples, I show how improvements both in sequencing chemistry and in software now allow obtaining consensus sequences directly in the field, with accuracy comparable with Sanger. Conversely, when analysing complex samples, sequencing reads cannot be collapsed for reducing the error rate. However, bioinformatic approaches exploiting increased read length largely compensate the higher error rate, resulting in high correlation between MinION and Illumina up to genus level, and a more marked sensitivity of MinION platform to detect spiked-in indicator species.

In conclusion, the results presented in this thesis show that bioinformatic pipelines for the analysis of MinION reads can largely mitigate platform issues, paving the way for this platform to become the gold-standard for barcoding in the near future.

Index

Introduction

Laying the foundations of barcoding

The concept of barcoding was first introduced by Hebert and colleagues, who proposed the adoption of genetic markers known as taxon ‘barcodes’ to aid in species identification; the mitochondrial cytochrome c oxidase subunit 1 (CO1) gene was first proposed as a global bioidentification system for animals [1]. The adoption of DNA sequences as a basis for taxonomic identification, through the definition of Operational Taxonomy Units (OTUs), aided the application of standardized protocols for the comparison of results among studies [2]. Subsequent analyses on a widespread group of taxa in the animal kingdom showed that more than 98% of congeneric species pairs showed greater than 2% CO1 sequence divergence (**Figure 1**), demonstrating the usefulness of the proposed barcoding system [3].

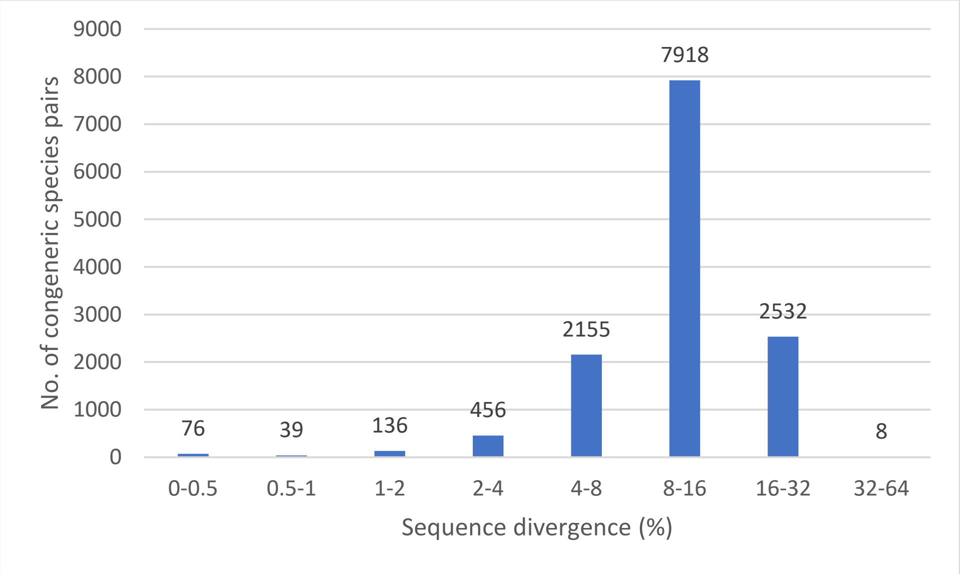


Figure 1: **CO1 sequence divergence in congeneric species pairs.** Data taken from [3].

In 2008, the international Barcode Of Life (iBOL) project was launched to transform biodiversity science by building the DNA barcode reference libraries, the sequencing facilities, the informatics platforms, the analytical protocols, and the international collaboration required to inventory and assess biodiversity (https://ibol.org/).

Different barcode marker genes were established across the tree of life, as the nuclear ribosomal internal transcribed spacer (ITS) gene for Fungi [4,5], or the 16S ribosomal

RNA gene for Bacteria [6-8]. Taxonomists involved in the classification of fungal and bacterial communities defined conventions based on sequence identity of the marker gene, setting the maximum divergence within an OTU at 2% and 3% respectively [2].

The sequencing of barcodes not only helps filling a gap in knowledge, but also offers a valuable source of information for decision making processes related to very diverse issues [5,9-11]. In the following sections, different applications of barcoding systems will be presented, together with the sequencing platforms currently used for this task. A glimpse at future perspectives will finally delineate the upcoming directions of barcoding.

Application of barcoding systems

Environmental changes and human activities are affecting the structural and functional role of ecosystems and may constitute a potential risk to human health [12]. The modification and destruction of natural habitats by humans have placed a wide variety of organisms at risk, resulting in what the scientific community refers to as the sixth great mass extinction [13]. While about 2 million species have been described to date, there are an estimated 5–30 million species in total on the planet [14]. This knowledge gap implies that our understanding of biodiversity is incomplete, and barcoding systems will have a fundamental role in speeding up the process of describing the complexity and diversity of nature, before the extinction of many species occurs [13]. Although the loss of biodiversity is global, the geographic patterns of species loss are non-random, with the most marked decline observed in tropical areas [13,15]. A necessary step to implement new conservation programs is training local science educators and conservationists in areas in which research funding and infrastructure are lacking [16].

Aside from enhancing our understanding of biodiversity [17], barcoding may help quickly identifying risks for human health, for example caused by food mislabelling [18,19], parasitic infections [20], invasive alien species [21,22] or shifts in the water microbiome composition [7,9]. In particular, seafood mislabelling may represent more than a consumer's fraud, and may affect food safety when toxic and unpalatable species enter the market by relabelling them as palatable species; moreover, globalization has led to an increased demand of seafood, creating incentives for seafood fraud [19]. Despite being one of the most abundant group of metazoan organisms, it is estimated that less than 4% of nematode species are currently known to science. Since easily distinguishable morphological characters are scarce in nematodes, their genetic identification is becoming increasingly important [20]. The World Health Organization estimates that worldwide infections with soil-transmitted nematodes cause a human annual disease burden of 3.8 million years lost to disabilities (YLD) [23]. The introduction of insect pests considered invasive alien species (IAS) into a non-native range threatens native plant health, and is estimated to have a negative impact on Canada's forests second only to wildfires in their effect [21]. The United Nations estimate that 1.8 billion people are still exposed to drinking water sources contaminated with faecal matter [24]. The current standard method for microbial examination of both drinking and bathing water requires the isolation and enumeration of organisms that indicate the presence of faecal contamination, such as *Escherichia coli* and *Enterococci*. Culture dependent approaches may require long incubation periods, and there is a demand for more rapid and comprehensive screening methods to detect faecal indicator organisms and putative pathogens in water samples [9].

The sequencing of bulk community samples extends the concept of barcoding, and is usually referred to as meta-barcoding [25] (**Figure 2**). In recognition of the need to consider the biological community as a whole, there has been a shift in emphasis from studies that focus on single indicator taxa to comparative studies across multiple taxa [25]. For example, marine researchers have proposed the use of the Autonomous Reef Monitoring Structure (ARMS), a standardized sampling tool that enables comprehensive documentation of marine biodiversity beyond standard indicator species [26]. ARMS are designed to mimic the structural complexity of coral reefs and are commonly deployed on the marine benthos for a length of time to allow marine organisms to colonize before subsequent retrieval and DNA sequencing (**Figure 3**).

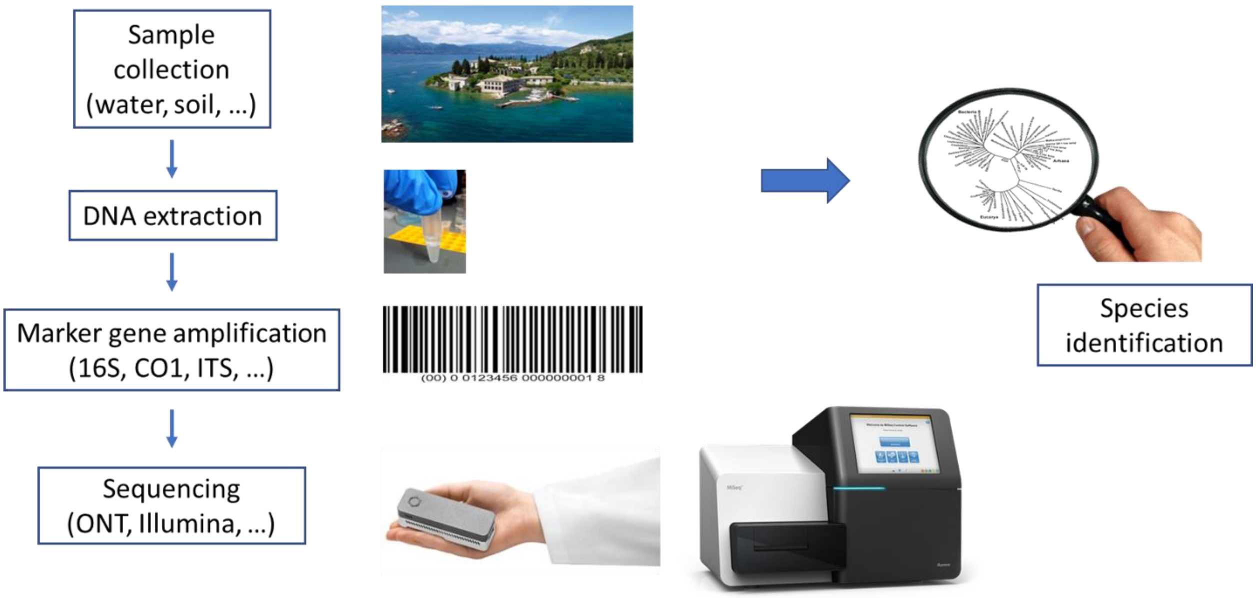

Sample collection (water, soil, ...)
DNA extraction
Marker gene amplification (16S, CO1, ITS, ...)
Sequencing (ONT, Illumina, ...)
(00) 0 0123456 000000001 8
Species identification

Figure 2: General overview of a meta-barcoding workflow. After sample collection, the DNA is extracted, and a marker gene is amplified and sequenced, in order to identify species present in the sample. 16S = 16S ribosomal RNA gene; CO1 = cytochrome oxidase subunit 1 gene; ITS = nuclear ribosomal internal transcribed spacer gene; ONT = Oxford Nanopore Technologies.

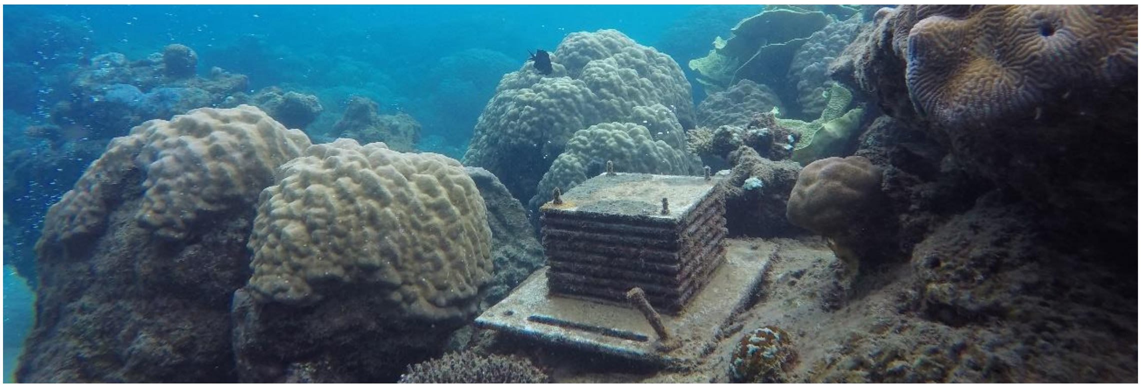


Figure 3: An Autonomous Reef Monitoring Structure. Taken from https://www.oceanarms.org/.

All these manuscripts prove that, since its conceptual definition at the beginning of the 21st century, barcoding has grown in importance, becoming a fundamental tool for monitoring the health of various ecosystems.

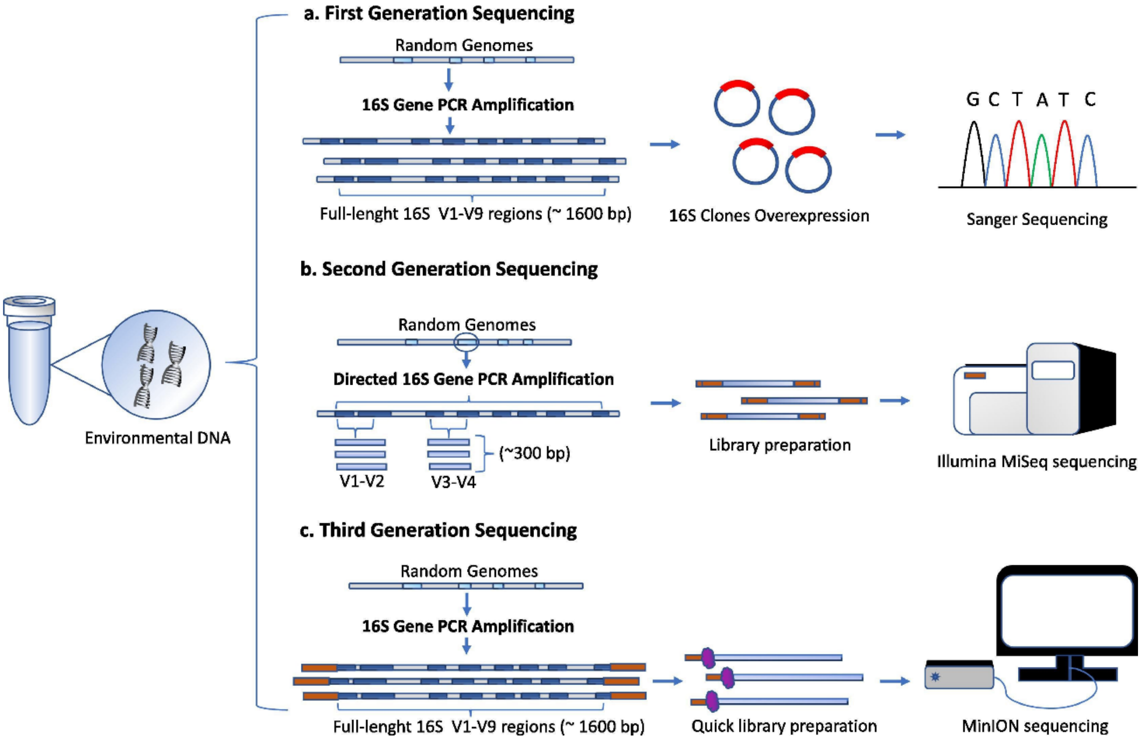
Sequencing platforms for barcoding

Barcoding has traditionally been performed with Sanger, known as first generation sequencing, and this platform still represents the gold-standard for studying one organism at a time [10,14]. The analysis of complex microbial communities with Sanger platform, without the need for cultivation, was first described by Pace [27]. However, this approach required a cloning step for analysing individual sequences, making it very cumbersome and time-consuming [8] (**Figure 4a**).

In recent years, DNA barcoding has taken advantage from the emergence of second-generation sequencing platforms, mainly represented by Illumina, that enabled the parallel generation of barcodes for hundreds of specimens and greatly simplified meta-barcoding studies for characterizing the diversity of entire ecosystems [25]. These platforms limit the fragment length of barcode sequences to a maximum of 600 bp when Illumina MiSeq is used, requiring the amplification of hypervariable regions within barcoding genes (**Figure 4b**). For large-scale monitoring projects that use Illumina HiSeq or NovaSeq, even shorter barcodes must be used [28]. The use of short sequences limits taxonomic resolution and phylogenetic information content, making them unsuitable to reach species-level classification [25]. Bioinformatic pipelines for Illumina meta-barcoding are now largely standardized, and typically include the merging of overlapping reads from the same fragment, clustering of merged reads to obtain a set of representative sequences with the corresponding counts and, finally, taxonomic assignment of representative sequences against a reference database [8] (**Figure 5**). Based on a plugin architecture, the QIIME2 framework is largely contributing to this standardization and increased reproducibility of the analyses [29].

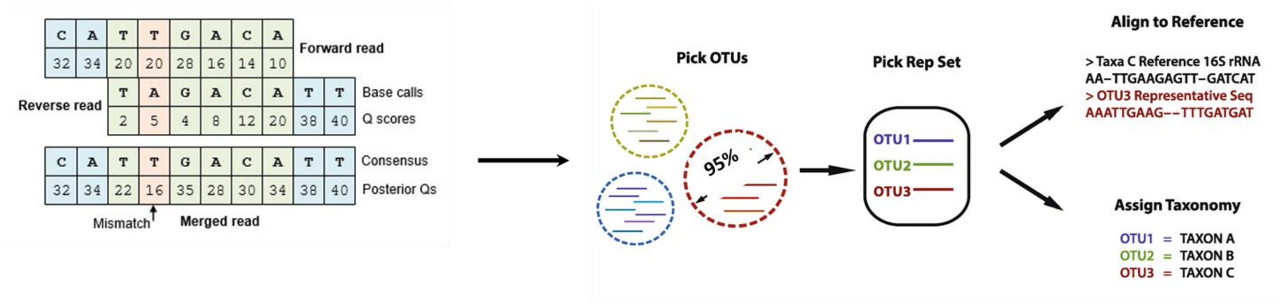
Sanger and Illumina sequencing platforms suffer of long turnaround time and the need for a dedicated infrastructure, greatly limiting their applicability in developing countries [28]. Moreover, complex regulations may impede biological research in biodiverse countries and can make it challenging to export material out of the country of origin [14]. More recently, Oxford Nanopore Technologies (ONT) developed the MinION, a portable and cheap third-generation sequencer, which allows the production of long sequencing reads directly in the field [10] (**Figure 6**). In principle, the MinION has the potential of overcoming the presented issues and become the gold-standard platform for barcoding

[16,28,30] (**Figure 4c**). However, MinION reads have high error rate that require *ad hoc* analysis pipelines. The development of novel bioinformatic pipelines, together with improvements in sequencing chemistry, represents a fundamental step to translate proof-of-principle workflows into standardized systems for on-site sequencing by professional users [10].



a. First Generation Sequencing
Random Genomes
16S Gene PCR Amplification
Full-lenght 16S V1-V9 regions (~ 1600 bp)
16S Clones Overexpression
G C T A T C
Sanger Sequencing
b. Second Generation Sequencing
Random Genomes
Directed 16S Gene PCR Amplification
(~300 bp)
V1-V2
V3-V4
Library preparation
Illumina MiSeq sequencing
Environmental DNA
c. Third Generation Sequencing
Random Genomes
16S Gene PCR Amplification
Full-lenght 16S V1-V9 regions (~ 1600 bp)
Quick library preparation
MinION sequencing

Figure 4. Sequencing platforms for barcoding. Depending on the sequencing technology adopted, a different library preparation for target sequencing is performed. As an example, the 16S sequencing of bacteria from environmental DNA is shown. Taken from [8].



	C	A	T	T	G	A	C	A			Forward read
	32	34	20	20	28	16	14	10			
Reverse read			T	A	G	A	C	A	T	T	Base calls
			2	5	4	8	12	20	38	40	Q scores
	C	A	T	T	G	A	C	A	T	T	Consensus
	32	34	22	16	35	28	30	34	38	40	Posterior Qs

Figure 5. Schematic overview of bioinformatic pipelines for Illumina meta-barcoding. Forward and reverse overlapping reads from the same fragment are merged, and clustered to obtain OTUs. Representative sequences from each OTU are then aligned to a reference sequenced and assigned a taxonomy.

Figure 6. The ONT MinION sequencer in the field. Taken from [14].

Bioinformatic approaches for barcoding with Nanopore sequencing

Nanopore sequencers identify DNA bases by measuring the changes in electrical conductivity generated while DNA strands pass through a biological pore and record the magnitude of the current in the nanopore in *fast5* format. A MinION flow-cell includes 2048 pores, connected to 512 channels, which are capable of sequencing in parallel [31]. A part from the sequencing chemistry, the base-caller, namely the software converting raw electric signal to a sequence of nucleotides, has a great impact on the sequencing accuracy [32]. Despite the great improvements occurred in the last years, MinION reads still suffer of error rate in the range of 5%-15% [10,33,34].

Compared with Illumina platform, there is a scarcity of bioinformatic approaches specifically developed for barcoding with MinION platform. However, some of the available approaches can be adapted to work with MinION reads, by adjusting software parameters [8]. The choice of the most suitable bioinformatic approach should consider many aspects, as the complexity of the sample under study, the quality of sequencing reads, the availability of *a-priori* knowledge and the requested computational resources.

Barcoding approaches for studying single samples aim to obtain a consensus sequence by integrating the information of multiple sequencing reads. These are either based on the alignment of reads from a resequencing experiment to a reference sequence, followed by variant calling and consensus sequence generation, or by *de novo* assembly of sequencing reads. The term *de novo* assembly is here used in a broad sense, since *de novo* assembly is usually meant as the production of longer contigs by joining partially overlapping reads [35]; while reads originating from amplicon sequencing experiments don't need to be assembled into longer contigs, but they only need to be collapsed to obtain a more accurate consensus sequence (**Figure 7**). This is the process of read error correction that is performed by assemblers such as Canu and Falcon [36,37] which, however, do not provide a single corrected consensus sequence. In particular, if some contaminant reads are generated due to nonspecific amplification, the corresponding corrected reads are going to be produced as well, making it nontrivial to pick one of those, representing the target amplification product.

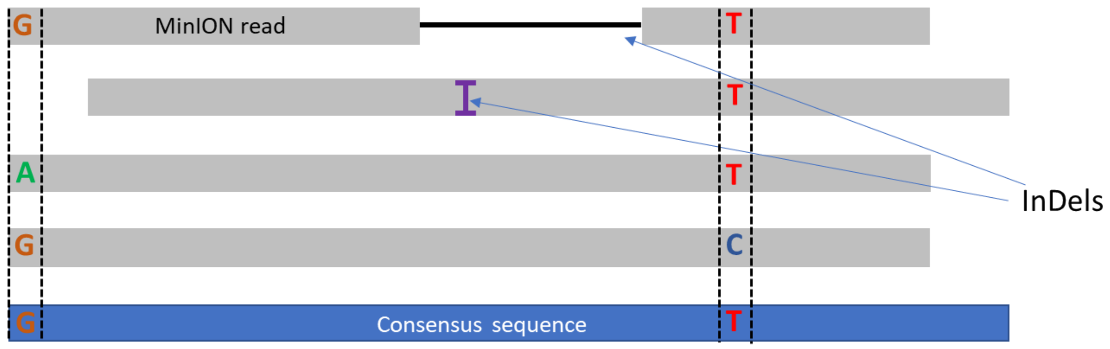


Figure 7. MinION reads from a single sample can be aligned and collapsed to obtain an accurate consensus sequence.

