UNIVERSITÀ
DI SIENA
1240

UNIVERSITÀ DEGLI STUDI DI SIENA

DIPARTIMENTO DI BIOTECNOLOGIE MEDICHE

DOTTORATO DI RICERCA IN BIOTECNOLOGIE MEDICHE
COORDINATORE: PROF. LORENZO LEONCINI
CICLO XXXIII

Implementation of a flexible Oxford Nanopore sequencing platform for microbial genomics

Relatore:
Dott. FRANCESCO SANTORO

Tesi di:
DAVID PINZAUTI

DAVID PINZAUTI

Anno Accademico 2020–2021

Table of Contents

Abstract 4

Chapter 1. General Introduction and aim of the thesis 5
- 1.1 Aim of the thesis 6
- 1.2 Nanopore Sequencing Applications for Bacterial Genomics 7

Chapter 2. DNA isolation methods for Nanopore sequencing of *Streptococcus mitis* genome 35
- 2.1 Conclusions 52

Chapter 3. Hybrid sequencing of *Enterococcus faecalis* isolates from periapical lesions 53
- 3.1 *Enterococcus faecalis* epidemiology 53
- 3.2 *Enterococcus faecalis* in endodontic infections and the role of saliva in its transmission 56
- 3.3 Comparative genomics of *Enterococcus faecalis* strains isolated from patients with apical lesions 79
- 3.4 References 92

Chapter 4. Sequencing of *Severe Acute Respiratory Syndrome Coronavirus* 2 97
- 4.1 Whole-Genome Sequence of SARS-CoV-2 Isolate Siena-1/2020 99
- 4.2 Genome Sequences of 10 SARS-CoV-2 Viral Strains Obtained by Nanopore Sequencing of Nasopharyngeal Swabs in Malta 104
- 4.3 Sequencing of 15 SARS-CoV-2 strains obtained at the beginning and at the end of the first pandemic wave in Siena 109
- 4.4 References 114

Chapter 5. Conclusions 116

Appendix 1 – Scientific Curriculum vitae 117

Abstract

Oxford Nanopore sequencing technology is slowly revolutionising the entire microbiology field. Its ease of use and cost effective approaches coupled with long reads sequencing represent an essential and powerful tool. In the present thesis I have implemented a sequencing platform based on the Oxford Nanopore technology, a flexible system suitable for both research and diagnostic fields. The first part of my work was dedicated to the optimization of a DNA extraction protocol capable of isolating high molecular weight (HMW) genomic DNA. In fact, Nanopore sequencing readouts are highly influenced by both the quality and the integrity of the genomic DNA. An enzymatic lysis based extraction protocol was optimized, recovering HMW DNA from two strains of *Streptococcus mitis* and generating multiple ultra long reads (i.e. >100 Kb in length), making it possible to achieve complete genome assemblies. As the extraction protocol was mainly optimized for Gram-positive bacteria, it is also suitable to lyse the thinner cell wall of Gram-negatives. Oxford Nanopore Whole Genome Sequencing (WGS) approaches have enabled the complete genome sequencing and assembly of 16 *Enterococcus faecalis* isolates from clinical dental samples. Sequencing data provided enough information enabling i) population studies, defining genomic clusters based on isolates homologies; ii) bacterial profiling, assessing antimicrobial resistance genes and virulence traits; iii) comparative analysis, identifying genomic rearrangements and homologies based on synteny blocks. Finally, the platform was used for the monitoring of the ongoing SARS-CoV-2 pandemic. We have proposed a 900 bp amplicon sequencing protocol, adapted from the ARTIC sequencing protocol (https://artic.network/), enabling a near-complete genome assembly of SARS-CoV-2 strains, helping in the detection of nucleotide changes and monitoring the circulating viral lineages. In conclusion, the Oxford Nanopore sequencing platform can bring several improvements in the microbiology field, allowing i) complete genome assembly, ii) rapid microbial profiling, and iii) helping in the monitoring of local or global outbreaks.

Chapter 1. General Introduction and aim of the thesis

The intricate story of adaptation and evolution of microorganisms is encoded in their genome. The study of genomic evolution requires a good understanding of the mechanisms driving biological events which shape the genomes, from small scale events (i.e. single nucleotide mutations, insertions and deletions) up to large scale chromosomal rearrangements, recombination, duplication, and gain or loss of genetic material.

Genomics field implies the study and understanding of the genome structure and all its properties. Employing nucleic acid sequencing technologies and computational analysis (bioinformatics), genomics is intended to sequence, assemble, and analyze the structures and functions of genomes. Since the release of the first complete bacterial genome, *Haemophilus influenzae* (L42023.1), in 1995, genomics has contributed to our understanding of infectious disease bringing essential improvements in our knowledge of pathogenesis, antibiotic resistance mechanisms and its relative spread all over communities, and host immune response. The evolution and release of novel sequencing technologies, incorporating revolutionary innovations to manage genome complexities, have enabled rapid and cost effective microbial sequencing, potentially achieving complete genome assemblies and changing the science of microorganisms.

1.1 Aim of the thesis

The main scope of this thesis was the development and optimization of an Oxford Nanopore based sequencing platform, a flexible system suitable for both research applications and clinical microbiology. Oxford Nanopore sequencing is relatively easy to implement, cost-effective, does not have a fixed run time and generates long reads. It is therefore suitable for rapid bacterial profiling, detecting antimicrobial resistance and virulence traits, achieving complete *de novo* genome assemblies, and allowing pathogen surveillance.

In the first chapter of this thesis, I reviewed the applications of nanopore sequencing for bacterial genomics, including an overview of the genome assembly algorithms, the applications for bacterial typing and characterization of antimicrobial resistance genotype. I then illustrate the optimization of an extraction protocol capable of isolating high molecular weight DNA from Gram-positive bacteria (Chapter 2). The platform was also used for the whole genome sequencing of *Enterococcus faecalis* isolates from clinical samples (Chapter 3). Sequencing data were used to evaluate antimicrobial resistance and virulence genes, performing genomic population studies and comparative genomic analysis. Finally, the sequencing station was applied to the management of the ongoing SARS-CoV-2 pandemic monitoring the local spread of viral variants (Chapter 4).

Chapter 1.2

Nanopore Sequencing Applications for Bacterial Genomics

David Pinzauti and Francesco Santoro

Abstract

The evolution of Next Generation Sequencing (NGS) technologies brought improvements in both clinical and research fields. Nanopore sequencing platforms, developed by Oxford Nanopore Technologies (ONT), enable single DNA molecule sequencing without prior DNA amplification, streaming reads in real-time. Key feature of this technology is the capability of sequencing long and ultra-long reads, as long as the native DNA, making it an attractive candidate for bacterial genomic applications. In this review, we evaluate the application of Nanopore sequencing in the bacterial genomics field as a promising candidate allowing i) complete *de novo* genome assembly, ii) bacterial identification at genus and species level, and iii) identification of antimicrobial resistance (AMR) genes.

KEYWORDS: Bacterial Genomics; *de novo* genome assembly; Bacterial Identification; Antimicrobial Resistance; nanopore sequencing

Introduction

Over the past decades Next Generation Sequencing (NGS) technologies have brought several advantages in both clinical and research fields. Since its release in early 2000s, Illumina sequencing technology has completely revolutionized the NGS field becoming the dominant technology for bacterial genomics and routine public health applications. Based on sequencing by synthesis approach, it enables low cost and accurate sequencing. However, major shortcomings are associated with: limited read length (300 bases maximum, more commonly 100-150 bases), uneven read depth, amplification bias, leading to incomplete genome assembly [1]. Driven by the increasing demand for improvements in terms of higher throughput, longer sequencing reads and faster workflows, a new generation of sequencing technologies was released. These technologies directly target single DNA molecules enabling (i) real-time sequencing, (ii) increased read length, (iii) reduced sequencing time from days to hours, (iv) elimination of PCR-amplification bias, and (v) high coverage of bacterial genomes. Currently, there are two main single molecule sequencing technologies available. The first long-read sequencer was the single-molecule real-time (SMRT) sequencing released in 2011 by Pacific Biosciences (PacBio). Briefly, this technology is able to detect the nucleotide incorporation by a DNA polymerase reaction. Nucleotides are labelled with fluorophores and, upon dNTP incorporation, the polymerase cleaves the fluorophore, which emits a light signal. The emitted light signal is recorded in real-time, through a laser and camera system, and associated to a specific nucleotide. Initial release was characterized by high error rate and a relatively short read length, however rapid advances increased the maximum sequencing length to 50 kb, further increasing read accuracy, enabling non-hybrid assembly of SMRT reads [2]. Recently, the release of a newer version, known as SMRT Sequel II system, enables the sequencing of high fidelity (HiFi) reads, improving the per base resolution with >99% accuracy [3]. Single, double stranded molecules are ligated with sequencing adapters, at which the polymerase binds. Through multiple polymerase passes a set of noisy sub-reads are generated, from whom the consensus reads are then derived. The Nanopore sequencing technology was developed and produced by the UK-based company Oxford Nanopore Technologies (ONT), and commercially released in 2014. Nanopore technology is currently the only sequencing technology based on DNA translocation through biological nanopores. Nanopore proteins are embedded in an electrical resistant polymer membrane where an ionic current flows: as native single stranded DNA (ssDNA) translocates through the nanopore, alteration in the current is measured. Shifts in the current voltage are characteristic of DNA sequences resulting in “squiggle” plots which are stored in a raw hdf5-based fast5 data format.

Voltage shifts are interpreted as k-mers (3-6 nucleotides) by the basecaller algorithm generating fastq files. Currently two sequencers are available and suitable for bacterial genomics: MinION and GridION devices. The MinION device was the first available Nanopore-based sequencer. Released in 2014, it is a USB-sized portable sequencer able to generate up to 50 Gb data (https://nanoporetech.com/). Later in release (2017), the GridION is a versatile bench-top sequencer allowing the sequencing of up to five flow cells contemporaneously. Nanopore sequencing brings several advantages compared to other platforms: (i) reduced costs, about 50$ per isolate if sequencing 24 samples on a single flow cell [4]; (ii) portability and easiness of use, with rapid and user-friendly library preparation protocols which allow sequencing outside of a laboratory environment due to MinION reduced size (https://nanoporetech.com/) [5]; (iii) real time sequencing and "read until", sequencing on a flow cell can go on up to 72 hours, but it can be stopped whenever a sufficient data amount is generated or certain organisms are detected, furthermore reads are immediately available for analysis providing immediate access to results; (iv) long read sequencing, since nanopore reads can be, in principle, as long as the DNA strands used for library preparation. Notably, there is a scientific community implementing methods to further extend the maximum length achieved (https://www.longreadclub.org/) (http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/) which currently achieved the longest reads reported of > 2 Mb. Long read sequencing has some limitations: specific extraction protocols are required for the isolation of High Molecular Weight (HMW) DNA and careful handling is needed to avoid DNA shearing. Moreover, it is challenging to obtain HMW DNA from some bacterial species, e.g. Gram positives with thick cell walls or *Mycobacteria* [6].

Nanopore sequencing is an interesting technology for many applications spacing from diagnosis to research fields. The capability of sequencing long and ultra-long reads makes it suitable for bacterial genomics, since longer reads facilitate *de novo* assembly processes. Due to its portability it can also be used as a rapid diagnostic tool to identify bacterial infections, manage outbreaks, or to identify drug resistant pathogens and track their spread in the population.

In this review, we evaluate the role of Nanopore sequencing focusing on bacterial genomics, giving insights about current advantages as well as limitations and challenges. Publicly available publications in PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) were reviewed, searching for papers matching to '*nanopore sequencing bacterial genome*' research key. As of 13/01/2021, 201 search results were available. A total of 30 papers were manually discarded because they were related to viral or plant genomics.

Bacterial genomics and *de novo* assembly

Genome assembly is the computational process of reconstructing the original genome sequence from a set of reads [7]. *De novo* assembly refers to a reconstruction based on the obtained sequence data, without consultation of previously resolved sequences from genome databases. *De novo* assembly is particularly suitable (i) when a reference genome is not available, (ii) to avoid biases from an imperfect reference, (iii) when the strain belongs to an unknown species, and (iv) to identify a bacterial strain from mixed samples (such as metagenomic samples). The presence of complete bacterial genomes within data banks can bring several benefits such as the possibility to make comparative genomics studies, to accurately perform genome annotation, to track the spread of mobile genetic elements, to monitor the pattern of drug resistance, and to understand microbial pathogenicity and evolution [8,9]. However, the assembly of complete genomes (i.e. a single, circular contiguous sequence for each replicon) is challenging because of bacterial genomes complexity. In fact, bacterial genomes contain long repetitive elements (transposases, duplicated genes) and can undergo structural variation, which contributes to maintain genome plasticity and stability through rearrangements, duplications or amplification of genetic material [10]. Koren et al [11] have proposed the classification of microbial genomes into three complexity classes based on the size and the number of repeated sequences. The rDNA operon, an ubiquitous large repeat with a length between 5 and 7 kb, was used as the basis to measure genome complexity. Class I genomes contain few repeats other than rDNA, while Class II genomes harbour many mid-scale repeats where rDNA is still the largest [11]. Finally, Class III genomes are characterized by large mobile genetic elements' associated repeats, segmental duplications, or large tandem arrays, larger than rDNA operon (> 7 kb in length) [11]. The localization of the repeats within the chromosome is an important determinant of the assembly quality: when repeats are localized in a relatively small region, such as a CRISPR element, the overall genome arrangement is not affected, while, when repeats are sparse, the assembly process is challenging and results in fragmented contigs whose order cannot be defined [12]. Sequencing reads shorter than the repetitive element are inherently incapable of resolving it, since they cannot span the entire repeated region. Short reads are however capable of generating correct, though not circular, assemblies of Class I genomes, while they cannot resolve Class II and III genomes [11]. Moreover, short sequencing reads are associated with computational challenges because they provide less information to determine structure and position of the repeats [12]. A correct and full assembly of microbial genomes can only be achieved by sequencing reads exceeding the length of the longest genomic repeats [2,8,11]. Nanopore technology is becoming an attractive and relatively cheap solution for *de novo* assembly. Long reads have more chances to overlap each other making much easier to assemble fragments in the right position. They have the potential to uniquely span the entire repeated elements anchoring both extremities [11,13,14].

To date (January 2021) 292,699 bacterial genomes are available in GenBank, of those only 21,406 (7.31%) are complete. Long reads technologies have the potential to close the gap between draft and complete genomes, the latter being much more informative. In the last five years, the use of nanopore sequencing technology has steadily increased, with more than 4,000 bacterial genomes deposited in the Sequence Read Archives in 2020 alone (Figure 1).

Early applications demonstrated the capability of long nanopore reads to help in the assembly of complete bacterial genomes [15,16,17,18,19]. Initial approaches were mainly focused on contigs scaffolding: previously sequenced and assembled Illumina fragmented contigs were pieced together using long reads which enabled the resolution of genomic architecture. The first proof-of-concept for nanopore-only *de novo* assembly was given by Loman and colleagues [13] who were able to assemble the genome of a reference *Escherichia coli* strain K-12 MG1655 relying solely on long nanopore reads. In the last few years, further improvements in assembly algorithms and sequencing technology enabled assembly of complete high contiguity genomes relying only on Nanopore data, resulting in a small but continuous increase in the number of publicly available complete genomes [5,6,20,21,22,23,24,25,26]. Few published papers showed the impact of long reads to fully resolve long repetitive element rearrangements [8,27]. The *Pseudomonas koreensis* genome assembly is particularly challenging since it harbors very long, nearly identical repeated regions of up to 70 kb [8]. Moreover the presence of repeats (at least 30 kb in length) shared between the chromosome and three plasmids further contributes to increase the complexity of the genome. Only long nanopore reads (N50 > 44 kb) which spanned the entire repeated element, provided unique sequences at both extremities correctly anchoring the repeated elements in the genome and could fully resolve the chromosomal arrangement. On the other hand, the presence in a genome of many smaller repeated elements, such as the insertion sequences (ISs) of about 1 kb in length, also represents a challenge for *de novo* assembly. Those genomes are in fact characterized by an increased complexity caused by rearrangements, deletions and, more rarely, duplications [9,28]. As an example, *Bordetella pertussis* harbors at least 300 copies of three different ISs, leading to complex genome rearrangements even in closely related isolates [9]. Sequencing reads shorter than IS length are incapable of solving the genomic architecture and their assembly generates several hundred contigs, or at least one contig per IS copy [9]. The inability of generating a closed chromosome hampers comparative genomic analyses among *B. pertussis* isolates. Complete and circular *B. pertussis* genomes were previously obtained combining long PacBio reads with short but accurate Illumina reads and optical mapping with Argus OpGen allowing the detection of large inversions centered on the replication origin and terminus [29]. However, employing a Nanopore based assembly pipeline Ring et al. [9] were able to fully resolve *B. pertussis* genome arrangement of 5 strains, detecting large-scale, inter- and intra-strain genomic rearrangements.

Despite leading advantages for a rapid and a complete *de novo* assembly, Nanopore reads still have a lower accuracy compared to short sequencing technologies. Raw nanopore reads are characterized by a relatively high-error rate, ranging between 5% and 15% [30]. Systematic random errors in form of homopolymers, insertions, deletions and alterations in the electrical signal due to chemical modifications are responsible for the associated lower accuracy. However, when a sufficient genome coverage is achieved, the assembly process is capable of removing random errors. A 20x depth of coverage with long reads (N50 of 7.8 kb) usually provides enough information to reconstruct the structure of entire bacterial chromosomes. A depth of coverage below the 20x threshold may produce worse assembly outcomes failing in the genome reconstruction and generating fragmented contigs. Increases in the read length (N50) can improve the assembly process even for genomes with low read depth: long reads in fact bring more information about the chromosomal arrangement, further reducing the number of assembled contigs within low covered genomes, suggesting the importance of coupling sufficient genome coverage together with long reads. The great majority of assembly errors (up to 95%) are single base insertions and deletions, while the remaining part is represented by nucleotide substitutions. Insertions and deletions in a read are most likely due to uncontrolled variations in the DNA translocation speed. Available genome polishing pipelines are able to increase the consensus quality, but single base errors persist. Usually increasing the genome coverage up to 100x brings improvements for both genome assembly and polishing, but coverage beyond 100x can also be associated with worse assembly outcomes: assembler tools in fact may not be able to completely understand and reconstruct the genomic architecture due to the high read variability. Polishing pipelines enable correction of insertions and deletions (reduction of ~86%), even though SNPs correction may be more complicated. Because sequencing errors are randomly distributed all over Nanopore reads, increased sequencing depth alone cannot completely compensate for the high error rate [4]. Despite not influencing the final chromosomal arrangement, insertions and deletions are associated with changes in protein annotation, creating shorter and incorrect predicted proteins by introducing premature stop codons or frameshift errors. The amount of interrupted ORFs can be easily evaluated with tools such as ideel (https://github.com/mw55309/ideel). The most efficient approach for error correction still remains the complementation with short but accurate reads. Hybrid assembly strategies in fact combine both long and short sequencing reads leading to complete and accurate genome assembly [16,17,22,30,31,32,33,34,35,36,37]. While long Nanopore reads provide information about the genome arrangement (scaffolding), short reads facilitate per-base error correction [4,30,38]. The theoretical costs of reagents for hybrid sequencing of a bacterial genome are relatively low, varying between $200 and $340 [30].

Assembly algorithms overview

Currently available assembler tools are based on two major classes of algorithms: Overlap Layout Consensus (OLC) and De Bruijn Graph (DBG) (Figure 2). DBG-based assemblers substring reads into shorter k-mers used to generate an assembly graph. Unique k-mers are connected together using sequencing reads to infer genome arrangement. DBG is typically the preferred choice for short reads assembly and it is normally associated with worse assembly quality for long, error-prone reads [39]. Two variants of DBG algorithm are available, both designed to fit with long noisy reads. The A-Bruijn variant, employed by the Flye assembler [40], has shown to be promising for long reads assembly. This algorithm tolerates the high error rate of long reads since it relies on approximate sequence matches, not on exact k-mer matching, and includes an error correction step. First, it combines error prone reads into disjointigs creating a repeat graph, and then it resolves repetitions in the graph to generate the final contig [14]. The second one is known as fuzzy Bruijn graph (FBG) which is embedded in the Redbean assembler (formerly Wtdbg2) [41]. It is based on inexact matches among long reads, allowing mismatches and gaps. Long sequencing reads are then chopped, merging similar segments to form a vertex. Vertices are finally connected based on segments' closeness on long reads, generating the consensus.

OLC-based assemblers are usually the preferred choice for long reads assemblies. In fact, sequencing reads are assembled without k-mer splitting, incorporating a consensus correction process [1]. Usually, OLC assemblers work in three stages: first, they create overlap graphs (O) from the sequencing reads. Then, they carry out a layout (L) of all the reads generating contigs, and finally infer the consensus (C) sequence by merging reads while correcting sequencing errors [1,42]. There are currently four main assembler tools based on OLC-algorithm: Canu, Miniasm, Raven, and Shasta. Canu [43] was released in 2015 as a fork of the Celera assembler. Its pipeline includes three stages: i) reads correction, ii) reads trimming (removing adapters and breaking chimeras), and iii) contigs assembly [14,39]. A genome set parameter, specifying an expected genome length, is required, representing a limitation in case of isolates of unknown length. Notably, read correction and trimming pipelines can also be used as a stand-alone module to improve Nanopore reads quality [4,9,39]. Miniasm assembler [44] is also based on an OLC algorithm, but lacks the consensus and error correction steps. Because it only concatenates reads, the final assembly error rate is comparable to the raw reads error rate. Improvement of Miniasm-based assembly quality can be obtained performing consensus polishing step [14]. The Raven tool [45] has an assembly pipeline similar to Miniasm, but implements a consensus step. Raven performs the genome assembly through reads overlapping, improving the assembly contiguity by removing repeat-induced false overlaps. Finally it performs a consensus step by polishing the assembly. Shasta assembler [46] was designed to be computationally efficient [14]. The assembly approach is not directly performed on sequencing reads but these are converted in sequences of markers. Those markers are randomly selected strings of nucleotides with a length of k (k=10 by default), the nucleotide sequence of a read is converted to the sequence of markers that occur in that read. This process reduces the length of the read and facilitates the alignment, since the number of possible markers is quite high. Overlaps among reads are then searched using a modified MinHash algorithm [46] and then an assembly graph is constructed. The position of a subset of short k-mers, used to find overlaps. An assembly graph is built from overlaps, then deriving the consensus sequence.

Each assembler tool has its own strengths and drawbacks in terms of assembly reliability, circularization efficiency and computational resources usage. In a recent benchmarking work using both simulated and real-life datasets, the Flye assembler was deemed to be the most reliable and accurate tool [14] (**Table 1**). In fact, Flye is optimal for both chromosome and plasmid rapid assemblies. Its major shortcomings are associated with the tendency to delete some bases (tens of bases) in the genome circularization process. Miniasm, despite not including a consensus step in its pipeline, produces fast and reliable results for genome architecture, requiring few computational resources, without the need for setting a genome length parameter. Raven is also reliable in terms of bacterial chromosome architecture, but it does not resolve plasmids and tends to lose hundreds of bases in the genome circularization process [14]. Canu is very reliable in resolving the architecture of both bacterial chromosomes and plasmids, users have the choice to adjust many parameters, which may lead to an improved assembly quality, and it also includes an error correction pipeline. On the other hand, it is quite slow and requires an intensive usage of computational resources [14,27,39]. Both Redbean and Shasta do not appear much suitable for bacterial genomics since they are not able to produce complete and reliable genome assemblies [14].

Polishing and error correction

The relatively high error rate of raw nanopore reads requires a post-assembly polishing to further improve the assembly accuracy [14,47]. Assembly polishing is usually mandatory when assembly pipelines lack a consensus step (e.g. Miniasm) but can likewise be used in combination with all the other tools. Roughly, assembly polishing works by remapping long reads on the raw assembly trying to increase its resemblance to the reads [39]. The Racon tool [48] was first released to complement the Miniasm assembly pipeline, but can be used for any long read assembler polishing. Racon is based on a graph-based polishing pipeline known as Partial Order Alignment (POA) approach. Briefly, sequencing reads are aligned to the raw assembly building up POA graphs. Then Racon finds the best alignment between the assembly and the POA graph building the consensus [39]. Notably, Racon allows assembly polishing using either Illumina or Nanopore reads.

In addition to graph-based approaches, neural networks pipelines are used for error correction. The Nanopolish tool [13] was the first neural network-based tool, using the raw current signal stored in fast5 files to improve consensus accuracy [39]. Reads are aligned to the raw assembly searching for positions where the assembly may differ generating a set of alternative candidates. Large modifications, such as multiple insertion or deletions are corrected using aligned basecalled reads. Then, every possible one-base modification, such as deletions, insertions, substitutions, is evaluated performing a correction based on reads raw signal. This process allows Nanopolish to observe several possible modifications [39], though in a time-consuming manner. Similar to Nanopolish, the ONT Medaka tool (https://nanoporetech.github.io/medaka/) improves consensus quality by using neural networks. Medaka allows error correction from a pileup file of Nanopore reads mapped against the raw assembly. The modest computational requirements and the comparable results achieved in rapid time make it a competitive tool. Notably, the release of a new Medaka version coupled with the availability of the new R10 pore will possibly enable the assembly of bacterial genome with a high consensus accuracy using only Nanopore long reads (https://nanoporetech.com/about-us/news/london-calling-clive-brown-and-team-plenary).

The involvement of polishing pipelines has further extended the quality of Nanopore-only based assemblies. However single-base errors (insertions and deletions) still persist, often requiring the integration with accurate sequencing reads to increase the per-base accuracy. Addition of short reads brings benefits in both assembly polishing and hybrid *de novo* assembly pipelines. Consequently, if short reads are available, hybrid assembly strategies usually represent the best solution for bacterial *de novo* genome assembly [4,9,38,49].

The tool Pilon [50] is specifically designed to perform assembly polishing relying on short reads. Pilon requires accurate reads, 75 bases or more in length, with a sufficient genome sequencing depth (at least 50x). Long nanopore reads can also be used by Pilon to polish the assembly but the final polished sequence may contain false corrections. Based on read alignment information Pilon is capable of improving consensus accuracy. It starts parsing the alignment accounting for reads covering each bases, and builds a pileup file containing information about possible variations. Then it analyses each possible variations in the draft assembly organizing them into four categories: confirmed, changed, ambiguous and unconfirmed. ‘Confirmed’ means that the majority of reads accounts for the same analyzed base, while ‘Changed’ means that reads support a nucleotide change. Variations are defined ‘Ambiguous’ when reads contain different possible alternatives bases, while they are called ‘Unconfirmed’ if no sufficient coverage is reached. Based on that classification, Pilon adapts the consensus performing per-base error correction, improving the quality.

On the other hand, hybrid assembly strategies combine both long and short reads to reconstruct the genomic architecture. Hybrid assembled genomes can be generated following two different strategies: a long-read-first or short-read-first approach. The long-read-first approach uses a *de novo* assembly with long reads (e.g. Canu) to create a scaffold, which is then followed by an error correction step using short and accurate reads (e.g. with Racon or Pilon polishers). Short-read-first approaches first assemble accurate contigs, which are then bridged together using long reads. The latter approach is implemented in the Unicycler tool [51]. Illumina reads are assembled using SPAdes (a DBG-based assembler) then mapped on a scaffold created assembling nanopore reads with Miniasm. Unicycler also includes polishing steps with Racon and Pilon, which use nanopore and Illumina reads, respectively [39]. Like Unicycler, the tool MaSuRCA [52] is another short-read-first hybrid assembler, which employs both DBG and OLC algorithms. The MaSuRCA pipeline is based on the generation of the so called ‘super-read’, reads usually longer than Illumina reads but with the same assessed quality. Starting from a user-defined k-mer value, the algorithm extends each k-mer at both 5’ and 3’ ends by adding nucleotides based on the reads similarities. Super reads extension continues as long as the added nucleotides are unambiguous (unique, without possible variations). The resulting ‘super-reads’ are usually longer than Illumina reads (400 bases or more), but their length is strictly dependent on genome complexity. Benefits of super reads are related to their length, which allows the construction of a reduced dataset that can be easily mapped with nanopore reads. Super-reads are then merged together generating ‘mega-reads’ under the guidance of nanopore reads, mega-reads are finally used to infer the assembly [53].

Unicycler is, to date, the only automated hybrid assembler, which usually yields very reliable results without further manual intervention [54], however it is sometimes unable to correctly assemble genomes with very high complexity. Moreover, when long reads are considerably more abundant than short reads, a long-reads-first approach is preferred [14]. In conclusion, if high quality Illumina and Nanopore reads are available, both hybrid assembly strategies are suitable, even if Unicycler has the advantages of wrapping together multiple steps and softwares and needs just a shallow coverage with long reads [14].

Nanopore sequencing for microbial identification and typing

Clinical microbiology aims to assess the presence of microbial pathogens in a sample to aid the clinician establishing an appropriate therapy. Identification of pathogens is a key step in this process and in most clinical microbiology laboratories is routinely performed by mass spectrometry directly on colonies grown in culture. However, this approach works properly only for organisms that can be cultivated *in vitro*, and has a relatively slow turnaround time (from 24 to 72 hours). In order to speed-up detection time and increase sensitivity, molecular methods can be implemented. Molecular methods have the potential to overcome culture-based limitations, giving a rapid pathogen identification within a few hours. Real time PCR (qPCR) tests show high sensitivity and allow quantification of bacteria at the species level. Ubiquitous genetic markers, such as ribosomal genes, are targeted by PCR assays, amplified and then sequenced to taxonomically classify bacteria (16S rRNA and 23S rRNA markers) or fungi (ITS1 and ITS2 markers). However qPCR tests are still affected by limitations: i) are not sensitive for uncommon or emerging microbial agents, ii) share biases related to specific primer choice, and iii) short fragment sequencing can be insufficient to assign taxonomy at species level [55], needing continuous update in molecular approaches [56]. The urgent need for rapid and accurate identification tools raised up the attention to NGS platforms, proposing Nanopore technology as one of the most attractive candidates for clinical molecular diagnostics [49,55,56,57,58,59,60,61,62,63,64,65,66]. Application of Nanopore technology in clinical diagnosis can drastically reduce identification times, giving rapid and accurate pathogen profiling. Moreover, it has the potential to correctly identify even atypical and slow growing pathogens. Employing Nanopore technology, Bialasiewicz and co-workers [57] were able to detect a rare, dog-bite associated infection, sustained by *Capnocytophaga canimorsus*, which is a common commensal bacteria of canine oral cavity, rarely causing human infection, but sometimes associated to septic shock in elderly patients. Since it is a fastidious and atypical human pathogen, it needs a prolonged incubation time that is associated with delays in diagnosis and proper therapy. Standard blood cultures in fact were not able to detect the pathogen at least until 6.25 days of incubation, associated in the meantime with worsening of patient's health. Real time Nanopore sequencing allowed a correct detection of the pathogen in rapid time (19 hours) driving the patient to correct and effective therapy.