Clustering software can be applied to solve the contamination issue, by grouping reads similar to each other, and providing a consensus sequence for each cluster. However, since most clustering tools have been developed for Illumina platform, they struggle with high error rate, and fail to provide accurate consensus sequences. In particular, VSEARCH [38] uses the center-star method for multiple sequence alignment and subsequent consensus sequence computation, by aligning all the sequences in a cluster to the centroid sequence. For this reason, InDels in the other sequences relative to the centroid sequence will have little or no impact on the consensus sequence. As an alternative to *de novo* assemblers, multiple sequence aligners as MAFFT [39] can be applied to align a set of reads to each other, but the aligned reads should be accurately preprocessed, in order to remove contaminant reads. In fact, MAFFT assumes that the input sequences are all homologous and, accordingly, all the letters in the input data are aligned [39]. Known bias of Nanopore reads, which tend to underestimate the length of homopolymers, should be taken into account as well [32]. Moreover, multiple sequence aligners may lack of a consensus module, providing a consensus sequence from the alignment, and tools as EMBOSS [40] should be applied afterwards. Finally, tools specialized at correcting residual errors, a process known as polishing, may be applied to improve the consensus sequence accuracy, with the most frequently used software being Racon [41], Nanopolish (https://github.com/jts/nanopolish), and Medaka (https://github.com/nanoporetech/medaka). Racon is a graph-based method which operates on base-called data. Similarly, Medaka operates on base-called data, but uses neural networks applied to a pileup of individual sequencing reads against a draft assembly. Conversely, Nanopolish calculates an improved consensus sequence for a draft genome assembly by directly exploiting information stored in raw *fast5* files.

Meta-barcoding approaches generally aim to obtain a taxonomic characterization of complex samples, consisting of a list of taxa detected in the sample and their relative abundance estimates. Given the difficulty of clustering noisy reads from a complex sample, correctly discriminating between sequencing errors and actual biological differences, the most frequently adopted approaches are based on the taxonomic classification of single reads [8]. At this aim, reads are either mapped or aligned to a reference database. Although the two terms are often used in an interchangeable way, mapping of sequences to taxa does not necessarily provide a base-to-base alignment and can be achieved using k-mer-based or FM-based indexing schemes. These strategies are implemented in tools such as Centrifuge [42] and Kraken 2 [43], which rely on indexed databases. To classify a sequence, Kraken 2 first extracts all k-mers in a sequence, namely all subsequences of fixed length 'k', and assigns them to the lowest common ancestor (LCA) of the nodes in the database that contain that k-mer. After all k-mers in a sequence are mapped, a weight is assigned to each node, corresponding to the number of k-mers that mapped to it. Finally, the score of each path in the classification tree is evaluated, and the query sequence is assigned the taxonomy of the node with the highest score [44] (**Figure 8**). Although usually quicker, these methods do not provide alignment identities, and their output may be less intuitive for an effective interpretation of the results. Conversely, aligners such as Blast [45], Minimap2 [46], LAST [47] and VSEARCH [38] may be used to align each sequence to a reference database, and the alignment file may be post-processed to retain alignments with a minimum length and percentage identity or to deal with multi-mapping reads. To classify a sequence, VSEARCH uses a fast heuristic based on k-mers shared by the query and target sequences, in order to quickly identify similar sequences. It then performs optimal global sequence alignment of the query with the most promising candidates [38].

Finally, the composition of the reference database is another critical aspect, since it strongly influences the percentage of sequences correctly assigned to different taxonomic levels [8]. For example, the most frequently used databases for classification of 16S bacterial sequences are NCBI 16S Bacterial, GreenGenes 13_8_99 and Silva

132_99_16S, which differ in the number of reference sequences by more than an order of magnitude (**Table 1**). This difference in size may largely affect both computational processing time and classification accuracy. In fact, classification of a taxon not yet reported in the database represents a challenging scenario for the taxonomic classifier; ideally, the classifier should be able to identify the nearest taxonomic lineage to which this taxon belongs, but no further [48].

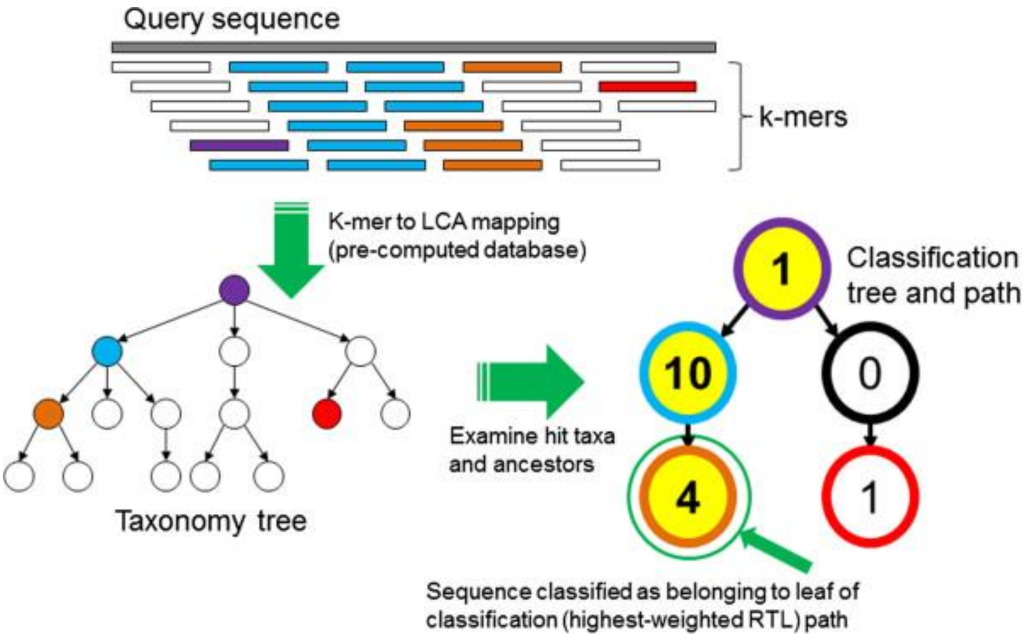


Figure 8. The Kraken 2 sequence classification algorithm. A query sequence is classified based on k-mers shared with nodes of a taxonomy tree. LCA = Lowest Common Ancestor. Taken from [44].

Table 1. Most frequently used databases for 16S meta-barcoding analyses.

Database name	**Number of sequences**
NCBI 16S Bacterial	20,160
GreenGenes 13_8_99	203,452
Silva_132_99_16S	369,953

The tools reported in this paragraph show that, although many software packages targeting specific tasks have been developed, bioinformatic pipelines for obtaining the desired results starting from raw data, and any best practice guidelines, are still lacking.

Aim of the thesis

The MinION platform has the potential to become the gold-standard for barcoding thanks to its portability and limited costs, but it produces sequencing reads with high error rate. Tools for performing specific tasks are available, however the most suitable ones and the optimal parameter settings for this application are yet to be identified. On top of that, the available tools need to be run sequentially, while streamlined pipelines providing results starting from raw data would be largely beneficial to non-expert users. This thesis describes bioinformatic approaches aimed at reducing the impact of error rate on species identification, for characterizing the taxonomy of single or complex samples with accuracy comparable with current state-of-the-art barcoding platforms.

Materials and Methods

Samples collection and data generation

The samples sequenced with ONT R7.3 chemistry were either provided by the Trento Science Museum (MUSE) or collected in a rainforest of central-south Tanzania. DNA extraction was performed as described in [49], and then PCR amplification was performed with different primer sets. In particular, for *Leptopelis vermiculatus*, primers Amp-P3 F 5′-CAATACCAAACCCCCTTRTTYGTWTGATC-3′ and Amp-P3 R 5′-GCTTCTCARATAATAAATATYAT-3′ [50] were used to target ~900 bp of CO1 gene. For *Rhynchocyon udzungwensis*, primers LCO1490 5'-GGTCAACAAATCATAAAGATATTGG-3' and HC02198 5′-TAAACTTCAGGGTGACCAAAAAATCA-3′ [51] were used to target ~710 bp of CO1 gene. For remaining vertebrates, primers 16S Sar 5'-CGCCTGTTTATCAAAAACAT-3 and 16S 5' -CCGGTTTGAACTCAGATCA-3' [52] were used to target ~600 bp of 16S gene. Library preparation was performed using MAP-006 kit, according to manufacturer's instructions.

The samples sequenced with ONT R9.4 chemistry were collected in Ulu Temburong National Park (Brunei, Borneo), during an expedition organized by Taxon Expeditions (https://taxonexpeditions.com/). DNA extraction was performed as described in [10], and then PCR amplification was performed with universal LCO1490 and HC02198 primers. Library preparation was performed using SQK-LSK108 kit, according to manufacturer's instructions.

A total of 9 water microbial samples were collected from River Tiber in Rome. Water filtration was performed using 0.45 µm filters by Istituto Superiore di Sanità (ISS). Bacterial DNA was extracted from filters using the DNeasy PowerWater Kit by QIAGEN, according to manufacturer's instructions. DNA samples coming from each condition (environmental sample, environment sample + spike-in$_{min}$, environmental sample + spile-in$_{max}$) were pooled together in groups of 3, to obtain 3 final pools. Each pool was subjected to library preparation and sequencing with ONT and Illumina platforms. For ONT sequencing, PCR primers 27F 5'-AGAGTTTGATCCTGGCTCAG-3' and 1492R 5'- GGTTACCTTGTTACGACTT-3' were used to amplify the full 16S gene (~1500 bp). Libraries were prepared using SQK-

RAB201 kit, according to manufacturer’s instruction, and sequenced on a R9.4 flow-cell. For ONT shotgun sequencing, whole genome metagenomic libraries were prepared using SQK-LSK108 kit and sequenced on a R9.4 flow-cell. For Illumina sequencing, PCR primers 341F 5’- CCTACGGGNGGCWGCAG-3’ and 785R 5’-GACTACHVGGGTATCTAATCC-3’ [53] were used to amplify the V3-V4 regions (~550 bp) of 16S gene. Libraries were prepared using the 16S metagenomic sequencing library preparation kit and sequenced on MiSeq instrument in 300 PE mode.

The environmental samples were supplemented with a spike-in of indicator species at two different concentrations by ISS. The exogenous bacteria contained in each spike-in, and the corresponding concentrations, are provided in **Table 2**.

Table 2. Indicator species spiked-in in the environmental sample.

Bacteria	**Source**	**Spike-in$_{min}$ (CFU/ml)**	**Spike-in$_{max}$ (CFU/ml)**
Escherichia coli	ATCC8739	10^2	10^5
Enterococcus faecalis	Clinical sample (urine)		
Enterobacter cloacae	Clinical sample (urine)		
Staphylococcus aureus	Clinical sample (urine)		
Salmonella infantis	Environmental sample		

All wet-lab protocols for data generation were performed by the wet-lab team of professor Delledonne’s laboratory.

Table 3. Summary of sequenced samples.

Sample ID	**Sampling location**	**Gene analysed**	**Sequencing platforms**
Leptopelis vermiculatus	Tanzania	16S	Sanger, ONT MinION R7.3 chemistry
Rieppeleon brachyurus	Tanzania	16S	
Sorex alpinus	Italy	16S	
Rhynchocyon udzungwensis	Tanzania	CO1	
Arthroleptis xenodactyloides	Tanzania	16S	
Snail1	Borneo	CO1	Sanger, ONT MinION R9.4 chemistry
Jap1	Borneo	CO1	
H36	Borneo	CO1	
H37	Borneo	CO1	
H42	Borneo	CO1	
H43	Borneo	CO1	
Colen1	Borneo	CO1	
Env. sample	Italy	16S	Illumina MiSeq 300PE, ONT MinION R9.4 chemistry
Env. Sample + [spike-in]$_{min}$	Italy	16S	
Env. Sample + [spike-in]$_{max}$	Italy	16S	
Env. Sample + [spike-in]$_{max}$	Italy	Whole genome	ONT MinION R9.4 chemistry

Nanopore reads preprocessing

Nanopore reads generated with sequencing chemistry R7.3 were base-called, demultiplexed and quality filtered online using the Metrichor Agent v2.23, a software officially provided by ONT, with the 2D Basecalling workflow. The *fastq* files were then extracted using poretools [54].

Nanopore reads generated with R9.4 sequencing chemistry were base-called, demultiplexed, quality checked and filtered with a set of scripts written in R and bash languages, reported in https://github.com/MaestSi/ONT_preprocessing repository. In brief, the user could modify a configuration file to define some options. Then, raw reads in *fast5* format were base-called offline using Guppy v4.2.2, a software officially provided by ONT, with parameters "--flowcell FLO-MIN106" and "--kit SQK_LSK108" or "--kit SQK-RAB201" for barcoding or meta-barcoding runs respectively. Base-called reads were demultiplexed by Guppy v4.2.2 by requiring the presence of indexes at both ends of the reads and adapters were trimmed with "--require_barcodes_both_ends --trim_barcodes" and specifying the barcoding kit "--barcode_kits EXP-PBC001" or "--barcode_kits SQK-RAB201" for barcoding and meta-barcoding runs respectively. Reads with quality < 7 and abnormal read length were discarded using NanoFilt v2.7.1 [55] with "cat $READS_FASTQ | NanoFilt -q7 -l $MINLENGTH --maxlength $MAXLENGTH". A quality report was produced using pycoQC v2.5.0.21 [56] with "pycoQC -f sequencing_summary.txt -b barcoding_summary.txt -o pycoQC_report.html --min_pass_qual 7".

Barcoding analysis with the *ONtoBar* pipeline

The first step of the *ONtoBar* pipeline performs de novo assembly of Nanopore reads using Loman's method, as described in [57]. Then, the assembled consensus sequence is aligned locally against the NCBI nucleotide (nt) database using BLAST v2.5.0+ [45], and the top Blast hit is retrieved. This sequence is used as a reference for the alignment of Nanopore reads with LAST v1060 [47]. Starting from the *bam* alignment file, a pileup file is created using Samtools v1.9 [58], and the frequency of each nucleotide per reference position is calculated using a custom-made script that parses the *pileup* file. This script then produces a new consensus sequence containing the most frequent nucleotide at each position. Finally, the obtained consensus is aligned against NCBI nt database using BLAST. All scripts for running the *ONtoBar* pipeline are reported in https://github.com/cianix/ONtoBAR repository.

Barcoding analysis with the *ONTrack* pipeline

The first step of the *ONTrack* pipeline uses VSEARCH v2.14.2 [38] to cluster reads at 70% identity and only reads in the most abundant cluster are retained for subsequent analysis in order to remove contaminant sequences. Of those, up to 200 reads are randomly sampled using Seqtk sample v1.3-r106 (https://github.com/lh3/seqtk) and aligned using MAFFT v7.458 [39] with parameters -localpair -maxiterate 1000, specific for iterative refinement, incorporating local pairwise alignment information. EMBOSS cons v6.6.0.0 (http://emboss.open-bio.org/rel/dev/apps/cons.html) is then used to retrieve a draft consensus sequence starting from the MAFFT alignment. The EMBOSS cons plurality parameter is set to the value obtained by multiplying the number of aligned reads by 0.15, in order to include a base in the draft consensus sequence if at least 15% of the aligned reads carriy that base. If less than 15% of the aligned reads carry the same base in a specific position, and a generic base (N) is included in the consensus sequence, the generic base is removed using a custom script. To polish the obtained draft consensus sequence, up to 200 reads are randomly sampled using Seqtk sample, with a different seed to the one used before, and mapped to the draft consensus sequence using Minimap2 v2.17-r941 [46]. The alignment file is filtered, sorted and compressed to the *bam* format using Samtools v1.9 [58]. Nanopolish index and nanopolish variants --consensus modules from Nanopolish v0.13.2 (https://github.com/jts/nanopolish) are used to obtain a polished consensus sequence. When the pipeline is run multiple times, the polished consensus sequences produced during each round are aligned with MAFFT, after setting the gap penalty to 0. The final consensus is retrieved based on the majority rule, namely, selecting the consensus sequence that was produced on the majority of times. PCR primers are trimmed from both sides of the consensus sequence using Seqtk trimfq. As a final step, the consensus sequence is aligned using BLAST v2.5.0+ against the NCBI nt database, which was downloaded locally; this step is optional and allows to retrieve, for each consensus sequence, the most similar sequences in the database, for taxonomical assignment purposes. The accuracy of MinION consensus sequences was evaluated by aligning the consensus sequences to the corresponding Sanger-derived reference sequences using BLAST. The percentage of MinION reads mapping to the Sanger sequence was evaluated by performing the alignment with Minimap2 and running samtools flagstat on the generated *bam* file. All scripts for running the *ONTrack* pipeline are reported in https://github.com/MaestSi/ONTrack repository.

Meta-barcoding analysis with the *MetONTIIME* pipeline

Reads in *fastq.gz* format were imported in qiime2 v.2020.8.0 using “qiime tools import”. Reads were then dereplicated using “qiime vsearch dereplicate-sequences” and “qiime vsearch cluster-features-de-novo” to obtain a set of representative sequences and the corresponding table with read counts. Representative sequences were then aligned to the reference database using “qiime feature-classifier classify-consensus-vsearch” or “qiime feature-classifier classify-consensus-blast”, and taxonomy tables and barplots describing the taxonomic classification at each taxonomic level were generated with “qiime taxa collapse” and “qiime taxa barplot”. All scripts for running the MetONTIIME pipeline are reported in https://github.com/MaestSi/MetONTIIME repository.

Nanopore shotgun metagenomics analysis

Preprocessed reads in fastq format were classified with Kraken v2.0.9-beta [43] using instruction: “$KRAKEN2 --db $DB –output $SAMPLE_NAME"_kraken2_output.txt" --report $SAMPLE_NAME"_kraken2_report.txt" --threads $THREADS $FASTQ”, using a database built on RefSeq complete bacterial genomes. A taxonomy pie chart was then generated using Krona [59] with instruction: “ktImportTaxonomy -q 2 -t 3 -s 4

$KRAKEN2_OUTPUT -o $KRONA_REPORT". All scripts for running the pipeline are reported in https://github.com/MaestSi/MetaKraken2.

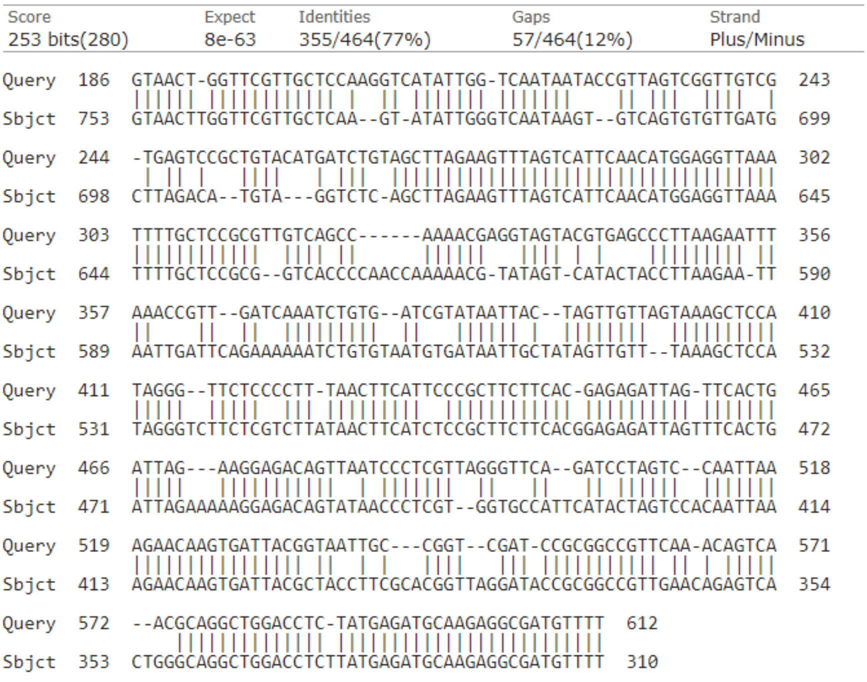
Illumina meta-barcoding analysis

Reads in *fastq.gz* format were imported in qiime2 v2020.8.0 using "qiime tools import" after generation of manifest.txt file, and PCR primers were trimmed with "qiime cutadapt trim-paired". Reads were then truncated (at 280 bp and 260 bp for forward and reverse reads respectively) and overlapping mates were merged with "qiime dada2 denoise-paired". A set of amplicon sequence variants (ASVs) was obtained, together with a feature table, describing the occurrence of ASVs in each sample. The database Silva_132_99_16S was then imported with "qiime tools import", and then "qiime feature-classifier extract-reads" and "qiime feature-classifier fit-classifier-naive-bayes" were used to train a Naïve-Bayes classifier on the V3-V4 region of the 16S gene. ASVs were then classified with "qiime feature-classifier classify-sklearn", and taxonomy tables and barplots were generated with "qiime taxa collapse" and "qiime taxa barplot". All scripts for running the pipeline are reported in https://github.com/MaestSi/QIIME2_Illumina repository.

Results

The early days of Nanopore sequencing: single-read error profile

To test the performances of the Oxford Nanopore Technologies (ONT) MinION sequencer at the early stages of development, I first focused on the 16S gene of the toad *Sclerophrys brauni*, by sequencing the amplicon both with ONT MinION (R7.3 chemistry) and with Sanger platforms. Exploiting the presence of a hairpin adapter linking the forward and the reverse strand of the same DNA molecule, the two sequences could be collapsed to obtain more accurate 2D reads. A total of 2,660 reads out of 51,273 (5.2%) were successfully collapsed and passed quality filtering set by the base-calling software. Alignment of a set of 2D PASS reads to the NCBI nt database showed that the error rate was too high to enable accurate species identification without implementing *ad hoc* bioinformatic strategies for error correction (**Figure 9**).



```
Score            Expect   Identities        Gaps             Strand
253 bits(280)    8e-63    355/464(77%)      57/464(12%)      Plus/Minus

Query  186  GTAACT-GGTTCGTTGCTCCAAGGTCATATTGG-TCAATAATACCGTTAGTCGGTTGTCG  243
            |||||| |||||||||||| |  || ||||||| |||||||    || |||  ||||  |
Sbjct  753  GTAACTTGGTTCGTTGCTCAA--GT-ATATTGGGTCAATAAGT--GTCAGTGTGTTGATG  699

Query  244  -TGAGTCCGCTGTACATGATCTGTAGCTTAGAAGTTTAGTCATTCAACATGGAGGTTAAA  302
             | || |   ||||   | |||  ||||||||||||||||||||||||||||||||||||
Sbjct  698  CTTAGACA--TGTA---GGTCTC-AGCTTAGAAGTTTAGTCATTCAACATGGAGGTTAAA  645

Query  303  TTTTGCTCCGCGTTGTCAGCC------AAAACGAGGTAGTACGTGAGCCCTTAAGAATTT  356
            ||||||||||||  |||| ||      ||||||   |||| | |    ||||||||| ||
Sbjct  644  TTTTGCTCCGCG--GTCACCCCAACCAAAAACG-TATAGT-CATACTACCTTAAGAA-TT  590

Query  357  AAACCGTT--GATCAAATCTGTG--ATCGTATAATTAC--TAGTTGTTAGTAAAGCTCCA  410
            ||    ||  ||  |||||||||  ||   |||||| |  ||||||||  ||||||||||
Sbjct  589  AATTGATTCAGAAAAAATCTGTGTAATGTGATAATTGCTATAGTTGTT--TAAAGCTCCA  532

Query  411  TAGGG--TTCTCCCCTT-TAACTTCATTCCCGCTTCTTCAC-GAGAGATTAG-TTCACTG  465
            |||||  |||||  ||| |||||||||  |||||||||||| |||||||||| |||||||
Sbjct  531  TAGGGTCTTCTCGTCTTATAACTTCATCTCCGCTTCTTCACGGAGAGATTAGTTTCACTG  472

Query  466  ATTAG---AAGGAGACAGTTAATCCCTCGTTAGGGTTCA--GATCCTAGTC--CAATTAA  518
            |||||   |||||||||||  | |||||||  ||   ||   || ||||||  |||||||
Sbjct  471  ATTAGAAAAAGGAGACAGTATAACCCTCGT--GGTGCCATTCATACTAGTCCACAATTAA  414

Query  519  AGAACAAGTGATTACGGTAATTGC---CGGT--CGAT-CCGCGGCCGTTCAA-ACAGTCA  571
            |||||||||||||||| ||  | |   ||||   ||| ||||||||||| || | |||||
Sbjct  413  AGAACAAGTGATTACGCTACCTTCGCACGGTTAGGATACCGCGGCCGTTGAACAGAGTCA  354

Query  572  --ACGCAGGCTGGACCTC-TATGAGATGCAAGAGGCGATGTTTT  612
                |||||||||||||| |||||||||||||||||||||||||
Sbjct  353  CTGGGCAGGCTGGACCTCTTATGAGATGCAAGAGGCGATGTTTT  310
```

Figure 9. Blast alignment of a Nanopore read to the top Blast hit from NCBI nt database

I then aligned this set of 2D PASS reads to the reference Sanger sequence and inspected in detail their error profile. First, I noticed that only 977 reads (36.7%) successfully aligned to the reference sequence, probably due to the higher error rate of unaligned reads. The mean error rate of Nanopore aligned reads was 17%, and consisted of mismatches

(8%), insertions (4%) and deletions (5%). Moreover, the error rate was found to be non-random, as shown by the spikes of low coverage in correspondence of homopolymeric stretches (**Figure 10**). This was due to a platform bias, and was already described in the literature [60]. Despite the high error rate of ONT aligned reads, the most frequent nucleotide at each reference coordinate was always the correct one. Although the correct nucleotide frequency decreased in correspondence of homopolymeric stretches, it remained the most frequent one (**Figure 11**). Therefore, the consensus sequence obtained by calling the most frequent nucleotide at each position had 100% match with the Sanger reference sequence (**Figure 12**). This result showed that only resequencing, namely the identification of variants with respect to a reference sequence, was feasible.

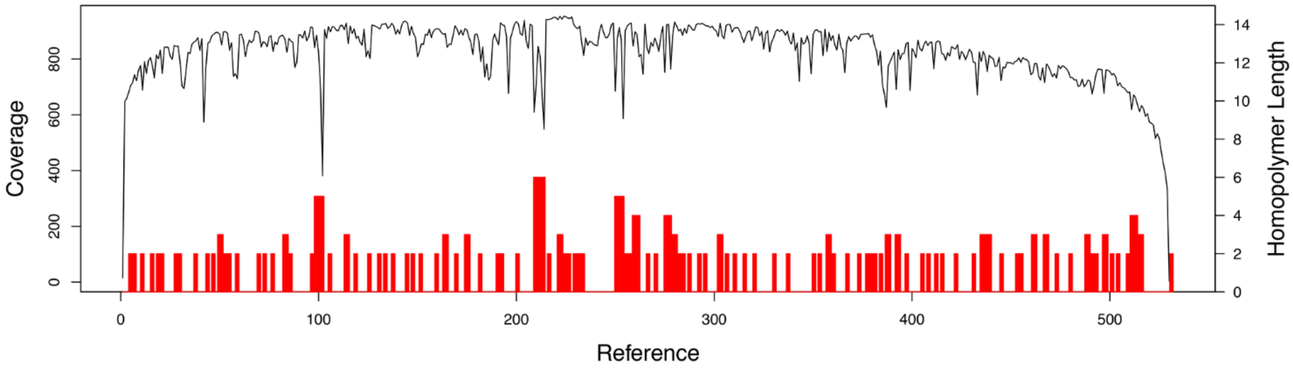


Figure 10. Spikes of low coverage in correspondence of homopolymer stretches.

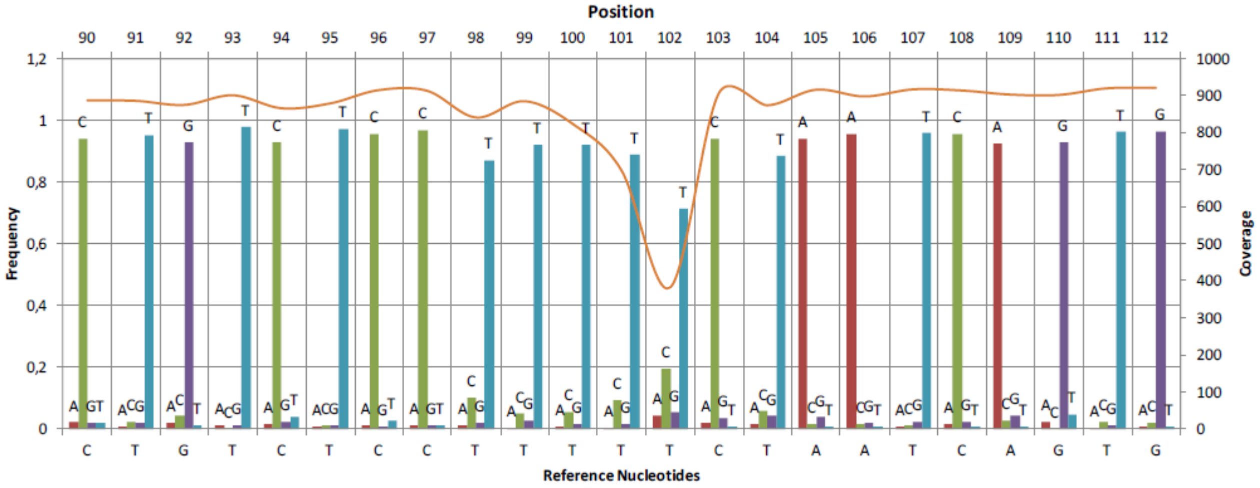
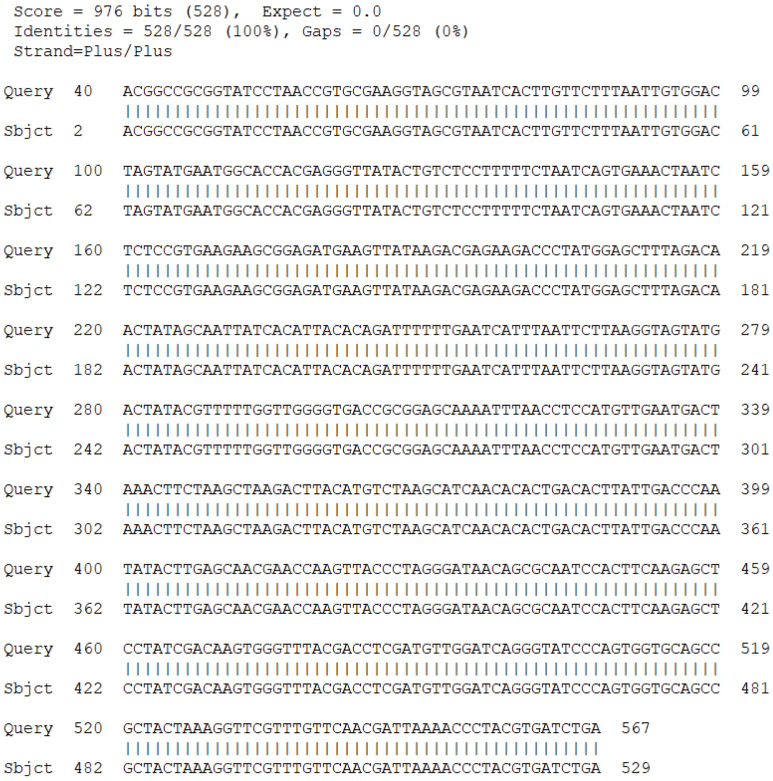


Figure 11. The most frequent nucleotide in Nanopore aligned reads is always the correct one.



```
 Score = 976 bits (528),  Expect = 0.0
 Identities = 528/528 (100%), Gaps = 0/528 (0%)
 Strand=Plus/Plus

Query  40   ACGGCCGCGGTATCCTAACCGTGCGAAGGTAGCGTAATCACTTGTTCTTTAATTGTGGAC  99
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2    ACGGCCGCGGTATCCTAACCGTGCGAAGGTAGCGTAATCACTTGTTCTTTAATTGTGGAC  61

Query  100  TAGTATGAATGGCACCACGAGGGTTATACTGTCTCCTTTTTCTAATCAGTGAAACTAATC  159
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  62   TAGTATGAATGGCACCACGAGGGTTATACTGTCTCCTTTTTCTAATCAGTGAAACTAATC  121

Query  160  TCTCCGTGAAGAAGCGGAGATGAAGTTATAAGACGAGAAGACCCTATGGAGCTTTAGACA  219
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  122  TCTCCGTGAAGAAGCGGAGATGAAGTTATAAGACGAGAAGACCCTATGGAGCTTTAGACA  181

Query  220  ACTATAGCAATTATCACATTACACAGATTTTTTGAATCATTTAATTCTTAAGGTAGTATG  279
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  182  ACTATAGCAATTATCACATTACACAGATTTTTTGAATCATTTAATTCTTAAGGTAGTATG  241

Query  280  ACTATACGTTTTTGGTTGGGGTGACCGCGGAGCAAAATTTAACCTCCATGTTGAATGACT  339
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  242  ACTATACGTTTTTGGTTGGGGTGACCGCGGAGCAAAATTTAACCTCCATGTTGAATGACT  301

Query  340  AAACTTCTAAGCTAAGACTTACATGTCTAAGCATCAACACACTGACACTTATTGACCCAA  399
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  302  AAACTTCTAAGCTAAGACTTACATGTCTAAGCATCAACACACTGACACTTATTGACCCAA  361

Query  400  TATACTTGAGCAACGAACCAAGTTACCCTAGGGATAACAGCGCAATCCACTTCAAGAGCT  459
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  362  TATACTTGAGCAACGAACCAAGTTACCCTAGGGATAACAGCGCAATCCACTTCAAGAGCT  421

Query  460  CCTATCGACAAGTGGGTTTACGACCTCGATGTTGGATCAGGGTATCCCAGTGGTGCAGCC  519
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  422  CCTATCGACAAGTGGGTTTACGACCTCGATGTTGGATCAGGGTATCCCAGTGGTGCAGCC  481

Query  520  GCTACTAAAGGTTCGTTTGTTCAACGATTAAAACCCTACGTGATCTGA  567
            ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  482  GCTACTAAAGGTTCGTTTGTTCAACGATTAAAACCCTACGTGATCTGA  529
```

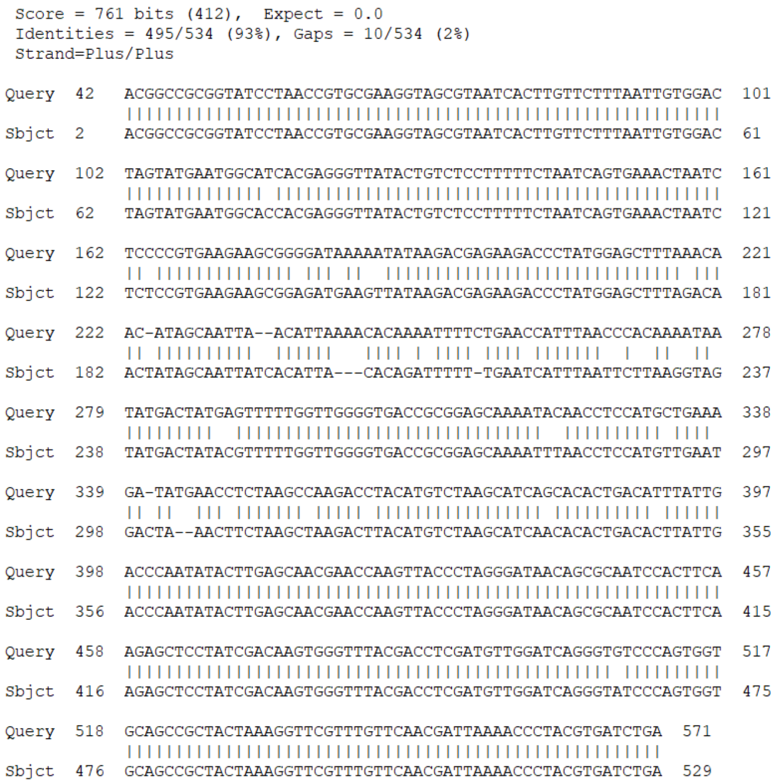
Figure 12. *S. brauni* consensus sequence generated using the Sanger as a reference completely matches the Sanger sequence.

Testing the robustness of a reference-guided assembly approach for barcoding

In the previous section, I showed that when aligning Nanopore reads obtained from resequencing of a species to the corresponding Sanger sequence, the generated consensus sequence completely matched the Sanger. I then wanted to test the robustness of this approach when using a reference sequence which is not identical to the Sanger sequence. At this aim, I first simulated *intra*-species variability by introducing a different number of variants randomly distributed in the Sanger sequence, totaling from 1% to 3% of nucleotides. Still, the obtained consensus sequences completely matched the Sanger sequence (**Table 4**).

Table 4. Reference-guided assembly is tolerant to randomly distributed variants in the reference sequence.

Variants introduced	Identity with Sanger
5 (1%)	100%
10 (2%)	100%
15 (3%)	100%

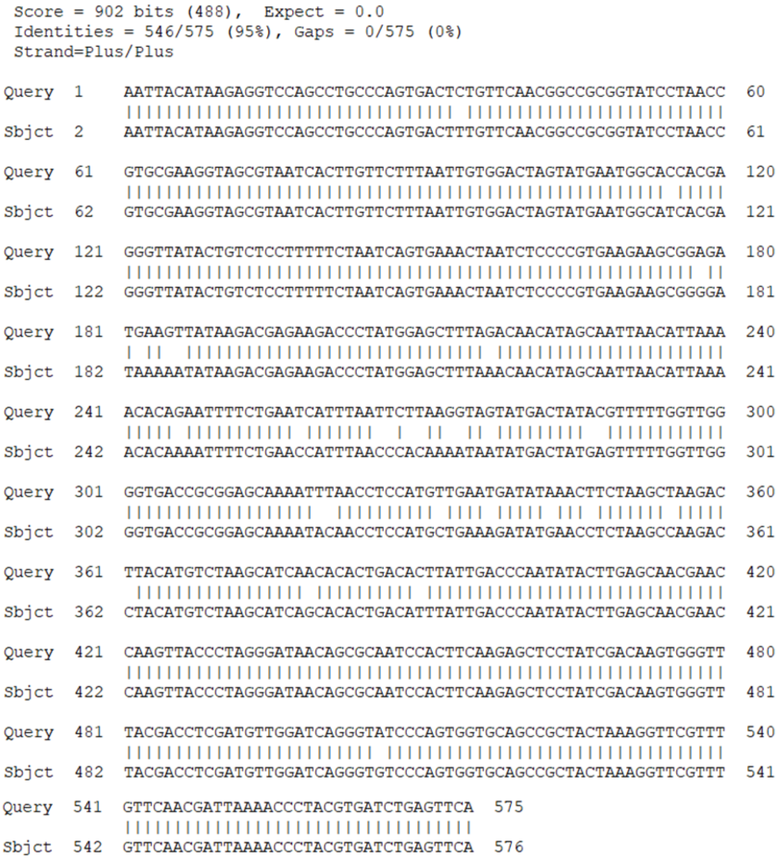


```
 Score = 761 bits (412),  Expect = 0.0
 Identities = 495/534 (93%), Gaps = 10/534 (2%)
 Strand=Plus/Plus

Query  42   ACGGCCGCGGTATCCTAACCGTGCGAAGGTAGCGTAATCACTTGTTCTTTAATTGTGGAC  101
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2    ACGGCCGCGGTATCCTAACCGTGCGAAGGTAGCGTAATCACTTGTTCTTTAATTGTGGAC  61

Query  102  TAGTATGAATGGCATCACGAGGGTTATACTGTCTCCTTTTTCTAATCAGTGAAACTAATC  161
            |||||||||||||| |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  62   TAGTATGAATGGCACCACGAGGGTTATACTGTCTCCTTTTTCTAATCAGTGAAACTAATC  121

Query  162  TCCCCGTGAAGAAGCGGGGATAAAAATATAAGACGAGAAGACCCTATGGAGCTTTAAACA  221
            || |||||||||||||| ||| ||  |||||||||||||||||||||||||||||| |||
Sbjct  122  TCTCCGTGAAGAAGCGGAGATGAAGTTATAAGACGAGAAGACCCTATGGAGCTTTAGACA  181

Query  222  AC-ATAGCAATTA--ACATTAAAACACAAAATTTTCTGAACCATTTAACCCACAAAATAA  278
            || ||||||||||  ||||||   |||| | |||| |||| |||||||  |  ||  || 
Sbjct  182  ACTATAGCAATTATCACATTA---CACAGATTTTT-TGAATCATTTAATTCTTAAGGTAG  237

Query  279  TATGACTATGAGTTTTTGGTTGGGGTGACCGCGGAGCAAAATACAACCTCCATGCTGAAA  338
            |||||||||  |||||||||||||||||||||||||||||||  |||||||||| |||| 
Sbjct  238  TATGACTATACGTTTTTGGTTGGGGTGACCGCGGAGCAAAATTTAACCTCCATGTTGAAT  297

Query  339  GA-TATGAACCTCTAAGCCAAGACCTACATGTCTAAGCATCAGCACACTGACATTTATTG  397
            || ||  ||| ||||||| ||||| ||||||||||||||||| |||||||||| ||||||
Sbjct  298  GACTA--AACTTCTAAGCTAAGACTTACATGTCTAAGCATCAACACACTGACACTTATTG  355

Query  398  ACCCAATATACTTGAGCAACGAACCAAGTTACCCTAGGGATAACAGCGCAATCCACTTCA  457
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  356  ACCCAATATACTTGAGCAACGAACCAAGTTACCCTAGGGATAACAGCGCAATCCACTTCA  415

Query  458  AGAGCTCCTATCGACAAGTGGGTTTACGACCTCGATGTTGGATCAGGGTGTCCCAGTGGT  517
            ||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct  416  AGAGCTCCTATCGACAAGTGGGTTTACGACCTCGATGTTGGATCAGGGTATCCCAGTGGT  475

Query  518  GCAGCCGCTACTAAAGGTTCGTTTGTTCAACGATTAAAACCCTACGTGATCTGA  571
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  476  GCAGCCGCTACTAAAGGTTCGTTTGTTCAACGATTAAAACCCTACGTGATCTGA  529
```

Figure 13. *S. brauni* Sanger sequence shares 93% identity to S. *pantherina* sequence from NCBI.

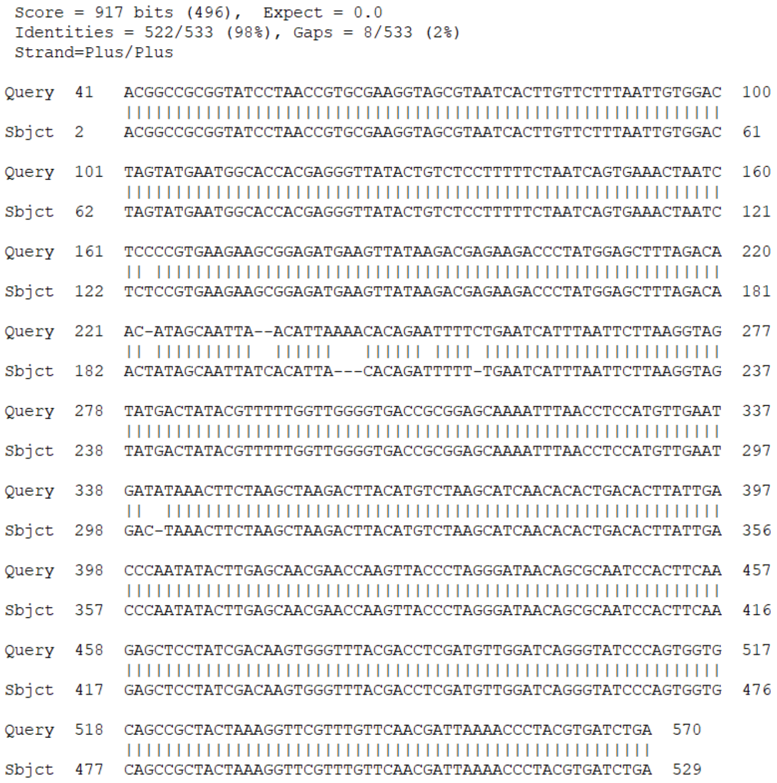
Then, I picked a sequence from NCBI database for the 16S gene of *S. pantherina*, a toad belonging to the same genus of *S. brauni*, but from a different species, and I aligned the Sanger sequence of *S. brauni* to it. The two sequences showed 93% identity.