In 2015, Kilianski and colleagues [60] for the first time proposed Nanopore sequencing as a novel tool enabling rapid and accurate bacterial and viral typing. Key feature is the capability of sequencing long reads: despite its associated high error rate, the lower accuracy of Nanopore reads can be compensated with longer fragments. There are two main approaches employing Nanopore sequencing technology: i) amplicon sequencing of ribosomal markers, or ii) real-time metagenomic sequencing.

The capability of sequencing long fragments enables Nanopore technology to target long amplicon ribosomal markers, such as the full length 16S rRNA (about 1,500 bp) [59] and the whole *rrn* operon (about 4,500 bp) [55,60,61]. In general "the longer the marker, the higher the taxonomical resolution" [55] which translates in accurate identification of pathogens even at the species level. Full-length 16S rRNA sequencing is designed to include V1 to V9 hypervariable regions within the 16S rRNA gene, while the whole *rrn* operon consists of the genes coding for 16S, 5S and 23S subunits plus the more variable internal transcribed spacer (ITS) regions. On the other side, Nanopore sequencing can also be used for real-time bacterial identification of Metagenomic samples. Metagenomics studies the set of microbial genomes within mixed communities that reside in environmental niches [25,62,63] or in human hosts [64,65,66]. This approach has the potential to overcome limitations associated with both culture-based and molecular-based methods, and is not influenced by prior knowledge of the organisms. The ONT-developed tool 'What's In My Pot?' (WIMP) [67] was specifically designed to taxonomically assign reads while they are generated: nanopore reads are aligned in real time against a curated NCBI RefSeq database containing bacteria, viruses, fungi and archaea sequences. Pilot studies suggested the potential of real-time metagenomics in clinical diagnosis [56,65,66]. Charalampous and coworkers [56] tested its ability to discriminate bacterial species in patients with suspected Lower Respiratory Infections (LRI). Nanopore technology was able to discriminate bacteria with a reported overall agreement of 96.6% to standard culture-based methods, within a reduced turnaround time of 6 hours. Extension of the sequencing time up to 48 h allows the generation of more data enabling i) genome reconstruction further improving pathogen identification, and ii) antimicrobial resistance profiling. Despite the rapid response, further improvements are still required. First, host cells depletion approaches are required to remove human nucleic acid contamination [66]; and secondly, lower sensitivity for pathogens with low titre require target enrichment [68].

Many groups are proposing Nanopore sequencing as a potent tool for outbreak surveillance [21,68,69,70,71]. Whole Genome Sequencing (WGS) has the capability of achieving fast pathogen identification, leading further details about Antimicrobial Resistance (AMR) genes, virulence factors, possible links between sources of contamination, real time outbreak monitoring, or patient-to-patient pathogen spread surveillance. Successful applications of Nanopore sequencing were reported during the Ebola virus outbreak [69] or during a novel *Neisseria meningitidis* strain outbreak in West Africa [70], thus demonstrating that Nanopore sequencing is an efficient tool for outbreak monitoring, enabling real time detection even in resource-limited settings.

Antimicrobial Resistance Pathogen Profiling

The increasing rate of AMR in bacterial pathogens represents one of the most serious public health threats of our century. Infections sustained by drug resistant pathogens are associated with i) an increased risk of worse clinical outcomes and death, ii) a great impact in the society due to higher health-care resources usage, and iii) the risk of AMR genes spread among the population. The role of the clinical microbiologist is to identify the etiologic agent of an infection and its antibiotic susceptibility to prescribe the most efficient treatment. Standard Antimicrobial Susceptibility Testing (AST) methods rely on isolation of the pathogen in pure culture and assessment of its antimicrobial susceptibility through phenotypic tests (e.g. determination of minimal inhibitory concentration or Kirby Bauer antibiogram) which are interpreted according to international guidelines. The process takes approximately 48-72 hours and may cause delays in administration of an appropriate therapy. Moreover, automated AST panels contain a limited number of antibiotics and do not identify the resistance mechanisms, essential for guiding treatment decisions, since *in vitro* activity does not always translate to in vivo activity [72]. The advent of PCR-based approaches (molecular profiling) has reduced identification times but fails in the detection of additional resistance mechanisms such as porin deletion, efflux pumps, DNA gyrase mutations, etc [72]. Rapid and accurate resistome profiling tools are urgently required to limit adverse clinical outcomes caused by inadequate empirical treatment and its selective pressure [66], to limit the spread and to monitor transmission of drug resistant pathogens [73].

Strongly influenced by improvements in NGS technologies, the application of Whole Genome Sequencing (WGS) to predict Antimicrobial Susceptibility profiles is slowly increasing. WGS has the potential to fully resolve chromosome and plasmids structure enabling in-depth identification of acquired resistance genes and/or chromosomal mutations [32,72]. The reported accuracy of WGS approaches is comparable to standard phenotypic methods but within a considerably reduced amount of time [72,74]. Identification of resistance genes occurs using computational methods through i) reads mapping or ii) genome assembly. The simplest approach is to map sequencing reads against a reference database of AMR genes or, alternatively, to match few bases (*k-mer*) in order to reduce the computational requirements. Reads mapping methods are mainly used to predict resistance profiles relying on the presence or absence of AMR genes, but fail to detect antibiotic resistance related to chromosomal mutations and are not sensitive for horizontally acquired genes [74]. On the other hand, in depth knowledge about antibiotic resistance mechanisms can be acquired through a *de novo* genome assembly approach [28,31,49]. *De novo* assembly allows the reconstruction of chromosomes and plasmids structures, helping in the detection of chromosomal mutations

[22,34] and horizontally transferred genes. The capability of sequencing plasmids represents an important feature of WGS-based profiling because bacterial plasmids are major vehicles for the transmission of resistance genes among bacteria [75]. However, since plasmids harbor multiple repeats of mobile elements where resistance genes are located [56], the use of short reads is not recommended, as they cannot fully resolve their structure [73]. Long Nanopore reads have the potential to fully resolve both chromosomes and plasmids architecture, correctly identifying AMR genes arrangements and their co-location inside mobile elements, but due to the higher error rate, they may lose accuracy in chromosomal mutation detection [73]. Judge and colleagues [76] in 2015 for the first time pointed out the attention on Nanopore sequencing for AMR profiling, and several papers have since been published on the same topic [22,28,31,32,34,49,66,72,73,74,75,77,78]. The "Antimicrobial Resistance Mapping Application" (ARMA) tool (https://nanoporetech.com/resource-centre/real-time-detection-antibiotic-resistance-genes-using-oxford-nanopore-technologies) allows read mapping against the Comprehensive Antibiotic Resistance Database (CARD) [79], identifying AMR genes in real time. Coupling WIMP and ARMA tools enable researchers and clinicians to identify bacterial taxon and their associated AMR profiles in a timely manner. One of the most successful applications of Nanopore sequencing is the prediction of AMR profiles in drug resistant *Klebsiella pneumoniae* strains [32,72,78]. *K. pneumoniae* is one of the most important life threatening pathogens, associated with nosocomial infections with a high mortality rate (up to 50%). This pathogen usually harbors multiple drug resistance genes, where the vast majority are plasmid-encoded accounting for a rapid dissemination [78]. Employing Nanopore technology, Tamma et al [72] have tested the potential of both real time and assembly based approaches to predict AST results. The achieved results were comparable to standard AST methods: the ARMA real time approach showed an overall accuracy of 77% (ranging between 30% and 100%), while the *de novo* genome assembly further extended the accuracy up to 92% [72] underlining the importance of genome structure reconstruction. The real time approach offers results streaming within 15 minutes and after 2 h most (>= 70%) of the resistance genes were already detected [72] but its lower accuracy affects the capability of detecting allelic variants. Extension of the run time enables generation of enough data to perform *de novo* assembly, increasing the accuracy of Nanopore approach mainly due to the capability of detecting allelic variants and chromosomal mutations. In both cases the resistance profile timing is shorter than standard AST: 8 hours for ARMA real-time and 14 hours for the assembly-based approach [72]. Several limitations affect Nanopore AMR profiling. First, it is not suitable for low-input procedures, requiring microbial DNA enrichment [80] or host DNA depletion to reduce the ratio of host DNA contamination. Secondly, performing a sequencing run per sample could be expensive,

introducing the requirement of samples multiplexing, which further extends turnaround time [66]. Third, allelic variants may be poorly distinguished due to sequencing errors, requiring short reads for a higher detection capability [66,78]. Notably, a gene may be present but fail to cause resistance due to poor expression, silencing or inactivation causing the failure of detection through Nanopore approaches and following requiring further analysis [66].

Nanopore WGS technology also offers the possibility of pathogen surveillance, accelerating pathogen profiling and guiding appropriate antibiotic therapy. Brinda and colleagues [81] have recently presented a computational pipeline to infer resistance/susceptibility to antibiotics based on the "genomic neighbor typing" approach. This method enables a rapid profiling by predicting the phenotype from sequencing data. Briefly, sequencing reads are mapped in real time against reference genomes with known phylogeny and phenotype, and then the probable phenotype of the sample is predicted based on the nearest neighbor, enabling drug resistance identification. This method was highly sensitive and specific when tested with genomes of *S. pneumoniae* and *N. gonorrhoeae*, enabling correct AMR profiling within a couple of minutes. In this approach, the definition of a correct reference database is essential to prevent false predictions, even in case of unknown isolates. For this reason, careful and accurate studies are required when building up the reference database, choosing the optimal genomes and implementing them with local samples. The time required for sample preparation is another limiting factor, since DNA isolation, host DNA depletion and library preparation can increase the handling time. In conclusion, approaches employing nanopore sequencing have the potential to completely transform pathogen profiling in clinical microbiology enabling rapid and real time species identification and providing data for administration of appropriate antibiotic therapy.

Concluding remarks

The advent of Nanopore technology has renewed the entire bacterial genomics field spacing from Whole Genome Sequencing to Bacterial and AMR genes profiling. Nanopore technology offers relatively easy and affordable solutions for both clinical and research areas, reducing sequencing costs and reducing times to achieve results. Key point of this technology is the capability of sequencing long reads, reaching hundreds of kilobases in length. Those long sequencing reads have enabled the assembly of complete bacterial genomes, furthermore providing improvements in the bacterial identification process. Because of the error prone nature of Nanopore reads, nowadays assembling high contiguous and high accurate bacterial genomes relying solely on Nanopore reads is not recommended due to indels errors. However, continuous improvements in both sequencing technology and basecalling algorithms will probably produce errorless nanopore-only assemblies (https://nanoporetech.com/about-us/news/london-calling-clive-brown-and-team-plenary).

Recently, the release of a new flow cell (R10) technology, coupled with algorithms improvements, have further increased sequencing reads quality, reducing biases related to homopolymer detection.

In conclusion, Nanopore sequencing is becoming one of the most promising and affordable sequencing technologies. Its employment in routine diagnosis may help clinicians in rapid and accurate pathogen profiling, giving insights about bacterial species and suggesting antibiotic sensitivity even in real time. In principle, the method can be easily extended to study different bacterial features (AMR genes, virulence factors, methylation profile), the main factors affecting the interaction with hosts and environments, and their evolution.

Authors' contributions

DP and FS conceived the manuscript, DP drafted the manuscript, FS contributed to draft and reviewed the manuscript. Both authors read and approved the final version of the manuscript.

Competing interests

The authors have declared no competing interests.

References

[1] Sohn JI, Nam JW. The present and future of de novo whole-genome assembly. Brief Bioinform 2018; 19:23-40.

[2] Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 2015; 23:110-20.

[3] Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 2019; 37:1155-1162.

[4] Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom 2017; 3:e000132.

[5] Elliott I, Batty EM, Ming D, Robinson MT, Nawtaisong P, de Cesare M, et al. Oxford Nanopore MinION Sequencing Enables Rapid Whole Genome Assembly of Rickettsia typhi in a Resource-Limited Setting. Am J Trop Med Hyg 2020; 102:408-414.

[6] Bouso JM, Planet PJ. Complete nontuberculous mycobacteria whole genomes using an optimized DNA extraction protocol for long-read sequencing. BMC Genomics 2019; 20:793.

[7] Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics 2010; 95:315-27.

[8] Schmid M, Frei D, Patrignani A, Schlapbach R, Frey JE, Remus-Emsermann MNP, Ahrens CH. Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. Nucleic Acids Res 2018; 46:8953-8965.

[9] Ring N, Abrahams JS, Jain M, Olsen H, Preston A, Bagby S. Resolving the complex Bordetella pertussis genome using barcoded nanopore sequencing. Microb Genom 2018; 4:e000234.

[10] Treangen TJ, Abraham AL, Touchon M, Rocha EP. Genesis, effects and fates of repeats in prokaryotic genomes. FEMS Microbiol Rev 2009; 33:539-71.

[11] Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 2013; 14:R101.

[12] Kingsford C, Schatz MC, Pop M. Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics 2010; 11:21.

[13] Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 2015; 12:733-5.

[14] Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Res 2019; 8:2138.

[15] Karlsson E, Lärkeryd A, Sjödin A, Forsman M, Stenberg P. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci Rep 2015; 5:11996.

[16] Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G, Blaxter M, Watson M. A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. Gigascience 2015; 4:60

[17] Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, et al. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics 2015; 16:327.

[18] Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif 2015; 3:1-8

[19] Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. Gigascience 2014; 3:22.

[20] Kuleshov KV, Margos G, Fingerle V, Koetsveld J, Goptar IA, Markelov ML, et al. Whole genome sequencing of Borrelia miyamotoi isolate Izh-4: reference for a complex bacterial genome. BMC Genomics 2020; 21:16.

[21] Taylor TL, Volkening JD, DeJesus E, Simmons M, Dimitrov KM, Tillman GE, et al. Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. Sci Rep 2019; 9:16350.

[22] Bainomugisa A, Duarte T, Lavu E, Pandey S, Coulter C, Marais BJ, Coin LM. A complete high-quality MinION nanopore assembly of an extensively drug-resistant Mycobacterium tuberculosis Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. Microb Genom 2018; 4:e000188.

[23] Passera A, Marcolungo L, Casati P, Brasca M, Quaglino F, Cantaloni C, Delledonne M. Hybrid genome assembly and annotation of Paenibacillus pasadenensis strain R16 reveals insights on endophytic life style and antifungal activity. PLoS One 2018; 13:e0189993.

[24] Tsurumaki M, Deno S, Galipon J, Arakawa K. Complete Genome Sequence of Halophilic Deep-Sea Bacterium Halomonas axialensis Strain Althf1. Microbiol Resour Announc 2019; 8:e00839-19.

[25] Somerville V, Lutz S, Schmid M, Frei D, Moser A, Irmler S, Frey JE, Ahrens CH. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. BMC Microbiol 2019; 19:143.

[26] Yu H, Taniguchi M, Uesaka K, Wiseschart A, Pootanakit K, Nishitani Y, et al. Complete Genome Sequence of Staphylococcus arlettae Strain P2, Isolated from a Laboratory Environment. Microbiol Resour Announc 2019; 8:e00696-19.

[27] Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol 2020; 38:701-707.

[28] Sydenham TV, Sóki J, Hasman H, Wang M, Justesen US; ESGAI (ESCMID Study Group on Anaerobic Infections). Identification of antimicrobial resistance genes in multidrug-resistant clinical Bacteroides fragilis isolates by whole genome shotgun sequencing. Anaerobe 2015; 31:59-64.

[29] Weigand MR, Peng Y, Loparev V, Batra D, Bowden KE, Burroughs M, et al. The History of Bordetella pertussis Genome Evolution Includes Structural Rearrangement. J Bacteriol 2017; 199:e00806-16.

[30] De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom 2019; 5:e000294.

[31] Parajuli P, Deimel LP, Verma NK. Genome Analysis of Shigella flexneri Serotype 3b Strain SFL1520 Reveals Significant Horizontal Gene Acquisitions Including a Multidrug Resistance Cassette. Genome Biol Evol 2019; 11:776-785.

[32] Ruan Z, Wu J, Chen H, Draz MS, Xu J, He F. Hybrid Genome Assembly and Annotation of a Pandrug-Resistant Klebsiella pneumoniae Strain Using Nanopore and Illumina Sequencing. Infect Drug Resist 2020; 13:199-206.

[33] Bayliss SC, Hunt VL, Yokoyama M, Thorpe HA, Feil EJ. The use of Oxford Nanopore native barcoding for complete genome assembly. Gigascience 2017; 6:1-6.

[34] Lemon JK, Khil PP, Frank KM, Dekker JP. Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates. J Clin Microbiol 2017; 55:3530-3543.

[35] Qi W, Colarusso A, Olombrada M, Parrilli E, Patrignani A, Tutino ML, Toll-Riera M. New insights on Pseudoalteromonas haloplanktis TAC125 genome organization and benchmarks of genome assembly applications using next and third generation sequencing technologies. Sci Rep 2019; 9:16444.

[36] Bouchez V, Baines SL, Guillot S, Brisse S. Complete Genome Sequences of Bordetella pertussis Clinical Isolate FR5810 and Reference Strain Tohama from Combined Oxford Nanopore and Illumina Sequencing. Microbiol Resour Announc 2018; 7:e01207-18.

[37] Nguyen SV, Muthappa DM, Hurley D, Donoghue O, McCabe E, Anes J, et al. Yersinia hibernica sp. nov., isolated from pig-production environments. Int J Syst Evol Microbiol 2019; 69:2023-2027.

[38] Molina-Mora JA, Campos-Sánchez R, Rodríguez C, Shi L, García F. High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. Sci Rep 2020; 10:1392.

[39] de Lannoy C, de Ridder D, Risse J. The long reads ahead: de novo genome assembly using the MinION. F1000Res 2017; 6:1083.

[40] Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 2019; 37:540-546.

[41] Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods 2020; 17:155-158.

[42] Cherukuri Y, Janga SC. Benchmarking of de novo assembly algorithms for Nanopore data reveals optimal performance of OLC approaches. BMC Genomics 2016; 17 Suppl 7:507.

[43] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017; 27:722-736.

[44] Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 2016; 32:2103-10.

[45] Vaser R, Šikić A. Raven: a de novo genome assembler for long reads. BioRxiv 2020.08.07.242461 [Preprint]. 2020.

[46] Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol 2020; 38:1044-1053.

[47] Goldstein S, Beka L, Graf J, Klassen JL. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. BMC Genomics 2019; 20:23.

[48] 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.

[49] Tan S, Dvorak CMT, Estrada AA, Gebhart C, Marthaler DG, Murtaugh MP. MinION sequencing of Streptococcus suis allows for functional characterization of bacteria by multilocus sequence typing and antimicrobial resistance profiling. J Microbiol Methods 2020; 169:105817.

[50] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014; 9:e112963.

[51] Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017; 13:e1005595.

[52] Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res 2017; 27:787-792.

[53] Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics 2013; 29:2669-77.

[54] Chen Z, Erickson DL, Meng J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. BMC Genomics 2020; 21:631.

[55] Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon. F1000Res 2018; 7:1755.

[56] Charalampous T, Richardson H, Kay GL, Baldan R, Jeanes C, Rae D, et al., Rapid Diagnosis of Lower Respiratory Infection using Nanopore-based Clinical Metagenomics. BioRxiv 387548 [Preprint]. 2018.

[57] Bialasiewicz S, Duarte TPS, Nguyen SH, Sukumaran V, Stewart A, Appleton S, et al. Rapid diagnosis of Capnocytophaga canimorsus septic shock in an immunocompetent individual using real-time Nanopore sequencing: a case report. BMC Infect Dis 2019; 19:660.

[58] Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, Kadavy DR, Rosenzweig CN, Minot SS. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. Gigascience 2015; 4:12.

[59] Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. Gigascience 2016; 5:4.

[60] Benítez-Páez A, Sanz Y. Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequencer. Gigascience 2017; 6:1-12.

[61] Kerkhof LJ, Dillon KP, Häggblom MM, McGuinness LR. Profiling bacterial communities by MinION sequencing of ribosomal operons. Microbiome 2017; 5:116.

[62] Brown BL, Watson M, Minot SS, Rivera MC, Franklin RB. MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience 2017; 6:1-10.

[63] Cummings PJ, Olszewicz J, Obom KM. Nanopore DNA Sequencing for Metagenomic Soil Analysis. J Vis Exp 2017;(130):55979.

[64] Sanderson ND, Street TL, Foster D, Swann J, Atkins BL, Brent AJ, et al. Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. BMC Genomics 2018; 19:714.

[65] Lin JH, Wu ZY, Gong L, Wong CH, Chao WC, Yen CM, et al. Complex Microbiome in Brain Abscess Revealed by Whole-Genome Culture-Independent and Culture-Based Sequencing. J Clin Med 2019; 8:351.

[66] Schmidt K, Mwaigwisya S, Crossman LC, Doumith M, Munroe D, Pires C, et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. J Antimicrob Chemother 2017; 72:104-114.

[67] Juul S, Izquierdo F, Hurst A, Dai X, Wright A, Kulesha E, Pettett R, Turner DJ. What's in my pot? Real-time species identification on the MinION™. BioRxiv 030742 [Preprint]. 2015.

[68] Lewandowski K, Xu Y, Pullan ST, Lumley SF, Foster D, Sanderson N, et al. Metagenomic Nanopore Sequencing of Influenza Virus Direct from Clinical Respiratory Samples. J Clin Microbiol 2019; 58:e00963-19.

[69] Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 2016; 530:228-232.

[70] Brynildsrud OB, Eldholm V, Bohlin J, Uadiale K, Obaro S, Caugant DA. Acquisition of virulence genes by a carrier strain gave rise to the ongoing epidemics of meningococcal disease in West Africa. Proc Natl Acad Sci U S A 2018; 115:5510-5515.

[71] Payne M, Octavia S, Luu LDW, Sotomayor-Castillo C, Wang Q, Tay ACY, et al. Enhancing genomics-based outbreak detection of endemic Salmonella enterica serovar Typhimurium using dynamic thresholds. Microb Genom 2019.

[72] Tamma PD, Fan Y, Bergman Y, Pertea G, Kazmi AQ, Lewis S, et al. Applying Rapid Whole-Genome Sequencing To Predict Phenotypic Antimicrobial Susceptibility Testing Results among Carbapenem-Resistant Klebsiella pneumoniae Clinical Isolates. Antimicrob Agents Chemother 2018; 63:e01923-18.

[73] Greig DR, Dallman TJ, Hopkins KL, Jenkins C. MinION nanopore sequencing identifies the position and structure of bacterial antibiotic resistance determinants in a multidrug-resistant strain of enteroaggregative Escherichia coli. Microb Genom 2018; 4:e000213.

[74] Schürch AC, van Schaik W. Challenges and opportunities for whole-genome sequencing-based surveillance of antibiotic resistance. Ann N Y Acad Sci 2017; 1388:108-120.

[75] Li R, Xie M, Dong N, Lin D, Yang X, Wong MHY, Chan EW, Chen S. Efficient generation of complete sequences of MDR-encoding plasmids by rapid assembly of MinION barcoding sequencing data. Gigascience 2018; 7:1-9.

[76] Judge K, Harris SR, Reuter S, Parkhill J, Peacock SJ. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. J Antimicrob Chemother 2015; 70:2775-8.

[77] Sakai J, Tarumoto N, Kodana M, Ashikawa S, Imai K, Kawamura T, Ikebuchi K, Murakami T, Mitsutake K, Maeda T, Maesaki S. An identification protocol for ESBL-producing Gram-negative bacteria bloodstream infections using a MinION nanopore sequencer. J Med Microbiol 2019; 68:1219-1226.

[78] Pitt ME, Nguyen SH, Duarte TPS, Teng H, Blaskovich MAT, Cooper MA, Coin LJM. Evaluating the genome and resistome of extensively drug-resistant Klebsiella pneumoniae using native DNA and RNA Nanopore sequencing. Gigascience 2020; 9:giaa002.

[79] McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 2013; 57:3348-57.

[80] Hyeon JY, Li S, Mann DA, Zhang S, Li Z, Chen Y, Deng X. Quasimetagenomics-Based and Real-Time-Sequencing-Aided Detection and Subtyping of Salmonella enterica from Food Samples. Appl Environ Microbiol 2018; 84:e02340-17.

[81] Břinda K, Callendrello A, Ma KC, MacFadden DR, Charalampous T, Lee RS, et al. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. Nat Microbiol 2020; 5:455-464.

Figure legends

Figure 1. Oxford Nanopore sequencing data deposited in the Sequencing Reads Archive by year

Oxford Nanopore sequencing data availability in the Sequencing Reads Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra). Total deposited data were retrieved for each year between 2015 and 2020 (from Jan 1st to Dec 31th) and are represented by red bars, sequencing data related to bacterial samples were retrieved applying a filter ("bacteria") in the search key and are represented in blue. Scale of y axis is logarithmic.

Figure 2. Comparison of assembler algorithms

Overview of the: **A**. De Bruijn Graph algorithm. **B**. Overlay Layout Consensus algorithm. See text for detailed explanation.

Tables

Table 1. Bacterial genome assembly tools

The table summarises advantages and disadvantages of each long read assembler for bacterial genome assembly [14]. Assembler tools were evaluated by evaluating different parameters, defining a value as high (+++), good (++), moderate (+), or low (-). The parameters taken into consideration are i) the core algorithm involved, De Bruijn Graph (**DBG**) or Overlay Layout Consensus (**OLC**), ii) the capability of achieving a complete genome assembly (**Reliability**), iii) the capability of reducing "indel" errors (**Accuracy**), iv) the capability of assembling circular chromosomes (**Contiguity**), v) the capability of assembling complete and circular plasmids (**Plasmids**), vi) computational requirements (**Resource usage**), vii) easiness of installation and running (**Ease of use**), and viii) the presence of additional running parameters (**Configurability**).

Figures and tables

Tables

Table 1 Bacterial genome assembly tools

Assembler	Algorithm	Reliability	Accuracy	Contiguity	Plasmids	Resource usage	Ease of use	Configurability	Pros	Cons
Canu	OLC	+++	++	+	+++	+++	+	+++	Highly reliable and accurate Easy to use, high configurability Reads error-correction stand alone module	Predicted genome size required Exceed circularisation Very slow Intensive computational resources usage
Flye	DBG	++	+++	++	+++	++	++	+	The most accurate and highly reliable Easy to use, moderately configurable Typically fast	Often deletes some bases in circularisation High memory usage
Miniasm	OLC	+	+	+++	++	+	++	++	Moderately reliable and accurate Exact circularisation (100% contiguity) Good at plasmid assembly Fast and low resource usage Ease of use, high	Consensus step not included Required polishing step Limitations for small plasmids

									configurability	
Raven	OLC	+++*	+	+	+	+	++	-	The most reliable for chromosome assembly (*) Moderately accurate Fast and low resources usage	Does not resolve plasmids Loses hundreds of bases in circularisation The least configurable
Redbean	DBG	-	+	-	-	+	++	++	Moderately accurate Fast and low resources usage Easy to use, highly configurable Good for low coverage genomes	Low reliability Erratic circularisation Fails plasmid assembly Predicted genome size
Shasta	OLC	+	-	-	-	-	++	++	Moderately reliable Very fast and low resources usage Easy of use, high configurability Mainly suitable for large genomes	Low accuracy Indels error size longer than 10 bp Erratic circularisation Fails plasmid assembly

Figures

Figure 1

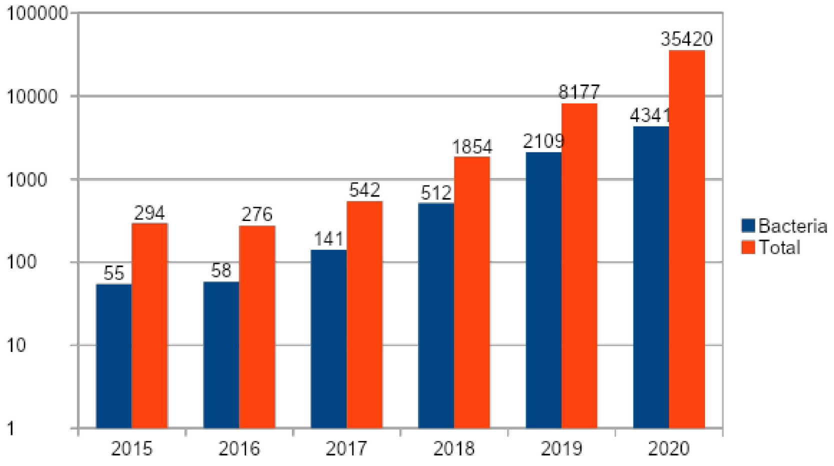
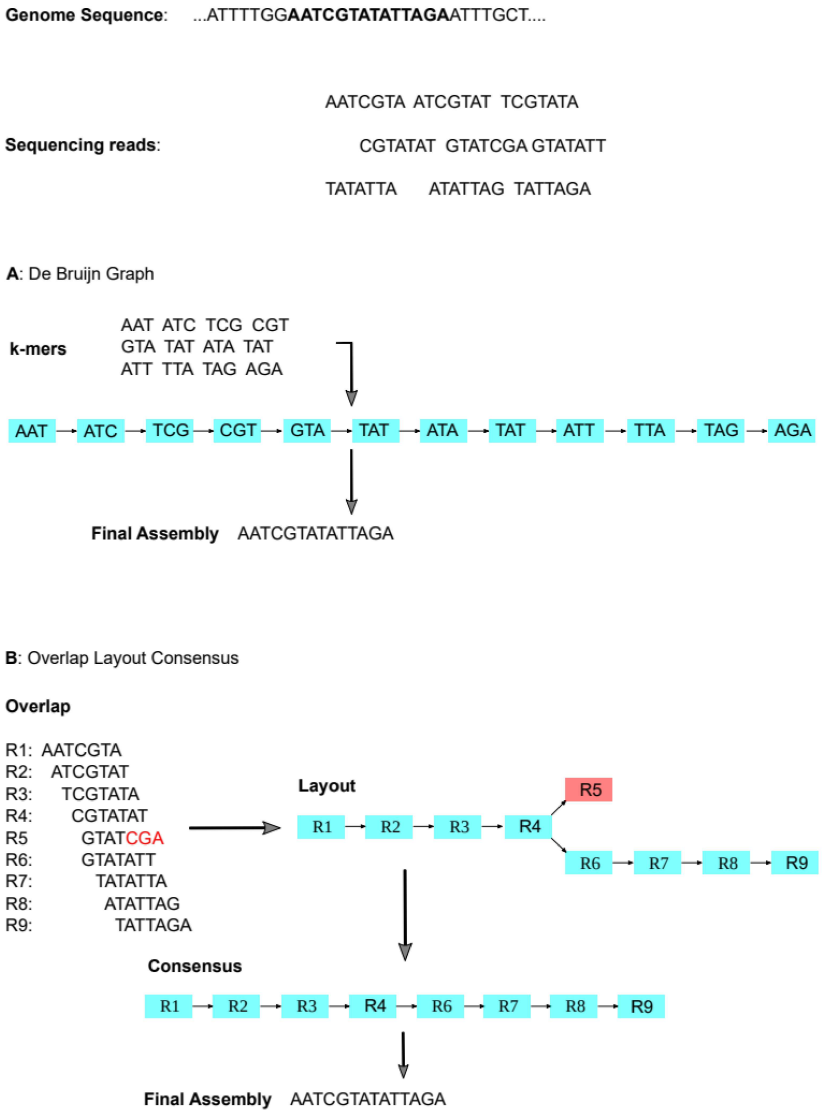


Figure 2



Chapter 2

DNA isolation methods for Nanopore sequencing of *Streptococcus mitis* genome

David Pinzauti, Francesco Iannelli, Gianni Pozzi and Francesco Santoro

Abstract

Streptococcus mitis is a Gram positive bacterium, member of the oral commensal microbiota. *S. mitis* can occasionally be the etiologic agent of diseases such as infective endocarditis, bacteraemia and septicaemia. The highly recombinogenic and repetitive nature of *S. mitis* genome makes it difficult to generate a complete genome sequence relying only on short reads data. Oxford Nanopore sequencing represents an optimal tool to overcome this limitation. Its capability of generating long reads enables the resolution of genomic repeated regions and makes it possible to achieve a complete genome sequence. However, because Nanopore sequencing is strongly influenced by genomic DNA quality and molecular weight, the DNA isolation step represents the first challenge for an optimal sequencing run. In the present paper we have compared three DNA extraction protocols, evaluating their capability of preserving genomic DNA integrity. Protocols were tested on two *S. mitis* strains: for both strains, the mechanical lysis based approach did not generate ultra-long reads (>100 kb), while enzyme-based approaches enabled the isolation of high molecular weight DNA allowing the generation of ultra-long reads and the reconstruction of a single, complete genome.

Introduction

Streptococcus mitis, formerly known as *Streptococcus mitior*, is a mesophilic α-hemolytic Streptococcus belonging to the Mitis group. It is a Gram positive, facultative anaerobe and catalase negative coccus, which has been generally considered a relatively benign oral streptococcus. As a member of the oral commensal flora, *S. mitis* is most commonly found in the throat, nasopharynx and mouth, but it can escape from this niche causing a variety of complications including endocarditis, bacteraemia and septicaemia (1). *S. mitis* has emerged as a significant pathogen in elderly, immunocompromised patients and in patients undergoing cytotoxic chemotherapy treatment. Moreover it is also an infrequent opportunistic pathogen of normal healthy infants and adults, implicated in a wide range of diseases from dental caries to bacterial infective endocarditis, bacteraemia, meningitis, eye infections, and pneumonia. Very little is understood about how exactly *S. mitis* causes this variety of diseases, although its ability to bind platelets has been directly implicated in the pathogenesis of infective endocarditis (1). The *S. mitis* genome harbours several virulence factors, homologues of many of the identified *Streptococcus pneumoniae* virulence factors such as autolysins, choline-binding proteins, IgA1 proteases, and cell-wall anchored adhesins. This physiological similarity has long been established suggesting a cross-relationship between them (2,3). Furthermore, this recurrence of virulence factors and significant structural similarities found in the capsule locus (2), raises important questions concerning the consequences for host-parasite relationships both for the commensals and for the pathogen *S. pneumoniae* (3).

The *Streptococcus mitis* genome is between 1.8 and 2.1 Mb in length encoding up to 2.277 genes, and has a median G+C content of 40%. To date, 141 *S. mitis* genomes are deposited in GenBank (https://www.ncbi.nlm.nih.gov/genome/genomes/530), of those only 9 are complete. Difficulties in the assembly process are mainly related to its highly recombinogenic and repetitive nature, making it difficult to generate a complete genome sequence. Oxford Nanopore technology can generate long and ultra-long sequencing reads, as long as the DNA fragments used for library preparation. Long reads enable resolution of genomic complexities, making it possible to obtain a complete genome assembly. However, because Nanopore sequencing is strongly influenced by genomic DNA quality and molecular weight, the DNA isolation step remains the first challenge for an optimal sequencing run.

In the present study we have tested three different DNA extraction protocols, evaluating their ability to isolate high molecular weight DNA prior to Oxford Nanopore sequencing. Two *S.*

mitis isolates, named S022-V3-A4 and S022-V7-A3, were used to assess protocols' efficiency. First, we have compared protocols yields in terms of genomic DNA quantity, purity and integrity, evaluating their capability of isolating high molecular weight DNA without compromising the purity. Then, we have evaluated their ability to generate long and ultra-long Nanopore reads, enabling complete *de novo* genome assembly. Lastly, relying on a hybrid assembly strategy, the original genomic architectures of both strains were reconstructed. Complete genomes were deposited in the GenBank database.

Materials and Methods

DNA extraction methods

Streptococcus mitis S022-V3-A4 and S022-V7-A3 were isolated from the saliva of a healthy subject treated with minocycline during the ANTIRESDEV project (4).

Both strains were inoculated 1:50 (vol:vol) from frozen starter cultures in 10 ml Tryptic Soy Broth (TSB) and incubated at 37° C in a water bath. Growth was monitored until an OD_{590} of 1.0 was achieved, corresponding to $3x10^7$ CFU/ml. Genomic DNA was extracted employing three different protocols respectively designed as CTAB, Raffinose, and TissueLyser. Two out of three protocols, CTAB and Raffinose respectively, rely on enzymatic lysis of the bacterial cell, inducing Protoplast formation and subsequent DNA purification. The third one is instead based on mechanical cell wall disruption using glass beads and the Qiagen TissueLyser device, followed by a DNA purification step employing AMPure XP beads (Beckman Coulter). A schematic overview of the extraction protocols workflows is presented in **Figure 1**.

The CTAB protocol was adapted from Current Protocols in Molecular Biology (5). Bacterial cultures were centrifuged at 6500 x *g* for 5 minutes and resuspended in 14.8 ml of Tris 10 mM-EDTA 1 mM (TE) buffer pH 8.0. Lysozyme (Sigma Aldrich) at a final concentration of 2.6 mg/ml was added to hydrolyse the cell wall and induce protoplast formation. The reaction was incubated for 1 hour at 37° C. Then 800 µl of 10% Sodium Dodecyl Sulphate (SDS) and Proteinase K (Sigma Aldrich) at a final concentration of 0.1 mg/ml were added, incubating the reaction for 30 minutes at 37° C. After the addition of 2 ml of 5M Sodium Chloride (NaCl) and 2 ml of CTAB (pre-heated at 65° C), the solution was incubated for 10 minutes at 65° C. DNA purification from cellular contaminants was performed by adding 1 volume of Sevag (Chloroform:Isoamyl alcohol, 24:1 vol:vol), the solution was centrifuged for 15 minutes at 6600 x *g* and the supernatant was recovered and transferred to a fresh tube using a Pasteur pipette. Sevag wash was repeated twice. The supernatant from the second wash was precipitated adding 0.6 volumes of ice-cold (-20° C) Isopropanol (Sigma Aldrich), and the solution was incubated for 30 minutes at -20° C. After a centrifugation (6600 x *g*, 15 minutes) the supernatant was discarded and the DNA pellet was finally re-suspended in 100 µl of physiologic solution (NaCl 0.9%) and stored at +4 °C.

The Raffinose protocol was adapted from the CTAB protocol itself but avoiding CTAB use, thought to have a detrimental effect on HMW DNA, and changing the initial lysis. In short, 10 ml of bacterial culture were centrifuged at 6600 x *g* for five minutes, discarding the supernatant. The bacterial pellet was washed with 10 ml of sterile TE buffer, spinned down again, and resuspended in 7.5 ml Raffinose buffer (50 mM Tris pH 8, 5 mM EDTA and 20 % Raffinose). Lysozyme at a final concentration of 2.6 mg/ml was added and the reaction was incubated 1 hour at 37° C (water bath). The solution was centrifuged at 6600 x *g* for 5 minutes, and sterile distilled water (dH_2O) was added to resuspend the pellet and osmotically lyse protoplasts. Proteinase K (0.1 mg/ml final concentration) and 400 µl of 10% SDS were then added incubating the solution for 30 minutes at 37° C. 1 ml of 5 M NaCl was added (10 minutes room temperature incubation) and two Sevag washes were performed. DNA precipitation and resuspension were performed as previously described.

The TissueLyser protocol relies on glass beads to mechanically disrupt the bacterial cell wall. The bacterial pellet was resuspended in 1 ml of sterile TE buffer and transferred into a clean 2 ml Eppendorf containing 0.04 gr of sterile glass beads (Sigma-Aldrich, Ø 150-212 µm). The cells were lysed with two passages of 2 minutes at 30 Hz frequency inside the TissueLyser device (Qiagen). The solution was then spinned down (5 min at 1800 x *g*) recovering the supernatant. DNA purification was performed by adding 0.4x AMPure XP beads (Beckman Coulter) and incubating for 15 minutes in a Rotator Mixer. By using a magnetic rack, the beads were pelleted and washed twice with freshly prepared 70% ethanol (EtOH). The pellet was finally resuspended in 100 µl NaCl 0.9%: after 15 minutes of incubation (37° C), beads were pelleted again and the supernatant was recovered.

DNA quantification

Extraction protocols were compared in terms of genomic DNA quantity, quality, and molecular weight. Extraction yields were measured using Qubit dsDNA BR assay kit (Thermo Fisher Scientific) in a Qubit 2.0 Fluorometer device (Invitrogen). The DNA concentration and the whole DNA amount recovered were measured for each extraction protocol. Genomic DNA purity grade was determined using a NanoPhotometer device (IMPLEN), evaluating the DNA absorbance ratios 260 nm/280 nm and 260 nm/230 nm. As a general rule, an optimal sample purity for Oxford Nanopore sequencing should be between 2.0-2.2 of the 260/230 ratio and around 1.8 of the 260/280 ratio. DNA integrity was finally determined using an Agarose gel electrophoretic assay: samples were loaded into a 0.8% Agarose gel and ran 4 hours at 3 V/cm in 0.5x Tris-Borate-EDTA (TBE) buffer.

Oxford Nanopore library preparation

Isolated DNA samples were subsequently used for sequencing library preparation employing the Ligation Sequencing kit (SQK-LSK108) and the Expansion Barcoding kit (EXP-NB103), manufactured by Oxford Nanopore Technologies (ONT). A unique barcode sequence was associated to each DNA sample, enabling samples multiplexing. Sequencing libraries were prepared following manufacturer's instructions (https://nanoporetech.com/), though introducing an initial size-selection step to reduce small DNA fragments contamination. Wide bore pipette tips were used and vortexing was avoided trying to further reduce DNA shearing. In short, for each extraction protocol 2.5 µg of genomic DNA were size-selected using 0.7x AMPure XP beads (Beckman Coulter). The reaction was incubated for 15 minutes in the Rotator Mixer and pelleted on a magnetic rack. Beads pellet was washed twice using freshly prepared 70% ethanol and eluted in 40 µl NaCl 0.9%. Size-selected samples were end-repaired using FFPE and Ultra II End-Prep kits from New England BioLabs (NEB). Barcodes were ligated using Blunt TA Master Mix (NEB) and mixed together into an equimolar pool of 1.2 µg DNA in 50 µl physiologic solution. Adapter proteins were ligated using Quick T4 Ligase (NEB), 30 minutes at room temperature incubation. In the end, 520 ng of sequencing library were loaded onto a new R9.4 flowcell. The run was performed using a MinION device.

Data Analysis

Oxford Nanopore native fast5 files were basecalled using stand-alone Guppy module (v. 2.1.3) and demultiplexed using Deepbinner (v. 0.2.0) (6). Sequencing readouts were analysed using NanoStat (v. 1.1.2) (7), while reads length distributions were plotted using the R-package ggplot2 (v. 2.2.1)(8). All tools were run using default parameters. Relying solely on Nanopore reads, a *de novo* genome assembly was performed by Unicycler tool (v. 0.4.7) (9) evaluating the capability of each extraction protocol to generate a complete genome assembly. Lastly, feeding Illumina reads to Unicycler, a hybrid assembly approach was also performed reconstructing the original genomic architecture of both S022-V3-A4 and S022-V7-A3 strains.

Results

DNA integrity is preserved using enzymatic lysis

Genomic DNA extractions yields are listed in **Table 1** while results from the gel electrophoretic assay are shown in **Figure 2**. For both *S. mitis* strains the CTAB protocol was the most efficient in terms of total DNA recovery, while the TissueLyser protocol had an overall lower yield. Qualitative spectrophotometric analysis showed that both the A260/A280 and A260/A230 ratios were within reasonable limits for all extraction approaches. The S022-V7-A3 DNA extracted with CTAB protocol showed low A260/A280 and A260/A230 ratios, possibly indicating the presence of residual proteins and ethanol, but this did not significantly affect sequencing results (see **Table 2**).

Agarose gel electrophoresis (**Figure 2**) showed that enzymatic lysis based methods were able to preserve the integrity of genomic DNA, which could be detected as a sharp high molecular weight band with little smear. To note, residual HMW in the loading well is also present. On the other hand, the TissueLyser protocol yielded a smear in the lane with no HMW band indicating DNA fragmentation.

Raffinose DNA extraction generates ultra-long nanopore reads

Nanopore sequencing was used as a final readout for the extraction protocols comparison. Results are shown in **Table 2**. Samples obtained from different extraction protocols were run in the same flow cell, the sequencing run was stopped after 4 hours, generating a total of ~870 Mb. Each extraction protocol yielded a 70x genome coverage, with the exception of the CTAB protocol for S022-V3-A4 which roughly achieved a 50x coverage. Statistical analysis suggests a similar mean read length across the three protocols, while the TissueLyser generated a higher median length possibly due to the two size selection steps performed (one during the extraction and the other before library preparation). NanoStat measured read length N50 values suggest that both enzymatic lysis-based approaches were able to preserve DNA integrity generating longer reads. In particular the Raffinose protocol achieved higher N50 values and was the only one capable of generating multiple ultra-long reads (i.e. >100 kb). Plotting the reads length distribution (**Figure 3**) further showed that TissueLyser generated more reads around 10 kb of length, but did not have a tail of longer reads which was instead present for both the enzymatic lysis method (more evident for the Raffinose protocol).

The capability of generating long reads can be translated into the prospect of assembling complete bacterial genomes. *De novo* genome assembly showed that both enzymatic lysis-based protocols generate long reads enough to solve *S. mitis* genomic complexities, making it possible to achieve a complete genome (**Table 3**). The TissueLyser protocol instead generated an incomplete, fragmented assembly: the S022-V3-A4 genome was assembled into two linear contigs (total length of 2.1 Mbp) where the longest measures 1,991,501 bp, while S022-V7-A3 genome was assembled into two linear contigs (total length 2.0 Mbp) where the longest measures 1,990,964 bp. Although faster, the TissueLyser has been proven to be a harsh process fragmenting genomic DNA, failing to achieve ultra-long reads and resulting in the impossibility to solve genomic complexities.

The original genome architecture of both *S. mitis* strains were finally reconstructed based on a hybrid assembly approach, involving both Illumina and Nanopore reads. Nanopore reads obtained from different extraction protocols were merged together achieving an overall 200x genome coverage (~400 Mb) for each *S. mitis* strains. Illumina reads were previously generated. *S. mitis* S022-V3-A4 was assembled into a complete chromosome of 2,086,958 bp, while S022-V7-A3 strain was assembled into a complete chromosome of 2,033,396 bp. Complete genome sequences were deposited in the GenBank database (**Table 4**).

Conclusions

In the present study we have tested three extraction approaches identifying protocols suitable for high molecular weight DNA isolation without compromising samples purity. Enzymatic lysis-based protocols were able to preserve genomic DNA integrity allowing the isolation of high molecular weight DNA, where best results were achieved by the Raffinose protocol. The isolation of high molecular weight DNA has been translated into the capability of sequencing long and ultra-long Nanopore sequencing reads (>100 kbp), which enable the resolution of *S. mitis* genomic complexities making it possible to achieve a complete genome assembly. On the other hand, the mechanical lysis method generated reads with a higher median length but with a maximum length of 59 kbp, resulting in a single open contig covering 98.2% of the whole genome, but has the advantage of being a quick and cheap method, which may be useful for less complex genomes. In conclusion, gentle enzymatic cell lysis is essential to preserve genomic DNA integrity which is the *sine qua non* for obtaining long reads allowing successful genome assembly.

Data availability

Streptococcus mitis S022-V3-A4 and S022-V7-A3 sequencing reads and genomic sequences are publicly available. Sequence Read Archive (SRA) and GenBank accessions numbers are reported in Table 4.

Table 1. *Streptococcus mitis* extraction results

	S022-V3-A4 Extraction Yields			**S022-V7-A3 Extraction Yields**		
Qubit Fluorometer:	**CTAB**	**Raffinose**	**TissueLyser**	**CTAB**	**Raffinose**	**TissueLyser**
ng/µl	250	192	98,6	562	214	25,8
µg total	25	19,2	9,86	56,2	21,4	2,58
Nanophotometer:						
A260/230	2.021	1.994	2.100	1.422	2.024	2.167
A260/280	2.185	2.026	1.909	1.656	2.473	2.167

Quantitative and Qualitative analysis of genomic DNA isolated using three different extraction protocols. Isolated DNA quantity values were measured by a Qubit 2.0 Fluorometer device assessing DNA concentration (ng/µl) and the total amount of extracted DNA (measured in a final resuspension volume of 100µl). DNA purity values were instead measured using Spectrophotometer device.

Table 2. *Streptococcus mitis* sequencing readout

S022-V3-A4 Sequencing Readout			
	CTAB	**Raffinose**	**TissueLyser**
Mean read length	4,416.70	5,271.30	5,027.60
Median read length	2,520	2,547	4,647
Number of reads	26,193	27,793	28,290
Read length N50	7,592	10,505	6,515
Total bases	115,687,056	146,505,005	142,231,432
Top 5 longest reads:			
1.	80,932	174,836	48,134
2.	79,082	128,776	44,248
3.	76,286	117,679	42,804
4.	75,212	114,489	40,531
5.	74,875	114,484	39,739

S022-V7-A3 Sequencing Readout			
	CTAB	**Raffinose**	**TissueLyser**
Mean reads length	4,133.30	4,293.80	3,933.10
Median reads length	2,170	2,112	3,285
Number of reads	31,539	38,579	46,404
Read length N50	7,661	8,461	5,144
Total bases	130,361,553	165,648,998	182,512,418
Top 5 longest reads:			
1.	105,257	122,493	58,740
2.	92,029	121,050	50,701
3.	81,057	102,516	46,516
4.	79,426	100,666	36,086
5.	78,773	96,686	35,467

Sequencing readouts from different DNA extraction protocols were generated using NanoStat tool (v. 1.1.2) (**7**). Starting from Guppy basecalled fastq reads, a statistical summary report was generated for each extraction protocol. Sequencing results were analysed in terms of reads length, reads number, total of sequenced bases, and longest reads.

Table 3. *Streptococcus mitis* genome assembly results

Strain	**CTAB**	**Raffinose**	**TissueLyser**
S022-V3-A4	2.087.020 bp	2.087.361 bp	2.106.883* bp
S022-V7-A3	2.033.037 bp	2.033.626 bp	2.003.277* bp

Long nanopore reads were assembled into contigs using the Unicycler tool (v. 0.4.7) (**9**). Uncomplete contigs are denoted by an asterisk (*).

Table 4. *Streptococcus mitis* data availability

S. mitis strain	Genome sequence	CTAB reads	Raffinose reads	TissueLyser reads
S022-V3-A4	CP047883.1	SRR13390533	SRR13390532	SRR13390531
S022-V7-A3	CP067992.1	SRR13390530	SRR13390535	SRR13390534

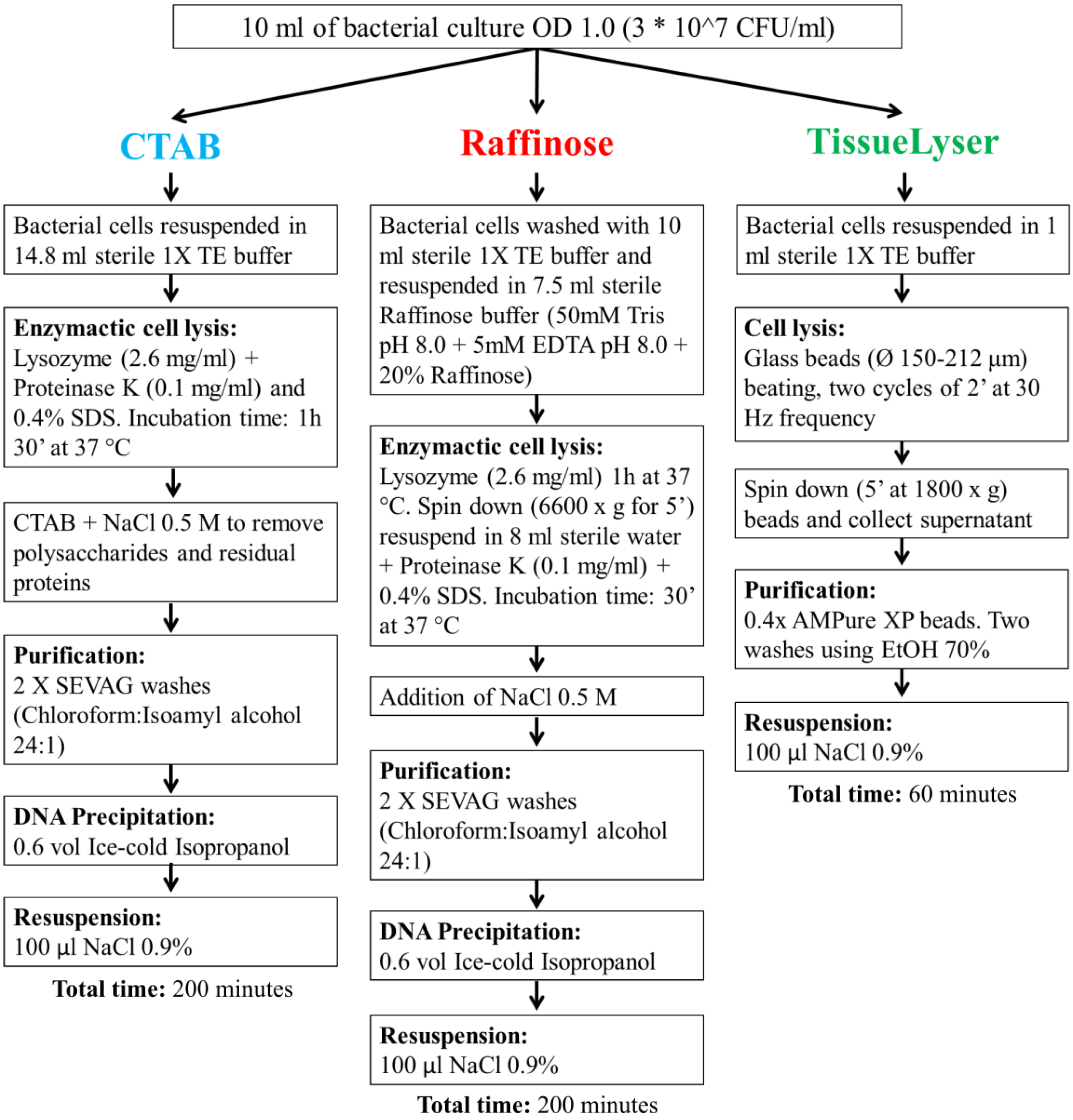


Figure 1. Extraction protocols workflows

10 ml of bacterial culture OD 1.0 (3 * 10^7 CFU/ml)
CTAB
Bacterial cells resuspended in 14.8 ml sterile 1X TE buffer
Enzymactic cell lysis: Lysozyme (2.6 mg/ml) + Proteinase K (0.1 mg/ml) and 0.4% SDS. Incubation time: 1h 30’ at 37 °C
CTAB + NaCl 0.5 M to remove polysaccharides and residual proteins
Purification: 2 X SEVAG washes (Chloroform:Isoamyl alcohol 24:1)
DNA Precipitation: 0.6 vol Ice-cold Isopropanol
Resuspension: 100 µl NaCl 0.9%
Total time: 200 minutes
Raffinose
Bacterial cells washed with 10 ml sterile 1X TE buffer and resuspended in 7.5 ml sterile Raffinose buffer (50mM Tris pH 8.0 + 5mM EDTA pH 8.0 + 20% Raffinose)
Enzymactic cell lysis: Lysozyme (2.6 mg/ml) 1h at 37 °C. Spin down (6600 x g for 5’) resuspend in 8 ml sterile water + Proteinase K (0.1 mg/ml) + 0.4% SDS. Incubation time: 30’ at 37 °C
Addition of NaCl 0.5 M
Purification: 2 X SEVAG washes (Chloroform:Isoamyl alcohol 24:1)
DNA Precipitation: 0.6 vol Ice-cold Isopropanol
Resuspension: 100 µl NaCl 0.9%
Total time: 200 minutes
TissueLyser
Bacterial cells resuspended in 1 ml sterile 1X TE buffer
Cell lysis: Glass beads (Ø 150-212 µm) beating, two cycles of 2’ at 30 Hz frequency
Spin down (5’ at 1800 x g) beads and collect supernatant
Purification: 0.4x AMPure XP beads. Two washes using EtOH 70%
Resuspension: 100 µl NaCl 0.9%
Total time: 60 minutes

Figure 2. Agarose gel electrophoresis of *Streptococcus mitis* DNAs

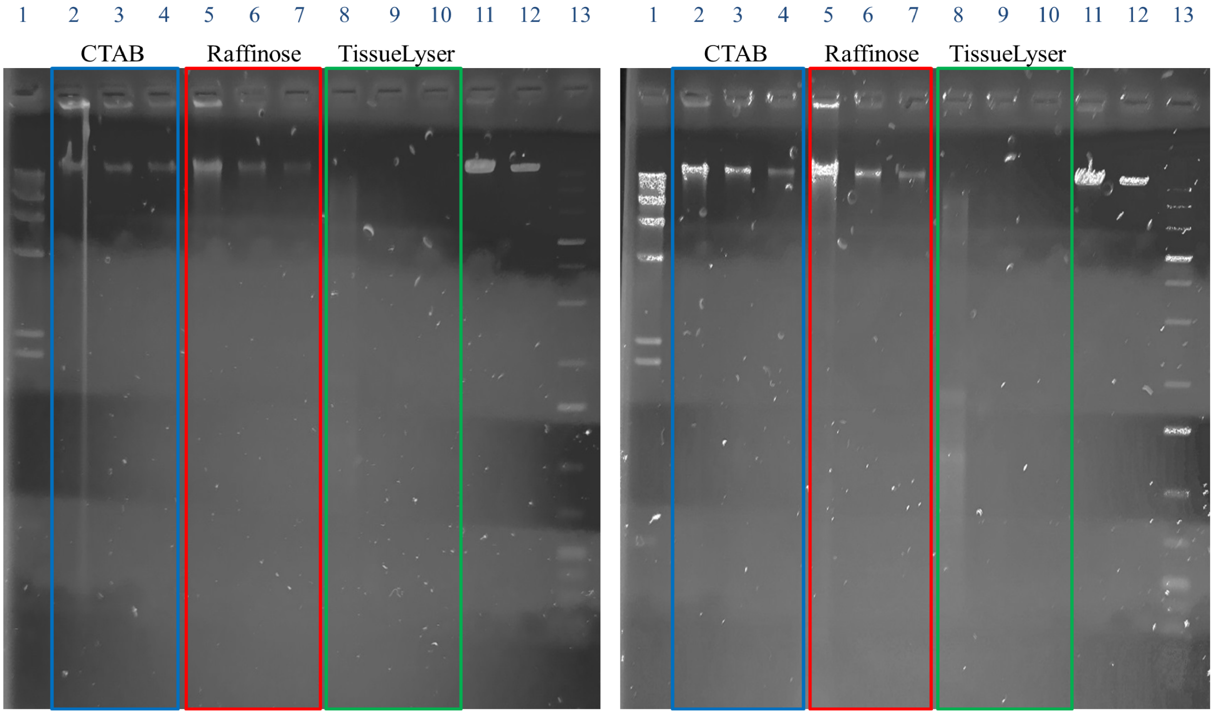


Figure 2. On the left are represented DNA samples extracted from *Streptococcus mitis* S022-V3-A4 strain, while on the right DNAs extracted from *Streptococcus mitis* S022-V7-A3 strain. Samples (1 µl) were loaded on a 0.8% agarose gel and ran 4 hours at 3 V/cm in 0.5x TBE buffer. CTAB extracted samples were loaded in lanes 2, 3 (1:10 dilution) and 4 (1:20 dilution); Raffinose in lanes 5, 6 (1:10 dilution) and 7 (1:20 dilution). TissueLyser samples were loaded in lanes 8, 9 (1:10 dilution) and 10 (1:20 dilution). DNA samples were run together with molecular markers Lambda (λ) DNA/HindIII (lane 1), GeneRuler™ 1 kb Plus DNA Ladder (lane 13), 100 ng and 30 ng (lanes 11-12) of λ DNA. Staining was performed 15' in Ethidium Bromide.