I then aligned Nanopore reads of *S. brauni* to the *S. pantherina* sequence from NCBI and called the consensus sequence. The obtained consensus sequence was realigned to *S. pantherina*, and they showed 95% identity (**Figure 13**). This value was slightly higher than the one found using S. *brauni* Sanger sequence, indicating that the reference sequence used for calling the consensus sequence introduced some bias. Accordingly, the consensus sequence was not identical to the Sanger, showing only 98% identity (**Figure 14**).



```
 Score = 902 bits (488),  Expect = 0.0
 Identities = 546/575 (95%), Gaps = 0/575 (0%)
 Strand=Plus/Plus

Query  1    AATTACATAAGAGGTCCAGCCTGCCCAGTGACTCTGTTCAACGGCCGCGGTATCCTAACC  60
            ||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct  2    AATTACATAAGAGGTCCAGCCTGCCCAGTGACTTTGTTCAACGGCCGCGGTATCCTAACC  61

Query  61   GTGCGAAGGTAGCGTAATCACTTGTTCTTTAATTGTGGACTAGTATGAATGGCACCACGA  120
            |||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  62   GTGCGAAGGTAGCGTAATCACTTGTTCTTTAATTGTGGACTAGTATGAATGGCATCACGA  121

Query  121  GGGTTATACTGTCTCCTTTTTCTAATCAGTGAAACTAATCTCCCCGTGAAGAAGCGGAGA  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct  122  GGGTTATACTGTCTCCTTTTTCTAATCAGTGAAACTAATCTCCCCGTGAAGAAGCGGGGA  181

Query  181  TGAAGTTATAAGACGAGAAGACCCTATGGAGCTTTAGACAACATAGCAATTAACATTAAA  240
            | ||  ||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  182  TAAAAATATAAGACGAGAAGACCCTATGGAGCTTTAAACAACATAGCAATTAACATTAAA  241

Query  241  ACACAGAATTTTCTGAATCATTTAATTCTTAAGGTAGTATGACTATACGTTTTTGGTTGG  300
            ||||| ||||||||||| |||||||  |  ||  || |||||||||  ||||||||||||
Sbjct  242  ACACAAAATTTTCTGAACCATTTAACCCACAAAATAATATGACTATGAGTTTTTGGTTGG  301

Query  301  GGTGACCGCGGAGCAAAATTTAACCTCCATGTTGAATGATATAAACTTCTAAGCTAAGAC  360
            |||||||||||||||||||  |||||||||| |||| ||||| ||| ||||||| |||||
Sbjct  302  GGTGACCGCGGAGCAAAATACAACCTCCATGCTGAAAGATATGAACCTCTAAGCCAAGAC  361

Query  361  TTACATGTCTAAGCATCAACACACTGACACTTATTGACCCAATATACTTGAGCAACGAAC  420
             ||||||||||||||||| |||||||||| ||||||||||||||||||||||||||||||
Sbjct  362  CTACATGTCTAAGCATCAGCACACTGACATTTATTGACCCAATATACTTGAGCAACGAAC  421

Query  421  CAAGTTACCCTAGGGATAACAGCGCAATCCACTTCAAGAGCTCCTATCGACAAGTGGGTT  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  422  CAAGTTACCCTAGGGATAACAGCGCAATCCACTTCAAGAGCTCCTATCGACAAGTGGGTT  481

Query  481  TACGACCTCGATGTTGGATCAGGGTATCCCAGTGGTGCAGCCGCTACTAAAGGTTCGTTT  540
            ||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||
Sbjct  482  TACGACCTCGATGTTGGATCAGGGTGTCCCAGTGGTGCAGCCGCTACTAAAGGTTCGTTT  541

Query  541  GTTCAACGATTAAAACCCTACGTGATCTGAGTTCA  575
            |||||||||||||||||||||||||||||||||||
Sbjct  542  GTTCAACGATTAAAACCCTACGTGATCTGAGTTCA  576
```

Figure 13. *S. brauni* consensus sequence generated using *S. pantherina* as a reference shares 95% identity with *S. pantherina* sequence from NCBI.



```
 Score = 917 bits (496),  Expect = 0.0
 Identities = 522/533 (98%), Gaps = 8/533 (2%)
 Strand=Plus/Plus

Query  41   ACGGCCGCGGTATCCTAACCGTGCGAAGGTAGCGTAATCACTTGTTCTTTAATTGTGGAC  100
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2    ACGGCCGCGGTATCCTAACCGTGCGAAGGTAGCGTAATCACTTGTTCTTTAATTGTGGAC  61

Query  101  TAGTATGAATGGCACCACGAGGGTTATACTGTCTCCTTTTTCTAATCAGTGAAACTAATC  160
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  62   TAGTATGAATGGCACCACGAGGGTTATACTGTCTCCTTTTTCTAATCAGTGAAACTAATC  121

Query  161  TCCCCGTGAAGAAGCGGAGATGAAGTTATAAGACGAGAAGACCCTATGGAGCTTTAGACA  220
            || |||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  122  TCTCCGTGAAGAAGCGGAGATGAAGTTATAAGACGAGAAGACCCTATGGAGCTTTAGACA  181

Query  221  AC-ATAGCAATTA--ACATTAAAACACAGAATTTTCTGAATCATTTAATTCTTAAGGTAG  277
            || ||||||||||  ||||||   |||||| |||| ||||||||||||||||||||||||
Sbjct  182  ACTATAGCAATTATCACATTA---CACAGATTTTT-TGAATCATTTAATTCTTAAGGTAG  237

Query  278  TATGACTATACGTTTTTGGTTGGGGTGACCGCGGAGCAAAATTTAACCTCCATGTTGAAT  337
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  238  TATGACTATACGTTTTTGGTTGGGGTGACCGCGGAGCAAAATTTAACCTCCATGTTGAAT  297

Query  338  GATATAAACTTCTAAGCTAAGACTTACATGTCTAAGCATCAACACACTGACACTTATTGA  397
            ||  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  298  GAC-TAAACTTCTAAGCTAAGACTTACATGTCTAAGCATCAACACACTGACACTTATTGA  356

Query  398  CCCAATATACTTGAGCAACGAACCAAGTTACCCTAGGGATAACAGCGCAATCCACTTCAA  457
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  357  CCCAATATACTTGAGCAACGAACCAAGTTACCCTAGGGATAACAGCGCAATCCACTTCAA  416

Query  458  GAGCTCCTATCGACAAGTGGGTTTACGACCTCGATGTTGGATCAGGGTATCCCAGTGGTG  517
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  417  GAGCTCCTATCGACAAGTGGGTTTACGACCTCGATGTTGGATCAGGGTATCCCAGTGGTG  476

Query  518  CAGCCGCTACTAAAGGTTCGTTTGTTCAACGATTAAAACCCTACGTGATCTGA  570
            |||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  477  CAGCCGCTACTAAAGGTTCGTTTGTTCAACGATTAAAACCCTACGTGATCTGA  529
```

Figure 14. *S. brauni* consensus sequence generated using *S. pantherina* as a reference shows 98% identity with *S. brauni* Sanger sequence.

Overall, these tests allowed me to conclude that when using the correct species as a reference, the obtained consensus sequence is 100% identical to the reference and to the Sanger, allowing to confirm the identified species. On the contrary, when using a different species as a reference, the obtained consensus sequence may contain some errors. In this case, the consensus sequence shows some differences with respect to the reference, but these differences may be underestimated due to reference bias, forcing the alignment of Nanopore reads to adapt to the reference. If differences are above a predefined threshold, usually set at 3%, it is possible to exclude that the organism belongs to the same species as the chosen reference.

A first attempt to *de novo* assembly

In the previous section I showed that a reference-guided approach is suitable to obtain accurate consensus sequences, given that a similar enough reference is available. On the contrary, *de novo* assembly does not rely on a reference sequence, but tries to join reads into larger contiguous contigs [35] (**Figure 15**). The high error rate of ONT reads made *de novo* assembly particularly challenging. Nonetheless, the group of professor Nick Loman first reported the *de novo* assembly of a complete bacterial genome using only Nanopore data [57]. This landmark work showed that, when using *ad hoc* developed bioinformatic tools, issues related to the high error rate could be largely overcome, yielding accurate and complete genome assemblies.

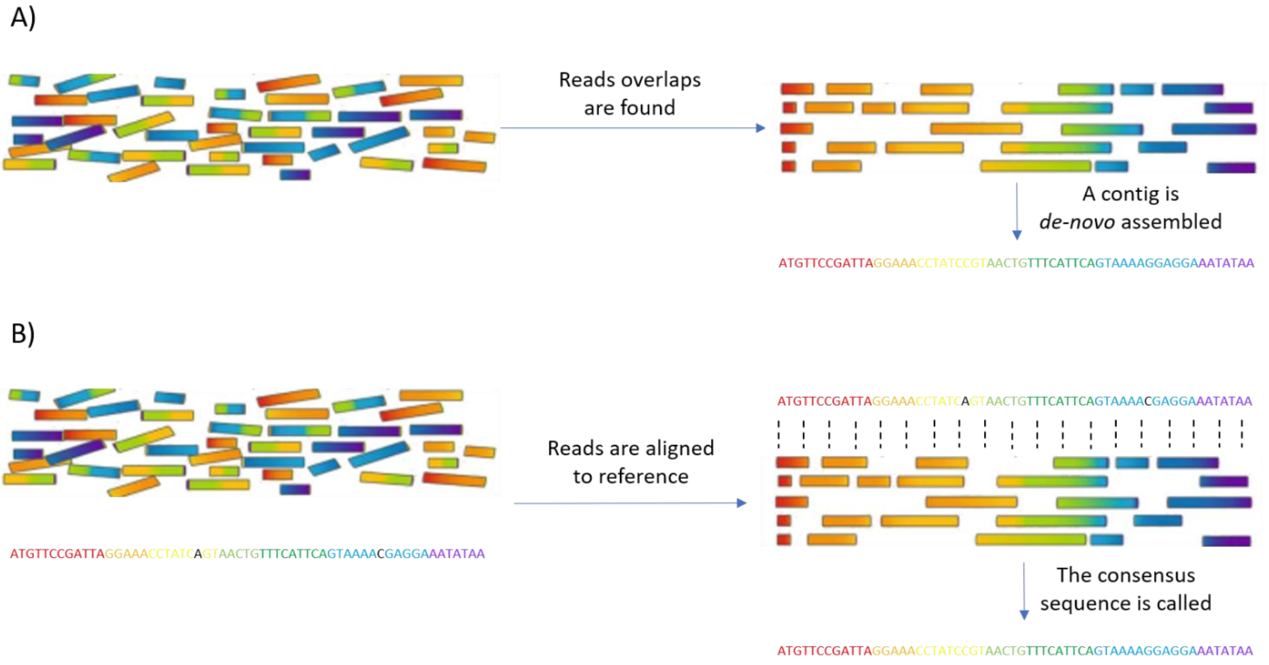
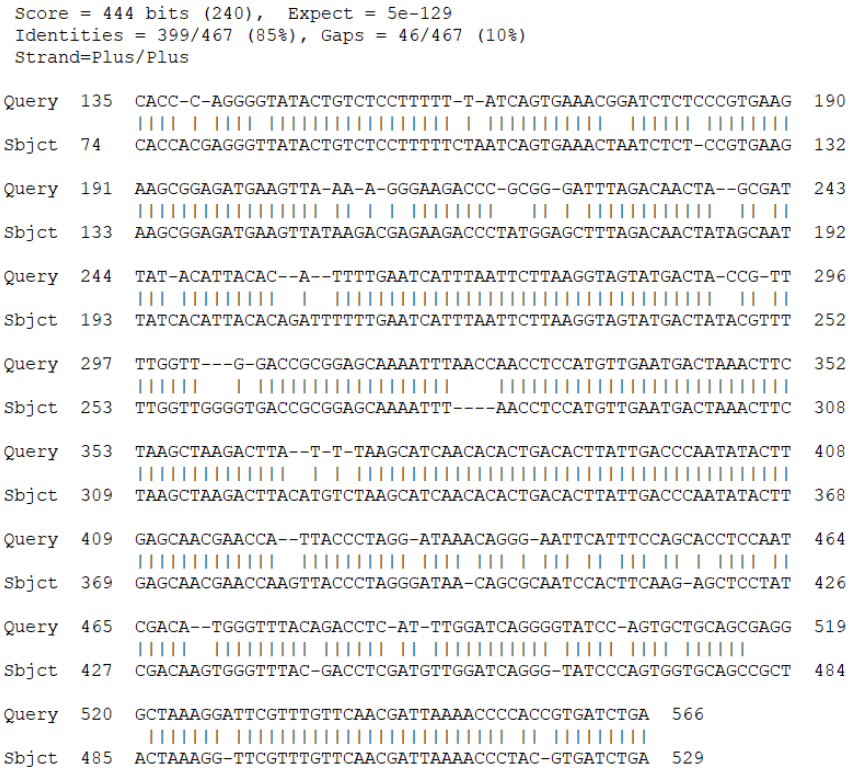

A)
Reads overlaps are found
A contig is *de-novo* assembled
ATGTTCCGATTAGGAAACCTATCCGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
B)
Reads are aligned to reference
ATGTTCCGATTAGGAAACCTATCAGTAACTGTTTCATTCAGTAAAACGAGGAAATATAA
ATGTTCCGATTAGGAAACCTATCAGTAACTGTTTCATTCAGTAAAACGAGGAAATATAA
The consensus sequence is called
ATGTTCCGATTAGGAAACCTATCCGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA

Figure 15. *De-novo* and reference-guided assembly strategies. In de-novo assembly, reads overlaps are found and a contig is assembled A); in reference-guided strategies, reads are first aligned to a reference sequence, and then the consensus sequence is called B).

As a preliminary test, I performed *de-novo* assembly of 2D PASS reads from the toad *S. brauni* 16S gene described in the previous section, using Loman's pipeline. The assembled contig was Blasted against the NCBI nt database, and it allowed to identify the most similar species in the database; however, some differences, mainly consisting of insertions and deletions were found, especially in homopolymeric regions (**Figure 16**).

Alignment to the Sanger sequence confirmed that insertions and deletions were errors in the assembled contig. This result showed that, despite *de novo* assembly of ONT reads was not mature yet for obtaining an accurate consensus sequence, it was accurate enough to identify candidate species in the NCBI nt database.



```
 Score = 444 bits (240),  Expect = 5e-129
 Identities = 399/467 (85%), Gaps = 46/467 (10%)
 Strand=Plus/Plus

Query  135  CACC-C-AGGGGTATACTGTCTCCTTTTT-T-ATCAGTGAAACGGATCTCTCCCGTGAAG  190
            |||| | |||| ||||||||||||||||| | |||||||||||  |||||| ||||||||
Sbjct  74   CACCACGAGGGTTATACTGTCTCCTTTTTCTAATCAGTGAAACTAATCTCT-CCGTGAAG  132

Query  191  AAGCGGAGATGAAGTTA-AA-A-GGGAAGACCC-GCGG-GATTTAGACAACTA--GCGAT  243
            ||||||||||||||||| || | | ||||||||   || | ||||||||||||  || ||
Sbjct  133  AAGCGGAGATGAAGTTATAAGACGAGAAGACCCTATGGAGCTTTAGACAACTATAGCAAT  192

Query  244  TAT-ACATTACAC--A--TTTTGAATCATTTAATTCTTAAGGTAGTATGACTA-CCG-TT  296
            ||| |||||||||  |  |||||||||||||||||||||||||||||||||||  || ||
Sbjct  193  TATCACATTACACAGATTTTTTGAATCATTTAATTCTTAAGGTAGTATGACTATACGTTT  252

Query  297  TTGGTT---G-GACCGCGGAGCAAAATTTAACCAACCTCCATGTTGAATGACTAAACTTC  352
            ||||||   | ||||||||||||||||||    |||||||||||||||||||||||||||
Sbjct  253  TTGGTTGGGGTGACCGCGGAGCAAAATTT----AACCTCCATGTTGAATGACTAAACTTC  308

Query  353  TAAGCTAAGACTTA--T-T-TAAGCATCAACACACTGACACTTATTGACCCAATATACTT  408
            ||||||||||||||  | | ||||||||||||||||||||||||||||||||||||||||
Sbjct  309  TAAGCTAAGACTTACATGTCTAAGCATCAACACACTGACACTTATTGACCCAATATACTT  368

Query  409  GAGCAACGAACCA--TTACCCTAGG-ATAAACAGGG-AATTCATTTCCAGCACCTCCAAT  464
            |||||||||||||  |||||||||| |||| ||| | ||| || ||| || | |||| ||
Sbjct  369  GAGCAACGAACCAAGTTACCCTAGGGATAA-CAGCGCAATCCACTTCAAG-AGCTCCTAT  426

Query  465  CGACA--TGGGTTTACAGACCTC-AT-TTGGATCAGGGGTATCC-AGTGCTGCAGCGAGG  519
            |||||  ||||||||| |||||| || ||||||||||| ||||| |||| ||||||
Sbjct  427  CGACAAGTGGGTTTAC-GACCTCGATGTTGGATCAGGG-TATCCCAGTGGTGCAGCCGCT  484

Query  520  GCTAAAGGATTCGTTTGTTCAACGATTAAAACCCCACCGTGATCTGA  566
             ||||||| ||||||||||||||||||||||||| || |||||||||
Sbjct  485  ACTAAAGG-TTCGTTTGTTCAACGATTAAAACCCTAC-GTGATCTGA  529
```

Figure 16. Blast alignment of a de-novo assembled contig to its top hit.

Combining *de novo* and reference-guided assembly approaches: the *ONtoBar* pipeline

In the previous sections I showed that the choice of the reference sequence used for performing a reference-guided assembly is critical and has an impact on consensus sequence accuracy. Moreover, I showed that the *de novo* assembled contig could aid in the process of identifying candidate species. Building up from these observations, a novel bioinformatic pipeline was developed, which could combine *de novo* and reference-guided assembly strategies to reduce both errors in homopolymeric regions and risks associated with arbitrary reference choice (**Figure 17**).

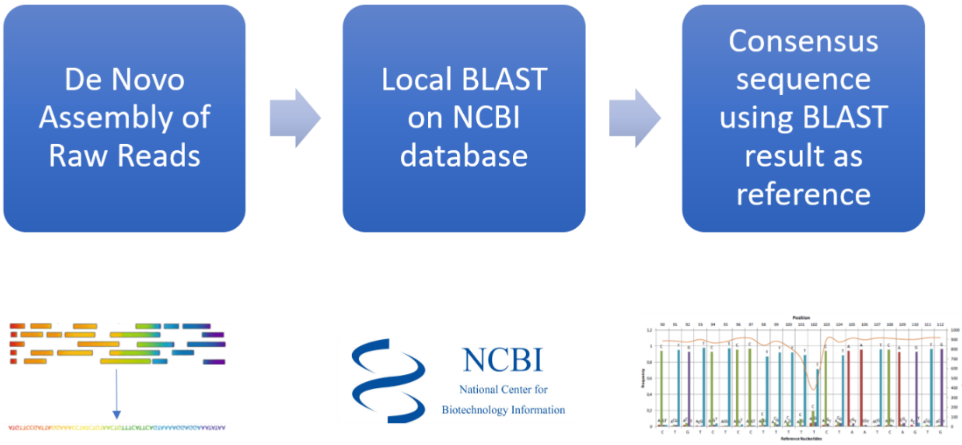

De Novo Assembly of Raw Reads
Local BLAST on NCBI database
Consensus sequence using BLAST result as reference
NCBI
National Center for Biotechnology Information

Figure 17. Schematic representation of a bioinformatic pipeline combining de novo and reference-guided assembly.

The combination of *de novo* and reference-guided assembly approaches was implemented in the *ONtoBar* pipeline [13]. In brief, ONT reads are *de novo* assembled exploiting Loman’s pipeline; the assembled contig is blasted against the NCBI nt database and the top hit is retrieved; the top hit is used as a reference sequence for performing a reference-guided assembly; finally, the *ONtoBar* consensus sequence is Blasted against the NCBI nt database and the final Blast hit is retrieved (**Figure 18**).

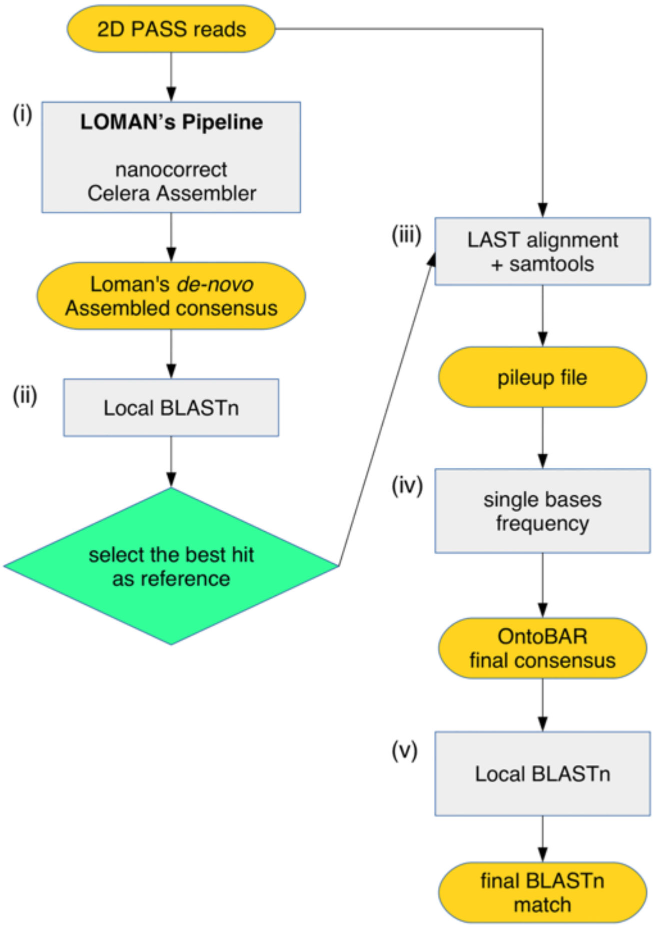


Figure 18. The *ONtoBar* pipeline. 2D PASS reads are *de novo* assembled using Loman's pipeline (i); the obtained preliminary consensus sequence is Blasted locally against the NCBI nt database and the top hit is retrieved (ii); the 2D PASS reads are then aligned to the top hit (iii) and a consensus sequence is produced (iv); the *ONtoBar* consensus sequence is finally Blasted against the NCBI nt database and the top hit is retrieved (v).

To test the discriminatory power of the *ONtoBar* pipeline compared to Loman's pipeline and Sanger sequencing, we performed five experiments in which different organisms and barcoding genes were sequenced. The same amplicons were sequenced with ONT MinION and with Sanger in parallel, to obtain a ground-truth reference. In particular, we examined the 16S gene of an amphibian (the big-eyed tree frog, *Leptopelis vermiculatus*), a squamate reptile (the beardless pygmy chameleon, *Rieppeleon brachyurus*), a mammal (the alpine shrew, *Sorex alpinus*), a wild frog (*Arthroleptis xenodactyloides*) and the CO1 gene of another mammal, the gray-faced sengi (*Rhynchocyon udzungwensis*). Nanopore sequencing produced 92,709 reads for each sample on average, and 40,657 reads (43.9%)

on average were collapsed by the base-calling software to obtain more accurate 2D reads (**Table 5**). Of those, only 31,112 reads (76.5%) on average passed the quality filtering step. All samples had a good number of 2D PASS reads except sample *A. xenodactyloides*, for whom the whole set of 2D reads was considered. ONT reads were analysed both with Loman's and *ONtoBar* pipelines and the similarity with NCBI reference was retrieved and compared to the similarity found with Sanger. On average, the absolute difference between Loman's pipeline and Sanger was 5.4%, while the absolute difference between *ONtoBar* pipeline and Sanger was 0.4%. This result showed that Nanopore sequencing with the *ONtoBar* pipeline had the same discriminative power of Sanger sequencing and could be effectively used for species identification. Moreover, the consensus sequence was always 100% identical to the Sanger, except in the case of *R. udzungwensis*, for whom an NCBI reference of another species was used.

Table 5. MinION sequencing data and identification of known species.

Sample species	**Total Reads**	**2D reads**	**2D PASS reads**	**Similarity with NCBI reference**			**NCBI reference**	**ONtoBar similarity with Sanger**
				Loman's	**ONtoBar**	**Sanger**		
Leptopelis vermiculatus	109,047	57,110	42,102	92%	98%	98%	*Leptopelis sp.*	100%
Rieppeleon brachyurus	97,080	16,760	8,026	92%	100%	100%	*R. brachyurus*	100%
Sorex alpinus	84,913	24,807	7,706	98%	99%	99%	*S. alpinus*	100%
Rhynchocyon udzungwensis	167,466	104,419	97,725	88%	99%	97%	*R. petersi*	98.22%
Arthroleptis xenodactyloides	5,039	187	2	97%	100%	100%	*A. xenodactyloides*	100%

A novel sequencing chemistry: new opportunities for accurate *de novo* assembly

In the last few years, continuous improvements in chemistry and software provided by ONT reflected in increased sequencing accuracy of 1D reads, that approached the accuracy of 2D reads. Kits based on 2D reads were then dismissed. To test the latest updates, we performed seven experiments in which the CO1 gene of different organisms from unknown species was sequenced. The same amplicons were sequenced with ONT MinION (R9.4 sequencing chemistry), producing 39,778 reads on average, and with Sanger in parallel, to obtain a ground-truth reference (**Table 6**). Despite requiring the presence of indexes at both sides of the read, thus excluding cross-contamination, we observed that samples H37 and Colen1 showed a high standard deviation in read length, pointing towards the presence of different amplification products. This was also evident when looking at read length distribution (**Figure 19**). We first analysed ONT reads with the *ONtoBar* pipeline. The similarity of the obtained consensus sequences with NCBI top hit was retrieved and compared to the similarity found with Sanger. This analysis showed that the *ONtoBar* pipeline was able to identify all samples as not present in the NCBI database. This was also confirmed by Sanger sequences, showing a similarity with NCBI top hit < 88% for all samples. When comparing the accuracy of the *ONtoBar* consensus sequence to the Sanger, I found out that two samples had 100% consensus accuracy, while the five remaining samples had slightly lower accuracy, ranging from 98.43% to 99.69%.

Table 6. MinION sequencing data and identification of unknown species.