Figure 3. Sequencing reads length distribution plots

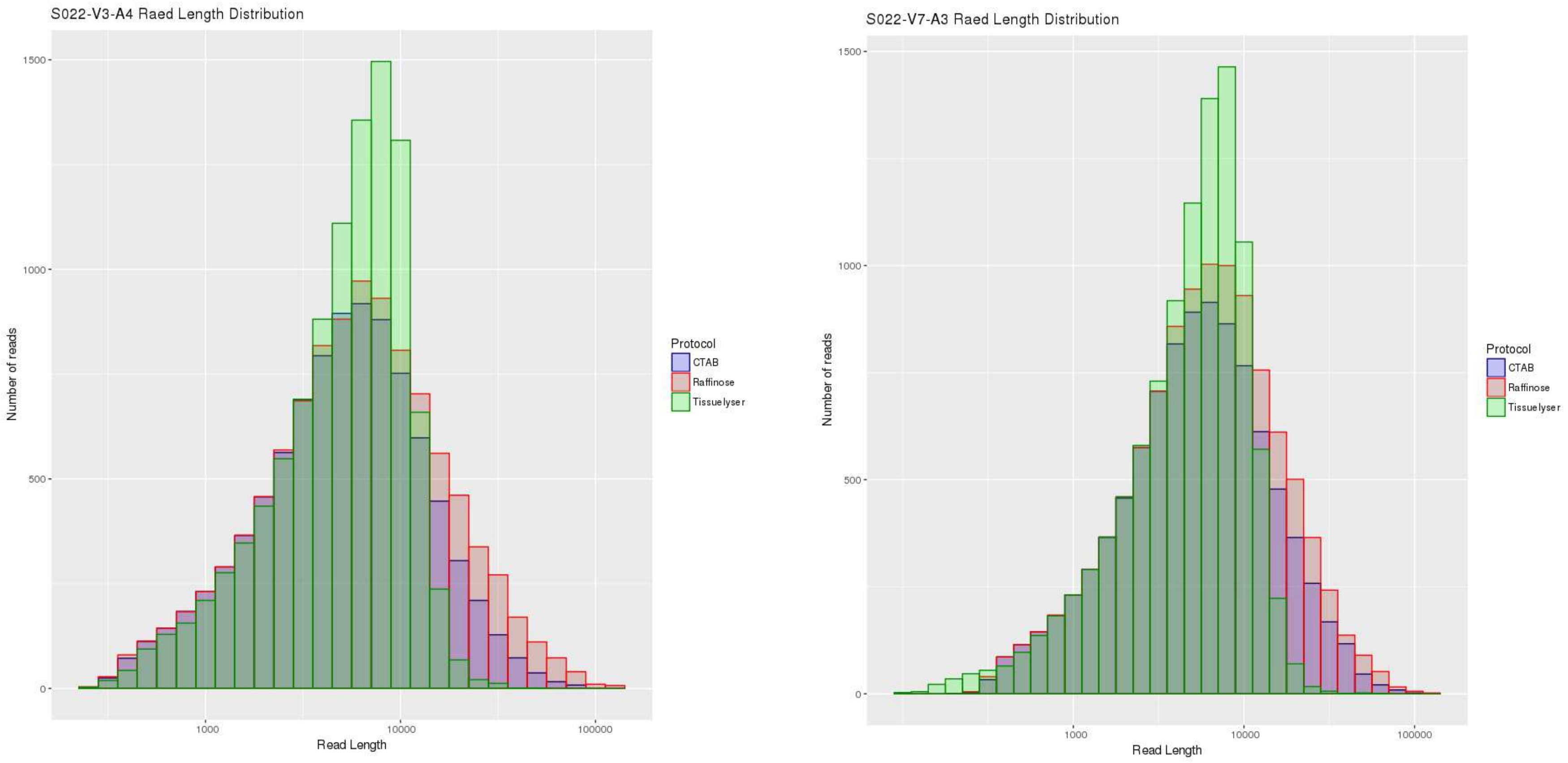


Figure 3. The read length distribution obtained by each extraction protocol was represented using the R-package ggplot2 (v. 2.2.1) (**8**). On the X-axis are plotted reads length values while on the Y-axis the reads number. Extraction protocols were represented using three different colours: blue for CTAB, red for Raffinose, and green for TissueLyser protocol.

Bibliography

1. Mitchell J. Streptococcus mitis: walking the line between commensalism and pathogenesis. Mol Oral Microbiol. 2011 Apr;26(2):89-98. doi: 10.1111/j.2041-1014.2010.00601.x

2. Kilian M, Riley DR, Jensen A, Bruggemann H, Tettelin H. Parallel evolution of Streptococcus pneumoniae and Streptococcus mitis to pathogenic and mutualistic lifestyles. MBio, 2014 Jul 22;5(4):e01490-14. doi: 10.1128/mBio.01490-14.

3. Sorensen UB, Yao K, Yang Y, Tettelin H, Kilian M. Capsular Polysaccharide Expression in Commensal *Streptococcus* Species: Genetic and Antigenic Similarities to *Streptococcus pneumoniae*. MBio, 2016 Nov 15;7(6). Pii: e01844. doi: 10.1128/mBio.01844-16.

4. Zaura E, Brandt BW, Teixeira de Mattos MJ, Buijs MJ, Caspers MPM, Rashid MU, et al. Same exposure but two radically different responses to antibiotics: resilience of the salivary microbiome versus long-term microbial shifts in feces. mBio 6(6):e01693-15. doi:10.1128/mBio.01693-15.

5. Wilson K. Preparation of genomic DNA from bacteria. Curr Protoc Mol Biol. 2001 Nov;Chapter 2:Unit 2.4. doi: 10.1002/0471142727.mb0204s56.

6. Wick RR, Judd LM, Holt KE. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. PLoS Comput Biol. 2018 Nov 20;14(11):e1006583. doi: 10.1371/journal.pcbi.1006583.

7. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018 Aug 1;34(15):2666-2669. doi: 10.1093/bioinformatics/bty149.

8. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

9. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017 Jun 8;13(6):e1005595. doi: 10.1371/journal.pcbi.1005595.

2.1 Conclusions

The first real challenge for an optimal Nanopore sequencing run is represented by the isolation of high quality genomic DNA. The possibility of isolating high molecular weight brings essential benefits for an optimal run. Because Oxford Nanopore technology has no fixed maximum reads length, it enables the sequencing of long and ultra long reads, as long as the DNA fragments used for the library preparation. Enzymatic lysis-based protocols achieve a gentle bacterial cell lysis, preserving the genome integrity without compromising the DNA quality.

We have successfully optimized an extraction protocol, called “Raffinose” protocol, capable of isolating high molecular weight DNA resulting in the sequencing of multiple ultra long reads. Despite being initially developed for though Gram-positive bacteria lysis, the protocol is also suitable for the lysis of thinner cell wall from Gram-negative bacteria.

Chapter 3. Hybrid sequencing of *Enterococcus faecalis* isolates from periapical lesions

3.1 *Enterococcus faecalis* epidemiology

The genus *Enterococcus* belongs to the family of *Enterococcaceae*, in the order of *Lactobacillales* of the phylum *Firmicutes* (https://lpsn.dsmz.de/genus/enterococcus).
The genus consists of 70 recognized species and 2 subspecies (https://lpsn.dsmz.de/search?word=Enterococcus). Due to morphological and biochemical similarities, the Enterococci were firstly considered as a part of the *Streptococcus* genus and classified as group D Streptococci (1899) (1). It was only in the 1970 that the genus *Enterococcus* was accepted and separated from Streptococci (2). Enterococci are non-spore-forming microorganisms with an ovoidal shape found as pairs, chains, or groups (3). They are facultative anaerobes whose metabolism relies on the fermentation of carbohydrates and the production of lactic acid as an end product (4). Most Enterococci are i) oxidase and catalase negative; ii) salt tolerant; iii) resistant to 40% bile; iv) esculin hydrolytic; and v) able to grow in the presence of sodium azide (1). Their genome ranges from 2.3 up to 5.4 Megabases, with a G+C content of 34-45 % and 2,154-5,107 predicted genes (1).
The genus *Enterococcus* comprises ubiquitous bacteria isolated in several different environments, from soil to surface and sea waters, but are also found in association with plants, in fermented food products, and as part of the normal gut microbiota (1). Enterococci are considered commensal organisms of the human gastrointestinal tract but have the capability of becoming pathogenic as causative agents of urinary tract infections (UTIs), bacteremia, endocarditis, soft tissue infections, dental diseases, burn and surgical wound infections, and infections of implanted medical devices (1). Moreover, multidrug resistant strains have rapidly emerged resulting in prolonged hospitalization time, increased treatment costs, higher chances of treatment failure and death. It has been reported that the incidence of enterococcal infections has increased since the late 1970s, becoming the 2nd most common pathogen of healthcare-associated infections (HAI) in both Europe and the United States (1). The major enterococcal species accounting for ~75% of the all typed enterococcal infections are *Enterococcus faecalis* and *Enterococcus faecium* (5).

To date (January 2021) the GenBank database hosts 1,857 genomes of *Enterococcus faecalis*, of which only 53 are complete while 1,804 are draft genome assemblies. The reported median total length of *E. faecalis* genome is 2.97 Megabases, encoding for 2,762 genes, with a median

G+C content of 37.4%. The first complete *E. faecalis* genome sequence was obtained from the vancomycin-resistant strain V583 and published in 2003 (6). A recent comparative genomic analysis performed on 78 Enterococcal isolates (7), defined the presence of 1,361 conserved genes composing the core-genome and suggesting that over half of the predicted ORFs in each genome were dispensable. A functional analysis of the core-genome reported that the 33.4% of core genes are mainly involved in nucleotide and amino acid transport, while very few genes are related to defence mechanism and mainly encoding for the ATP-binding cassette (ABC) transport system conferring multidrug resistance (7). On the contrary, the biggest pan-genome comprises the subset of multiple defence mechanisms-related genes.

The success of *E. faecalis* as a pathogen resides in its capability of expressing a range of virulence factors and surviving in a hostile, antimicrobial-rich environment due to its intrinsic and acquired resistance mechanisms. Virulence genes confer the ability to evade host immune response, the capacity of binding extracellular matrix, host cells, or inert materials and the capacity of forming biofilms (8). *E. faecalis* expresses a range of microbial surface components recognizing adhesive matrix molecules (MSCRAMMs) promoting cell adhesion and initiating the infection. Among MSCRAMMS, the most studied are the Ace adhesin, promoting collagen-binding and enhancing heart valve colonization (9,10); the *ebp* genes (endocarditis- and biofilm-associated pilus) encoding for a pilus implicated in initial adherence and biofilm formation, but also linked to the pathogenesis of endocarditis and UTIs (11,12); and the Enterococcal surface protein (Esp), which contributes to cell adhesion in urethral and abiotic surfaces colonization. The synthesis of secreted lytic proteins, such as the Cyl Cytolysin (13, 14), the GelE gelatinase (15), and the SerE Serine protease (16), induces host cell damage further promoting the infection. The Cyl protein has also been associated with bacteriocin activity, damaging other Gram-positive bacteria (14) and facilitating the *E. faecalis* predominance in polymicrobial environments. A quorum sensing system, encoded by the *fsr* operon, plays an important role in *E. faecalis* virulence, regulating the expression of surface proteins and biofilm formation (16). On the other hand, resistance to antimicrobial compounds further contributes to *E. faecalis* survivability in environments subject to selective pressures with antibiotics. With the only exception of ampicillin, *E. faecalis* is intrinsically resistant to virtually all cephalosporins, penicillins and carbapenems. Such resistance is related to the production of low affinity penicillin binding proteins (PBPs), which weakly bind to β-lactams (17). *E. faecalis* shares intrinsic low-level resistance to aminoglycosides: by limiting the drug intake through changes in outer membrane permeability or increasing excretion by efflux, low-level resistance is achieved. Notably, the combination of cell wall active agents such as penicillins or glycopeptides together with aminoglycosides results in bactericidal activity (bactericidal synergism) (17). The presence of chromosomally encoded ATP-binding cassette

(ABC) efflux pumps confer intrinsical resistance to lincosamides and streptogramins, pumping antibiotics out of the cell. *E. faecalis* is also intrinsically resistant to trimethoprim-sulfamethoxazole (18) drugs combination: Enterococci have the unusual capability of absorbing folate from the environment bypassing the folate synthesis pathway targeted by trimethoprim-sulfamethoxazole (18). Glycopeptides are usually effective against *E. faecalis*: however resistant phenotypes were characterized mediated by the *van* operons, but are usually rare (19). Although *E. faecalis* is susceptible to fluoroquinolones and linezolid, acquired resistant phenotypes have been reported. Acquired resistance occurs through sporadic mutation events or through the acquisition of new genetic material via horizontal gene transfer (20).

Horizontal gene transfer plays an important role in genomes evolution (21), allowing the acquisition of new resistance phenotypes, virulence factors, and bacteriocins from organisms in the same environments. Mobile genetic elements (MGEs) are regions of DNA able to move throughout the genome and are responsible for that exchange. The repertoire of mobile genes is defined mobilome, the mobile genome (22). The most frequent Enterococcal horizontal gene transfer occurs via conjugation. Moreover while phage-mediated transduction has been also reported, natural transformation has never been observed (23). The majority of MGEs in Enterococci are Conjugative Transposons (CTns), also known as Integrative and Conjugative Elements (ICEs) (1, 24), which contain genetic information to mediate their own transfer within and between cells and are also able to comobilize plasmids, transposons, and large chromosomal fragments. The first known CTn to carry antimicrobial resistance was identified in *E. faecalis* by Clewell and colleagues in 1981 (25, 26): the transposon Tn*916* harbours the *tet*(M) gene conferring tetracycline resistance. Other conjugative elements identified in Enterococci are mostly associated with Macrolide, Lincosamide, Streptogramin B, and Glycopeptide resistance (*vanB2* type) (27). The acquisition of genetic material in Enterococci occurs also through the transfer of conjugative plasmids (28). Pheromone-responsive plasmids (PRPs) are a major source of antimicrobial resistance spread and are capable of mobilizing large chromosomal regions via the formation of a plasmid-chromosome cointegrate (29, 30). Among PRPs, pCF10 and pAD1 plasmids are the most studied: pCF10 plasmids are mostly vehicles for antibiotic resistance (29), whereas pAD1 plasmids carry Cytolysin, Bacteriocins, Hemolysins, and UV light resistance (31). Non-pheromone-responsive plasmids (NRPRs) confer resistance to Macrolides, Aminoglycosides, and Glycopeptides. Because NPRPs and PRPs can coexist, recombination events forming hybrid plasmids represent a potential problem in the spread of multidrug resistance: the PRE25 plasmid, identified in a foodborne *E. faecalis* isolate, carries resistance to 12 antimicrobial compounds (32), while the Inc18-PRP hybrid plasmids are associated to dissemination of *vanA* resistance (33).

Chapter 3.2

Enterococcus faecalis in endodontic infections and the role of saliva in its transmission

Carlo Gaeta, David Pinzauti, Crystal Marruganti,
Andrea Fabbro, Gianni Pozzi, Simone Grandini,
and Francesco Santoro

Introduction

The role of bacteria in the initiation and progression of endodontic infections has been widely demonstrated (1). Bacterial invasion of the pulp chamber of the tooth can follow different pathways (*e.g.* cavities, dental cracks, or endo-periodontal lesions) (2).

The microorganisms involved in the primary infections of the pulp consist mainly in anaerobic bacteria with a small proportion of facultative anaerobes (2,3,4); indeed, the microflora associated with primary endodontic infections showed a wide inter-individual variability in terms of composition and dominant species (5). *Enterococcus faecalis* is an anaerobic Gram-positive coccus that normally inhabits the gastrointestinal tract and the vagina; it also constitutes one of the main causes of nosocomial infections worldwide (6,7). Several studies demonstrated that the prevalence of E. *faecalis* is higher in secondary rather than in primary infections (8) due to its ability to withstand prolonged periods of nutrient deficiency, thus persisting as a pathogen within the root canal (9). *E. faecalis* prevalence in persistent infections, as assessed by culture and molecular methods, can reach up to 80% (2,10). On the other hand, Next Generation Sequencing (NGS) molecular approaches recently shed light on the multi-species etiology of endodontic infections and highlighted a low detection rate of *E. faecalis* presence in cases of Post-treatment Apical Periodontitis (11). Moreover, a long-held assumption is that microorganisms found in the root canal space are partially derived from those colonizing the oral cavity (12). Although *E. faecalis* is not a normal inhabitant of the oral cavity, it can be transiently found in saliva (13) in relation to the patient’s periodontal status, oral hygiene habits or the consumption of specific foods (*e.g.,* cheese, vegetables) (14,15).

The purpose of the present study was to evaluate the association of *E. faecalis* with the different forms of pulpal and periapical infections, as well as the association between its presence in the canal and in saliva, using a combination of cultural and molecular approaches. We also aimed at evaluating the efficacy of the endodontic chemo-mechanical treatment in the eradication of *E. faecalis* in the root canal system.

Materials and methods

Patient population and data collection

Sixty patients that needed endodontic treatments were recruited from July to November 2020 at the Department of Endodontics, School of Dentistry, University of Siena. The teeth included in the study were classified in one of the following five clinical-radiographic conditions: healthy tooth (CVT); healthy treated tooth (HTT) tooth with irreversible pulpitis (IP); necrotic tooth (N); tooth with post-treatment apical periodontitis (R). Exclusion criteria are the follows: subjects with teeth with probing depth > 4 mm, with caries or restoration beyond the cemento-enamel junction that are not perfectly isolable by rubber dam and patients who had received antibiotic treatment within the preceding 3 months. The Human Research Ethics Committee of the Siena University Hospital approved a protocol (n. P1EF) describing the sample collection for this investigation, and all patients signed an informed consent form for their participation in this research. For each patient, data on age, gender, state of pulp, pain, sensitivity to percussion, painfulness, history of pain, swelling of periodontal tissues, mobility, presence of periodontal pockets, were collected. Pathological and pharmacological anamnestic data were also recorded. The state of the pulp was assessed by cold testing and classified the dental element into: healthy vital (CVT), with normal response cold; irreversible pulpitis (IP), with altered response to cold; pulp necrosis (N), with no response to cold. The periapical index (PAI) was used as a reference index to assess periapical bone alterations. Two expert endodontist observers evaluated the presence/absence of peri-apical lesions. Teeth with a PAI of 1 and 2 were considered without periapical lesion, while teeth with a PAI of 3, 4 and 5 were considered with periapical lesion. The data were inserted in Case Report Forms (CRF) and appropriately conserved, associating each patient to an enrollment number.

Sampling and clinical procedures

Root canal and saliva samples were collected as previously described (16). Before carrying out the isolation of the field with the rubber dam, saliva samples were taken from each patient on the oral floor, lingual body and on the crown of the affected tooth, using three sterile paper cones (ISO 40, 02). Cones were resuspended in 100 µl of PBS/10% glycerol and stored at -70°C until processing. Plaque around the affected tooth was gently removed using scalers and cleaned surfaces were brushed with pumice. The tooth was isolated by a rubber dam and eventual carious processes were removed. The tooth crown and the dam were then disinfected with 30% hydrogen peroxide and with 5.25 % sodium hypochlorite followed by sodium thiosulphate 5% to deactivate the action of antiseptics and prevent an altered sampling of the canal system. As a sterility control, three sterile paper cones (ISO 40, 02) were rubbed on the crown of the tooth and on the surrounding areas. The access cavity was performed using sterile cutters and root canal patency was established with minimal instrumentation, where possible, and without the use of chemically active irrigation. In case of retreatment, coronal guttapercha was removed with sterile Gates Glidden burs, while the apical one was removed with K-file, Hedström-file or both, avoiding the use of chemical solvents. Irrigation with sterile saline was performed to remove any remaining treatment materials before sample collection. Once reached the presumed working length based on intraoral X-ray and apex locator signaling, a pre-treatment sample was taken using 10 K-file in the canal and transferred in PBS/10% glycerol. An additional microbial sampling was performed by introducing two sterile paper cones (ISO 15, 02) on the entire canal length and keeping them in position for 60 seconds. When the canal was dry, an additional sterile paper cone moistened in sterile saline was used, to ensure the acquisition of the sample. In multi-rooted teeth, a single root canal was chosen, based on the evaluation of the presence of periapical radiolucency and/or exudation. The shaping and cleaning of the root canal system was then carried out until the Reciproc 25 reached the apex. In the case of apex greater than 25 the Reciproc 40/50 was used. After performing the final rinse (2 minutes with EDTA 17% followed by 5 minutes with 5.25% NaOCl) the canals were dried by using sterile paper cones. Then three sterile paper cones were brought to canal length, moistened with sterile saline, kept in position for 60 seconds and introduced into the transport medium. Filling of the root canal was finally performed together with coronal reconstruction.

Isolation and identification of *Enterococci*

Ten µl of PBS/10% glycerol from each sample were plated on the differential and selective medium Bile Esculin Azide Agar (BEA) (Remel, sodium azide 0.25 g/l; OX bile 10.0 g /l; esculin 1.0 g /l; pancreatic digest of casein 17.0 g/l; yeast extract 5.0 g/l; ferric ammonium citrate 0.5 g/l; sodium chloride 5.0 g/l; meat peptone 3.0 g/l, agar 15.0 g/l) and on Brain Heart Infusion (BHI) agar containing 5% horse blood, plates were incubated at 37° C in the presence of 5% CO_2 for 48 hours; plates were monitored daily for the presence of microbial growth and for the formation of a black halo due to hydrolyzation of esculin to glucose and esculetin, the latter reacts with iron ion and produces a black pigment. Black colonies were isolated on both BEA and BHI agar/blood and identified with a latex agglutination test (Oxoid™ Streptococcal Grouping Kit, Thermo Fisher). Group D colonies were then identified on a MALDI Biotyper (Bruker Daltonics) and by ribosomal RNA operon sequencing (17). Colonies identified as Enterococci were frozen at -70°C in BHI/10% glycerol.

High molecular weight DNA extraction

Strains were plated on BHI agar/blood, incubated overnight at 37°C and checked for purity. About ten colonies were inoculated in BHI broth and starter cultures of exponentially growing bacteria (OD590 of 0.3-0.4) were frozen at -70°C with glycerol at a final concentration of 10%. Bacteria were inoculated 1:50 (vol:vol) from starter cultures in 10 ml of BHI broth and incubated at 37° C until an OD590 of 1.0 was reached. Samples were then centrifuged at 6600 x *g* for 5 minutes. Bacterial pellets were washed with 10 ml sterile 1X TE buffer (Tris 10 mM-EDTA 1 mM) and resuspended in 7.5 ml of Raffinose buffer (50 mM Tris pH 8, 5 mM EDTA, 20 % Raffinose). Lysozyme (Sigma-Aldrich) at a final concentration of 2.6 mg/ml was added, incubating 1 hour at 37° C, to induce protoplasts formation. The solution was then centrifuged at 6600 x *g* for 5 minutes and the supernatant was discarded. The pellet was resuspended in 8 ml of distilled water (dH2O), Proteinase K (Sigma-Aldrich) at a final concentration of 0.1 mg/ml and 10% SDS (400 µl) were added to induce protoplast lysis. The reaction was incubated for 30 minutes at 37° C. Following, 1 ml of 5M NaCl was added, incubating 10 minutes at room temperature. Extracted DNA was purified from cellular contaminants by adding 1 volume of Sevag (Chloroform:Isoamyl alcohol, 24:1 vol:vol): the solution was centrifuged for 15 minutes at 6600 x *g* and, by using a Pasteur pipette, the supernatant was transferred into a clean tube and mixed again with an equal volume of Sevag. The supernatant from the second wash was precipitated by adding 0.6 volumes of ice-cold Isopropanol (Sigma-Aldrich) and incubated 30 minutes at -20° C. Samples were centrifuged (6600 x *g*, 15 minutes) and the DNA pellet was resuspended in 100 µl of saline. Genomic DNA was quantified using a Qubit 2.0 fluorometer (Invitrogen) and a NanoPhotometer device (Implen).

Sequencing and bioinformatic analysis

Whole genome sequencing (WGS) was performed employing Oxford Nanopore technology. Following manufacturers' instruction, the sequencing library was prepared using ligation sequencing kit (SQK-LSK108) and barcode expansion kits (EXP-NBD104/114), for sample multiplexing. The sequencing run was performed on the GridION x5 platform (Oxford Nanopore Technologies). Samples were also sequenced with Illumina technology at MicrobesNG (Birmingham, UK) (https://microbesng.com/) which performed library preparation and sequencing of paired end 250 bp reads on a HiSeq2500. Genomes were *de novo* assembled using Unicycler (v 0.4.7) (18), with both Nanopore and Illumina reads as an input. Phylogenetic relationships among sequenced genomes were explored using PopPUNK

(v. 1.1.5) (19). Virulence factors were determined using the tool ABRicate (v. 1.0.1) (https://github.com/tseemann/abricate): genomic sequences were compared against the VFDB (20) database, a comprehensive virulence factors database for bacterial pathogens. Both PopPUNK and Abricate were run using default parameters.

Statistical analysis

We included 66 patients and 79 samples in the present protocol. Given *E. faecalis* detection rates in primary and secondary endodontic infections of 2% and 71% respectively and setting alpha= 0.05, by including 66 patients we obtained a power of 99% and an actual alpha of 0.0270. Data were analyzed with SPSS 17.0 (SPSS Inc., Chicago, IL, USA). In the bivariate analysis, Fisher's exact test was used to assess the statistical significance of the association between the presence of *E. faecalis* in root canals and endodontic diseases, saliva, PAI score, type of coronal restoration (direct or indirect), quality of restoration (satisfactory or unsatisfactory). McNemar test was used to test the efficacy of the chemomechanical instrumentation in the eradication of *E. faecalis*. Logistic regression analysis was conducted to assess the effects of the above independent variables in a multivariate model in which the presence of *E. faecalis* in root canals (1=yes; 2=no) was the independent variable. The cutoff point for statistical significance was set at 0.05.

Results

Patient characteristics

A total of 66 patients, mean age 56 ± 13, 36 men and 30 females, that needed endodontic treatment, were enrolled in this study as shown in Table 1. Seventy-nine samples were taken and eleven were excluded at various stages because of sampling or laboratory errors. Of the 68 teeth examined, 7 samples were in the CVT group (9.23%), 8 in the HTT group (12.3%), 13 in the IP group (20%), 18 in the N group (27.69%) and 22 in the R group (30.77%), two were null due to rubber dam contamination. In 21 teeth a radiologically visible lesion (PAI> 2) was present, of those, 9 were in the R group (43%), 9 in the N group (43%) and 3 in the IP group (14%). The overall prevalence of *E. faecalis* in saliva was not significantly different among the five groups and ranged from 16.7% (1 out of 6) in the CVT group to 44.5% (8 out of 18) in the N group. *E. faecalis* was more commonly found in the root canal of patients in the N (36.8%, 7/19 samples), R (33.3%, 7/21 samples) and HTT (33.3% 3/9 samples) groups, even if the difference with the CVT and IP groups was not statistically significant. *E. faecalis* was found in root canal after the endodontic treatment in only 3 patients, in which it was also present in the saliva and in the root canal before treatment.

The presence of *E. faecalis* in saliva and root canals is associated with apical lesions

We used Fisher exact test in bivariate analyses to explore whether the presence of *E. faecalis* in the different sampling sites was associated with clinical parameters (Table 2). A significant positive association ($P<0.05$) was found between the presence of *E. faecalis* in both saliva and root canals and the presence of a radiographic lesion (chi-squared 4.357 and 6.129, respectively). A higher PAI score was also associated with the presence of *E. faecalis* in root canals ($P<0.05$; chi-squared 8.097). A previous indirect restoration was significantly associated to the presence of *E. faecalis* in root canals both before ($P<0.05$; chi-squared 5.756) and after ($P<0.001$; chi-squared 12.138) endodontic treatment. The chemo-mechanical instrumentation was able to remove *E. faecalis* in root canals ($P<0.05$). Finally, the presence of *E. faecalis* in saliva was associated with its presence in the root canal ($P<0.001$; chi-squared 25.867). When performing logistic regression analysis to model the factors influencing the presence of *E. faecalis* in root canals, we found that the main determinant was the presence of *E. faecalis* in saliva (Table 3; odds ratio 34, 95% confidence interval 5.4–214.7; $P<0.0001$) followed by the presence of a radiologically visible lesion (odds ratio 5.7, 95% confidence interval 1–33.5; $P=0.049$).

Phylogenetic relationships of *E. faecalis* isolates

Sixteen strains isolated in the present study were completely sequenced by nanopore sequencing. The assembled genomic sequences were analyzed with the PopPUNK tool (19) to identify the phylogenetic relationships between the different isolates. Figure 2 shows the identified phylogenetic relationships: isolates obtained from saliva and before endodontic treatment in the same patient are extremely correlated with each other. Of the two post-treatment isolates, one is related to the saliva isolates and pre-treatment of the same subject, while the other seems less related. It is worth to note that two different isolates from the saliva of the same subject are also related to each other, indicating that *E. faecalis* can persist in the mouth.