Sample ID (unknown species)	Gene	Total Reads	Mean bp read Length (SD)	Sanger similarity with NCBI top hit		NCBI top hit	OntoBar consensus similarity with Sanger
				ONtoBar	Sanger		
Snail1	CO1	26,295	687 (31)	88%	88%	P. flavocarinata	99.69%
Jap1	CO1	79,654	688 (27)	86%	85%	B. aeruginosa	98.87%
H36	CO1	22,148	688 (27)	85%	86%	Roraima sp.	100%
H37	CO1	97,571	381 (211)	85%	87%	Macronychus sp.	98.43%
H42	CO1	59,604	689 (24)	86%	85%	Hydrophilidae sp.	99.69%
H43	CO1	96,181	683 (67)	87%	87%	Stenelmis sp.	100%
Colen1	CO1	43,368	351 (180)	87%	88%	A. ventricosa	99.63%

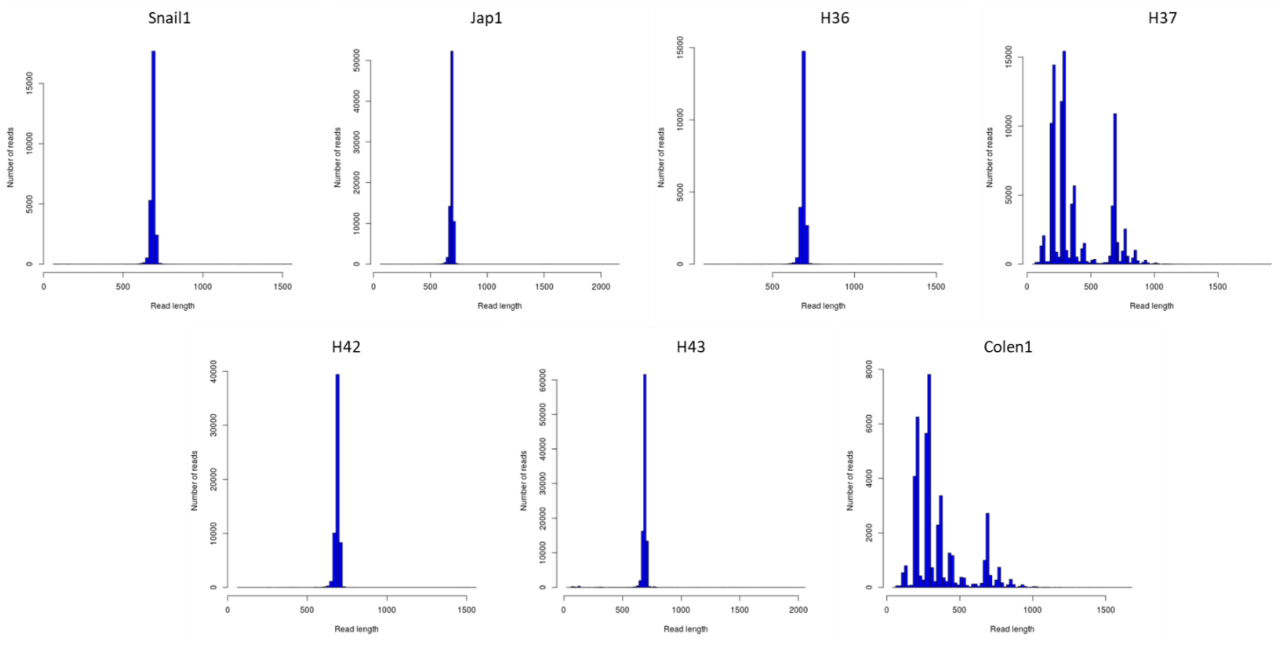


Figure 19. Read length distribution shows multiple peaks for samples H37 and Colen1.

This result confirmed that, despite higher quality data, the lack of a suitable reference sequence for novel species may result in errors in the consensus sequence produced with a reference-guided assembly approach. I therefore set out to develop a novel *de novo* assembly pipeline, which could perform accurate species identification and consensus sequence generation. After some preliminary tests and literature search, it was evident that software traditionally used for the de novo assembly of ONT reads were not suitable to assemble reads originated from amplicon sequencing, since they are designed to produce longer contigs by bridging together partially overlapping reads [36]. I then decided to implement a strategy based on multiple sequence alignment, generation of a draft consensus sequence and polishing of residual errors with a specialized tool. Since in a multiple sequence alignment all reads are aligned to each other, it is fundamental to remove unwanted reads prior to this step. Alignment of Nanopore reads to Sanger sequences confirmed that the presence of multiple peaks in read length for samples H37 and Colen1 was due to different amplification products, as only 23.92% and 17.89% of reads mapped to the Sanger sequence respectively (**Table 7**).

Table 7. Percentage of reads mapping to Sanger before and after filtering.

Sample ID (unknown species)	%Reads mapping to Sanger	%Filtered reads mapping to Sanger	
		After filtering by length	After filtering by length + clustering
Snail1	99.88%	99.97%	100%
Jap1	99.92%	99.96%	100%
H36	99.87%	99.95%	100%
H37	23.92%	96.24%	99.98%
H42	99.92%	99.97%	100%
H43	96.60%	97.65%	99.96%
Colen1	17.89%	98.66%	99.98%

As a first approach for reducing the number of unwanted reads, I filtered reads based on their length, retaining only those in the range 630 bp-730 bp. This filtering largely increased the percentage of reads mapping to Sanger (**Table 7**) and reflected in unimodal read length distribution plots (**Figure 20**).

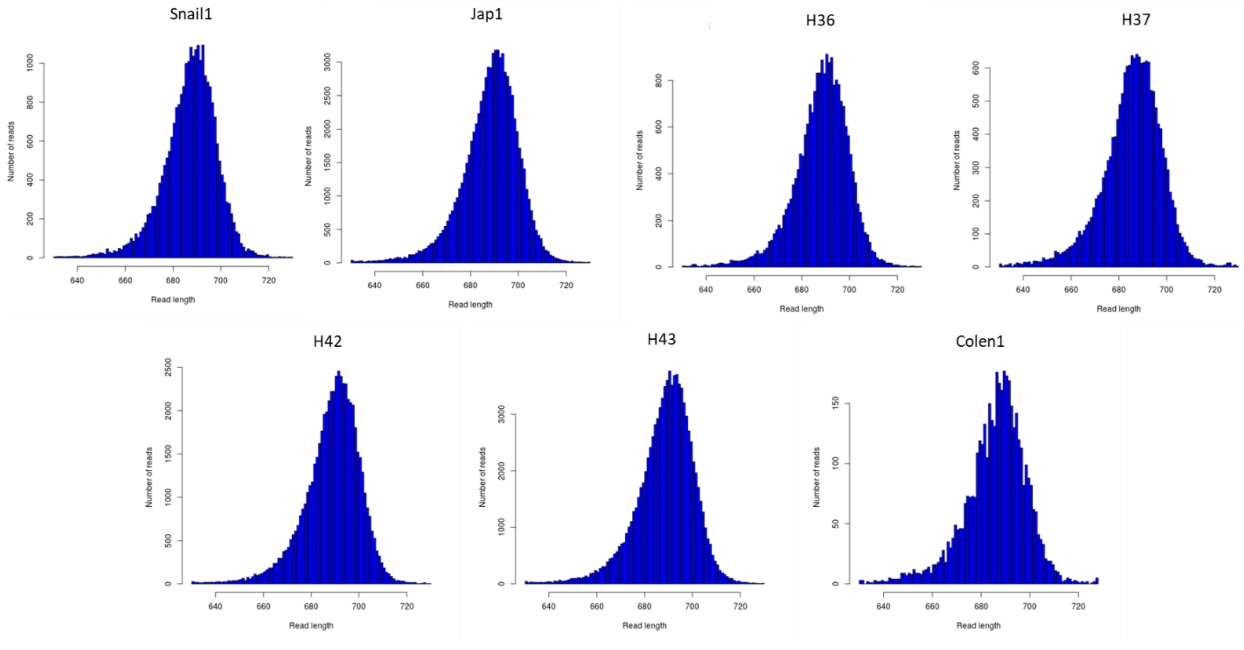


Figure 20. Read length distribution after filtering by length.

However, filtering by length may not be enough in case the unwanted reads have the same length as the target reads. For example, sample H37 still had 3.76% reads that did not map to the Sanger sequence after filtering by length. For this reason, I implemented a preliminary clustering step which aggregates reads similar to each other, possibly associated with the target amplification product, and saves to separate files all reads which are not assigned to the most abundant cluster. This approach is based on the assumption that contaminants would be less abundant than the target amplification product. However, in cases contaminants are more abundant than the target amplification product, all reads in the most abundant cluster could be discarded and the process could be repeated starting from the identified subset of reads. To test the impact of the preliminary clustering step, I reproduced three scenarios with different amounts of contaminant reads, subsampling sets of mapped and unmapped reads for sample H37. In particular, I simulated low, medium and high contamination levels, corresponding to 25%, 35% and 45% contamination levels respectively. The obtained consensus sequence, with or without the preliminary clustering step, was aligned to the Sanger, and the average alignment identity was calculated. To obtain more robust accuracy estimates, I repeated the experiment in triplicate, and calculated average accuracy values. This result showed that, while in the absence of clustering a higher contamination level is associated with lower draft consensus accuracy, the preliminary clustering completely neutralized the impact of contaminants (**Figure 21**).

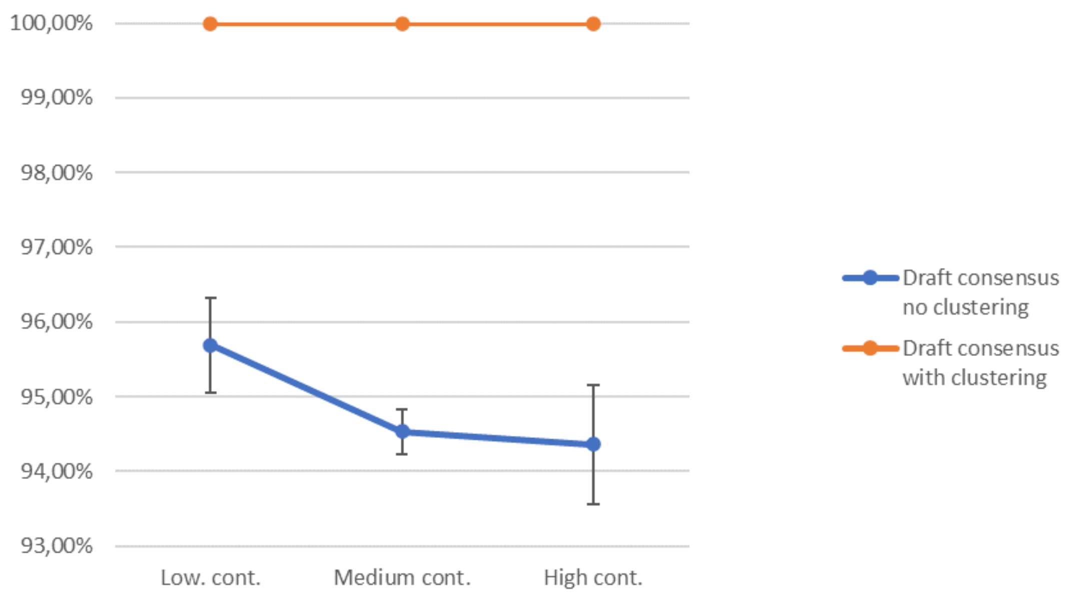


Figure 21. Impact of contaminants on draft consensus sequence accuracy. For each contamination level, the average draft consensus sequence accuracy across three replicates is reported; error bars represent standard deviation.

I then inspected the impact of contamination on the polished consensus sequences. This result showed that polishing contributes reducing the impact of contaminants, since it requires the alignment of reads to the draft consensus sequence. Still, polishing alone does not allow to reach the performances of the preliminary clustering step in presence of contaminants (**Figure 22**). Overall, these tests showed that the preliminary clustering step is able to neutralize the impact of contaminants, by greatly increasing the percentage of reads mapping to Sanger (**Table 7**).

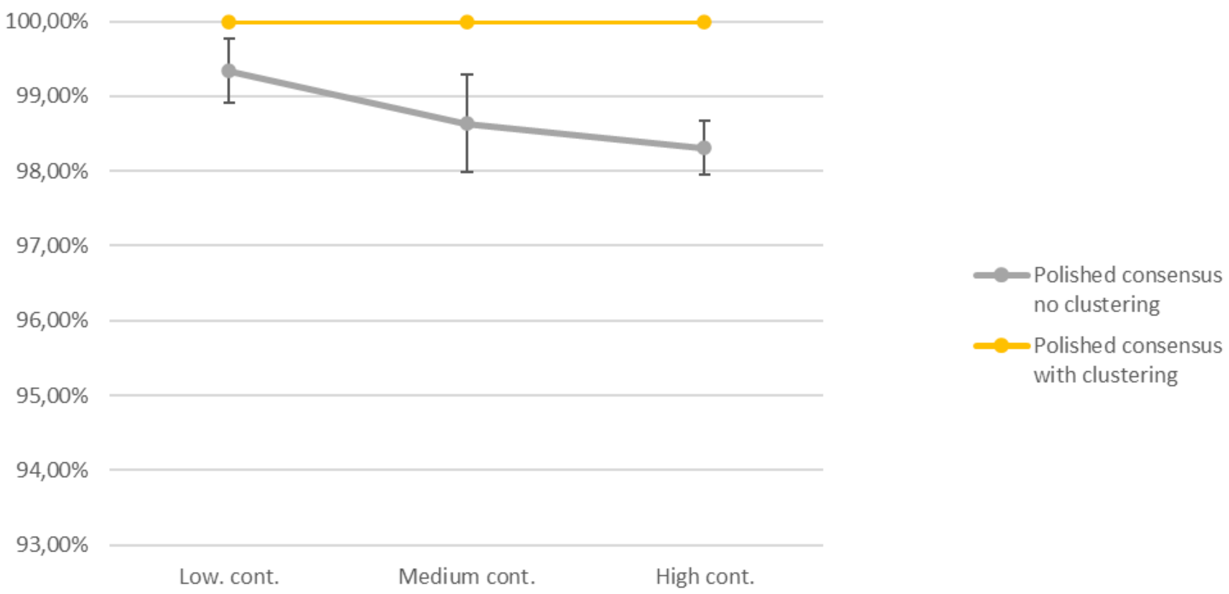


Figure 22. Impact of contaminants on polished consensus sequence accuracy. For each contamination level, the average polished consensus sequence accuracy across three replicates is reported; error bars represent standard deviation.

I then performed a downsampling analysis to identify the minimum number of reads for obtaining accurate consensus sequences. For each sample, I considered sets of reads ranging in size from 10 to 5,000, produced a consensus sequence, and evaluated the accuracy by comparing it to the Sanger. This result showed that consensus accuracy reached a plateau with 200 reads, thus making unnecessary a deeper sequencing coverage (**Figure 23**).

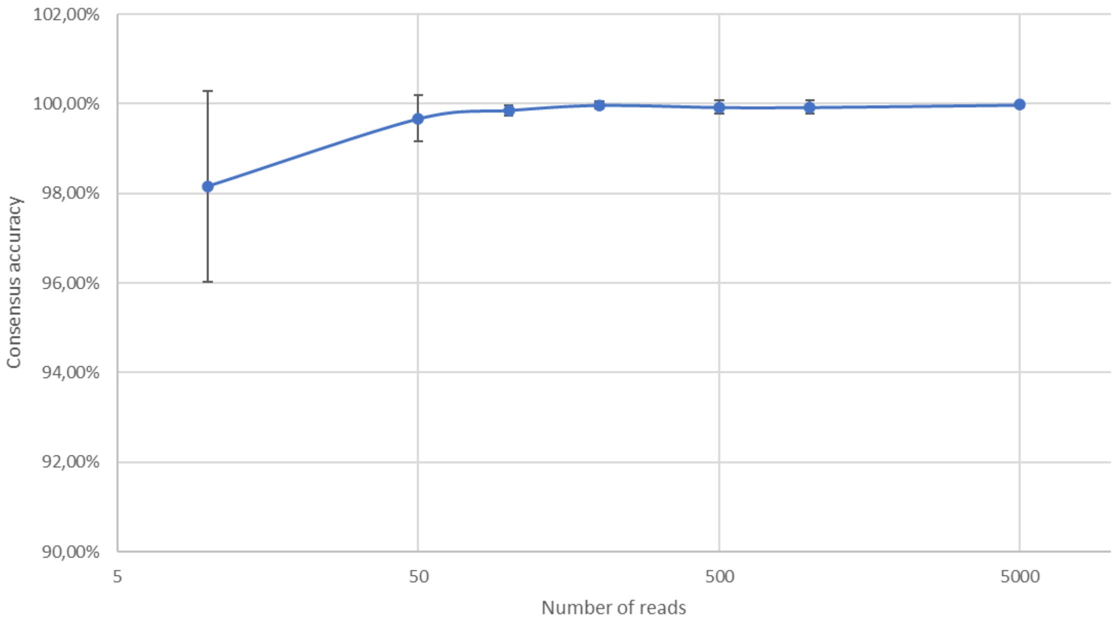


Figure 23. The impact of downsampling on consensus sequence accuracy. For each number of reads, the average polished consensus sequence accuracy across seven samples is reported; error bars represent standard deviation.

The *ONTrack* pipeline

Based on the results reported in the previous section, I developed a novel *de novo* assembly pipeline named *ONTrack* [10] (**Figure 24**). The first step of the pipeline clusters MinION sequencing reads, and only reads in the most abundant cluster are further considered for the analysis. Up to 200 reads are then randomly subsampled and a draft consensus sequence is obtained. Another set composed of up to 200 reads is randomly subsampled and used for the polishing of the consensus sequence. To test the accuracy of the *ONTrack* pipeline, I reanalysed the samples described in the previous section. Since the pipeline randomly subsamples two sets of 200 reads, I compared the consensus sequence accuracy across three iterations of the pipeline. For each iteration, the average consensus accuracy ranged from 99.95% to 100%. I then picked a final consensus accuracy for each sample, by selecting the consensus sequence which was produced on the majority of times. The *ONTrack* final consensus sequence was 100% identical to the Sanger for all samples (**Table 8**). This value represents a 0.53% increase in accuracy with respect to the *ONtoBar* pipeline, thus highlighting that a *de novo* assembly pipeline could overcome most sequencing errors by collapsing information included in multiple reads of the same gene, without relying on external information (**Table 9**).

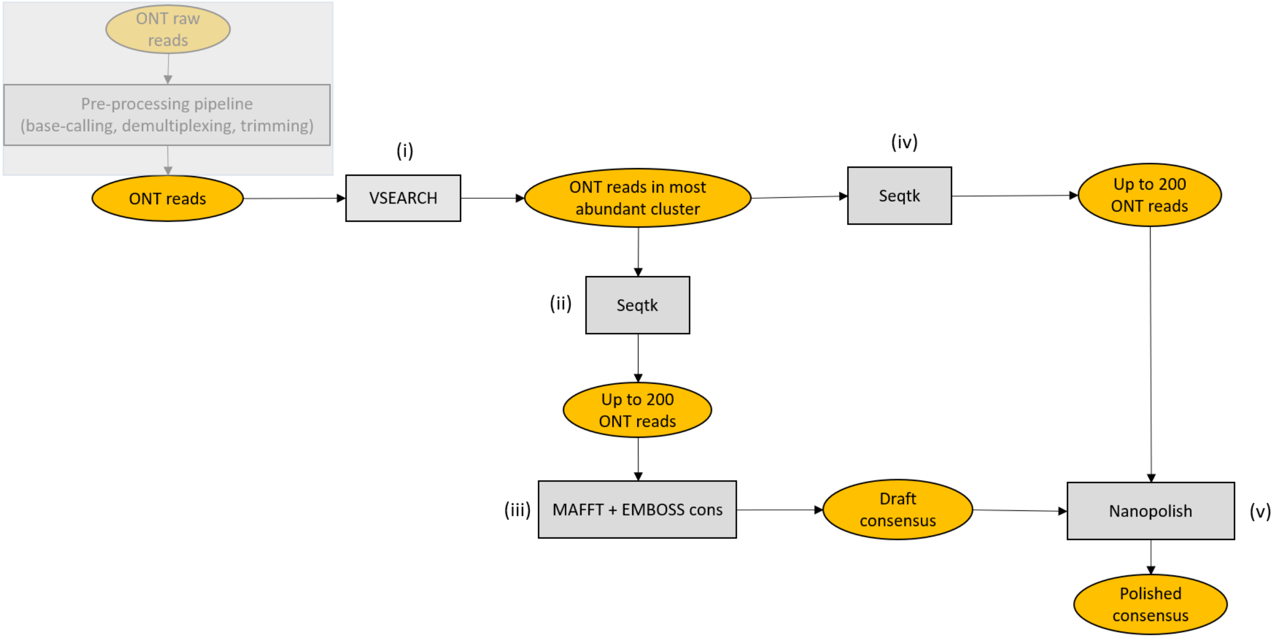


Figure 24. The *ONTrack* pipeline. ONT MinION reads are clustered to remove contaminants and only reads in the most abundant cluster are further considered (i); a set composed of up to 200 reads is randomly subsampled (ii) and used for obtaining a draft consensus sequence (iii); another set composed of up to 200 reads is randomly subsampled (iv) and used for polishing the draft consensus sequence (v).

Table 8. *ONTrack* consensus accuracy across three iterations.

Sample ID **(unknown species)**	***ONTrack* consensus accuracy read set 1**	***ONTrack* consensus accuracy read set 2**	***ONTrack* consensus accuracy read set 3**	***ONTrack* final consensus accuracy**
Snail1	100%	100%	100%	100%
Jap1	100%	100%	100%	100%
H36	100%	100%	100%	100%
H37	99.83%	100%	100%	100%
H42	100%	100%	99.85%	100%
H43	100%	100%	100%	100%
Colen1	100%	100%	99.81%	100%
Average	99.98%	100%	99.95%	100%

Table 9. Consensus accuracy of *ONtoBar* and *ONTrack* pipelines

Sample ID **(unknown species)**	**Consensus accuracy**	
	ONtoBar	***ONTrack***
Snail1	99.69%	100%
Jap1	98.87%	100%
H36	100%	100%
H37	98.43%	100%
H42	99.69%	100%
H43	100%	100%
Colen1	99.63%	100%
Average	99.47%	100%

Meta-barcoding with Nanopore sequencing

As shown in the previous section, the error rate of Nanopore reads can be largely overcome by collapsing multiple reads of the same genomic region. However, in case complex samples are analysed, each read, in principle, may belong to a different organism with biological differences in the amplified region. This hampers the possibility of performing a *de novo* assembly and requires alternative methods for the taxonomic characterization of a complex community. Compared to gold-standard methods for meta-barcoding, such as Illumina sequencing with the MiSeq platform [4,5,11], Nanopore reads trade sequencing accuracy for increased mappability, allowed by the longer reads lengths (**Figure 25**).

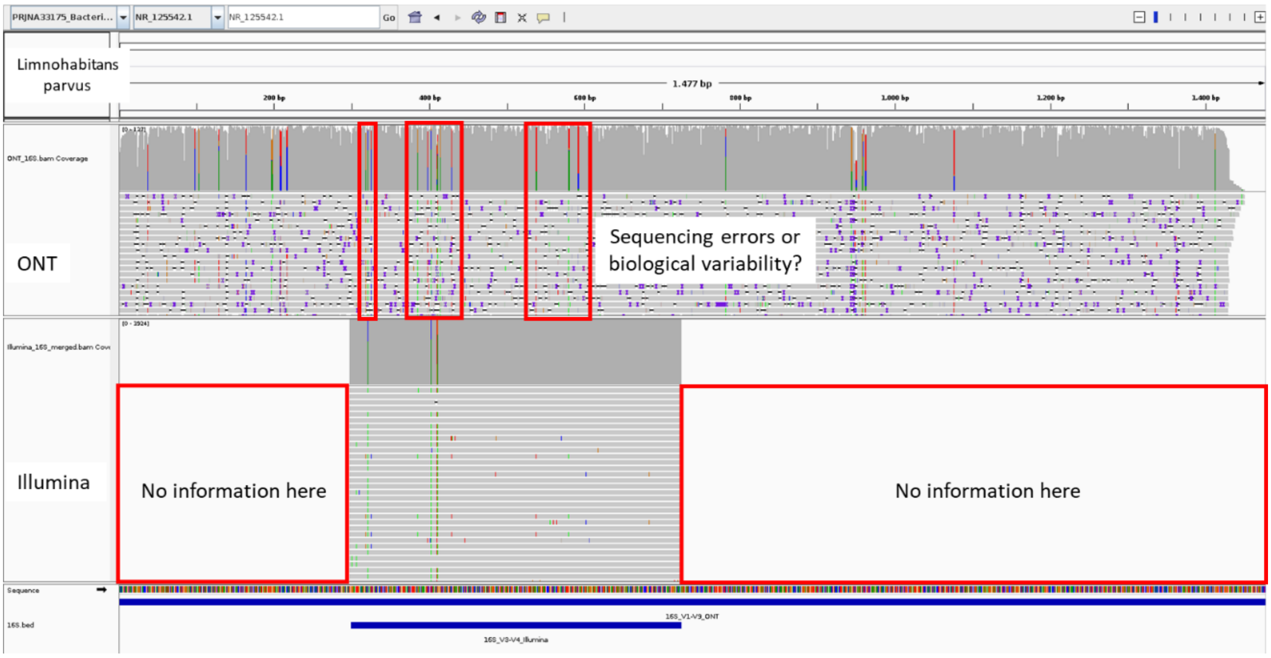


Figure 25. Nanopore and Illumina reads from a meta-barcoding experiment. The alignment of ONT and Illumina reads to species *L. parvus* is shown as an example.

Since no general guideline was available for the analysis of Nanopore meta-barcoding reads, I first tried to cluster Nanopore reads into Operational Taxonomic Units (OTUs), similarly to what is usually performed with Illumina reads [4,5,11]. Despite requiring 90% clustering identity, to account for the error rate, I observed that the number of clusters increased almost linearly with the number of reads, never reaching a plateau (**Figure 26**). This test showed that OTU-picking approaches are not suitable for the analysis of error-prone Nanopore reads, and that approaches based on the analysis of single-reads should be preferred instead. As a second test, I aligned each read to the NCBI 16S Bacterial database and inspected the alignment identity distribution of the top hits. The distribution showed an average alignment identity of 92% (SD=5%) (**Figure 27**). Moreover, a second peak at low alignment identity was detected. This result showed that species-level identification may not be reliable for all alignments. In fact, most of the alignments showed an identity lower than 97%, a threshold that is frequently used by OTU-picking approaches for species identification [5]. While stringent identity filtering may help reducing the number of misassigned species, it would result in discarding most sequencing reads, and possibly introducing some biases in the estimated taxonomic composition. In fact, reads originating from species whose sequence is unavailable in the database would be filtered out, without retaining any information at higher taxonomic levels. This issue may be overcome by performing a consensus taxonomy assignment [61]. This approach is based on the retrieval, for each read, of multiple top hits surviving the filtering criteria, and the evaluation of their taxonomy assignment; based on that, the read would be classified at species level, if all hits agree at species level, or at a higher level, if the top hits agree only at a higher taxonomic level (**Figure 28**).

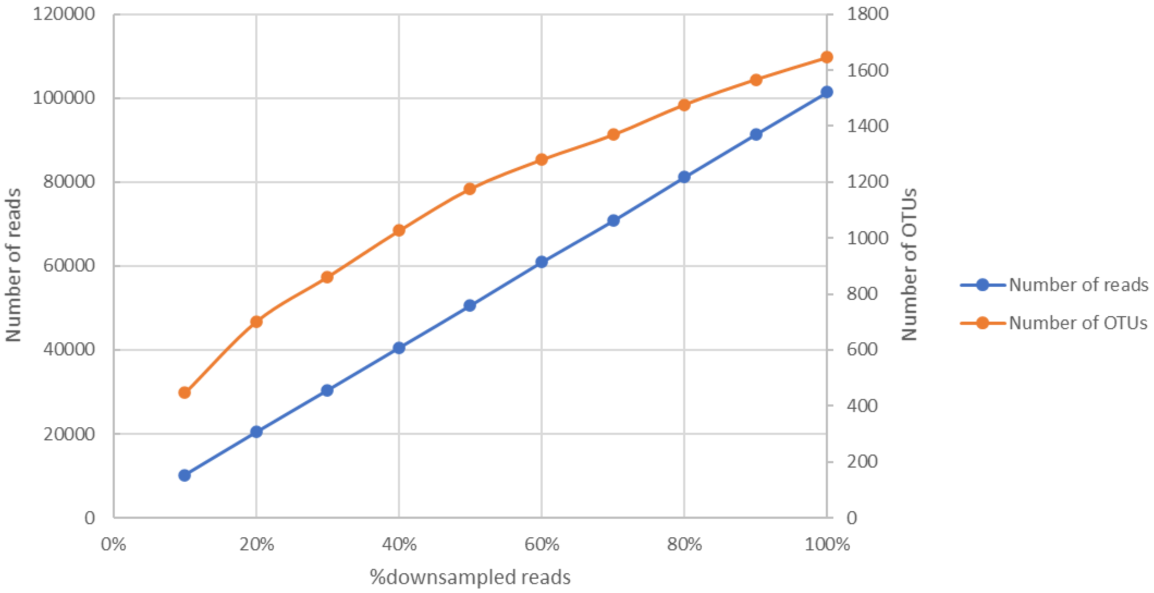
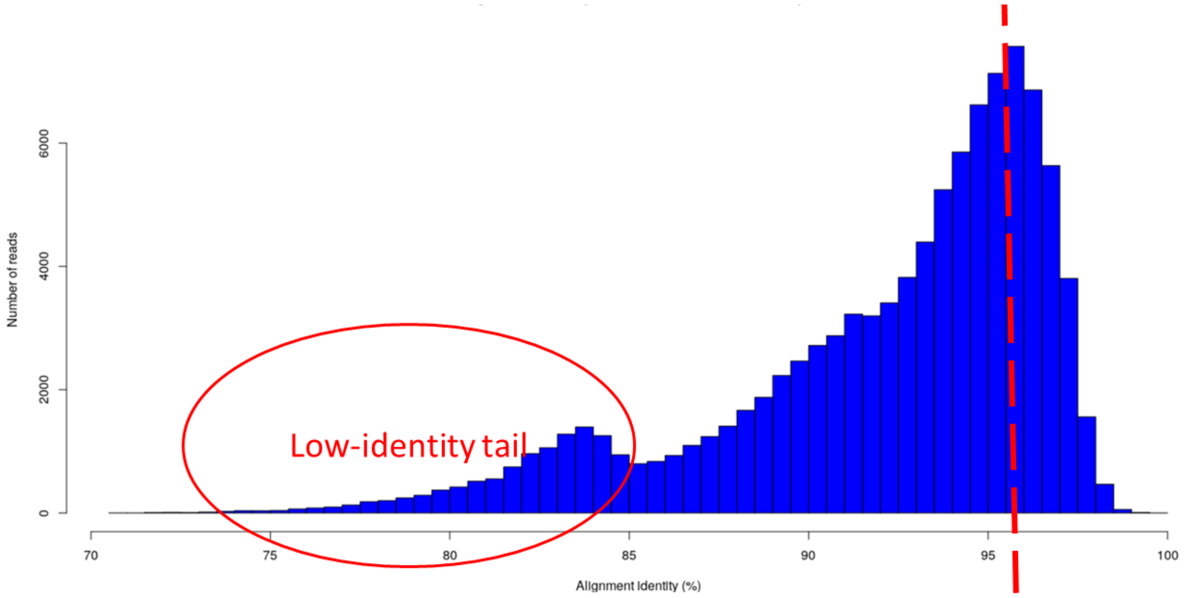
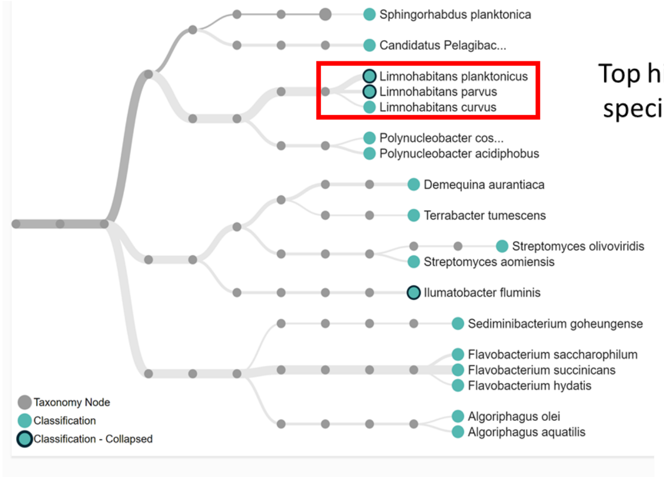


Figure 26. The number of OTUs increases almost linearly with the number of reads.



97% identity threshold frequently used for species identification

Figure 27. Single-read alignment identity distribution from a meta-barcoding experiment.



Top hits for a read belong to 3 species from the same genus

Assign taxonomy to genus level

Figure 28. Consensus taxonomy assignment. As an example, a read is assigned at genus level, since 3 top hits belong to different species from the same genus.

ONT developed an online workflow for the meta-barcoding analysis of a microbial community based on the full 16S gene. The workflow is called *EPI2ME 16S* and performs alignment of each read to the NCBI 16S Bacterial database and consensus taxonomy assignment. Despite very useful for having a first impression, the workflow is not flexible, as it does not allow the user to change alignment parameters and database. In particular, the implemented database is much smaller than other frequently used databases, containing approximately 5.5% of sequences included in the Silva database (**Table 1**).

The *MetONTIIME* pipeline

Despite enabling very quick analyses, the NCBI 16S Bacterial database may not be adequate for capturing the whole complexity of a microbial community. Moreover, the impossibility to use a more complete database hampers effective comparisons with previous analyses based on another database, due to naming inconsistencies and missing taxa. For this reason, I developed *MetONTIIME*, a meta-barcoding pipeline for analysing ONT data in *QIIME2* environment (**Figure 29**). The *MetONTIIME* pipeline performs alignment of single reads and consensus taxonomy assignment as the *EPI2ME 16S* workflow, but it also allows the user to specify a custom database and to tune alignment parameters.

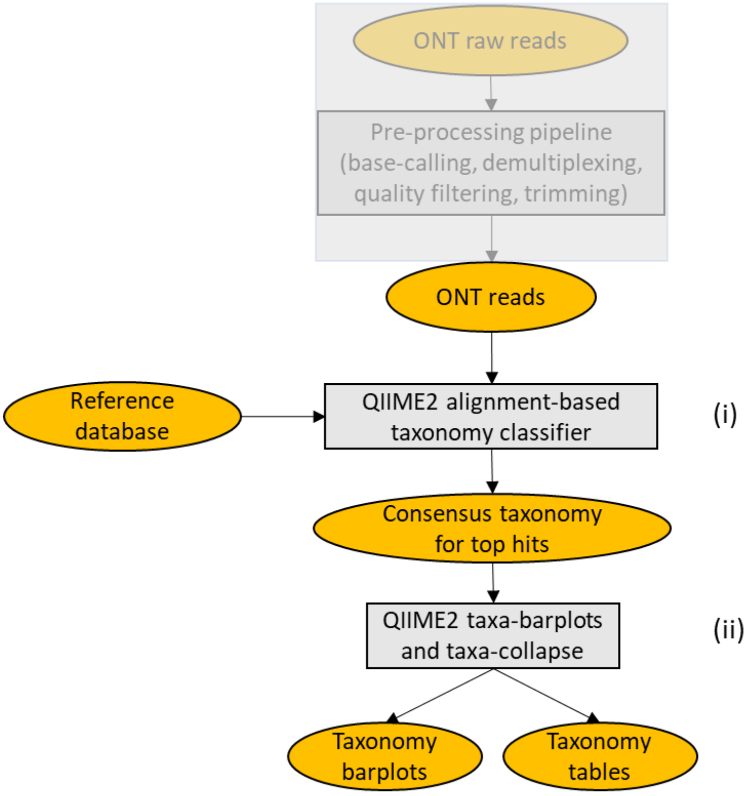

ONT raw reads
Pre-processing pipeline (base-calling, demultiplexing, quality filtering, trimming)
ONT reads
Reference database
QIIME2 alignment-based taxonomy classifier
(i)
Consensus taxonomy for top hits
QIIME2 taxa-barplots and taxa-collapse
(ii)
Taxonomy barplots
Taxonomy tables

Figure 29. The *MetONTIIME* pipeline. ONT reads are aligned to a reference database and consensus taxonomy assignment is performed (i); then, taxonomy barplots and tables are produced (ii).

For validating the pipeline, we first generated Nanopore sequencing data (R9.4 sequencing chemistry) from an environmental sample obtained filtering the polluted water of river Tiber (**Table 10**). As a first test, I compared the taxonomic classifications obtained using the *EPI2ME 16S* workflow and the *MetONTIIME* pipeline with the same database and parameters. The two pipelines showed high correlation at genus level (R^2=0.99), thus indicating that the *MetONTIIME* pipeline can be used in place of the *EPI2ME 16S* workflow with very similar results (**Figure 30**).

Table 10. Sequencing data generated for the validation of the *MetONTIIME* pipeline.

Sample name	**Sequencing platform**	**Genomic target**	**Number of filtered fragments**	**Mean bp fragment length (SD)**
Env. sample	Illumina	V3-V4 16S	217,085	415 (14)
	Nanopore	Full 16S	53,013	1,412 (41)
Env. Sample + [spike-in]$_{min}$	Illumina	V3-V4 16S	253,270	415 (13)
	Nanopore	Full 16S	49,314	1,414 (39)
Env. Sample + [spike-in]$_{max}$	Illumina	V3-V4 16S	256,460	415 (14)
	Nanopore	Full 16S	59,168	1,416 (41)
	Nanopore	WGS	451,469	2,641 (3,237)

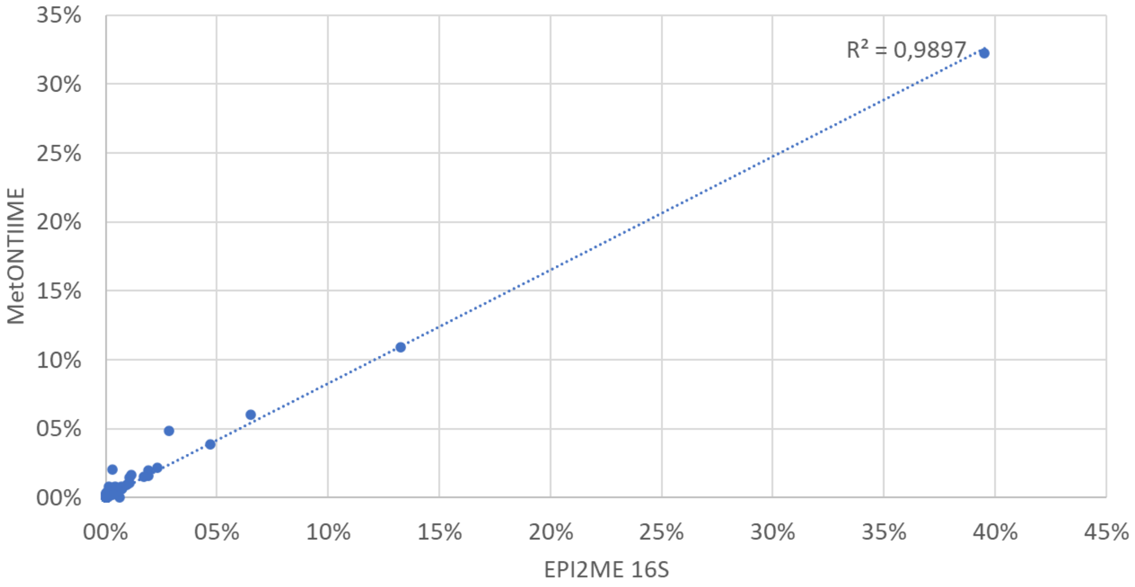


Figure 30. Correlation at genus level between the *EPI2ME 16S* and the *MetONTIIME* pipelines. Relative abundance of reads at genus level is shown.

Nanopore meta-barcoding compared to gold-standard methods

For further validating Nanopore meta-barcoding with the *MetONTIIME* pipeline, we generated meta-barcoding data from the same environmental sample with a platform that is currently considered the gold-standard for meta-barcoding analyses, namely Illumina data with the MiSeq platform in 300 PE configuration (**Table 10**) [8]. I processed Illumina data in QIIME2 environment and performed taxonomic classification using Silva database, producing taxonomy tables at different taxonomic levels [61] (**Figure 31**).

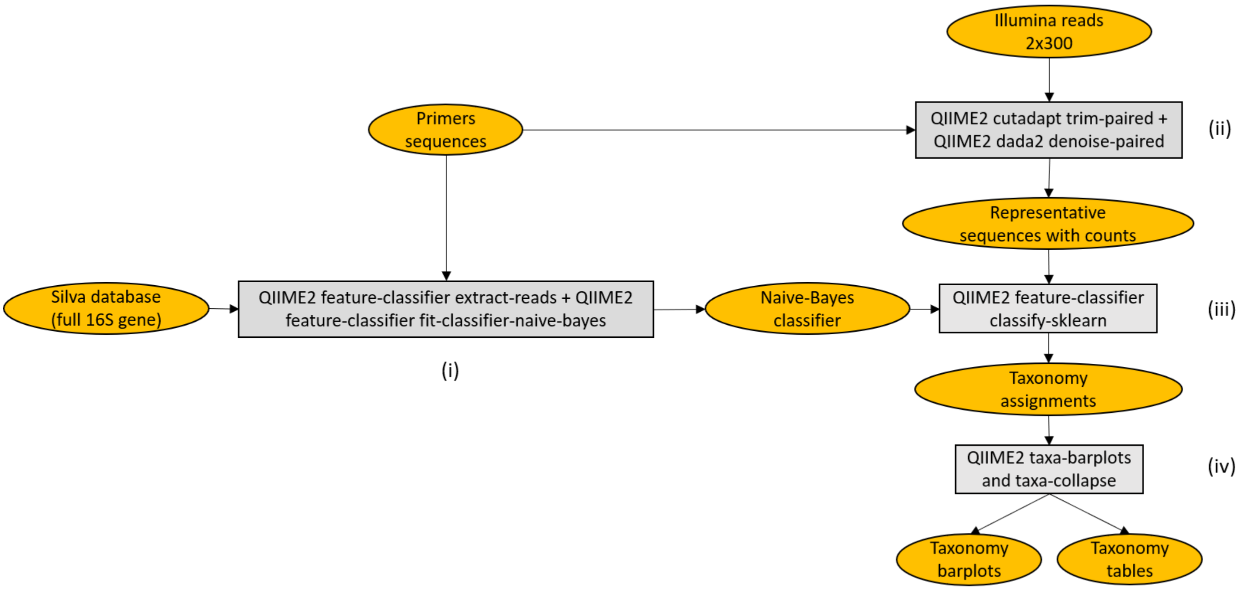

Illumina reads 2x300
Primers sequences
QIIME2 cutadapt trim-paired + QIIME2 dada2 denoise-paired
(ii)
Representative sequences with counts
Silva database (full 16S gene)
QIIME2 feature-classifier extract-reads + QIIME2 feature-classifier fit-classifier-naive-bayes
Naive-Bayes classifier
QIIME2 feature-classifier classify-sklearn
(iii)
(i)
Taxonomy assignments
QIIME2 taxa-barplots and taxa-collapse
(iv)
Taxonomy barplots
Taxonomy tables

Figure 31. Analysis pipeline for Illumina meta-barcoding data. A Naïve-Bayes classifier is trained on the Silva database, after extracting the region corresponding to the V3-V4 region of the 16S gene amplified by PCR primers (i); Illumina reads are trimmed, denoised and merged, to obtain representative sequences with associated counts (ii); the representative sequences are then assigned a taxonomy with the trained classifier (iii); barplots and taxonomy tables are finally produced.

I then analysed Nanopore data from the same sample using the *MetONTIIME* pipeline with the Silva database, and performed comparisons between the relative abundances found by the two platforms at phylum, genus, and species levels. Analysis at phylum level showed good correlation between the two platforms (R^2=0.74), with 84.4% of phyla identified by both platforms. When considering only phyla with >0.5% abundance, the percentage of shared phyla decreased to 77.8%, with 2 phyla detected only by Illumina; these 2 phyla were also detected by Nanopore platform, but with an abundance <0.5%

(**Figure 32**). Analysis at genus level showed again good correlation between the two platforms (R^2=0.71), with 46.3% of genera identified by both platforms. When considering only genera with >0.5% abundance, the percentage of shared genera increased to 58.0%. All 6 proprietary Illumina genera were also detected by Nanopore platform, but with an abundance <0.5%; while only 5 of the 7 Nanopore proprietary genera were also detected by Illumina platform, but with an abundance < 0.5% (**Figure 33**).

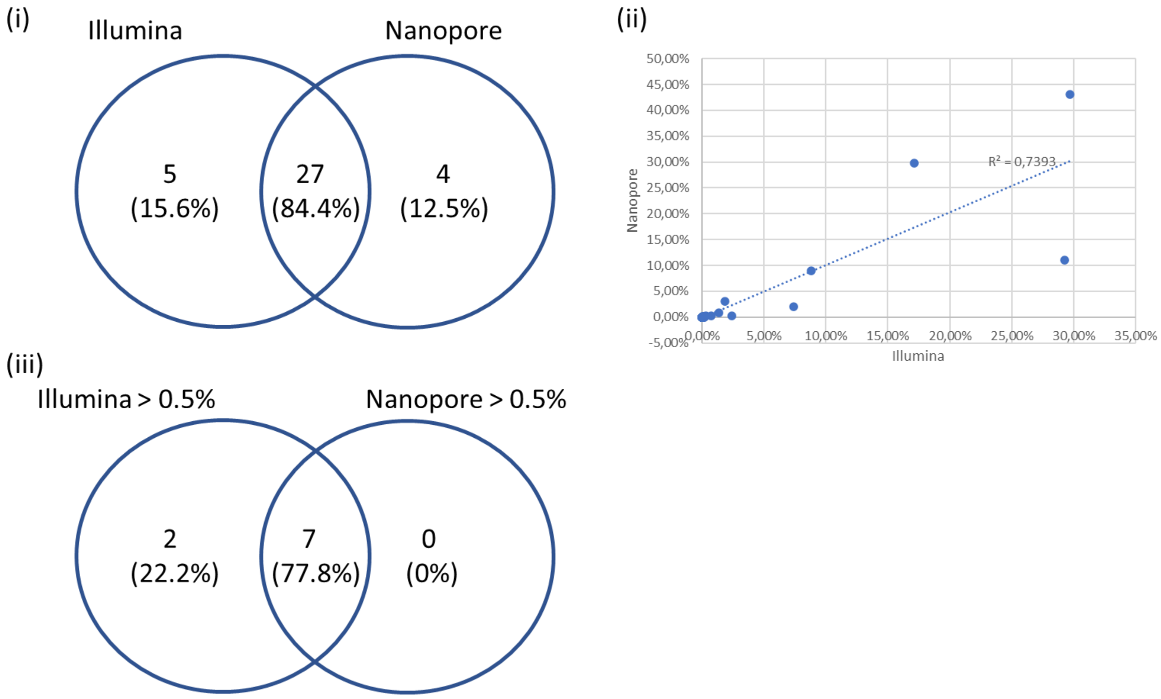


Figure 32. Comparison between Illumina and Nanopore meta-barcoding at phylum level. Number of phyla identified by the two platform (i); correlation between relative abundance of phyla identified by the two platform (ii); number of phyla with relative abundance > 0.5% identified by the two platform (iii).