Virulence factors of salivary and endodontic *E. faecalis* strains

We then analyzed the presence of virulence factors in our 16 isolates. Virulence factors were grouped in: i) adherence factors, promoting cell adhesion and colonization; ii) genes involved in biofilm formation; iii) exoenzymes, capable of degrading various substrates; iv) a quorum sensing system, regulating the expression of several virulence factors and triggering biofilm formation; v) toxins, and vi) capsular polysaccharide. Virulence factors are summarized in Table 4, while genes involved in the capsule biosynthesis are reported in Table 5. Among seven putative adhesion factors only two, *ebp* genes (21) and *efaA* (22) gene, were identified in all the 16 isolates. The collagen adhesin gene *ace* (23) and the *esp* gene (24, 25), coding for the Enterococcal surface protein Esp, were present in seven isolates. The *asa1* gene (25), which encodes the Aggregation substance Asa1, was identified only in one isolate, while two harbour the EF0485 gene (26). The three adhesion genes (26) *fss1*, *fss2*, and *fss3*, which code for hypothetical fibrinogen binding proteins, were present in 14, 4 and 6 isolates, respectively. The capability of *E. faecalis* to form biofilm structure is well known. Endocarditis and biofilm-associated pilus (*ebp*) (21) and the *srtC* sortase (27) necessary for its extracellular localization, are associated to the production of biofilm and were present in all isolates; the enterococcal surface protein gene *esp* (24,27), also associated to biofilm production, was present in 7 isolates. Furthermore, all the 16 isolates harboured the *bopD* sugar-binding transcriptional regulator (28) putatively associated with enhanced biofilm formation. *E. faecalis* is also capable of producing several exoenzymes, which have the capability of degrading a broad spectrum of substrates. The presence of hyaluronidase enzymes (EF0818 and EF3023 genes) (9) was confirmed in all the 16 isolates. A bile acid hydrolase, encoded by the *cbh* gene (29), was identified in 10 out of 16 samples; 6 samples harbour a glycosyl hydrolase (29) homologous to *xylS*, 9 out of 16 samples harbour the *nuc-1* gene (30), responsible for the production of a nuclease protein, which exhibits sequence similarity to staphylococcal nuclease that hydrolyze nucleic acids. Eight isolates harbour the *gelE* gene (25,31,32,33) responsible for the production of Gelatinase exoenzyme a well studied enzyme responsible for the lysis of gelatin, collagen, hemoglobin, and other substrates promoting tissue damage. Eight out of 16 isolates harbour the *sprE* gene (34,35) encoding for a Serine protease found to promote adhesion to dentin.

Two isolates harbour the Cyl operon (25,34,35), which codes for a toxin (Cytolysin) capable of lysing erythrocytes, leukocytes, macrophages, and Gram-positive bacteria causing dentinal and periapical tissue damage furthermore facilitating the *E. faecalis* instauration in dental infections compared to other pathogens. Four out of 16 isolates harbour an intact *fsrABC* operon (32,33,36), an important quorum sensing system in *E. faecalis* which regulates the expression of biofilm formation (*bopD* gene) along with other virulence factors such as *gelE* and *sprE*. Three isolates only harboured the *fsrC* gene coding for a sensor histidine kinase. A metal binding protein encoded by the *psaA* gene (29) was also identified in 6 samples. The *cps* operon (37,38,39) consists of 11 genes, but only 7 are essential for the capsule production: *cpsC*, *cpsD*, *cpsE*, *cpsG*, *cpsI*, *cpsJ*, and *cpsK*. Three capsule types can be defined: CPS1 type includes *cpsA* and *cpsB* genes only, CPS2 type includes all the 11 genes, while CPS5 type consists of all the genes except *cpsF*. Only CPS2 and CPS5 types are capable of expressing a capsular polysaccharide. Among the 16 analyzed *E. faecalis* isolates, 12 out of 16 samples were CPS1, two were CPS2 type and two CPS5.

Discussion

The aim of the present study was to evaluate the presence of *E. faecalis* in root canal and saliva samples obtained in patients with different pulpal and peri-apical conditions using a cultural approach. We found *E. faecalis* both in irreversible pulpitis and vital tooth with no symptoms and although our statistical model shows significant association between these conditions ($P<0.05$) the risk to find these bacteria in the healthy and initial pulpal infection is relatively rare (Odds ratio<0.05). We used as clinical control a vital tooth that needed endodontic treatment for prosthetic reasons. *E. faecalis* was detected in this case only one time possibly due to the field contamination during clinical procedures being saliva positive to culture. Although the actual beliefs in endodontics exclude the possibility of finding bacteria in healthy pulps, results of some studies suggest that bacteria could have alternative ways to contaminate this space (40) resulting in line with our data. We found *E. faecalis* in high percentage also in case of irreversible pulpitis (20%) confirming its particular attitude to invade root vascular system (41), despite not usually isolated from deep carious lesions and canal with this clinical condition (42). This data could be explained by the fact that in the current literature used primary and secondary endodontic infections as diagnostic classification not distinguishing between irreversible pulpitis and necrosis making difficult the microbiota differentiation in first and late stage of infection. The prevalence of *E. faecalis* tends to increase with a statistics significance ($P<0.05$) in the case of pulp necrosis (50% of cases) demonstrating its bent to invade the pulp chamber in the late phase of infection, although previous findings showed a detection rate ranged from 2 to 18% of cases (43,44). Gomes et al. (16) demonstrated high prevalence of *E. faecalis* in necrotic tooth (82%) showing a very high variability in the detection rate of this microorganism. This phenomenon could be due to different detection methods, geographic and numerosity differences in population samples. Our study confirms the ability of *E. faecalis* to colonize previously treated teeth (30.77%) although seems not to be associated with the development of periapical lesions and PAI score as shown in previous studies (45). However, once a time this data demonstrates the capacity of this bacteria to survive without nutrients in hostile surrounding for a long time. In present study *E. faecalis* was detected in 32 saliva samples (49%) in contrast with the percentage resulting from recent studies that ranged from 10 to 35% (30). Recently Wang et al (46,47) showed high prevalence in saliva samples although this microorganism is not a normal commensal of oral microbiome but its presence in various foods could influence the permanence in the mouth. Interestingly we found a strong association with *E. faecalis* in saliva and in root canal indifferently from a clinical condition as statistically confirmed by Fisher's test (p=0,000). We also found a strong association between the presence of *E. faecalis* in saliva and in pretreatment sample as confirmed by logistic regression test confirming the studies of Kaufman and Wang. In our study, we used culture on a selective medium since we wanted to investigate the role of *E. faecalis* as a pathogen, this also allowed the molecular characterization of the isolated strains and guaranteed the vitality of the bacteria in the infected root canal. Many recent studies, based on Next Generation Sequencing (NGS), have demonstrated that multiple microbial communities can be associated with endodontic infection, somehow lessening the role of *E. faecalis* in the etiology of persistent root canal infections. It is also worth to note that molecular methods are unable to differentiate between dead or alive bacteria inside root canals, considering that DNA can persist up to one year (48), moreover DNA extraction protocols may be insufficient for the lysis of gram positive cell wall, resulting in underestimation of some genera. The sequenced isolates obtained from different samples of the same patient showed a close phylogenetic relationship, possibly indicating that *E. faecalis* can infect the root canals from saliva. Although the source of *E. faecalis* in the oral cavity is not clear, we found that the same strain of *E. faecalis* is able to persist in the saliva of an individual for at least four months, thus indicating that the colonization is not transient. In conclusion, here we demonstrate a role of *E. faecalis* as a pathogen causing periapical lesions and we identify salivary carriage as a risk factor for developing such infections.

Tables

Table 1. Descriptive statistics of patients' characteristics

Variable		**Mean±SD/Proportion**	**95% Confidence Interval**	
			Lower	**Upper**
Age		56.32±13.93	53.20	59.44
Gender (*females*)		45.56%	35.04	56.50
Groups				
	N	26.67%	15.56	41.77
	CVT	11.11%	4.58	24.53
	IP	17.77%	8.97	32.17
	R	33.33%	20.90	48.60
	HTT	11.11%	4.58	24.53
Lesion		29.11%	19.43	40.41
PAI score				
	1	60.52%	48.99	70.99
	2	9.21%	4.40	18.27
	3	25%	16.43	36.10
	4	3.94%	1.25	11.72
	5	1.31%	0.18	9.01
Position *(posterior)*		63.63%	52.48	73.49
Improper restorations		90.78%	82.18	95.46
Restoration type (*indirect*)		8.86%	4.35	17.12
Saliva +		32.43%	22.86	43.73
Canal pre-treatment +		23.37%	15.33	33.95
Canal post-treatment +		3.94%	1.35	10.97
Rubber dam +		2.56%	0.70	8.87

Abbreviations: SD, standard deviation; N, necrotic tooth; CVT, healthy vital tooth; IP, irreversible pulpitis; R, post-treatment apical periodontitis; HTT, healthy treated tooth; PAI score, periapical index score. Saliva +, proportion of saliva samples positive to *E. faecalis*; Canal pre-treatment, proportion of samples in the canal before treatment positive to *E. faecalis*; Canal post-treatment +, proportion of samples in the canal after treatment positive to *E. faecalis*; Rubber dam +, proportion of rubber dam samples positive to *E. faecalis*.

Table 2. Association between clinical and microbiological variables.

Variable/ *E. faecalis* in sample	**Saliva (yes/no)**	X^2	**PreCanal (yes/no)**	X^2	**PostCanal (yes/no)**	X^2
Group						
N	6/10	2.199	6/11	6.640	1/16	1.146
CVT	1/6		0/8		0/7	
IP	4/9		1/12		1/12	
R	7/14		7/14		1/20	
HTT	1/6		2/6		0/9	
Lesion						
present	18/26	4.537*	9/13	6.129 *	30/42	2.083
absent	5/24		8/46		0/3	
PAI score						
1	11/33	3.141	6/39	8.097 *	1/44	7.504
2	3/4		1/6		0/7	
3	8/10		8/10		1/18	
4	1/2		1/2		1/2	
5	0/1		0/1		0/1	
Proper restoration						
Yes	1/6	1.216	0/7	2.080	0/7	0.337
No	22/41		15/49		3/62	
Type of restoration						
Direct	19/48	3.744	12/55	5.756 *	1/67	12.138**
Indirect	4/2		4/3		2/5	
Position						
Anterior	12/13	4.290	8/17	2.118	1/26	0.017
Posterior	11/35		8/39		2/44	
Canal pre-treatment						
present	12/2	25.867**	/	/	3/13	13.00**
absent	9/47		/		0/55	
Canal post-treatment						
present	3/0	6.537*	/	/	/	/
absent	20/48		/		/	

The first column presents different categorical variables, the following columns detail the presence/absence of *E. faecalis* in each of the samples (saliva, root canal before treatment, root canal after treatment) among teeth belonging to the different categories. *$p<0.05$, **$p<0.001$

Table 3. Best model of the logistic multivariate regression analysis assessing the impact of the presence of *E. faecalis* in the saliva (Saliva) on its presence in the canal before the endodontic treatment (Canal pre-treatment).

Best model (AIC=47.2; AUC=0.889)

LR chi2	***p*-value**	**pseudo R^2**
28.02	0.000**	0.408

					95% CI	
Canal pre-treatment	**OR**	**SE**	**z**	***p*-value**	**Lower**	**Upper**
Saliva	34.085	32.011	3.76	0.000**	5.409	214.775
Lesion	5.692	5.146	1.92	0.049*	1.067	33.482
Improper	4.995	7.726	1.04	0.298	0.240	103.56
cons	0.004	0.009	-2.75	0.006	0.001	0.214

Abbreviations: AIC, Aikaike information criterion; AUC, area under the curve; LR chi2, likelihood ratio chi-squared test; OR, odds ratio; SE, standard error; CI, confidence interval. *$p<0.05$, **$p<0.001$

Table 4. Virulence factors of *Enterococcus faecalis* isolates.

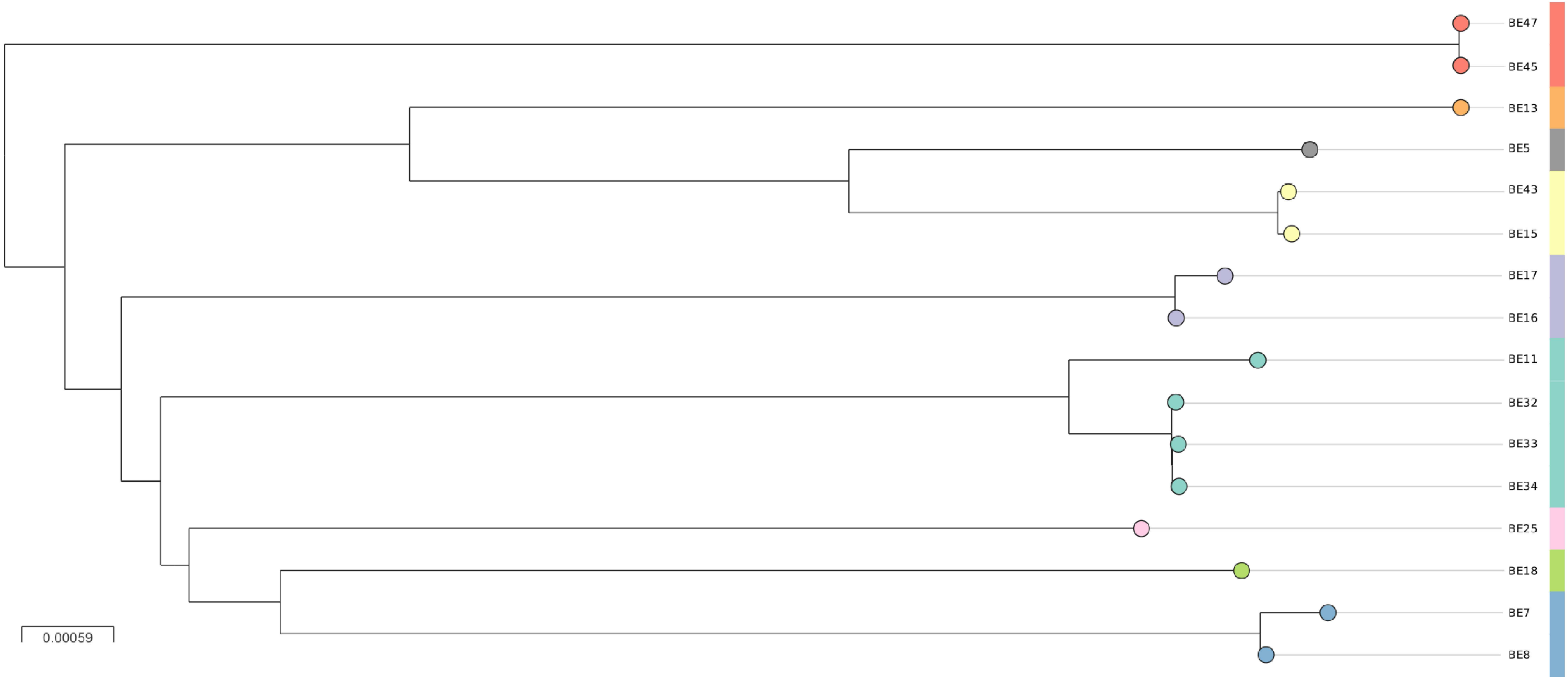
Virulence factors	Putative function	BE5	BE7	BE8	BE11	BE13	BE15	BE16	BE17	BE18	BE25	BE32	BE33	BE34	BE43	BE45	BE47
ace	Collagen adhesin protein				x						x	x	x	x		x	x
asa1	Aggregation substance Asa1	x															
bopD	Sugar-binding transcriptional regulator (biofilm formation)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
cbh	Bile salt hydrolase	x	x	x	x	x	x					x	x	x	x		
Cyl Operon	Cytolysin production		x	x													
ebpABC + *srtC*	Endocarditis and biofilm-associated pilus + *srtC* sortase	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
EF0485	Aggregation substance		x	x													
EF0818	Polysaccharide lyase family 8 (hyaluronidase)				x	x	x	x	x	x	x	x	x	x	x	x	x
EF3023		x	x	x	x	x		x	x	x	x	x	x	x			

efaA	Endocarditis specific antigen	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
esp	Enterococcal surface protein	x	x	x	x							x	x	x			
fsrA	Quorum sensing system					x		x	x	x							
fsrB						x		x	x	x							
fsrC		x				x	x	x	x	x					x		
fss1	*E. faecalis* surface protein fibrinogen binding protein	x	x	x	x	x		x	x	x	x	x	x	x		x	x
fss2			x	x		x				x							
fss3		x			x		x								x	x	x
gelE	Gelatinase	x				x	x	x	x	x	x				x		
gls24-like	General stress response protein	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
hydrolase	Glycosyl hydrolase (*xylS* homolog)		x	x	x							x	x	x			
nuc-1	Nuclease (homolog)		x	x	x			x	x	x		x	x	x			
psaA	Metal binding protein		x	x	x							x	x	x			
sprE	Serine protease V8 family	x				x	x	x	x	x	x				x		

Table 5. Presence of capsule genes and predicted capsular type.

Capsule genes	BE5	BE7	BE8	BE11	BE13	BE15	BE16	BE17	BE18	BE25	BE32	BE33	BE34	BE43	BE45	BE47
cpsA	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
cpsB	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
cpsC		x	x		x				x							
cpsD		x	x		x				x							
cpsE		x	x		x				x							
cpsF		x	x													
cpsG		x	x		x				x							
cpsH		x	x		x				x							
cpsI		x	x		x				x							
cpsJ		x	x		x				x							
cpsK		x	x		x				x							
Capsule type	**1**	**2**	**2**	**1**	**5**	**1**	**1**	**1**	**5**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

Figure 1. Phylogenetic relationships among *E. faecalis* isolates. Different colours indicate different genetic clusters. The sample number and the sampling site are indicated on the right. S: salivary isolate, PRE: pre-treatment isolate, POST: post-treatment isolate.



References

1. Korzen BH, Krakow AA, Green DB. Pulpal and periapical tissue responses in conventional and monoinfected gnobiotic rats. Oral Surg Oral Med Oral Pathol. 1974 May; 37(5):783-802.

2. Siqueira JF Jr, Rôças IN. Diversity of endodontic microbiota revisited. J Dent Res. 2009 Nov;88(11):969-81. doi: 10.1177/0022034509346549.

3. Fabricius L, Dahlén G, Ohman AE, Möller AJ. Predominant indigenous oral bacteria isolated from infected root canals after varied times of closure. Scand J Dent Res. 1982 Apr;90(2):134-44. doi: 10.1111/j.1600-0722.1982.tb01536.x.

4. Munson MA, Pitt-Ford T, Chong B, Weightman A, Wade WG. Molecular and cultural analysis of the microflora associated with endodontic infections. J Dent Res. 2002 Nov;81(11):761-6. doi: 10.1177/0810761.

5. Sakamoto, M., Rôças, I. N., Siqueira, J. F., & Benno, Y. (2006). Molecular analysis of bacteria in asymptomatic and symptomatic endodontic infections. Oral Microbiology and Immunology, 21(2), 112–122. doi: 10.1111/j.1399-302X.2006.00270.x

6. Murray BE. The life and times of the Enterococcus. Clin Microbiol Rev. 1990 Jan;3(1):46-65. doi: 10.1128/cmr.3.1.46.

7. Hammerum AM. Enterococci of animal origin and their significance for public health. Clin Microbiol Infect. 2012 Jul;18(7):619-25. doi: 10.1111/j.1469-0691.2012.03829.x.

8. Rôças IN, Siqueira JF Jr, Santos KR. Association of Enterococcus faecalis with different forms of periradicular diseases. J Endod. 2004 May;30(5):315-20. doi: 10.1097/00004770-200405000-00004.

9. Kayaoglu G, Ørstavik D. Virulence factors of Enterococcus faecalis: relationship to endodontic disease. Crit Rev Oral Biol Med. 2004 Sep 1;15(5):308-20. doi: 10.1177/154411130401500506.

10. Gomes BP, Pinheiro ET, Sousa EL, Jacinto RC, Zaia AA, Ferraz CC, de Souza-Filho FJ. Enterococcus faecalis in dental root canals detected by culture and by polymerase chain reaction analysis. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2006 Aug;102(2):247-53. doi: 10.1016/j.tripleo.2005.11.031.

11. Siqueira JF Jr, Antunes HS, Rôças IN, Rachid CT, Alves FR. Microbiome in the Apical Root Canal System of Teeth with Post-Treatment Apical Periodontitis. PLoS One. 2016 Sep

30;11(9):e0162887. doi: 10.1371/journal.pone.0162887.

12. Sedgley C, Nagel A, Dahlén G, Reit C, Molander A. Real-time quantitative polymerase chain reaction and culture analyses of Enterococcus faecalis in root canals. J Endod. 2006 Mar;32(3):173-7. doi: 10.1016/j.joen.2005.10.037.

13. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. Defining the normal bacterial flora of the oral cavity. J Clin Microbiol. 2005 Nov;43(11):5721-32. doi: 10.1128/JCM.43.11.5721-5732.2005.

14. Sedgley CM, Lennan SL, Clewell DB. Prevalence, phenotype and genotype of oral enterococci. Oral Microbiol Immunol. 2004 Apr;19(2):95-101. doi: 10.1111/j.0902-0055.2004.00122.x.

15. Razavi A, Gmür R, Imfeld T, Zehnder M. Recovery of Enterococcus faecalis from cheese in the oral cavity of healthy subjects. Oral Microbiol Immunol. 2007 Aug;22(4):248-51. doi: 10.1111/j.1399-302X.2006.00349.x.

16. Gomes BP, Endo MS, Martinho FC. Comparison of endotoxin levels found in primary and secondary endodontic infections. J Endod 2012; 38:1082–6.

17. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon. F1000Res. 2018 Nov 6;7:1755. doi: 10.12688/f1000research.16817.2.

18. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017 Jun 8;13(6):e1005595. doi: 10.1371/journal.pcbi.1005595.

19. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019 Feb;29(2):304-316. doi: 10.1101/gr.241455.118.

20. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. Nucleic Acids Res. 2016 Jan 4;44(D1):D694-7. doi: 10.1093/nar/gkv1239.

21. La Rosa SL, Montealegre MC, Singh KV, Murray BE. Enterococcus faecalis Ebp pili are important for cell-cell aggregation and intraspecies gene transfer. Microbiology (Reading). 2016 May;162(5):798-802. doi: 10.1099/mic.0.000276.

22. Preethee T, Kandaswamy D, Hannah R. Molecular identification of an Enterococcus faecalis endocarditis antigen efaA in root canals of therapy-resistant endodontic infections. J Conserv Dent. 2012 Oct;15(4):319-22. doi: 10.4103/0972-0707.101886.

23. Hubble TS, Hatton JF, Nallapareddy SR, Murray BE, Gillespie MJ. Influence of Enterococcus faecalis proteases and the collagen-binding protein, Ace, on adhesion to dentin. Oral Microbiol Immunol. 2003 Apr;18(2):121-6. doi: 10.1034/j.1399-302x.2003.00059.x.

24. Toledo-Arana A, Valle J, Solano C, Arrizubieta MJ, Cucarella C, Lamata M, Amorena B, Leiva J, Penadés JR, Lasa I. The enterococcal surface protein, Esp, is involved in Enterococcus faecalis biofilm formation. Appl Environ Microbiol. 2001 Oct;67(10):4538-45. doi: 10.1128/aem.67.10.4538-4545.2001. PMID: 11571153; PMCID: PMC93200.

25. Kiruthiga A, Padmavathy K, Shabana P, Naveenkumar V, Gnanadesikan S, Malaiyan J. Improved detection of esp, hyl, asa1, gelE, cylA virulence genes among clinical isolates of Enterococci. BMC Res Notes. 2020 Mar 20;13(1):170. doi: 10.1186/s13104-020-05018-0.

26. Jamet A, Dervyn R, Lapaque N, Bugli F, Perez-Cortez NG, Blottière HM, Twizere JC, Sanguinetti M, Posteraro B, Serror P, Maguin E. The Enterococcus faecalis virulence factor ElrA interacts with the human Four-and-a-Half LIM Domains Protein 2. Sci Rep. 2017 Jul 4;7(1):4581. doi: 10.1038/s41598-017-04875-3.

27. Nallapareddy SR, Singh KV, Sillanpää J, Garsin DA, Höök M, Erlandsen SL, Murray BE. Endocarditis and biofilm-associated pili of Enterococcus faecalis. J Clin Invest. 2006 Oct;116(10):2799-807. doi: 10.1172/JCI29021.

28. Hufnagel M, Koch S, Creti R, Baldassarri L, Huebner J. A Putative Sugar-Binding Transcriptional Regulator in a Novel Gene Locus in Enterococcus faecalis Contributes to Production of Biofilm and Prolonged Bacteremia in Mice. J Infect Dis. 2004;189:420–430. doi: 10.1086/381150.

29. McBride SM, Fischetti VA, Leblanc DJ, Moellering RC Jr, Gilmore MS. Genetic diversity among Enterococcus faecalis. PLoS One. 2007 Jul 4;2(7):e582. doi: 10.1371/journal.pone.0000582. PMID: 17611618; PMCID: PMC1899230.

30. Hu Y, Meng J, Shi C, Hervin K, Fratamico PM, Shi X. Characterization and comparative analysis of a second thermonuclease from Staphylococcus aureus. Microbiol Res. 2013 Mar 30;168(3):174-82. doi: 10.1016/j.micres.2012.09.003.

31. Su YA, Sulavik MC, He P, Makinen KK, Makinen PL, Fiedler S, Wirth R, Clewell DB. Nucleotide sequence of the gelatinase gene (gelE) from Enterococcus faecalis subsp liquefaciens. Infect Immun. 1991;59:415–420. doi: 10.1128/IAI.59.1.415-420.1991.

32. Qin X, Singh KV, Weinstock GM, Murray BE. Effects of Enterococcus faecalis fsr genes on production of gelatinase and a serine protease and virulence. Infect Immun. 2000 May;68(5):2579-86. doi: 10.1128/iai.68.5.2579-2586.2000.

33. Engelbert M, Mylonakis E, Ausubel FM, Calderwood SB, Gilmore MS. Contribution of gelatinase, serine protease, and fsr to the pathogenesis of Enterococcus faecalis endophthalmitis. Infect Immun. 2004 Jun;72(6):3628-33. doi: 10.1128/IAI.72.6.3628-3633.2004.

34. Booth MC, Bogie CP, Sahl HG, Siezen RJ, Hatter KL, Gilmore MS. Structural analysis and proteolytic activation of Enterococcus faecalis cytolysin, a novel lantibiotic. Mol Microbiol. 1996 Sep;21(6):1175-84. doi: 10.1046/j.1365-2958.1996.831449.x.

35. Shankar N, Coburn P, Pillar C, Haas W, Gilmore M. Enterococcal cytolysin: activities and association with other virulence traits in a pathogenicity island. Int J Med Microbiol. 2004 Apr;293(7-8):609-18. doi: 10.1078/1438-4221-00301.

36. Hashem, Y.A., Amin, H.M., Essam, T.M. et al. Biofilm formation in enterococci: genotype-phenotype correlations and inhibition by vancomycin. Sci Rep 7, 5733 (2017). https://doi.org/10.1038/s41598-017-05901-0

37. Hancock LE, Gilmore MS. The capsular polysaccharide of Enterococcus faecalis and its relationship to other polysaccharides in the cell wall. Proc Natl Acad Sci U S A. 2002 Feb 5;99(3):1574-9. doi: 10.1073/pnas.032448299

38. Thurlow LR, Thomas VC, Hancock LE. Capsular polysaccharide production in Enterococcus faecalis and contribution of CpsF to capsule serospecificity. J Bacteriol. 2009 Oct;191(20):6203-10. doi: 10.1128/JB.00592-09

39. Thurlow LR, Thomas VC, Fleming SD, Hancock LE. Enterococcus faecalis capsular polysaccharide serotypes C and D and their contributions to host innate immune evasion. Infect Immun. 2009 Dec;77(12):5551-7. doi: 10.1128/IAI.00576-09

40. Widmer C, Skutas J, Easson C, Lopez JV, Torneck C, Flax M, Sayin TC. Culture-independent Characterization of the Microbiome of Healthy Pulp. J Endod. 2018 Jul;44(7):1132-1139.e2. doi: 10.1016/j.joen.2018.03.009.

41. Nishio Ayre W, Melling G, Cuveillier C, Natarajan M, Roberts JL, Marsh LL, Lynch CD, Maillard JY, Denyer SP, Sloan AJ. Enterococcus faecalis Demonstrates Pathogenicity through Increased Attachment in an Ex Vivo Polymicrobial Pulpal Infection. Infect Immun. 2018 Apr 23;86(5):e00871-17. doi: 10.1128/IAI.00871-17.

42. Rôças IN, Siqueira JF Jr. Characterization of microbiota of root canal-treated teeth with posttreatment disease. J Clin Microbiol. 2012 May;50(5):1721-4. doi: 10.1128/JCM.00531-12.

43. Łysakowska ME, Ciebiada-Adamiec A, Sienkiewicz M, Sokołowski J, Banaszek K. The cultivable microbiota of primary and secondary infected root canals, their susceptibility to

antibiotics and association with the signs and symptoms of infection. Int Endod J. 2016 May;49(5):422-30. doi: 10.1111/iej.12469.

44. Mahmoudpour A, Rahimi S, Sina M, Soroush MH, Shahi S, Asl-Aminabadi N. Isolation and identification of Enterococcus faecalis from necrotic root canals using multiplex PCR. J Oral Sci. 2007 Sep;49(3):221-7. doi: 10.2334/josnusd.49.221.

45. Kaufman B, Spångberg L, Barry J, Fouad AF. Enterococcus spp. in endodontically treated teeth with and without periradicular lesions. J Endod. 2005 Dec;31(12):851-6. doi: 10.1097/01.don.0000164133.04548.26.

46. Wang QQ, Zhang CF, Chu CH, Zhu XF. Prevalence of Enterococcus faecalis in saliva and filled root canals of teeth associated with apical periodontitis. Int J Oral Sci. 2012 Mar;4(1):19-23. doi: 10.1038/ijos.2012.17.

47. Gold OG, Jordan HV, van Houte J. The prevalence of enterococci in the human mouth and their pathogenicity in animal models. Arch Oral Biol. 1975 Jul;20(7):473-7. doi: 10.1016/0003-9969(75)90236-8.