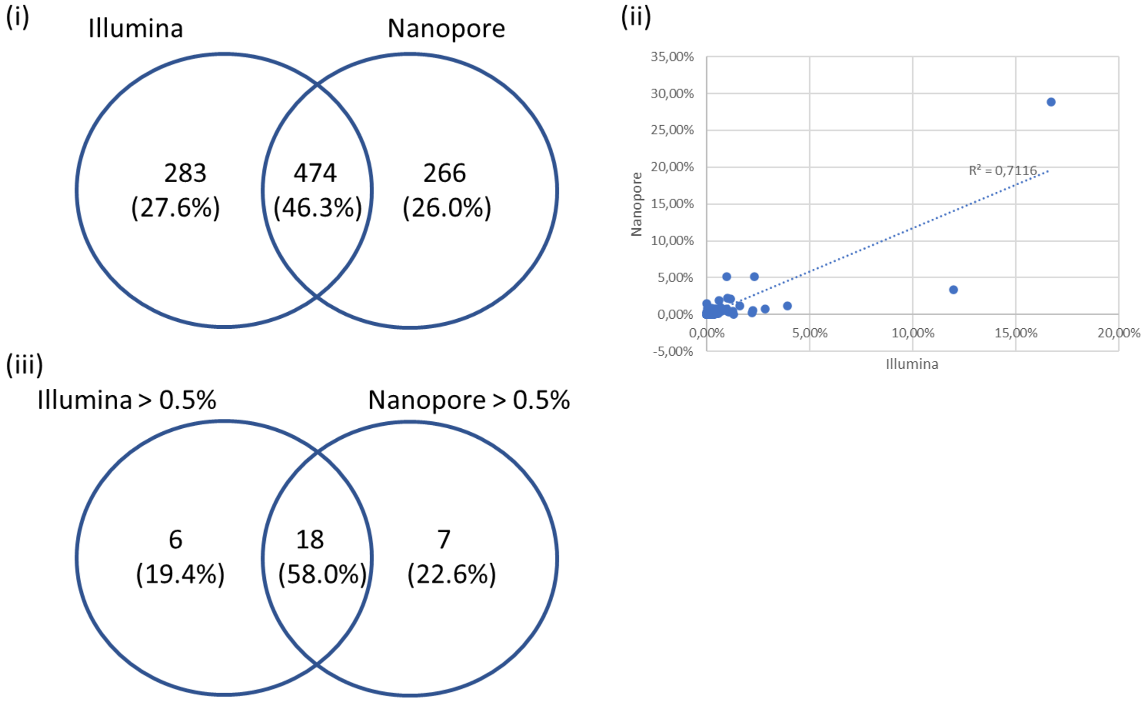


Figure 33. Comparison between Illumina and Nanopore meta-barcoding at genus level. Number of genera identified by the two platform (i); correlation between relative abundance of genera identified by the two platform (ii); number of genera with relative abundance >0.5% identified by the two platform (iii).

Conversely, analysis at species level showed low correlation between the two platforms (R^2=0.18), with only 25.1% of species identified by both platforms. When considering only species with >0.5% abundance, the percentage of shared species decreased to 12.0%. Only 3 of the 6 proprietary Illumina species were also detected by Nanopore platform, but with an abundance <0.5%; while only 6 of the 16 Nanopore proprietary species were also detected by Illumina platform, but with an abundance <0.5% (**Figure 34**).

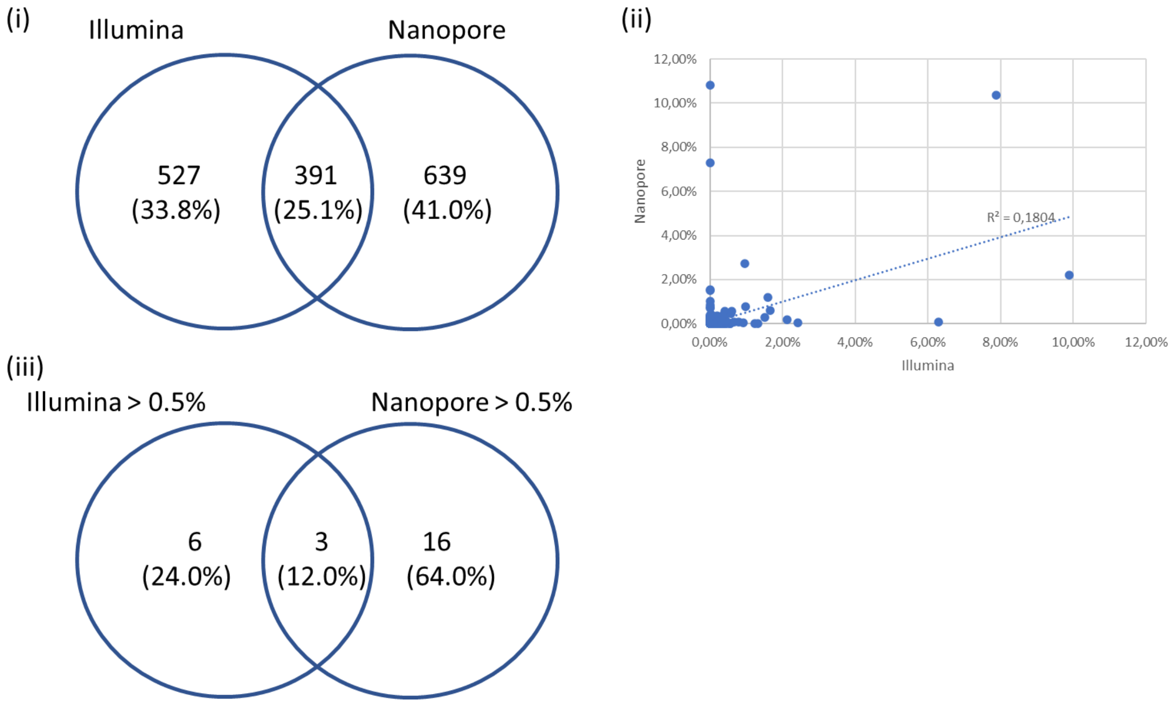


Figure 34. Comparison between Illumina and Nanopore meta-barcoding at species level. Number of species identified by the two platform (i); correlation between relative abundance of species identified by the two platform (ii); number of species with relative abundance >0.5% identified by the two platform (iii).

The analysis of indicator species

Overall, the results presented in the previous section indicated that Illumina and Nanopore platforms had good concordance up to genus level but showed marked differences at species level. Following the legislative decree of February 2nd 2001, No. 31 and the Directive 2006/7/EC of February 15th [62,63], the water quality assessment in Italy should rely on the culture-based analysis of 5 indicator species. For this reason, I first inspected if indicator species could be detected by either platform in the environmental sample. This analysis showed that only 2 and 3 genera out of 5 could be detected with Illumina and Nanopore platform, respectively. Moreover, Illumina could not detect any of the indicator species, while Nanopore detected only one (**Figure 35**).

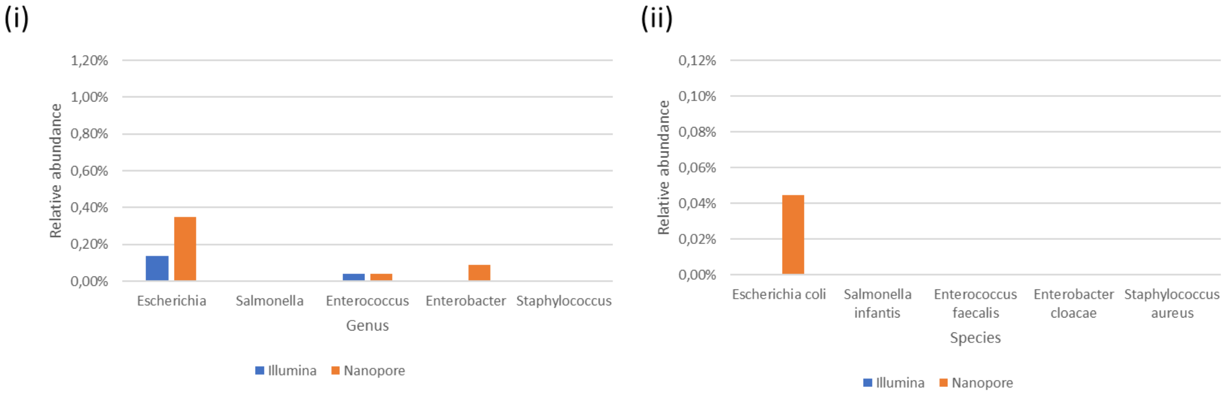


Figure 35. Analysis of indicator species with Illumina and Nanopore platforms. Analysis is performed at genus (i) and at species (ii) levels.

This result led us to question if indicator species were not found in the environmental sample either because they were not present, or because of some technological issues. For this aim, we added to the environmental sample a spike-in composed of the indicator species at two different concentrations (min and max) and we sequenced them with the two platforms (**Table 10**).

Moreover, we sequenced with Nanopore the environmental sample with spike-in at max concentration using a shotgun metagenomics approach (**Table 10**). I analysed meta-barcoding data from the spiked-in samples as described previously, and Nanopore shotgun data with a pipeline based on Kraken2 classifier [43] (**Figure 36**).

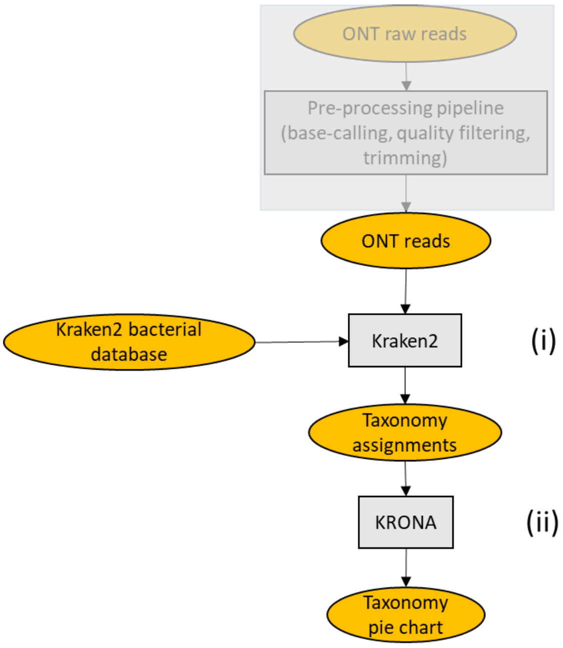

ONT raw reads
Pre-processing pipeline (base-calling, quality filtering, trimming)
ONT reads
Kraken2 bacterial database
Kraken2
(i)
Taxonomy assignments
KRONA
(ii)
Taxonomy pie chart

Figure 36. Analysis pipeline for Nanopore shotgun metagenomics data. ONT reads are classified with Kraken2 using the standard Kraken 2 database for bacteria (i); taxonomy assignments are then used by Krona to produce taxonomy pie chart (ii).

The analysis of indicator species at genus level showed that, after the addition of the spike-in, Illumina could detect up to 3 of the 5 taxa, while Nanopore could detect up to 4 of the 5 taxa (**Figure 37**); moreover, an increase in the average percentage of reads assigned to the 5 genera after the addition of the spike-in was observed with both platforms. Conversely, the analysis of indicator species at species level showed that Illumina was not able to detect any of the indicator species, while Nanopore could detect up to 3 of the 5 taxa (**Figures 38**); moreover, Nanopore showed an increase in the average percentage of reads assigned to the 5 species after the addition of the spike-in (**Figure 39**). Shotgun metagenomics analysis showed that all 5 genera were present in the environmental sample with the addition of spike-in at maximum concentration, while no reads were assigned to *S. infantis*.

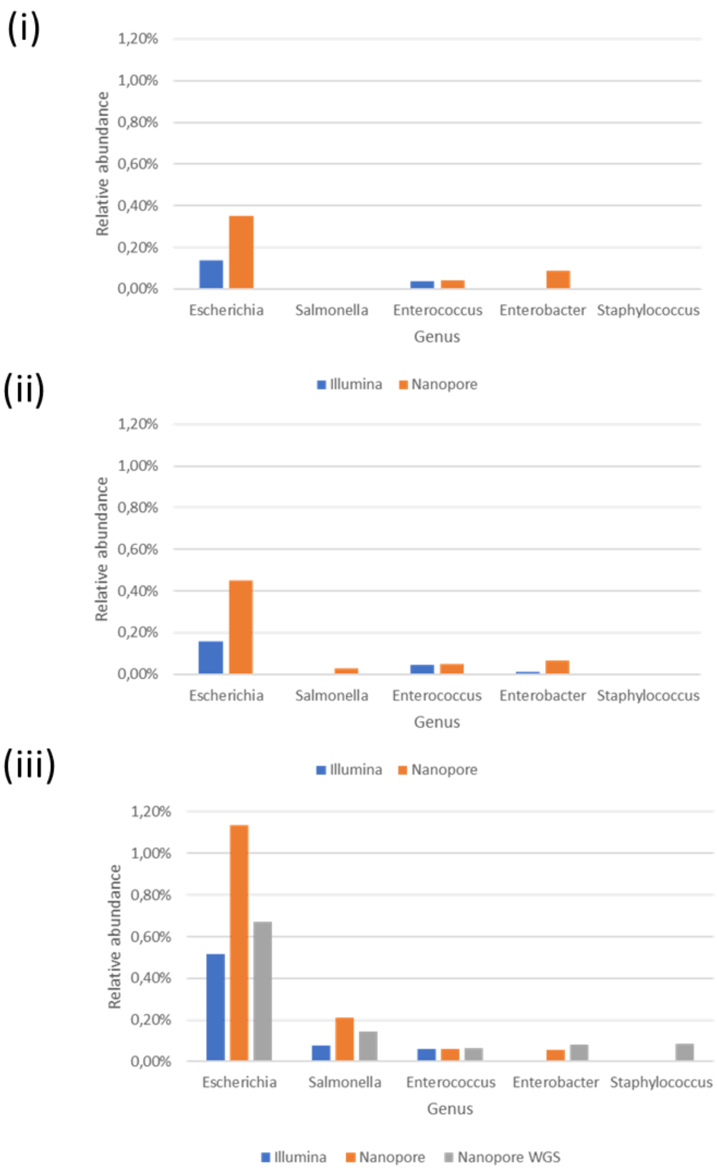


Figure 37. Analysis of indicator species at genus level. Relative abundance of genera corresponding to indicator species in the environmental sample (i), after addition of spike-in$_{min}$ (ii) and after addition of spike-in$_{max}$ (iii).

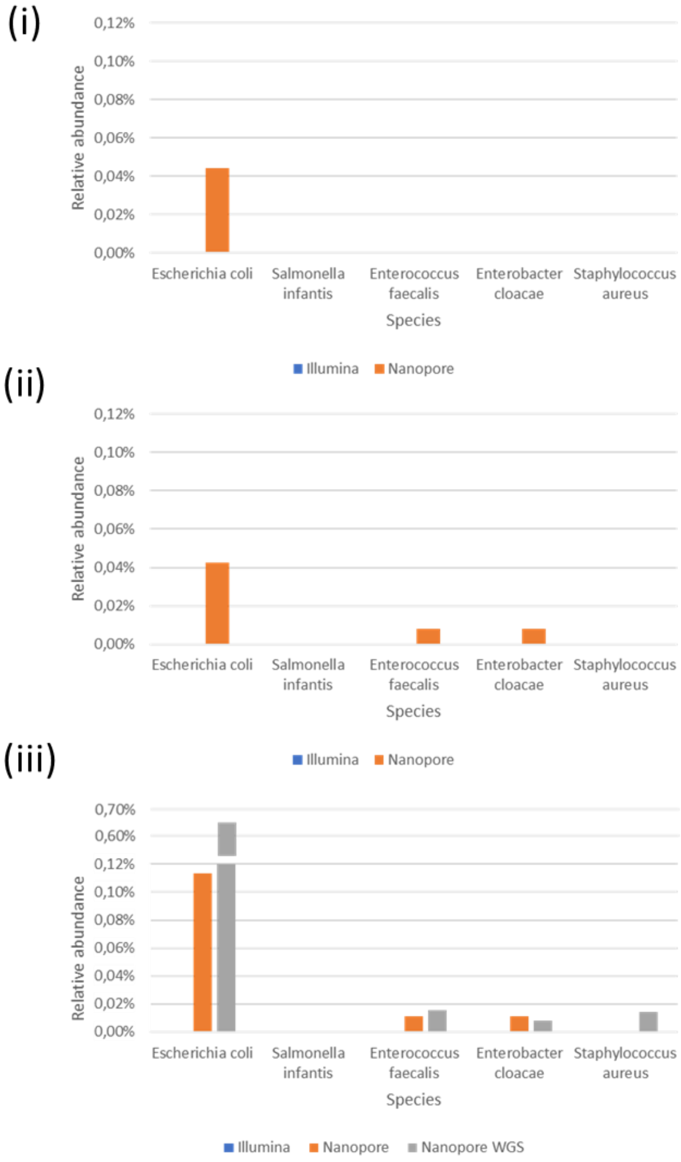


Figure 38. Analysis of indicator species at species level. Relative abundance indicator species in the environmental sample (i), after addition of spike-in$_{min}$ (ii) and after addition of spike-in$_{max}$ (iii).

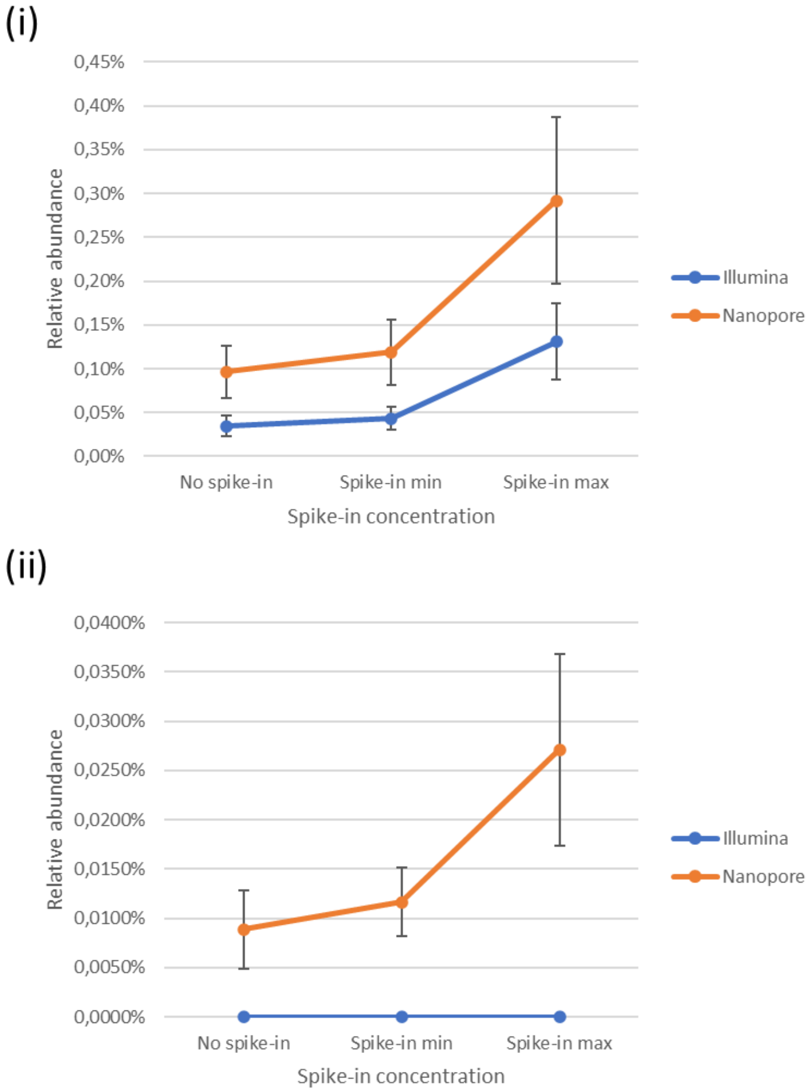


Figure 39. Relative abundance of indicator species after spike-in addition. The average relative abundance of indicator species at genus (i) and species (ii) level is shown; error bars represent the standard error.

Discussion

In this thesis I have shown that barcoding with Nanopore sequencing has almost reached accuracy comparable with Illumina and Sanger platforms. Moreover, the bioinformatic analysis of Nanopore reads does not need a big computing infrastructure and takes only 10 minutes per sample on an ASUS laptop with 16 GB RAM, i7 processor and SSD disk. The whole analysis does not require internet connection, since in the last few years ONT has provided software for sequencing and base-calling that can work offline, and databases for comparing consensus sequences to known species can be downloaded locally [10]. This has been possible thanks to a steady improvement in both the sequencing chemistry and software performances. Until a few years ago, Nanopore reads showed higher error rate and could not be used for accurate *de novo* assembly. Only resequencing was possible, and alignment of the reads to the reference sequence could either confirm the species, if none or few variants were identified, or suggest a different species was sequenced. Moreover, reads were base-called online, thus requiring a good internet connection [13]. After our first demonstration of barcoding in the field [13], the advent of a new sequencing chemistry and of a new generation of offline base-callers made *de novo* assembly in the field for species identification feasible [14,64]. However, none of these works proposed an automated and user-friendly bioinformatic pipeline providing consensus sequences in a streamlined way, as they either adopted sample-dependent bioinformatic processing [14], or they relied on an external database for final error correction [64]. The novel barcoding pipeline described in this work could address issues which may arise in field settings, as possible contamination or limited computing resources [10]. All recent works describing barcoding in the field showed that consensus sequences generated with Nanopore sequencing have reached high quality and have the potential to replace Illumina and Sanger platforms [7,10,20,26,30,65-67]. The newest R10.3 chemistry is expected to enable the final step for consensus accuracy to reach Sanger quality, due to the presence of pores with a longer barrel and a dual reader head, which allow better consensus accuracy in homopolymer regions [65,68,69]. Moreover, smaller MinION flow cells (Flongles) were recently made available, suitable for experiments that do not require a massive throughput, thus substantially reducing sequencing costs for small datasets. Since a modest number of reads per sample is required for accurate barcoding, multiple samples could be multiplexed in a single run and still fit Flongle specifications (1 Gbp), thus bringing also costs on the same level of Sanger platforms [10].