48. Brundin M, Figdor D, Johansson A, Sjögren U. Preservation of bacterial DNA by human dentin. J Endod. 2014 Feb;40(2):241-5. doi: 10.1016/j.joen.2013.08.025.

3.3 Comparative genomics of *Enterococcus faecalis* strains isolated from patients with apical lesions

Endodontic treatment failures are mainly associated with inadequate treatment procedures, which are not capable of eradicating bacteria from the root canal system. Due to its innate resistance to antimicrobial and disinfection agents as well as the expression of virulence factors (i.e. biofilm formation), *Enterococcus faecalis* is the leading cause of treatment failures causing persistent inflammation and healing impairment. Accurate bacterial profiling approaches are essential to drive specific and effective treatment. The possibility of performing Whole Genome Sequencing achieving complete bacterial genomes, has enabled the study of genomes through comparative genomic analysis. Comparative genomics aim to understand the structure and the function of the genomic fragments by comparison among the different genomes and with well studied, reference genomes.

In the present study, the 16 *Enterococcus faecalis* isolates, previously sequenced and assembled (Chapter 3.2), were subjected to a comparative genomic analysis. The scope of the study was the definition of a MLST profile, identifying sequence types collected from patients, the characterization of antimicrobial resistance genes and of mobile genetic elements. The identification of synteny blocks have enabled a comparative analysis among the isolates, identifying unique or shared genomic regions as well as genomic rearrangements.

Materials and Methods

Sample collection and whole genome sequencing were described in Chapter 3.2. WGS yields are listed in Table 1: for each genome were measured the total length and the corresponding G+C content, evaluating the presence of plasmids. A rapid annotation was performed using the tool Prokka (v 1.14.5) (34) assessing the number of putative genes. A multilocus sequence typing (MLST) analysis was performed using the webtool MLST 2.0 (v. 2.0.4) available from the Center for Genomic Epidemiology (CGE) (https://www.genomicepidemiology.org/). Through the identification of seven housekeeping genes was possible to determine the genetic relatedness of the *E. faecalis* strains (35). Results are reported in Table 2. The presence of antimicrobial resistance genes (AMR) was determined using the tool Abricate (v 1.0.1) (https://github.com/tseemann/abricate); results are reported in Table 3. In short, the tool compares genomic sequences against a curated database. For such analysis both CARD (36) and ARGANNOT (37) databases were used. A comparative analysis was finally performed using the tool Sibelia (v. 3.0.7) (38). Sibelia enables the detection of small to large scale rearrangements and the identification of shared regions by decomposing input genomes into synteny blocks, blocks of non-overlapping sequences that exhibit conserved features across the input genomes. The pipeline relies on iterative de Bruijn graphs: starting from a fixed *k*-mer value, the pipeline is executed for a range of successive and increasing values of *k*-mer sizes, extending the analysis to the whole genome. At each iteration, a different set of blocks is generated and placed as a node into a tree structure, where the root of the tree corresponds to the whole genome. Using the tool Circos (v 0.69-8) (39), synteny blocks were represented into a visual hierarchical structure.

Results

The assembled *Enterococcus faecalis* genomes have a G+C content ranging from 37.35% to 37.68%. The average genome size was 2.87 ± 0.12 Mb, with 2,750 ± 179 predicted CDS. Plasmids were identified in 11 out of 16 isolates comprising from 1 up to 5. Plasmid content is concordant within genomic clusters: samples belonging to the same cluster share the same plasmids (Table 1). Nevertheless, few exceptions exist. A total of three and four plasmids were respectively reconstructed for the related BE16 and BE17 isolates. However only two of them, the 19.7 Kb and 5.1 Kb in length plasmids, were identified in both the isolates. BE16 harbours a 102 Kb in length plasmid, which shares 100% homology with the 56.2 Kb and 45.3 Kb plasmids identified in the BE17 isolate. The 102 Kb "fusion" plasmid harbours two replication initiator (*rep*) genes, the same encoded by the BE17 plasmids. In order to investigate whether this was a sequencing or assembly artifact, Nanopore sequencing reads were mapped to the plasmid sequence using minimap2 (v 2.17) (40) and visually inspected with Tablet (41). A total of 533 reads (mean length 18,922 bp) and 315 reads (mean length 21,852 bp) were respectively mapped for BE16 and BE17 isolates: BE16 reads were found to span continuously the entire length of the plasmid, while BE17 reads were found interrupted in the putative plasmids junction sites, suggesting that the reads belong to two distinct plasmid structures. The same result was achieved when mapping Illumina reads. These data suggest that the 56.2 Kb and 45.3 Kb plasmids can become fused in a larger 102 Kb element. Different plasmid content was also identified between BE7 and BE8 samples. A linear 30.4 Kb extrachromosomal DNA molecule was assembled in BE but not in the related BE8 strain. The Nanopore sequencing reads of isolate BE8 were then mapped to the sequence using minimap2 tool (v 2.17) (40): because none of the reads were successfully mapped, it was further confirmed that BE8 does not harbour the extrachromosomal DNA molecule. Similarly, the isolates BE32, BE33, and BE34 harbour a linear 29 Kb in length extrachromosomal molecule, which was not identified in BE11 isolate belonging to the same genomic cluster. Using the webtool PHASTER (42) both the elements were identified as a putative Enterococcal phage, sharing homologies with *Enterococcus* phage EF62phi (CP002495), a pseudotemperate phage which belongs to *Podoviridae* family (43). The phage is characterized by extrachromosomal replication via *repB* (*DnaD*) gene and harbours a toxin-antitoxin system which is thought to maintain the temperate status (43). A visual comparison is represented in Figure 1 using clinker tool (v 0.0.19) (44).

MLST results are extremely concordant with genomic clusters inferred by PopPUNK. Each genomic cluster was successfully assigned with a specific sequence type (ST). Eight different STs were identified (Table 2). However, BE15 and BE45 could not be typed by MLST profiling: the presence of a single nucleotide change in the *aroE* gene generated a new ST.

Antimicrobial resistance genes and their putative resistant phenotype are reported in Table 3, distinguishing between clinically relevant resistance (3-A) and not (3-B) according to EUCAST clinical breakpoints (https://www.eucast.org/). Results achieved employing both CARD and ARGANNOT databases are concordant. All isolates harbour *efrAB* and *lsaA* genes encoding for efflux pumps conferring multiple resistance to clinically relevant antibiotics (Table 3). None of the isolates encode Vancomycin resistance genes. Nine isolates harbour a Tn*916* conjugative transposon (45) carrying a *tetM* gene which confers Tetracycline resistance. A Tn*916*-like transposon, carrying *tetM* and *tetL* genes, which confer resistance to Tetracycline, was identified in BE5 isolate, inserted into a 62 kb plasmid. The isolates BE11, BE32, BE33, and BE34 harbor a mobile genetic element (Tn*6000*-like) (46) encoding for *tetS* gene which confers Tetracycline resistance. The BE13 isolate harbours a Tn*916* transposon which lacks the *tetM* gene: a putative Bacitracin ATP-binding cassette transporter is instead inserted between ORF13 and ORF9 and putatively conferring Bacitracin resistance (47). A schematic comparison of the identified Tn*916*-like mobile elements is represented in Figure 2. A ~30 Kb genetic element was identified in BE7 and BE8 isolates. This genetic element is flanked by two IS1216E-family transposases, it harbours multiple resistance genes and a Toxin-Antitoxin locus. Identified resistance genes are putatively linked to Aminoglycosides, Chloramphenicol, Erythromycin, Lincosamides, Macrolides, and Streptogramins resistance. A schematic representation of the genetic element is illustrated in Figure 3.

The study of microbial genomes architecture revealed the presence of conserved gene order and content among genomes under the study via syntenic blocks identification. Employing the *Enterococcus faecalis* OG1RF (NC_017316.1) as reference genome, comparative genomics reveals high genomic rearrangements and the presence of unique regions from one genome to another. Visual representation of genomic architecture was achieved using Circos tool (v 0.69-8) (39). Three genomic clusters at time were compared to the reference. This choice was purely aesthetic: in fact increasing the number of input genomes results in an increased number of synteny blocks, generating extremely piled images. Identified synteny blocks were labelled using two different colors depending on their orientation: the green label represents syntenic blocks identified in the positive strand, while the red ones in the negative. Unique regions were instead represented as an empty space (light green shadow) in the chromosome. Synteny blocks shared between one or more input chromosomes were linked together using ribbons (randomly coloured). While Circos provides graphical representation of genomic rearrangements, Sibelia generates an interactive html-based diagram showing synteny blocks coordinates and enabling an in-depth characterization of shared syntenic blocks or unshared features. The BE isolates genomes share an average 96.10 ± 1.4% homologies with *E. faecalis* OG1RF reference genome. Unique regions identified were mainly associated with signal and metabolic cellular processes.

Conclusions

From a previous study (Chapter 3.2), we have identified the presence of *E. faecalis* in patients' saliva, stressing out the importance of *E. faecalis* identification and characterization as a risk factor for dental diseases. The use of Oxford Nanopore Sequencing, coupled with the accuracy of Illumina reads, enabled complete and accurate genome assemblies, revealing the presence of a multitude of putative antimicrobial resistance genes as well as the identification of virulence traits promoting colonization and infection. An in-depth analysis highlighted the presence of mobile genetic elements carrying AMR genes, responsible for a dissemination of resistance phenotypes among microbial populations.

In conclusion, the use of rapid and cost effective sequencing approaches has the capability of performing rapid genomic population studies as well as to identify antimicrobial resistance genes, virulence factors and genetic elements.

Tables

Table 1. *Enterococcus faecalis* Whole Genome Sequencing results

Sample	Genome length (bp)	Plasmid(s) and bacteriophage(s)	GC content	Genes (CDS)	Cluster
BE5	2,879,031	74,135, 62,694, 13,461, 5,143	37.57%	3002 (2928)	⬛
BE7	2,947,520	30,410^{+}	37.35%	2906 (2828)	🟦
BE8	2,863,584		37.45%	2734 (2660)	🟦
BE11	2,913,763	43,800	37.46%	2854 (2779)	🟩
BE13	2,995,083		37.56%	2933 (2857)	🟧
BE15	2,918,125	38,295, 30,214, 28,909, 5,143, 3,265	37.53%	2984 (2910)	🟨
BE16	2,851,025	102,346*, 19,782, 5,166	37.54%	2936 (2861)	🟪
BE17	2,846,841	56,232, 45,378, 19,755, 5,166	37.54%	2936 (2864)	🟪
BE18	2,841,493	76,311, 45,462	37.5%	2857 (2783)	🟢
BE25	2,761,658		37.68%	2644 (2571)	⬜
BE32	2,913,940	43,766, 29,025^{++}	37.46%	2909 (2836)	🟩
BE33	2,910,762	43,757, 29,013^{++}	37.46%	2923 (2850)	🟩
BE34	2,913,954	43,480, 29,025^{++}	37.46%	2910 (2837)	🟩
BE43	2,919,632	38,307, 30,227, 28,952, 5,143, 3,265	37.53%	2971 (2898)	🟨
BE45	2,745,951		37.6%	2697 (2625)	🟥
BE47	2,757,740		37.6%	2702 (2630)	🟥

+ = bacteriophage, absent in BE8 isolate; ++ = bacteriophage, absent in BE11 isolate; * = sum of pls1 (56 kbp) and pls2 (45 kbp) BE17 sample.

Table 2. *Enterococcus faecalis* MLST profiles

MLST genes	**BE5**	**BE7**	**BE8**	**BE11**	**BE13**	**BE15**	**BE16**	**BE17**	**BE18**	**BE25**	**BE32**	**BE33**	**BE34**	**BE43**	**BE45**	**BE47**
aroE	3	7	7	9	23	23	21	21	29	27	9	9	9	23	90	90
gdh	2	5	5	3	20	2	50	50	34	5	3	3	3	2	10	10
gki	65	3	3	1	25	11	5	5	37	5	1	1	1	11	86	86
gyd	7	1	1	7	3	7	10	10	2	1	7	7	7	7	5	5
pstS	39	1	1	23	7	11	30	30	17	30	23	23	23	11	83	83
xpt	4	7	7	16	2	4	2	2	23	20	16	16	16	4	22	22
yqiL	2	6	6	7	2	2	1	1	6	3	7	7	7	2	85	85
Sequence type	**326**	**16**	**16**	**55**	**72**	**NA**	**260**	**260**	**100**	**173**	**55**	**55**	**55**	**NA**	**699**	**699**

NA: BE15 and BE43 isolates harbour a single nucleotide change in the *aroE_23* gene (G346A) and were not associated with a specific sequence type. The nearest STs are 219, 25, and 239.

Table 3-A. Clinically relevant antimicrobial resistance genes

AMR genes	BE5	BE7	BE8	BE11	BE13	BE15	BE16	BE17	BE18	BE25	BE32	BE33	BE34	BE43	BE45	BE47	Predicted resistance
ANT(6)-Ia		x	x														A
APH(3')-IIIa		x	x														A
efrAB	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	F, M, S
ErmB		x	x														L, M, S
lsaA	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	L, M, O, S, T
lsaE		x	x														L, M, O, S, T
spw		x	x														A
str	x	x	x														S

A: Aminoglycosides; F: Fluoroquinolones; L: Lincosamides; M: Macrolides; O: Oxazolidinones; S: Streptogramins; T: Tetracyclines.

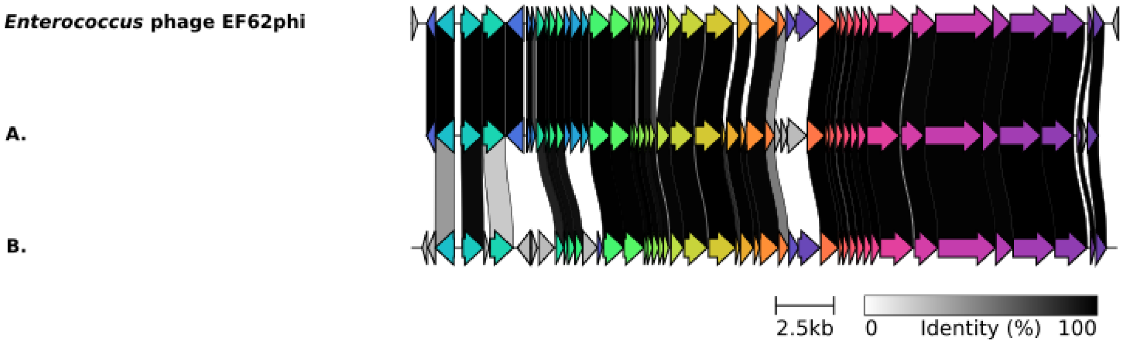
Table 3-B. Additional antimicrobial resistance genes

AMR genes	BE5	BE7	BE8	BE11	BE13	BE15	BE16	BE17	BE18	BE25	BE32	BE33	BE34	BE43	BE45	BE47	Predicted resistance
catA8		x	x														C
dfrE	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	D
dfrG		x	x														D
emeA	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Ac
tet(L)	x																T
tetM	x	x	x	x		x				x	x	x	x	x			T
tetS				x							x	x	x				T
SAT-4		x	x														St

Ac: Acriflavine; C: Chloramphenicol; D: Diaminopyrimidine; St: Streptothricin; T: Tetracyclines.

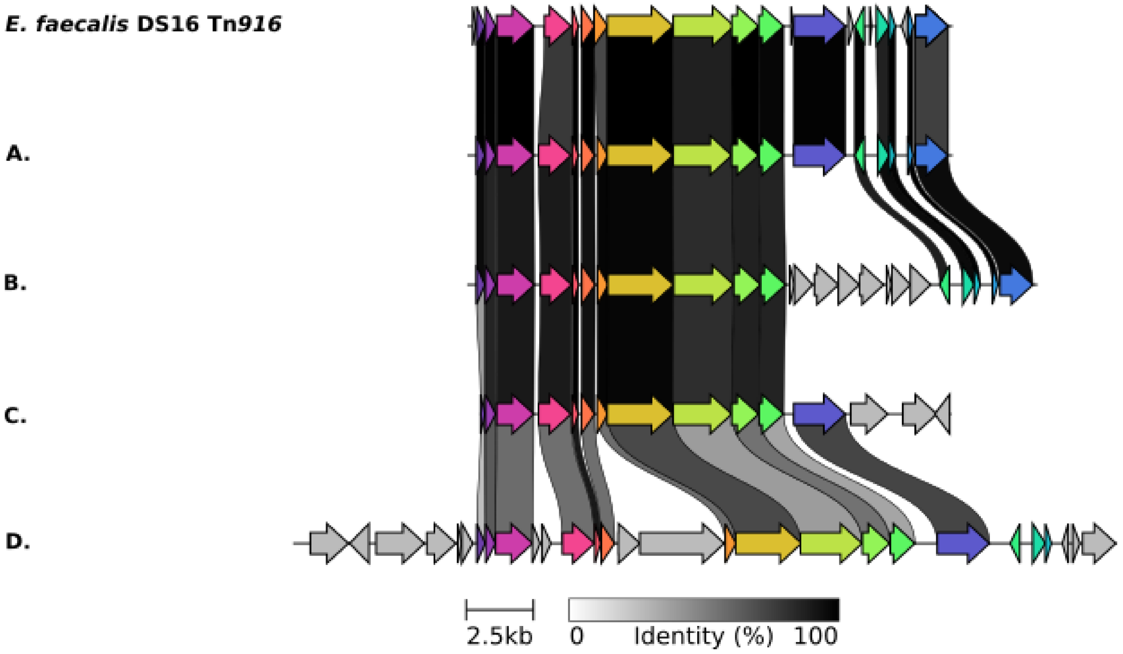
Figures

Figure 1. Comparison of Enterococcal phages.



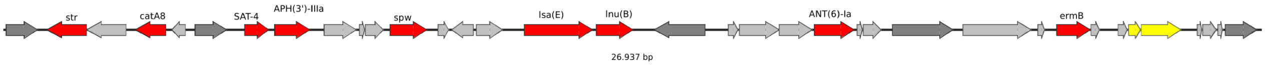
Visual comparison of putative Enterococcal phages identified in BE32, BE33, and BE34 (**A.**) and in BE7 (**B.**) isolates. The *Enterococcus* phage EF62phi (top) (CP002495) was used as a reference. Homology is represented as a bar between genes (arrows): the darker the colour, the higher the homology.

Figure 2. Comparison of Tn*916*-like elements.



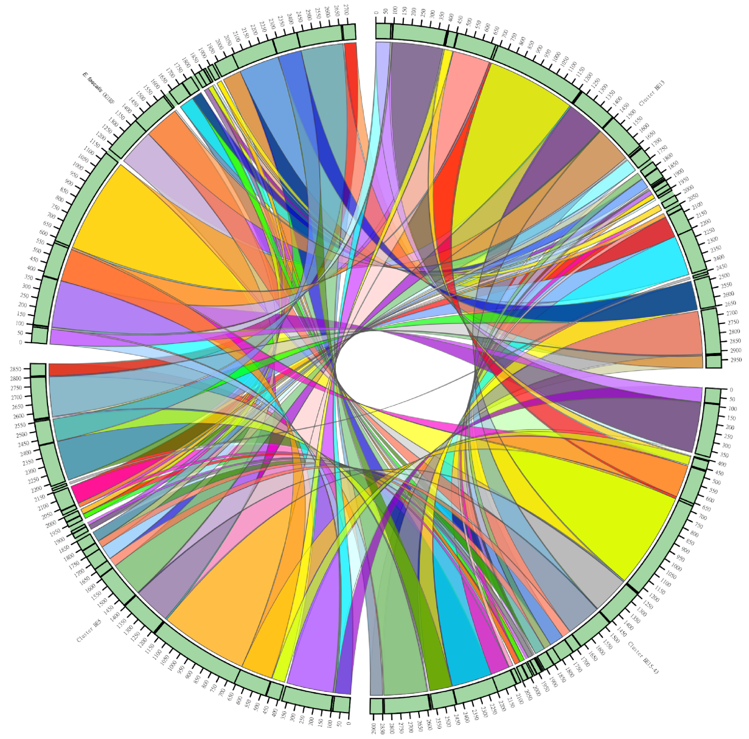
Comparison of the identified Tn*916*-like mobile genetic elements. **A:** Tn*916* conjugative transposon carrying *tetM* gene (purple), identified in isolates BE7, BE8, BE11, BE15, BE25, BE32, BE33, BE34 and BE43. **B:** Tn*916*-like mobile genetic element identified in BE13 isolate. A putative Bacitracin ABC transporter is inserted between ORF13 and ORF9, in place of the *tetM* gene (purple). **C:** Tn*916*-like identified in BE5 isolate, carrying *tetM* gene (purple) and *tetS* gene. **D:** BE11, BE32, BE33, and BE34 Tn*6000*-like element carrying the *tetS* gene.

Figure 3. BE7-BE8 Mobile Genetic Element



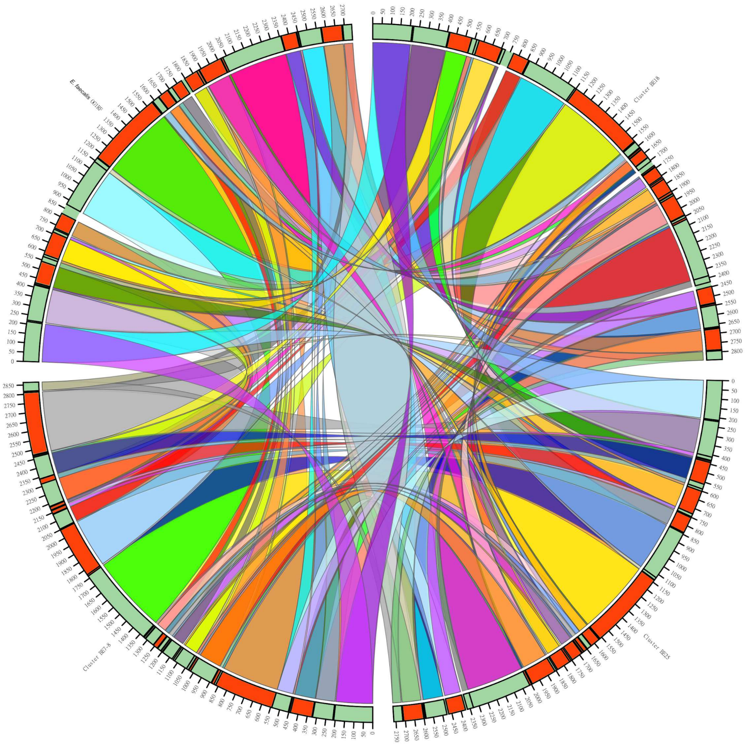
Schematic representation of the ~30 Kb mobile element identified in BE7 and BE8 isolates. Antimicrobial resistance genes are represented using red arrows, while the Toxin-Antitoxin locus is represented using yellow arrows. Dark-grey arrows represent insertion sequences (IS).

Figure 4. Synteny blocks of isolates BE5, BE13, BE15 and BE43



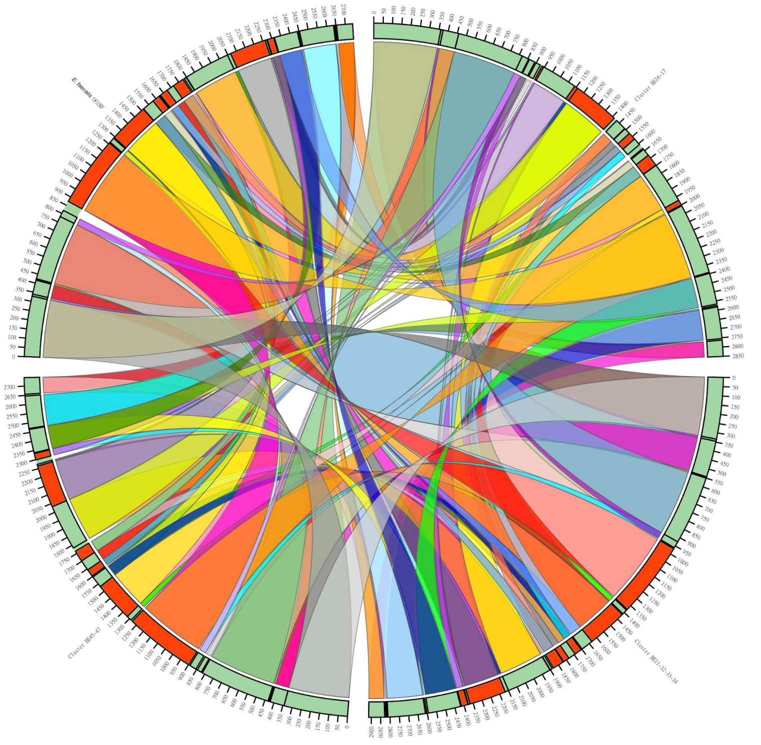
A total of 32 synteny blocks were identified among the input chromosomes. Those blocks account for a 94.66% BE13 genome coverage highlighting a 5.34% unique traits. The BE15-43 cluster has a 96.40% of coverage (3.6% uniqueness), while the BE5 cluster has a 95.21% genome coverage (4.79% uniqueness). The *E.faecalis* OG1RF has a 95.10% coverage, suggesting a 4.9% unique region. Interestingly, all the synteny blocks were identified with the same orientation (positive strand).

Figure 5. Synteny blocks of isolates BE7, BE8, BE18 and BE25



A total of 37 synteny blocks were identified, covering a 94.26% of BE18 genome. On the other hand, BE25 and BE7-8 clusters have respectively 96.67% and 97.28% of genome coverage, highlighting 3.33% and 2.72% of uncovered (unique) regions. Shared synteny blocks cover the 95.21% of *E. faecalis* OG1RF genome.

Figure 6. Synteny blocks of isolates BE11, BE16, BE32, BE33, BE34, BE45 and BE47



A total of 29 synteny blocks were identified, accounting for a 96.36% genome coverage of BE16 cluster (3.64% unique sequences). The cluster BE11-32-33-34 has 97.90% of covered genome (2.1% uniqueness), while the BE45-47 cluster has 96.2% genome coverage (3.8% uniqueness). Lastly, the *E.faecalis* OG1RF has a 96.29% coverage.

3.4 References

1. García-Solache M, Rice LB. The Enterococcus: a Model of Adaptability to Its Environment. Clin Microbiol Rev. 2019 Jan 30;32(2):e00058-18. doi: 10.1128/CMR.00058-18.

2. Kalina AP. 1970. The taxonomy and nomenclature of enterococci. Int J Syst Evol Microbiol 20:185–189. doi:10.1099/00207713-20-2-185

3. Schleifer KH, Kilpper-Bälz R. 1984. Transfer of Streptococcus faecalis and Streptococcus faecium to the genus Enterococcus nom. rev. as Enterococcus faecalis comb. nov. and Enterococcus faecium comb. nov. Int J Syst Evol Microbiol 34:31–34. doi:10.1099/00207713-34-1-31

4. Švec P, Franz CMAP. 2014. The genus Enterococcus, p 175–211. In Holzapfel WH, Wood BJB (ed), Lactic acid bacteria: biodiversity and taxonomy. John Wiley & Sons, Ltd, Chichester, England. doi:10.1002/9781118655252.ch15

5. Weiner LM, Webb AK, Limbago B, Dudeck MA, Patel J, Kallen AJ, Edwards JR, Sievert DM. 2016. Antimicrobial-resistant pathogens associated with healthcare-associated infections: summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2011-2014. Infect Control Hosp Epidemiol 37:1288–1301. doi:10.1017/ice.2016.174.

6. Paulsen IT, Banerjei L, Myers GS, Nelson KE, Seshadri R, Read TD, et al. Role of mobile DNA in the evolution of vancomycin-resistant Enterococcus faecalis. Science. 2003;299(5615):2071–4.

7. He Q, Hou Q, Wang Y, Li J, Li W, Kwok L-Y, Sun Z, Zhang H, and Zhong Z. Comparative genomic analysis of Enterococcus faecalis: insights into their environmental adaptations. BMC Genomics 19, 527 (2018). https://doi.org/10.1186/s12864-018-4887-3

8. Mohamed JA, Huang DB. Biofilm formation by enterococci. J Med Microbiol. 2007 Dec;56(Pt 12):1581-1588. doi: 10.1099/jmm.0.47331-0

9. Rich RL, Kreikemeyer B, Owens RT, LaBrenz S, Narayana SV, Weinstock GM, Murray BE, Hook M. 1999. Ace is a collagen-binding MSCRAMM from Enterococcus faecalis. J Biol Chem 274:26939–26945. doi:10.1074/jbc.274.38.26939.

10. Singh KV, Nallapareddy SR, Sillanpaa J, Murray BE. 2010. Importance of the collagen adhesin Ace in pathogenesis and protection against Enterococcus faecalis experimental endocarditis. PLoS Pathog 6:e1000716. doi:10.1371/journal.ppat.1000716

11. Telford JL, Barocchi MA, Margarit I, Rappuoli R, Grandi G. 2006. Pili in gram-positive pathogens. Nat Rev Microbiol 4:509–519. doi:10.1038/nrmicro1443

12. Nallapareddy SR, Singh KV, Sillanpää J, Garsin DA, Höök M, Erlandsen SL, Murray BE. 2006. Endocarditis and biofilm-associated pili of Enterococcus faecalis. J Clin Invest 116:2799–2807. doi:10.1172/JCI29021

13. Haas W, Shepard BD, Gilmore MS. 2002. Two-component regulator of Enterococcus faecalis cytolysin responds to quorum-sensing autoinduction. Nature 415:84–87. doi:10.1038/415084a

14. Chow JW, Thal LA, Perri MB, Vazquez JA, Donabedian SM, Clewell DB, Zervos MJ. 1993. Plasmid-associated hemolysin and aggregation substance production contribute to virulence in experimental enterococcal endocarditis. Antimicrob Agents Chemother 37:2474–2477. doi:10.1128/AAC.37.11.2474.