If barcoding with Nanopore platform has almost reached maturity, meta-barcoding is yet to be optimized, due to the impossibility to reduce the error rate by collapsing the information coming from multiple reads of the same sequence. Preliminary proof-of-concept studies showed the feasibility of obtaining high-accuracy consensus sequences from meta-barcoding experiments by integrating Unique Molecular Identifiers (UMIs) in PCR amplicons [68] or by performing Rolling Circle Amplification (RCA) [70,71]. These approaches enable the grouping of amplification products originating from the same molecule, either exploiting the presence a molecule specific UMI or physical contiguity of copies amplified by RCA. After reads from the same amplification product are grouped, bioinformatic pipelines similar to those developed for barcoding could be applied on each read bin. However, these approaches are yet to be optimized, as they imply laborious wet-lab protocols and they require very high sequencing throughput compared to the number of obtained consensus sequences [68]. Currently, Nanopore reads from meta-barcoding experiments are generally classified by aligning single reads to a reference database [9,19,25,72]. In fact, approaches based on *de novo* OTU picking, namely clustering of reads belonging to the same species, are not suitable to the analysis of Nanopore reads from complex samples due to the high error rate, and their implementation should be evaluated carefully [8,73]. A few manuscripts recently described a low-identity clustering step in their meta-barcoding pipeline, mainly aimed at reducing computational time of the analysis [7,18]; however, the implementation of this approach should be evaluated carefully, since the analysis of Nanopore reads with inappropriate OTU clustering tool could provide a completely incorrect picture of the diversity of the sample [8]. The implementation of a closed-reference OTU picking strategy with Nanopore reads was recently described [74,75]. These approaches combine alignment of reads to a reference database, grouping of reads based on the taxonomy of the top hit and consensus calling. While this approach should work well for low complexity samples, it may struggle with samples including closely related species, or species not included in the database. Despite a general agreement on the recommended approach, there is no consensus on the tool used for the alignment, software parameters, and the adopted database. The most extensively used pipeline is the cloud-based data analysis *EPI2ME 16S* workflow provided by ONT [8]. The main limitations of this workflow are identified in the impossibility of changing the reference database and requiring a stable internet connection for the analysis, limiting portability and meaningful comparisons with analyses based on another database. The novel meta-barcoding pipeline presented in this work addresses both these issues and provides evidence that the increased read length of Nanopore reads alleviates their higher error rate. In fact, results showed good concordance up to genus level with the gold-standard method for meta-barcoding [4,5,8,11], and proved to be even more sensitive to detect an increase in the average relative abundance of five spiked-in indicator species. Differences between Illumina and Nanopore-based meta-barcoding abundances may also be explained by taxa-specific amplification efficiency of different PCR primers, and by the low taxonomic resolution at the species level for some bacterial groups [8]. Still, not all indicator species could be detected by Nanopore meta-barcoding analysis. For example, *S. aureus* could not be detected at genus and species level, while it was detected with a shotgun metagenomics approach based on Nanopore sequencing. Again, this may be either due to PCR amplification-related issues or to high similarity between the 16S gene of *S. aureus* and other species in the database. Similarly, *S. infantis* was not detected at species level by Nanopore meta-barcoding analysis. However, this species was not detected by the shotgun metagenomics analysis as well, pointing towards issues related to DNA extraction. Finally, differences in relative abundance obtained comparing meta-barcoding to shotgun metagenomics approaches may be due to the variable number of copies of 16S gene in bacterial genomes [8].

In conclusion, the results presented in this work show that the development of suitable bioinformatic pipelines is contributing to make barcoding and meta-barcoding analyses with Nanopore platform more accurate and reliable, enabling on-site analysis of data generated with a portable genomics laboratory. In particular, this work shows that monitoring of biodiversity and identification of risks for human health with MinION-based DNA barcoding now provides results comparable in accuracy to gold-standard sequencing platforms. The small turnaround time from sample to result will aid in the transition to real-time environmental monitoring, providing quick and accurate information for decision making processes.

Bibliography

1. Hebert, P.; Cywinska, A.; Ball, S.; deWaard, J. Biological identifications through DNA barcodes. *Proc Biol Sci* **2003**, *270*, doi:https://doi.org/10.1098/rspb.2002.2218.
2. Ratnasingham, S.; Hebert, P. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS One* **2013**, *8*.
3. Hebert, P.; Ratnasingham, S.; deWaard, J. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci* **2003**, *270*, S96-S99.
4. Bálint, M.; Schmidt, P.; Sharma, R.; Thines, M.; Schmitt, I. An Illumina metabarcoding pipeline for fungi. *Ecol. Evol.* **2014**, *4*, 2642-2653.
5. Rocchi, S.; Valot, B.; Reboux, G.; Millon, L. DNA metabarcoding to assess indoor fungal communities: Electrostatic dust collectors and Illumina sequencing. *J. Microbiol. Methods* **2017**, *139*, 107-112.
6. Rotimi, A.; Pierneef, R.; Reva, O. Selection of marker genes for genetic barcoding of microorganisms and binning of metagenomic reads by Barcoder software tools. *BMC Bioinformatics* **2018**, *19*.
7. Davidov, K.; Iankelevich-Kounio, E.; Yakovenko, I.; Koucherov, Y.; Rubin-Blum, M.; Oren, M. Identification of plastic-associated species in the Mediterranean Sea using DNA metabarcoding with Nanopore MinION. *Sci Rep.* **2020**, *10*.
8. Santos, A.; van Aerle, R.; Barrientos, L.; Martinez-Urtaza, J. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Comput Struct Biotechnol J.* **2020**, *18*, 296-305.
9. Acharya, K.; Khanal, S.; Pantha, K.; Amatya, N.; Davenport, R.; Werner, D. A comparative assessment of conventional and molecular methods, including MinION nanopore sequencing, for surveying water quality. *Sci Rep.* **2019**, *9*.
10. Maestri, S.; Cosentino, E.; Paterno, M.; Freitag, H.; Garces, J.; Marcolungo, L.; Alfano, M.; Njunjić, I.; Schilthuizen, M.; Slik, F., et al. A Rapid and Accurate MinION-Based Workflow for Tracking Species Biodiversity in the Field. *Genes (Basel)* **2019**, *10*.
11. Sato, Y.; Mizuyama, M.; Sato, M.; Minamoto, T.; Kimura, R.; Toma, C. Environmental DNA metabarcoding to detect pathogenic Leptospira and associated organisms in leptospirosis-endemic areas of Japan. *Sci Rep.* **2019**, *9*.
12. Mancini, L.; Marcheggiani S; Puccinelli, C.; Lacchetti, I.; Carere, M.; Bouley, T. Global environmental changes and the impact on ecosystems and human health. *La Sanità tra Scienza e Tecnologia* **2017**, *3*, 98-105.
13. Menegon, M.; Cantaloni, C.; Rodriguez-Prieto, A.; Centomo, C.; Abdelfattah, A.; Rossato, M.; Bernardi, M.; Xumerle, L.; Loader, S.; Delledonne, M. On site DNA barcoding by nanopore sequencing. *PLoS ONE* **2017**, *12*.
14. Pomerantz, A.; Peñafiel, N.; Arteaga, A.; Bustamante, L.; Pichardo, F.; Coloma, L.; Barrio-Amorós, C.; Salazar-Valenzuela, D.; Prost, S. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* **2018**, *7*.
15. Schilthuizen, M.; Lim, J.; van Peursen, A.; Alfano, M.; Jenging, A.; Cicuzza, D.; Escoubas, A.; Escoubas, P.; Grafe, U.; Ja, J., et al. Craspedotropis gretathunbergae, a new species of Cyclophoridae (Gastropoda: Caenogastropoda), discovered and described on a field course to Kuala Belalong rainforest, Brunei. *Biodivers Data J.* **2020**, *8*.
16. Watsa, M.; Erkenswick, G.; Pomerantz, A.; Prost, S. Portable sequencing as a teaching tool in conservation and biodiversity research. *PLos Biol* **2020**, *18*, doi:https://doi.org/10.1371/journal.pbio.3000667.

17. Schilthuizen, M.; van Oostenbrugge, W.; Visser, S.; van der Meer, M.; Delval, R.; Dias, C.; Köster, H.; Maarschall, R.; Peeters, R.; Venema, P., et al. Ptomaphagus thebeatles n. sp., a previously unrecognized beetle from Europe, with remarks on urban taxonomy and recent range expansion (Coleoptera: Leiodidae). *Contributions to Zoology* **2020**.

18. Voorhuijzen-Harink, M.; Hagelaar, R.; van Dijk, J.; Prins, T.; Kok, E.; Staatsa, M. Toward on-site food authentication using nanopore sequencing. *Food Chem X.* **2019**, *30*.

19. Ho, J.; Puniamoorthy, J.; Srivathsan, A.; Meier, R. MinION sequencing of seafood in Singapore reveals creatively labelled flatfishes, confused roe, pig DNA in squid balls, and phantom crustaceans. *Food Control* **2020**, *112*.

20. Knot, I.; Zouganelis, G.; Weedall, G.; Wich, S.; Rae, R. DNA Barcoding of Nematodes Using the MinION. *Front. Ecol. Evol.* **2020**, *8*.

21. Milián-García, Y.; Young, R.; Madden, M.; Bullas-Appleton, E.; Hanner, R. Optimization and validation of a cost-effective protocol for biosurveillance of invasive alien species. *Ecol. Evol.* **2021**, *11*, 1999-2014.

22. Boykin, L.; Sseruwagi, P.; Alicai, T.; Ateka, E.; Mohammed, I.; Stanton, J.; Kayuki, C.; Mark, D.; Fute, T.; Erasto, J., et al. Tree Lab: Portable genomics for Early Detection of Plant Viruses and Pests in Sub-Saharan Africa. *Genes (Basel)* **2019**, *10*, 632.

23. Organization, W.H. Global Health Estimates 2016: Disease burden by Cause, Age, Sex, by Country and by Region, 2000–2016. . Availabe online: https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html (accessed on

24. UN. United Nations Sustainable Development Goals. Availabe online: https://sustainabledevelopment.un.org/ (accessed on

25. Krehenwinkel, H.; Pomerantz, A.; Henderson, J.; Kennedy, S.; Lim, J.; Swamy, V.; Shoobridge, J.; Patel, N.; Gillespie, R.; Prost, S. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *Gigascience* **2019**, 10.1093/gigascience/giz006, doi:10.1093/gigascience/giz006.

26. Chang, J.; Ip, Y.; Bauman, A.; Huang, D. MinION-in-ARMS: Nanopore Sequencing to Expedite Barcoding of Specimen-Rich Macrofaunal Samples From Autonomous Reef Monitoring Structures. *Front. Mar. Sci.* **2020**, *7*.

27. Pace, N. A molecular view of microbial diversity and the biosphere. *Science* **1997**, *276*, 737-740.

28. Krehenwinkel, H.; Pomerantz, A.; Prost, S. Genetic Biomonitoring and Biodiversity Assessment Using Portable Sequencing Technologies: Current Uses and Future Directions. *Genes (Basel)* **2019**, *10*.

29. Bolyen, E.; Rideout, J.; Dillon, M.; Bokulich, N.; Abnet, C.; Al-Ghalith, G.; Alexander, H.; Alm, E.; Arumugam, M.; Asnicar, F., et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* **2019**, *37*, 852-857.

30. Seah, A.; Lim, M.; McAloose, D.; Prost, S.; Seimon, T. MinION-Based DNA Barcoding of Preserved and Non-Invasively Collected Wildlife Samples. *Genes (Basel)* **2020**, *18*.

31. Lu, H.; Giordano, F.; Ning, X. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics* **2016**, *14*, 265-279.

32. Maestri, S.; Maturo, M.; Cosentino, E.; Marcolungo, L.; Iadarola, B.; Fortunati, E.; Rossato, M.; Delledonne, M. A Long-Read Sequencing Approach for Direct Haplotype Phasing in Clinical Settings. *Int J Mol Sci* **2020**, *21*.

33. Jain, M.; Koren, S.; Miga, K.; Quick, J.; Rand, A.; Sasani, T.; Tyson, J.; Beggs, A.; Dilthey, A.; Fiddes, I., et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **2018**, *36*, 338-345.

34. Vereecke, N.; Bokma, J.; Haesebrouck, F.; Nauwynck, F.; Boyen, F.; Pardon, B.; Theuns, S. High quality genome assemblies of Mycoplasma bovis using a taxon-specific Bonito basecaller for MinION and Flongle long-read nanopore sequencing. *BMC Bioinformatics* **2020**, *21*.
35. Lischer, H.; Shimizu, K. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics* **2017**, *18*.
36. Koren, S.; Walenz, B.; Berlin, K.; Miller, J.; Bergman, N.; Phillippy, A. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722-736.
37. Chin, C.; Peluso, P.; Sedlazeck, F.; Nattestad, M.; Concepcion, G.; Clum, A.; Dunn, C.; O'Malley, R.; Figueroa-Balderas, R.; Morales-Cruz, A., et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **2016**, *13*, 1050-1054.
38. Rognes, T.; Flouri, T.; Nichols, B.; Quince, C.; Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **2016**, *4*.
39. Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066.
40. Rice, P.; Longden, I.; A, B. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **2000**, *16*, 276-277.
41. Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **2017**, *27*, 737-746.
42. Mitsuhashi, S.; Kryukov, K.; Nakagawa, S.; Takeuchi, J.; Shiraishi, Y.; Asano, K.; Imanishi, T. A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Sci. Rep.* **2017**, *7*.
43. Wood, D.; Lu, J.; Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **2019**, *20*.
44. Wood, D.; Salzberg, S. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **2014**, *15*.
45. Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D. Basic local alignment search tool. *J. Mol. Biol* **1990**, *215*, 403-410.
46. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *191*.
47. Kiełbasa, S.; Wan, R.; Sato, K.; Horton, P.; Frith, M. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **2011**, *21*, 487-493.
48. Bokulich, N.; Kaehler, B.; Rideout, J.; Dillon, M.; Bolyen, E.; Knight, R.; Huttley, G.; Caporaso, G. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **2018**, *6*.
49. Deelen, J.; Beekman, M.; Uh, H.; Helmer, Q.; Kuningas, M.; Christiansen, L.; Kremer, D.; van der Breggen, R.; Suchiman, H.; Lakenberg, N., et al. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging Cell* **2011**, *10*, 686-698.
50. San Mauro, D.; Gower, D.; Oommen, O.; Wilkinson, M.; Zardoya, R. Phylogeny of caecilian amphibians (Gymnophiona) based on complete mitochondrial genomes and nuclear RAG1. *Mol Phylogenet Evol.* **2004**, *33*, 413-427.
51. Folmer, O.; Black, M.; Hoeh, W.; Lutz, R.; Vrijenhoek, R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol.* **1994**, *3*, 294-299.
52. Schneider D, L.L., Dierl W, Wink M. Androconial Hairbrushes of the Syntomis (Amata) phegea (L.) Group (Lepidoptera, Ctenuchinae): A Synapomorphic Character Supported

by Sequence Data of the Mitochondrial 16S rRNA Gene. *Zeitschrift für Naturforschung C* **2014**, *54*.

53. Klindworth, A.; Pruesse, E.; Schweer, T.; Peplies, J.; Quast, C.; Horn, M.; Glöckner, F. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. . *Nucleic Acids Res.* **2013**, *41*.
54. Quinlan, A.; Loman, N. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **2014**, *30*.
55. De Coster, W.; D'Hert, S.; Schultz, D.; Cruts, M.; Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **2018**, *34*.
56. Leger, A.; Leonardi, T. pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software* **2019**, *4*.
57. Loman, N.; Quick, J.; Simpson, J. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* **2015**, *12*, 733-735.
58. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Subgroup, G.P.D.P. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079.
59. Ondov, B.; Bergman, N.; Phillippy, A. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **2011**, *12*.
60. Jain, M.; Fiddes, I.; Miga, K.; Olsen, H.; Paten, B.; Akeson, M. Improved data analysis for the MinION nanopore sequencer. *Nat Methods* **2015**, *12*, 351-356.
61. Bolyen, E.; Rideout, J.; Dillon, M.; Bokulich, N.; Abnet, C.; Al-Ghalith, G.; Alexander, H.; Alm, E.; Arumugam, M.; Asnicar, F., et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol.* **2019**, *37*, 852-857.
62. Implementation of Directive 98/83/EC concerning the quality of water intended for human consumption. 2001.
63. Directive 2006/7/EC of February 15th 2006 (management of bathing water quality). 2006.
64. Srivathsan, A.; Baloğlu, B.; Wang, W.; Tan, W.; Bertrand, D.; Ng, A.; Boey, E.; Koh, J.; Nagarajan, N.; Meier, R. A MinION™-based pipeline for fast and cost-effective DNA barcoding. *Mol Ecol Resour* **2018**.
65. Chang, J.; Ip, Y.; Ng, C.; Huang, D. Takeaways from Mobile DNA Barcoding with BentoLab and MinION. *Genes (Basel)* **2020**, *11*.
66. Freitag, H.; Molls, C.; Bouma, A.; Garces, J.; Rossato, M.; Cosentino, E.; Delledonne, M. Additional new species of Grouvellinus Champion 1923 (Insecta, Coleoptera, Elmidae) discovered by citizen scientists and DNA barcoded in the field applying a novel MinION-based workflow. *Journal of Natural History* **2020**, *53*.
67. Blanco, M.; Greene, L.; Rasambainarivo, F.; Toomey, E.; Williams, R.; Andrianandrasana, L.; Larsen, P.; Yoder, A. Next-generation technologies applied to age-old challenges in Madagascar. *Conservation Genetics* **2020**, *21*, 785-793.
68. Karst, S.; Ziels, R.; Kirkegaard, R.; Sørensen, E.; McDonald, D.; Zhu, Q.; Knight, R.; Albertsen, M. Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *bioRxiv* **2020**.
69. Tytgat, O.; Gansemans, Y.; Weymaere, J.; Rubben, K.; Deforce, D.; Van Nieuwerburgh, F. Nanopore Sequencing of a Forensic STR Multiplex Reveals Loci Suitable for Single-Contributor STR Profiling. *Genes (Basel)* **2020**, *11*.
70. Li, C.; Chng, K.; Boey, E.; Ng, A.; Wilm, A.; Nagarajan, N. INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* **2016**, *5*.
71. Baloğlu, B.; Chen, Z.; Elbrecht, V.; Braukmann, T.; MacDonald, S.; Steinke, D. A workflow for accurate metabarcoding using nanopore MinION sequencing. *BioRxiv*

2020, https://doi.org/10.1101/2020.05.21.108852, doi:https://doi.org/10.1101/2020.05.21.108852.

72. Ibironke, O.; McGuinness, L.; Lu, S.; Wang, Y.; Hussain, S.; Weisel, C.; Kerkhof, L. Species-level evaluation of the human respiratory microbiome. *GigaScience* **2020**, *9*.

73. Ma, X.; Stachler, E.; Bibby, K. Evaluation of Oxford Nanopore MinION^{TM} Sequencing for 16S rRNA Microbiome Characterization. *BioRxiv* **2017**, https://doi.org/10.1101/099960, doi:https://doi.org/10.1101/099960.

74. Neuenschwander, S.; Terrazos Miani, M.; Amlang, H.; Perroulaz, C.; Bittel, P.; Casanova, C.; Droz, S.; Flandrois JP; Leib, S.; Suter-Riniker, F., et al. A Sample-to-Report Solution for Taxonomic Identification of Cultured Bacteria in the Clinical Setting Based on Nanopore Sequencing. *J Clin Microbiol.* **2020**, *58*.

75. Benítez-Páez, A.; Hartstra, A.; Nieuwdorp, M.; Sanz, Y. Strand-wise and bait-assisted assembly of nearly-full rrn operons applied to assess species engraftment after faecal microbiota transplantation. *BioRxiv* **2020**, https://doi.org/10.1101/2020.09.11.292896, doi:https://doi.org/10.1101/2020.09.11.292896.

Acknowledgments

Giunto alla fine di questo intenso percorso, non posso che voltarmi e riconoscere l'impatto determinante che tante persone hanno avuto sulla mia crescita personale e professionale.
Il primo ringraziamento va di diritto al Professor Delledonne, mia guida durante questi tre anni di esplorazione del mondo della Genomica. La passione e la dedizione che mette quotidianamente nel suo lavoro lo rendono un Capo sempre presente e disponibile al confronto. Certo, in questi anni mi risulta di avergli procurato qualche mal di testa di troppo; dopotutto, "La ricerca non è bella se non è litigarella", dicevano gli antichi (?), e allora non resta che essere riconoscente per le possibilità di incontro e scontro, diretto e schietto, senza ricorso ad ampollosi giri di parole.

Un ringraziamento sentito va a Marzia R, per essere stata una seconda guida, in grado di svolgere un gran lavoro di mediazione nei momenti più critici e per avere sempre avuto una parola di incoraggiamento, sostegno e comprensione. Un grazie a Marzia V, per il modo di fare così materno e rassicurante. E grazie al dott. Salviati, perché con le sue competenze trasversali e la sua passione per la divulgazione, è sempre in grado di dipingere scenari remoti.

Un grazie a tutti i membri del Centro di Genomica Funzionale, con cui ho condiviso tanto: a Barbara G., per lo spirito critico dato dall'esperienza internazionale, a Barbara I., per sapere ascoltare (e sapersi anche esprimere in "ciaobellese", all'occorrenza), a Chiara, per la spontaneità e l'entusiasmo (d'altronde sei giovane come l'acqua…), a Cri, per ricordarmi che oltre al lavoro c'è di più (vado a prendere Sofi!), a Denise, per essere così naturale e trasparente, ad Elena, per essere sempre sul pezzo, a Luca D.A., per le *soft-skills* come biografo degli *Arthemis,* a Luca M., per avere la rara capacità di non parlare mai a sproposito, a Giovanni, per essere un pacato-ma-esuberante intrattenitore, a Giulia, per la forza di volontà e la sensibilità (non ho detto empatia!), a Manu, per l'esplosione di vitalità che porta con sé, a Marta, per essere così combattiva ed energica, a Martina, perché sa prendere il meglio da ognuno di noi (deformazione professionale da studiosa di pangenoma?), a Massimiliano, perché con l'utilizzo di tempi verbali desueti seppe rallegrarmi di frequente nonostante le *camorrìe*, a Stefania, per lo sguardo attento e informato sulla realtà, a Vale, per essere così alla mano e così "prof" al tempo stesso. Un grazie particolare anche a Salvo, Rob e Luciano, cari amici anche a distanza di qualche anno dal loro passaggio al DDLab.

Grazie alla mia famiglia: mio padre, mia madre, mia sorella e le mie nonne, che ogni fine settimana mi hanno riaccolto a casa con rinnovato entusiasmo, come se non mi avessero visto per anni. Il vostro pieno sostegno morale e materiale ha significato molto per me.

Grazie poi a tutti i miei amici di vecchia data: Teo, Jacopo, Giarra, Maria, Miriam, Andrea, Chiara e Luca. I momenti di svago trascorsi insieme sono stati davvero fondamentali per ricaricare le batterie tra una settimana e la successiva.

Infine, un ringraziamento speciale alla ragazza che sta al mio fianco da diversi anni, Alessia. Mi hai sempre sostenuto con attenzioni e premure costanti, condividendo la scelta di intraprendere questo percorso, nonostante ciò significasse non potersi frequentare quotidianamente. La tua personalità, data da una combinazione di intraprendenza, solarità e genuinità, non smette per me di essere uno stimolo a superare i miei limiti.