15. Thurlow LR, Thomas VC, Narayanan S, Olson S, Fleming SD, Hancock LE. 2010. Gelatinase contributes to the pathogenesis of endocarditis caused by Enterococcus faecalis. Infect Immun 78:4936–4943. doi:10.1128/IAI.01118-09

16. Qin X, Singh KV, Weinstock GM, Murray BE. 2000. Effects of Enterococcus faecalis fsr genes on production of gelatinase and a serine protease and virulence. Infect Immun 68:2579–2586. doi:10.1128/IAI.68.5.2579-2586.2000.

17. Hollenbeck BL, Rice LB. Intrinsic and acquired resistance mechanisms in enterococcus. Virulence. 2012 Aug 15;3(5):421-33. doi: 10.4161/viru.21282.

18. Murray BE. 1990. The life and times of the Enterococcus. Clin Microbiol Rev 3:46–65. doi:10.1128/CMR.3.1.46.

19. Miller WR, Munita JM, Arias CA. Mechanisms of antibiotic resistance in enterococci. Expert Rev Anti Infect Ther. 2014 Oct;12(10):1221-36. doi: 10.1586/14787210.2014.956092.

20. Roberts M. 1994. Epidemiology of tetracycline resistance determinants. Trends Microbiol 2:353–357. doi:10.1016/0966-842X(94)90610-6

21. Burrus V, and Waldor M.K. (2004) Shaping bacterial genomes with integrative and conjugative elements. Res Microbiol 155: 376-386.

22. Frost LS, Leplae R, Summers AO, and Toussaint A. (2005) Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 3: 722-732.

23. Clewell DB, Weaver KE, Dunny GM, Coque TM, Francia MV, Hayes F. 2014. Extrachromosomal and mobile elements in enterococci: transmission, maintenance, and

epidemiology, p 309–420. In Gilmore MS, Clewell DB, Ike Y, Shankar N (ed), Enterococci: from commensals to leading causes of drug resistant infection. Massachusetts Eye and Ear Infirmary, Boston, MA.

24. Schwarz FV, Perreten V, Teuber M. 2001. Sequence of the 50-kb conjugative multiresistance plasmid pRE25 from Enterococcus faecalis RE25. Plasmid 46:170–187. doi:10.1006/plas.2001.1544.

25. Clewell DB. 1981. Plasmids, drug resistance, and gene transfer in the genus Streptococcus. Microbiol Rev 45:409–436.

26. Franke AE, Clewell DB. 1981. Evidence for a chromosome-borne resistance transposon (Tn916) in Streptococcus faecalis that is capable of "conjugal" transfer in the absence of a conjugative plasmid. J Bacteriol 145:494–502

27. Hegstad K, Mikalsen T, Coque TM, Werner G, Sundsfjord A. 2010. Mobile genetic elements and their contribution to the emergence of antimicrobial resistant Enterococcus faecalis and Enterococcus faecium. Clin Microbiol Infect 16:541–554. doi:10.1111/j.1469-0691.2010.03226.x

28. Jensen LB, Garcia-Migura L, Valenzuela AJS, Løhr M, Hasman H, Aarestrup FM. 2010. A classification system for plasmids from enterococci and other Gram-positive bacteria. J Microbiol Methods 80:25–43. doi:10.1016/j.mimet.2009.10.012.

29. Hirt H, Manias DA, Bryan EM, Klein JR, Marklund JK, Staddon JH, Paustian ML, Kapur V, Dunny GM. 2005. Characterization of the pheromone response of the Enterococcus faecalis conjugative plasmid pCF10: complete sequence and comparative analysis of the transcriptional and phenotypic responses of pCF10-containing cells to pheromone induction. J Bacteriol 187:1044–1054. doi:10.1128/JB.187.3.1044-1054.2005.

30. Manson JM, Hancock LE, Gilmore MS. 2010. Mechanism of chromosomal transfer of Enterococcus faecalis pathogenicity island, capsule, antimicrobial resistance, and other traits. Proc Natl Acad Sci U S A 107:12269–12274. doi:10.1073/pnas.1000139107.

31. Clewell DB. 2007. Properties of Enterococcus faecalis plasmid pAD1, a member of a widely disseminated family of pheromone-responding, conjugative, virulence elements encoding cytolysin. Plasmid 58:205–227. doi:10.1016/j.plasmid.2007.05.001.

32. Schwarz FV, Perreten V, Teuber M. 2001. Sequence of the 50-kb conjugative multiresistance plasmid pRE25 from Enterococcus faecalis RE25. Plasmid 46:170–187. doi:10.1006/plas.2001.1544.

33. Freitas AR, Novais C, Tedim AP, Francia MV, Baquero F, Peixe L, Coque TM. 2013.

Microevolutionary events involving narrow host plasmids influences local fixation of vancomycin-resistance in Enterococcus populations. PLoS One 8:e60589. doi:10.1371/journal.pone.0060589

34. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. doi: 10.1093/bioinformatics/btu153.

35. Ruiz‑Garbajosa P, Bonten MJM, Robinson DA, Top J, Nallapareddy SR, Torres C, et al. Multilocus sequence typing scheme forEnterococcus faecalis reveals hospital‑adapted genetic complexes in a background of high rates of recombination. J Clin Microbiol, 44 (2006), pp. 2220-2228

36. McArthur AG, Waglechner N, Nizam F, et al. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother. 2013;57(7):3348-3357. doi:10.1128/AAC.00419-13

37. Minkin I, Patel A, Kolmogorov M, Vyahhi N, and Pham S. Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In Algorithms in Bioinformatics, pp. 215-229. Springer Berlin Heidelberg, 2013.

38. Minkin I, Patel A, Kolmogorov M, Vyahhi N, and Pham S. Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In Algorithms in Bioinformatics, pp. 215-229. Springer Berlin Heidelberg, 2013.

39. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009 Sep;19(9):1639-45. doi: 10.1101/gr.092759.109.

40. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34:3094-3100. doi:10.1093/bioinformatics/bty191

41. Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform. 2013 Mar;14(2):193-202. doi: 10.1093/bib/bbs012.

42. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016 Jul 8;44(W1):W16-21. doi: 10.1093/nar/gkw387.

43. Brede DA, Snipen LG, Ussery DW, Nederbragt AJ, Nes IF. Complete genome sequence of the commensal Enterococcus faecalis 62, isolated from a healthy Norwegian infant. J Bacteriol. 2011 May;193(9):2377-8. doi: 10.1128/JB.00183-11.

44. Gilchrist CLM, Chooi YH. Clinker & clustermap.js: Automatic generation of gene cluster

comparison figures. Bioinformatics. 2021 Jan 18:btab007. doi: 10.1093/bioinformatics/btab007.

45. Flannagan, S. E., Zitzow, L. A., Su, Y. A., & Clewell, D. B. (1994). Nucleotide sequence of the 18-kb conjugative transposon Tn916 from Enterococcus faecalis. Plasmid, 32(3), 350-354. DOI: 10.1006/plas.1994.1077.

46. Michael S.M. Brouwer, Peter Mullany, Adam P. Roberts, Characterization of the conjugative transposon Tn6000 from Enterococcus casseliflavus 664.1H1 (formerly Enterococcus faecium 664.1H1), FEMS Microbiology Letters, Volume 309, Issue 1, August 2010, Pages 71–76. doi:10.1111/j.1574-6968.2010.02018.x

47. Manson JM, Keis S, Smith JM, Cook GM. Acquired bacitracin resistance in Enterococcus faecalis is mediated by an ABC transporter and a novel regulatory protein, BcrR. Antimicrob Agents Chemother. 2004 Oct;48(10):3743-8. doi: 10.1128/AAC.48.10.3743-3748.2004.

Chapter 4. Sequencing of *Severe Acute Respiratory Syndrome Coronavirus 2*

In late December 2019, a cluster of suspicious cases of pneumonia, caused by an unknown etiologic agent, was reported in the city of Wuhan, Hubei-province (China). The causative agent was quickly identified as a virus belonging to the genus *Betacoronavirus*, *Coronaviridae* family, named *Severe acute respiratory syndrome-related coronavirus 2* (SARS-CoV-2) (1). Since its first identification in China, the virus spread rapidly and uncontrolled all over the world and was declared pandemic on 11 March 2020 by the World Health Organization (WHO). To date (January 2021), SARS-CoV-2 infected over 100 millions people worldwide resulting in over 2 millions deaths.

Betacoronavirus are a group of enveloped, positive-sense, single strand RNA viruses characterized by a large genome ranging from 27 kb to 32 kb in length. The genome encodes two large polyproteins (ORF1a and ORF1b), which undergo proteolysis generating non structural proteins (nsps) of various functions such as viral proteases and the RNA-dependent RNA-polymerase (RdRP), and four conserved structural proteins (the small envelope protein (E), the spike protein (S), the matrix protein (M), and the nucleocapsid protein (N)) (2,3). While RNA viruses are generally characterized by a high error rate during the replication process resulting in quasispecies (population of viruses characterized by genomic mutations that resides in the same host), coronaviruses are instead characterized by an approximately 10-fold lower mutation rate due to an intrinsic proofreading activity mediated by a nonstructural protein (nsp14). In fact, it has been estimated that SARS-CoV-2 undergoes 33 genomic mutations/year (4). Those mutations are used by scientists to assign lineages or clades to each strain, essential to study and track the spread of the virus.

Since the declaration of pandemic, the urgent need of SARS-CoV-2 genomic data was immediately clear to the entire scientific community. Whole Genome Sequencing (WGS) approaches have the potential to enable a rapid SARS-CoV-2 profiling, helping in the surveillance of circulating viral strains, and understanding the transmission dynamics. In the struggle of standardise sequencing procedures, the ARTIC Network (https://artic.network/ncov-2019) released an open-source sequencing protocol involving Oxford Nanopore sequencing technology, which was later adapted for other sequencing platforms such as Illumina, PacBio and Ion Torrent. In short, the protocol relies on direct amplification of the virus using tiled, multiplexed primers generating 400 bp amplicons characterized by a 50 bp overlapping region. The global WGS data are deposited in publicly-accessible repositories like the GISAID EpiCov (https://www.gisaid.org/) database.

The GISAID database combines genetic sequences and related clinical, epidemiological, and geographical metadata enabling global phylogenetic analysis and helping to understand how the virus evolved and spread during the pandemic. To date, more than 620,000 SARS-CoV-2 sequences are available.

In this chapter I will illustrate: (i) the sequencing of the first SARS-CoV-2 viral isolate obtained in Tuscany by the combined use of direct RNA sequencing and of a new amplicon scheme (chapter 4.1); (ii) the sequencing of the first 10 viral strains from Malta using the Artic primer scheme directly from clinical samples (chapter 4.2); (iii) the sequencing of 15 viral strains obtained at the beginning and at the end of the first pandemic wave in Siena (chapter 4.3).

Chapter 4.1

Whole-Genome Sequence of SARS-CoV-2 Isolate Siena-1/2020

Maria Grazia Cusi, David Pinzauti, Claudia Gandolfo, Gabriele Anichini, Gianni Pozzi, Francesco Santoro

AMERICAN SOCIETY FOR MICROBIOLOGY
Microbiology Resource Announcements

Check for updates

Whole-Genome Sequence of SARS-CoV-2 Isolate Siena-1/2020

Maria Grazia Cusi,[a,b] David Pinzauti,[a] Claudia Gandolfo,[a] Gabriele Anichini,[a] Gianni Pozzi,[a] Francesco Santoro[a,b]

[a]Department of Medical Biotechnologies, University of Siena, Siena, Italy
[b]UOC Microbiologia e Virologia, Azienda Ospedaliera Universitaria Senese, Siena, Italy

ABSTRACT The complete genome sequence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) isolate Siena-1/2020 was obtained by Nanopore sequencing, combining the direct RNA sequencing and amplicon sequencing approaches. The isolate belongs to the B1.1 lineage, which is prevalent in Europe, and contains a mutation in the spike protein coding sequence leading to the D614G amino acid change.

Here, we report the complete genome sequence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) isolate Siena-1/2020, belonging to the genus *Betacoronavirus* in the family *Coronaviridae*. The virus was isolated at the University Hospital of Siena (Tuscany, Italy) in April 2020, from a nasopharyngeal swab collected on 1 March 2020, and seeded on Vero E6 cells. This research was carried out according to the principles of the Helsinki Declaration, with reference to the document BIOBANK-MIU-2010, approved by the Siena University Hospital Ethics Committee with amendment no. 1, on 17 February 2020, regarding general data protection and regulation (GDPR).

After 3 days, cytopathic effect appeared on the cells, and the culture medium was collected and frozen at −80°C. Since this was the first SARS-CoV-2 viral isolate in our region, we decided to sequence it. Total RNA was isolated using the NucliSens easyMAG system (bioMérieux, Italy). Viral RNA was sequenced using both the direct RNA and amplicon sequencing approaches on a MinION instrument (Oxford Nanopore Technologies [ONT], UK).

Direct RNA sequencing was performed using the SQK-RNA002 kit (ONT). Briefly, about 300 ng of total RNA was ligated to the reverse transcriptase (RT) adapter, and the first strand was retrotranscribed using SuperScript III reverse transcriptase (Thermo Fisher); sequencing adapters were then ligated to the cDNA-RNA hybrid, and the library was loaded onto a flow cell (R9.4.1).

Amplicon sequencing was performed based on a modification of the Artic Network protocol (https://www.protocols.io/view/ncov-2019-sequencing-protocol-v2-bdp7i5rn); primers were designed using Primal Scheme (1) to generate 39 amplicons of about 900 bp with an overlap of about 50 bp (Table 1). About 100 ng of total RNA was reverse-transcribed using the SuperScript VILO reverse transcriptase kit (Thermo Fisher) following the manufacturer's instructions and then amplified in two multiplex PCRs using PrimeSTAR GXL polymerase (TaKaRa). The samples were barcoded, pooled, and adapter ligated following the ONT ligation-based sequencing protocol. The sequencing run was managed by MinKNOW v19.12.5, enabling live base calling. For data analysis, all tools were run with default parameters unless otherwise specified. Sequencing reads were demultiplexed using Guppy v3.6.1 and then filtered using the guppyplex command of the ARTIC environment to include only reads between 700 and 1,500 bases long (https://github.com/artic-network/artic-ncov2019). Amplicon reads were mapped to the reference genome Wuhan Hu-1 (GenBank accession no. MN908947) with minimap2 v2.17 (2) and indexed using SAMtools v1.9 (3). Primer sequences were trimmed

Citation Cusi MG, Pinzauti D, Gandolfo C, Anichini G, Pozzi G, Santoro F. 2020. Whole-genome sequence of SARS-CoV-2 isolate Siena-1/2020. Microbiol Resour Announc 9:e00944-20. https://doi.org/10.1128/MRA.00944-20.

Editor Simon Roux, DOE Joint Genome Institute

Copyright © 2020 Cusi et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Francesco Santoro, santorof@unisi.it.

Received 26 August 2020
Accepted 9 September 2020
Published 24 September 2020

TABLE 1 Primers used for amplification of the SARS-CoV-2 genome

Primer	Nucleotide sequence	Pool	Length (no. of nucleotides)	Start[a]	End[a]
CoV-2_1_L	ACCAACCAACTTTCGATCTCTTGT	1	24	30	54
CoV-2_1_R	ATGCACTCAAGAGGGTAGCCAT	1	22	867	845
CoV-2_2_L	AGTGGTGTTACCCGTGAACTCA	2	22	763	785
CoV-2_2_R	ACCTTCGGAACCTTCTCCAACA	2	22	1600	1578
CoV-2_3_L_v2	GGCTGTGTGTTCTCTTATGTTGGT	1	24	1487	1510
CoV-2_3_R	ACAATCCCTTTGAGTGCGTGAC	1	22	2414	2392
CoV-2_4_L	TTTGGCTTTGTGTGCTGACTCT	2	22	2319	2341
CoV-2_4_R	AGCAGAAGTGGCACCAAATTCC	2	22	3166	3144
CoV-2_5_L	GATTGTGAAGAAGAAGAGTTTGAGCC	1	26	3067	3093
CoV-2_5_R	CAGCGATCTTTTGTTCAACTTGCT	1	24	3878	3854
CoV-2_6_L	TCGCACAAATGTCTACTTAGCTGT	2	24	3771	3795
CoV-2_6_R	ACCGAGCAGCTTCTTCCAAATT	2	22	4658	4636
CoV-2_7_L	ACAACTGTAGCGTCACTTATCAACA	1	25	4549	4574
CoV-2_7_R	AGCATCTTGTAGAGCAGGTGGA	1	22	5359	5337
CoV-2_8_L	ACTTCTATTAAATGGGCAGATAACAACTG	2	29	5257	5286
CoV-2_8_R	AGCCACCACATCACCATTTAAGT	2	23	6172	6149
CoV-2_9_L	CCATATCCAAACGCAAGCTTCG	1	22	6019	6041
CoV-2_9_R	GCCTCTAGACAAAATTTACCGACACT	1	26	6903	6877
CoV-2_10_L	AAACCGTGTTTGTACTAATTATATGCCTT	2	29	6747	6776
CoV-2_10_R	ACTGTAGTGACAAGTCTCTCGCA	2	23	7694	7671
CoV-2_11_L	GCTTTTGCAAACTACACAATTGGAAT	1	26	7592	7618
CoV-2_11_R	GCAGCACTACGTATTTGTTTTCGT	1	24	8463	8439
CoV-2_12_L	GCGCAGGTAGCAAAAAGTCACA	2	22	8365	8387
CoV-2_12_R	TGATCTTTCACAAGTGCCGTGC	2	22	9241	9219
CoV-2_13_L	TGCTCATGGATGGCTCTATTATTCAA	1	26	9128	9154
CoV-2_13_R	GAGCCTTTGCGAGATGACAACA	1	22	9977	9955
CoV-2_14_L	GTGATGTGCTATTACCTCTTACGCA	2	25	9848	9873
CoV-2_14_R	CAGCAGCGTACAACCAAGCTAA	2	22	10688	10666
CoV-2_15_L	CAACTGGAGTTCATGCTGGCAC	1	22	10556	10578
CoV-2_15_R	GTCCACACTCTCCTAGCACCAT	1	22	11394	11372
CoV-2_16_L	TGTCTGGTTTTAAGCTAAAAGACTGTGT	2	28	11285	11313
CoV-2_16_R	ATCACCATTAGCAACAGCCTGC	2	22	12181	12159
CoV-2_17_L	GGCAACCTTACAAGCTATAGCCT	1	23	12078	12101
CoV-2_17_R	CCTACAAGGTGGTTCCAGTTCTG	1	23	12907	12884
CoV-2_18_L	GGAGGTAGGTTTGTACTTGCACTG	2	24	12793	12817
CoV-2_18_R	CGTCCTTTTCTTGGAAGCGACA	2	22	13621	13599
CoV-2_19_L	ACAGGCACTAGTACTGATGTCGT	1	23	13509	13532
CoV-2_19_R	TGGGTGGTATGTCTGATCCCAA	1	22	14328	14306
CoV-2_20_L	CAAAGCCTTACATTAAGTGGGATTTGT	2	27	14224	14251
CoV-2_20_R	GGTGCGAGCTCTATTCTTTGCA	2	22	15108	15086
CoV-2_21_L	AGGATCAAGATGCACTTTTCGCA	1	23	15004	15027
CoV-2_21_R	AGTAAGGTCAGTCTCAGTCCAACA	1	24	15858	15834
CoV-2_22_L	TGCATCTCAAGGTCTAGTGGCT	2	22	15749	15771
CoV-2_22_R	GCGTTTCTGCTGCAAAAAGCTT	2	22	16648	16626
CoV-2_23_L	CGATAATGTTACTGACTTTAATGCAATTGC	1	30	16535	16565
CoV-2_23_R	GTGCAGGTAATTGAGCAGGGTC	1	22	17458	17436
CoV-2_24_L	TCTTTGATGAAATTTCAATGGCCACA	2	26	17350	17376
CoV-2_24_R	GCTTCTTCGCGGGTGATAAACA	2	22	18275	18253
CoV-2_25_L	TGGCATACCTAAGGACATGACCT	1	23	18167	18190
CoV-2_25_R	ACCAATGTCGTGAAGAACTGGG	1	22	19038	19016
CoV-2_26_L	TGATGAACTGAAGATTAATGCGGCT	2	25	18938	18963
CoV-2_26_R	GCAGCAATGTCCACACCCAAAT	2	22	19862	19840
CoV-2_27_L	ACACAAAAGTTGATGGTGTTGATGT	1	25	19714	19739
CoV-2_27_R	GGTTGCCACGCTTGACTAGATT	1	22	20678	20656
CoV-2_28_L	TCTGTAGTTTCTAAGGTTGTCAAAGTGA	2	28	20553	20581
CoV-2_28_R	AAAGACATAACAGCAGTACCCCTTAA	2	26	21443	21417
CoV-2_29_L_v2	CAAACCACGCGAACAAATAG	1	20	21297	21316
CoV-2_29_R_v2	CGAAAAACCCTGAGGGAGAT	1	20	22225	22206
CoV-2_30_L	AAACAGGGTAATTTCAAAAATCTTAGGGAA	2	30	22105	22135
CoV-2_30_R	TGTGCTACCGGCCTGATAGATT	2	22	22996	22974
CoV-2_31_L	AACAATCTTGATTCTAAGGTTGGTGGT	1	27	22876	22903
CoV-2_31_R	TGCTGCATTCAGTTGAATCACCA	1	23	23813	23790
CoV-2_32_L	ACCCACAAATTTTACTATTAGTGTTACCAC	2	30	23703	23733
CoV-2_32_R	TGCACTTCAGCCTCAACTTTGT	2	22	24537	24515

(Continued on next page)

TABLE 1 (Continued)

Primer	Nucleotide sequence	Pool	Length (no. of nucleotides)	Start[a]	End[a]
CoV-2_33_L	TGCACAAGCTTTAAACACGCTT	1	22	24426	24448
CoV-2_33_R	GCAGCAGGATCCACAAGAACAA	1	22	25324	25302
CoV-2_34_L	CTAGGTTTTATAGCTGGCTTGATTGC	2	26	25213	25239
CoV-2_34_R	ACATGTTCAACACCAGTGTCTGT	2	23	26075	26052
CoV-2_35_L	GGGAATCTGGAGTAAAAGACTGTGT	1	25	25969	25994
CoV-2_35_R	AATGACCACATGGAACGCGTAC	1	22	26857	26835
CoV-2_36_L	TGGATCACCGGTGGAATTGCTA	2	22	26744	26766
CoV-2_36_R	GTGTTTTACGCCGTCAGGACAA	2	22	27612	27590
CoV-2_37_L	ACGAGGGCAATTCACCATTTCA	1	22	27511	27533
CoV-2_37_R	ACTGCCAGTTGAATCTGAGGGT	1	22	28351	28329
CoV-2_38_L	AGAGTATCATGACGTTCGTGTTGT	2	24	28219	28243
CoV-2_38_R	GCTTCTTAGAAGCCTCAGCAGC	2	22	29045	29023
CoV-2_39_L	TGCTTGACAGATTGAACCAGCT	1	22	28940	28962
CoV-2_39_R	TTCTCCTAAGAAGCTATTAAAATCACATGG	1	30	29866	29836

[a] Nucleotide positions relative to the Wuhan Hu-1 reference genome.

from the aligned reads using BAMClipper v1.1.1 (4). Clipped reads were then merged with direct RNA sequencing reads with the −cat command of the Linux environment. Finally, Medaka v0.12.1 (https://github.com/nanoporetech/medaka) was used to build the consensus and call the single nucleotide variants. The reference genome Wuhan Hu-1 was edited using a Perl script, selecting variants with a quality score cutoff of 35 (https://github.com/CDCgov/SARS-CoV-2_Sequencing/tree/master/protocols/CDC-Comprehensive); nucleotide variations were also visually inspected using Tablet (5). We could not sequence the nucleotides corresponding to positions 1 and 2 of the Wuhan Hu-1 genome; therefore, we obtained a 29,901-bp viral genome with an average GC content of 37.97% and a mean depth of coverage of 1,153.64×, as calculated by the SAMtools −coverage function (3).

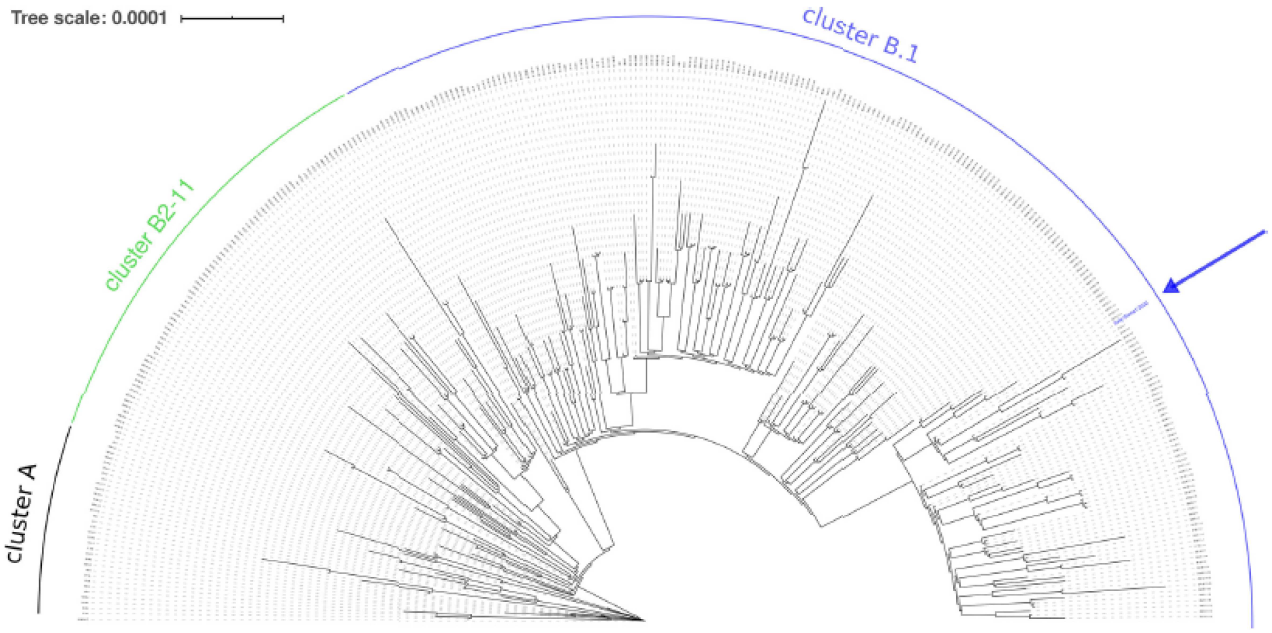


FIG 1 Phylogenetic tree of SARS-CoV-2 genomes. The tree was generated with Pangolin v1.14 and visualized using Interactive Tree Of Life (iTOL) (8). A total of 322 viral genomes are displayed, including the genomes selected by Pangolin software as representatives for the genetic diversity of SARS-CoV-2. As of 30 July 2020, two major clusters (A and B) were identified. Cluster B was subdivided into 11 clusters (B1 to B11); of those, the most represented is cluster B1 (covered by the blue arch), which comprises most of the lineages identified and sequenced in Europe. Cluster B1.1 is a large subcluster characterized by the mutation of three consecutive nucleotides at position 28881. The blue arrow indicates the position of the Siena-1/2020 isolate in the phylogenetic tree.

Phylogenetic analysis performed with Pangolin v1.14 (6) assigned strain Siena-1/2020 to the B1.1 lineage, which is associated with the Italian SARS-CoV-2 outbreak and includes isolates circulating in Europe (Fig. 1). Compared to the reference genome Wuhan Hu-1, Siena-1/2020 harbors 5 single nucleotide changes (at positions 241, 3037, 14408, 19839, and 23403) and mutations of 3 consecutive nucleotides (GGG→AAC) at position 28881. Among the 5 single nucleotide changes, the one at position 23403 causes a change in the predicted amino acid sequence of the spike (S) protein (D614G), which is now the most common variant worldwide (7).

Data availability. The genome sequence of SARS-CoV-2 Siena-1/2020 has been deposited in GenBank under the accession no. MT531537. The version described in this paper is the second version. The raw Nanopore reads were deposited in the Sequence Read Archive under BioProject accession no. PRJNA658490 with accession no. SRX8982904 (direct RNA sequencing) and SRX8982905 (amplicon sequencing).

ACKNOWLEDGMENTS

We thank all the laboratory personnel involved in SARS-CoV-2 diagnostics at the University Hospital of Siena for their tireless commitment during the coronavirus disease 2019 (COVID-19) pandemic.

REFERENCES

1. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, Oliveira G, Robles-Sikisaka R, Rogers TF, Beutler NA, Burton DR, Lewis-Ximenez LL, de Jesus JG, Giovanetti M, Hill SC, Black A, Bedford T, Carroll MW, Nunes M, Alcantara LC, Jr, Sabino EC, Baylis SA, Faria NR, Loose M, Simpson JT, Pybus OG, Andersen KG, Loman NJ. 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc 12:1261–1276. https://doi.org/10.1038/nprot.2017.066.
2. *Li H. 2018. minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
4. Au CH, Ho DN, Kwong A, Chan TL, Ma ESK. 2017. BAMClipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing. Sci Rep 7:1567. https://doi.org/10.1038/s41598-017-01703-6.
5. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform 14:193–202. https://doi.org/10.1093/bib/bbs012.
6. Rambaut A, Holmes EC, Hill V, O'Toole Á, McCrone JT, Ruis C, Du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv https://doi.org/10.1101/2020.04.17.046086.
7. *Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W,* Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC, Sheffield COVID-19 Genomics Group. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182:812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043.
8. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.

Chapter 4.2

Genome Sequences of 10 SARS-CoV-2 Viral Strains Obtained by Nanopore Sequencing of Nasopharyngeal Swabs in Malta

Manuele Biazzo, Silvia Madeddu, Elfath Elnifro, Tessabella Sultana, Josie Muscat, Christian A. Scerri, Francesco Santoro, David Pinzauti

AMERICAN SOCIETY FOR MICROBIOLOGY
Microbiology® Resource Announcements

Genome Sequences of 10 SARS-CoV-2 Viral Strains Obtained by Nanopore Sequencing of Nasopharyngeal Swabs in Malta

Manuele Biazzo,[a] Silvia Madeddu,[a] Elfath Elnifro,[b] Tessabella Sultana,[b] Josie Muscat,[b] Christian A. Scerri,[c] Francesco Santoro,[d] David Pinzauti[d]

[a]The BioArte Limited, San Gwann, Malta
[b]Medical Laboratory Services within St James Hospital, Sliema, Malta
[c]Department of Physiology and Biochemistry, University of Malta, Msida, Malta
[d]Department of Medical Biotechnologies, University of Siena, Siena, Italy

ABSTRACT The genome sequences of 10 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) strains from Sliema, Malta, were obtained by Nanopore sequencing using the amplicon sequencing approach developed by the Artic Network. The assembled genomes were analyzed with Pangolin software and assigned to the B.1 lineage, which is widely circulating in Europe.

We present the genome sequences of 10 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; genus *Betacoronavirus*, family *Coronaviridae*) viral strains retrieved from Sliema, Malta. Because of a relatively high incidence of SARS-CoV-2 infection in Malta, we decided to sequence these viral strains to contribute to the local and global epidemiology of SARS-CoV-2. Rapid sequencing is important for surveillance of circulating viral strains and for prompt contact tracing (1). Viral RNA was obtained from nasopharyngeal swabs collected at the St. James Hospital (Sliema, Malta) in August and September 2020 (Table 1). The research described in this work was performed in adherence to the Declaration of Helsinki; no specific authorization was issued by the University of Malta Institutional Review Board (IRB), since samples were anonymized and no human data were used. Swabs were placed in 3 ml ImproViral viral preservative medium (VPM) (Improve Medical). Viral RNA was extracted with a MagCore viral nucleic acid extraction kit (RBC Bioscience; catalog number MVN400-04) using a MagCore Plus II automated nucleic acid extractor (RBC Bioscience) and was then tested by quantitative PCR (qPCR) to assess the presence of SARS-CoV-2 RNA using the GrandPerformance SARS-CoV-2 kit (TATAA Biocenter, Sweden) targeting the RNA-dependent RNA polymerase gene. Viral RNA from qPCR-positive samples was sequenced using the multiplex PCR amplicon sequencing approach developed by the ARTIC Network (https://www.protocols.io/view/ncov-2019-sequencing-protocol-v2-bdp7i5rn?version_warning=no&step=16.3). Primers for 98 overlapping amplicons were used in two multiplex PCRs to amplify the whole viral genome, except for the conserved 5′ and 3′ untranscribed regions (UTRs). The sequencing library was prepared using the native barcode kit EXP-NBD104 from Oxford Nanopore Technologies (ONT) and sequenced on an R9.4 flow cell using a MinION MK1B device (ONT). The sequencing run was managed with MinKNOW v19.12.5, disabling live base calling. Raw Nanopore reads were base called using the GridION Guppy module v4.0.11, enabling the high-accuracy mode. Base-called reads were analyzed following the ARTIC Network pipeline (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html). All tools were run using default parameters unless otherwise specified. In total, 2,645,399 reads were obtained (mean, 264,540; range, 73,135 to 468,322 per sample).

Table 1 summarizes the information for each sample, including genome coverage

Citation Biazzo M, Madeddu S, Elnifro E, Sultana T, Muscat J, Scerri CA, Santoro F, Pinzauti D. 2021. Genome sequences of 10 SARS-CoV-2 viral strains obtained by Nanopore sequencing of nasopharyngeal swabs in Malta. Microbiol Resour Announc 10:e01375-20. https://doi.org/10.1128/MRA.01375-20.

Editor Simon Roux, DOE Joint Genome Institute

Copyright © 2021 Biazzo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Francesco Santoro, santorof@unisi.it.

Received 2 December 2020
Accepted 6 January 2021
Published 28 January 2021

TABLE 1 Data for each Malta/Sliema SARS-CoV-2 isolate

Strain name	Sampling date (day/mo/yr)	Symptoms[a]	qPCR threshold cycle[b]	Genome size (no. of bases) /coverage (%)	GC content (%)	Gapped regions[c]	No. of sequenced reads	Avg sequencing depth (×)	GenBank accession no.	SRA accession no.
Malta/Sliema_1/2020	19/08/2020	NA	22.4	29,782/99.6	37.98	1–54, 29837–29903	451,721	5,700	MW079418	SRR12778135
Malta/Sliema_2/2020	04/09/2020	–	29.5	29,782/99.6	37.98	1–54, 29837–29903	381,291	4,800	MW079419	SRR12778134
Malta/Sliema_3/2020	14/08/2020	+	33.3	29,782/94	37.86	1–54, 1313–1596, 19276–19576, 21147–21387, 27512–28105, 28757–29007, 28757–29007, 29837–29903	178,878	2,200	MW079420	SRR12778133
Malta/Sliema_4/2020	07/09/2020	–	26.6	29,782/99.6	37.98	1–54, 29837–29903	387,947	4,900	MW079421	SRR12778132
Malta/Sliema_5/2020	20/08/2020	–	21.6	29,782/99.6	37.98	1–54, 29837–29903	468,322	5,900	MW079422	SRR12778131
Malta/Sliema_6/2020	07/09/2020	–	23.7	29,782/98.6	37.96	1–54, 16187–16487, 29837–29903	146,139	2,500	MW079423	SRR12789615
Malta/Sliema_7/2020	05/09/2020	NA	25.6	29,782/95.6	37.94	1–54, 2248–2851, 5260–5287, 16187–16444, 16805–17087, 29837–29903	199,902	2,900	MW079424	SRR12789614
Malta/Sliema_8/2020	07/09/2020	+	23.5	29,611/93.5	37.92	1–54, 7672–7968, 8636–8913, 9558–9806, 16187–16444, 17431–17697, 24146–24416, 29666–29903	73,135	1,200	MW079425	SRR12789613
Malta/Sliema_9/2020	07/09/2020	+	24.3	29,611/98.3	37.92	1–54, 12235–12439, 29666–29903	173,524	2,800	MW079426	SRR12789612
Malta/Sliema_10/2020	07/09/2020	+	25.2	29,782/89.4	38.06	1–54, 2432–2850, 4427–4658, 5260–5287, 7390–7651, 8636–8913, 9246–9806, 17431–17706, 20119–20200, 23833–24721, 29837–29903	184,540	2,200	MW079427	SRR12789611

[a] +, present; –, absent; NA, information not available.

[b] Obtained with the GrandPerformance SARS-CoV-2 kit (TATAA Biocenter, Sweden), targeting the RdRP gene.

[c] With reference to the Wuhan Hu-1 genome. All samples miss regions 1 to 54 and 29837 to 29903, which are not covered by the primers.

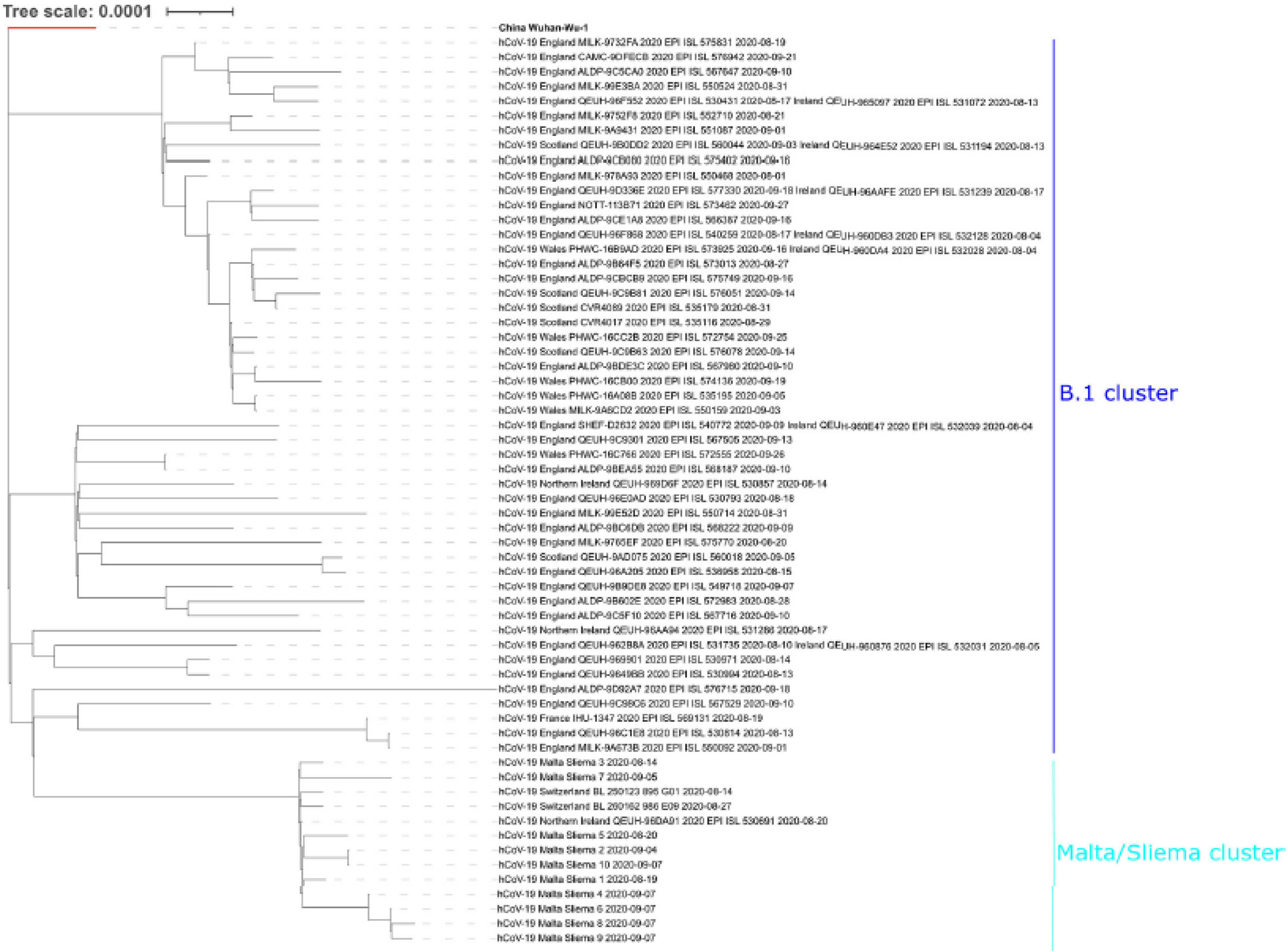

Tree scale: 0.0001
China Wuhan-Wu-1
hCoV-19 England MILK-9732FA 2020 EPI ISL 575831 2020-08-19
hCoV-19 England CAMC-9DFECB 2020 EPI ISL 576942 2020-09-21
hCoV-19 England ALDP-9C5CA0 2020 EPI ISL 567647 2020-09-10
hCoV-19 England MILK-99E3BA 2020 EPI ISL 550524 2020-08-31
hCoV-19 England QEUH-96F552 2020 EPI ISL 530431 2020-08-17 Ireland QEUH-965097 2020 EPI ISL 531072 2020-08-13
hCoV-19 England MILK-9752F8 2020 EPI ISL 552710 2020-08-21
hCoV-19 England MILK-9A9431 2020 EPI ISL 551087 2020-09-01
hCoV-19 Scotland QEUH-9B0DD2 2020 EPI ISL 560044 2020-09-03 Ireland QEUH-964E52 2020 EPI ISL 531194 2020-08-13
hCoV-19 England ALDP-9CB080 2020 EPI ISL 575402 2020-09-16
hCoV-19 England MILK-978A93 2020 EPI ISL 550468 2020-08-01
hCoV-19 England QEUH-9D336E 2020 EPI ISL 577330 2020-09-18 Ireland QEUH-96AAFE 2020 EPI ISL 531239 2020-08-17
hCoV-19 England NOTT-113B71 2020 EPI ISL 573462 2020-09-27
hCoV-19 England ALDP-9CE1A8 2020 EPI ISL 566387 2020-09-16
hCoV-19 England QEUH-96F868 2020 EPI ISL 540259 2020-08-17 Ireland QEUH-960D83 2020 EPI ISL 532128 2020-08-04
hCoV-19 Wales PHWC-16B9AD 2020 EPI ISL 573925 2020-09-16 Ireland QEUH-960DA4 2020 EPI ISL 532028 2020-08-04
hCoV-19 England ALDP-9B64F5 2020 EPI ISL 573013 2020-08-27
hCoV-19 England ALDP-9CBCB9 2020 EPI ISL 575749 2020-09-16
hCoV-19 Scotland QEUH-9C9B81 2020 EPI ISL 576051 2020-09-14
hCoV-19 Scotland CVR4089 2020 EPI ISL 535179 2020-08-31
hCoV-19 Scotland CVR4017 2020 EPI ISL 535116 2020-08-29
hCoV-19 Wales PHWC-16CC2B 2020 EPI ISL 572754 2020-09-25
hCoV-19 Scotland QEUH-9C9B63 2020 EPI ISL 576078 2020-09-14
hCoV-19 England ALDP-9BDE3C 2020 EPI ISL 567980 2020-09-10
hCoV-19 Wales PHWC-16CB00 2020 EPI ISL 574136 2020-09-19
hCoV-19 Wales PHWC-16A08B 2020 EPI ISL 535195 2020-09-05
hCoV-19 Wales MILK-9A8CD2 2020 EPI ISL 550159 2020-09-03
hCoV-19 England SHEF-D2632 2020 EPI ISL 540772 2020-09-09 Ireland QEUH-960E47 2020 EPI ISL 532039 2020-08-04
hCoV-19 England QEUH-9C9301 2020 EPI ISL 567505 2020-09-13
hCoV-19 Wales PHWC-16C766 2020 EPI ISL 572555 2020-09-26
hCoV-19 England ALDP-9BEA55 2020 EPI ISL 568187 2020-09-10
hCoV-19 Northern Ireland QEUH-969D6F 2020 EPI ISL 530857 2020-08-14
hCoV-19 England QEUH-96E0AD 2020 EPI ISL 530793 2020-08-18
hCoV-19 England MILK-99E52D 2020 EPI ISL 550714 2020-08-31
hCoV-19 England ALDP-9BC6D8 2020 EPI ISL 568222 2020-09-09
hCoV-19 England MILK-9765EF 2020 EPI ISL 575770 2020-08-20
hCoV-19 Scotland QEUH-9AD075 2020 EPI ISL 560018 2020-09-05
hCoV-19 England QEUH-96A205 2020 EPI ISL 536958 2020-08-15
hCoV-19 England QEUH-9B9DE8 2020 EPI ISL 549718 2020-09-07
hCoV-19 England ALDP-9B602E 2020 EPI ISL 572983 2020-08-28
hCoV-19 England ALDP-9C5F10 2020 EPI ISL 567716 2020-09-10
hCoV-19 Northern Ireland QEUH-96AA94 2020 EPI ISL 531288 2020-08-17
hCoV-19 England QEUH-962B8A 2020 EPI ISL 531735 2020-08-10 Ireland QEUH-960876 2020 EPI ISL 532031 2020-08-05
hCoV-19 England QEUH-969901 2020 EPI ISL 530971 2020-08-14
hCoV-19 England QEUH-9649B8 2020 EPI ISL 530994 2020-08-13
hCoV-19 England ALDP-9D82A7 2020 EPI ISL 576715 2020-09-18
hCoV-19 England QEUH-9C98C6 2020 EPI ISL 567529 2020-09-10
hCoV-19 France IHU-1347 2020 EPI ISL 569131 2020-08-19
hCoV-19 England QEUH-96C1E8 2020 EPI ISL 530814 2020-08-13
hCoV-19 England MILK-9A673B 2020 EPI ISL 560092 2020-09-01
hCoV-19 Malta Sliema 3 2020-08-14
hCoV-19 Malta Sliema 7 2020-09-05
hCoV-19 Switzerland BL 250123 895 G01 2020-08-14
hCoV-19 Switzerland BL 260162 986 E09 2020-08-27
hCoV-19 Northern Ireland QEUH-96DA91 2020 EPI ISL 530691 2020-08-20
hCoV-19 Malta Sliema 5 2020-08-20
hCoV-19 Malta Sliema 2 2020-09-04
hCoV-19 Malta Sliema 10 2020-09-07
hCoV-19 Malta Sliema 1 2020-08-19
hCoV-19 Malta Sliema 4 2020-09-07
hCoV-19 Malta Sliema 6 2020-09-07
hCoV-19 Malta Sliema 8 2020-09-07
hCoV-19 Malta Sliema 9 2020-09-07
B.1 cluster
Malta/Sliema cluster

FIG 1 Phylogenetic analysis of Malta/Sliema SARS-CoV-2 genomes. The phylogenetic tree was generated using IQ-TREE v1.6.12 and visualized using iTOL (6). A total of 63 viral genomes are displayed, including (i) 10 genomes from Malta, (ii) 2 related genomes, Switzerland BL 2501123 and BL 2601162, (iii) 50 genomes from GISAID randomly selected among samples obtained in Europe in August and September 2020, and (iv) the SARS-CoV-2 reference Wuhan Hu-1 genome (red branch) as an outgroup. All genomes except Hu-1 belong to the B.1 cluster (blue line). The Malta/Sliema genomes cluster together with Switzerland BL 2501123 and BL 2601162 and Northern Ireland QEUH-96DA9 (light blue line), since they share the 15 nucleotide variations reported in the text. For each viral genome, the strain name and place and date of sampling are indicated.

and accession numbers of raw and assembled data. Four out of 10 strains (Sliema_1, Sliema_2, Sliema_4, and Sliema_5) were assembled into a near-complete genome. The six remaining samples presented gaps (from 1.7 to 10.6% of the total genome length) in the assembled sequences, associated with low or absent amplicon coverage. Compared to the reference Wuhan Hu-1 genome, a set of 15 nucleotide variations was identified, shared among all Malta/Sliema strains. Single nucleotide variations were at positions 241, 2416, 3037, 8371, 9430, 14408, 15477, 18395, 23403, 23730, 25563, 26319, 28854, and 29044, and a mutation of 3 consecutive nucleotides (AGA → TTT) was at position 20622. Variations were confirmed by visual inspection using Tablet (2). Phylogenetic analysis with the Pangolin tool v2.0.8 (3) assigned the Malta/Sliema samples to lineage B1. IQ-TREE v1.6.12 (4) was used to reconstruct a phylogenetic tree as follows: a set of representative SARS-CoV-2 genomes was downloaded from the GISAID database (https://www.gisaid.org/), including samples (i) isolated in Europe, (ii) assigned to the B1 lineage, and (iii) collected between August 1 and September 30. To date (13 October 2020), a total of 14,990 sequences have been downloaded. A total of 50 representative genomes were randomly selected using the tool seqtk (v 1.3) (https://github.com/lh3/seqtk). The samples Switzerland/BL/250123_895_G01/2020-08-14 and Switzerland/BL/260162_986_E09/2020-08-27 were added to the phylogenetic analysis, since BLAST analysis of the *Betacoronavirus* NCBI resource indicated high similarity with the Sliema genomes. Using the cat command from the Linux environment, representative genomes and Sliema samples were merged and then aligned to the reference Wuhan Hu-1 genome using the MAFFT algorithm v7.471 (5). The resulting alignment was used by IQ-TREE to reconstruct the phylogenetic tree with an ultrafast bootstrap value of 1,500 (6). The tree file was finally visualized using the Interactive Tree of Life (iTOL) Web tool (7) (Fig. 1). The Malta genomes all belonged to a separate subcluster, indicating that viral strains circulating in Malta in September 2020 were genetically homogenous. Genomic epidemiology is essential for monitoring the emergence and spread of SARS-CoV-2 variants.

Data availability. The Nanopore reads and genome sequences of the SARS-CoV-2 Malta/Sliema isolates are publicly available. The Sequence Read Archive (SRA) and GenBank accession numbers are listed in Table 1.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, Curran MD, Parmar S, Caller LG, Caddy SL, Khokhar FA, Yakovleva A, Hall G, Feltwell T, Forrest S, Sridhar S, Weekes MP, Baker S, Brown N, Moore E, Popay A, Roddick I, Reacher M, Gouliouris T, Peacock SJ, Dougan G, Török ME, Goodfellow I. 2020. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis 20:1263–1272. https://doi.org/10.1016/S1473-3099(20)30562-4.
2. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform 14:193–202. https://doi.org/10.1093/bib/bbs012.
3. Rambaut A, Holmes EC, Hill V, O'Toole Á, McCrone JT, Ruis C, Du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv https://doi.org/10.1101/2020.04.17.046086.
4. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
5. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066. https://doi.org/10.1093/nar/gkf436.
6. Hoang DT, Chemomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35:518–522. https://doi.org/10.1093/molbev/msx281.
7. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.

4.3 Sequencing of 15 SARS-CoV-2 strains obtained at the beginning and at the end of the first pandemic wave in Siena

The sequencing protocol developed in (5) was initially tested and validated on the Siena-1/2020 SARS-CoV-2 strain isolated from cell culture. The next step was the study of its suitability and accuracy for direct use on RNA from nasopharyngeal swabs, without the need of culture. The possibility of generating sufficient genomic data directly from the swabs represents an important goal to reduce time prior SARS-CoV-2 profiling and helping in the local surveillance, tracking lineages spread.

Fourteen SARS-CoV-2 positive nasopharyngeal swabs, collected between March and June 2020 at the University Hospital of Siena (Italy), were selected. A strain seeded on Vero E6 cells (Italy/Siena-2/2020) was also introduced as a positive control. The Nanopore sequencing libraries were prepared as previously described (5), performing the sequencing run on a GridION x5 device. Sequencing yields are reported in Table 1.

Ten isolates were successfully reconstructed into a near-complete genome (99.7%) missing only the 5' and 3' UTR regions, which are not covered by the primers. The five remaining samples were instead characterized by the presence of one or more gaps (from 4% up to 12% of the total genome length) resulting in an incomplete genome. The phylogenetic analysis, performed by Pangolin tool v 2.0.8 (6), showed that the majority of SARS-CoV-2 isolates (9 out of 15) belong to B.1 lineage, a large European lineage associated to the Italian outbreak and characterized by four single nucleotide mutations in position 241, 3037, 14408, and 23403. Two isolates were assigned to B.1.1 lineage, a sub-cluster of the B.1 shearing the same changes and additionally harbouring a 3-consecutive nucleotide mutation in position 28881. The isolates Italy/Siena_18 and Italy/Siena_19 were assigned to a UK lineage B.1.98: single nucleotide changes at positions 241, 2416, 3037, 14408, 22021, 23403, 25552, 25563, 27649 and 29362 were identified. The Italy/Siena-2 isolate (positive control) was assigned to the B.1.97 lineage, another English lineage which harbours additional mutations in the ORF1ab polyprotein (positions 13568 and 15017). The remaining Italy/Siena_9 isolate was associated with the UK lineage B.1.1.35; this isolates is characterized by a 3-nucleotide change in position 28801, furthermore including a 9 bp deletion at position 685 (non coding region). A phylogenetic representation was obtained using IQ-TREE v1.6.12 (7). From the GISAID database (https://www.gisaid.org/) were downloaded representative SARS-CoV-2 genomes accounting for samples (i) isolated in Europe, (ii) assigned to B.1 lineage, and (iii) collected between March 1 and May 31: a total of 15,737 sequences were downloaded. A set of 50 representative genomes were randomly selected using the tool seqtk v 1.3

(https://github.com/lh3/seqtk) and, using the cat command, were merged with the 15 Siena isolates. Genomes were then aligned to the reference Wuhan Hu-1 genome using MAFFT algorithm v 7.471 (8) and used by IQ-TREE to reconstruct the phylogenetic tree with an ultrafast bootstrap of 1,500 (9). The tree was finally visualized using the Interactive Tree of Life (iTOL) web tool (10) (Figure 1).

In conclusion, the sequencing protocol has proven to be suitable also for nasopharyngeal swabs, generating near-complete SARS-CoV-2 genomic sequences. Its application during the first wave in Siena has enabled the identification of circulating SARS-CoV-2 lineages. The screening of local samples collected from outpatients and from COVID units showed no significant genomic differences among circulating lineages. Few optimizations in the protocol are still required: the lack of sufficient genome coverage or uneven amplicon synthesis may result in the impossibility to achieve complete genome sequences. Moreover further primer optimizations are also required, increasing their affinity for uncovered regions such as the 21444-22135 region.

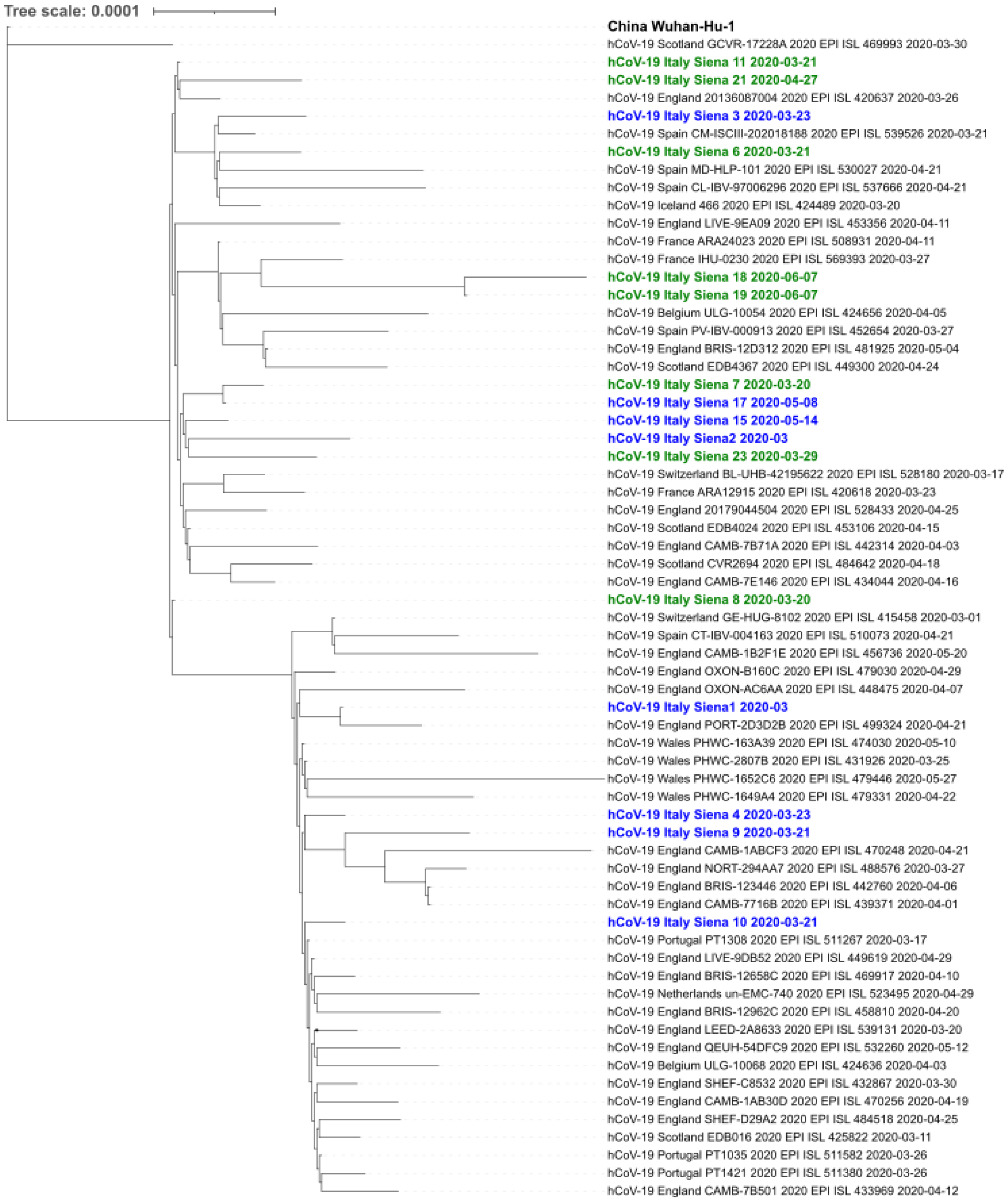
Table 1 Sequencing results

Strain name	Sampling date	Source	Genome size (coverage)	GC %	Gapped regions[a]	Lineage	Sequenced reads	Average sequencing depth	GISAID	GenBank Accession	Raw data accession
Italy/Siena-1/2020	2020-03	VeroE6	29,901 (99.99%)	37.97	1-2	B.1.1	40,374	1,153.64x	EPI_ISL_582123	MT531537	PRJNA658490
Italy/Siena-2/2020	2020-03	VeroE6	29,806 (99.67%)	38.0	1-30; 29836-29903	B.1.97	23,223	600x	EPI_ISL_58353	MW134558	SRR12847874
Italy/Siena_10/2020	2020-03-21	COVID unit	29,803 (99.66%)	38.01	1-33; 29836-29903	B.1.1	5,478	130x	EPI_ISL_58354	MW134559	SRR12847873
Italy/Siena_11/2020	2020-03-21	Outpatient	29,830 (99.75%)	38.02	1-4; 29836-29903	B.1	974,819	24,000x	EPI_ISL_58355	MW134560	SRR12847867
Italy/Siena_18/2020	2020-06-07	Outpatient	29,786 (99.60%)	37.99	1-30; 29836-29903	B.1.98	5,227	130x	EPI_ISL_583956	MW134561	SRR12847868
Italy/Siena_19/2020	2020-06-07	Outpatient	29,840 (99.78%)	37.99	1-27; 29867-29903	B.1.98	675,665	18,000x	EPI_ISL_583957	MW134562	SRR12847866
Italy/Siena_21/2020	2020-04-27	Outpatient	29,833 (99.76%)	38.0	1-3; 29836-29903	B.1	4,803	110x	EPI_ISL_583958	MW134563	SRR12847865
Italy/Siena_4/2020	2020-03-23	ICU	29,829 (99.75%)	38.0	1-30; 29866-29903	B.1.1	40,638	1,000x	EPI_ISL_583959	MW134564	SRR12847864
Italy/Siena_6/2020	2020-03-21	Outpatient	29,781 (99.59%)	38.0	1-53; 29836-29903	B.1	6,596	160x	EPI_ISL_583960	MW134565	SRR12847863
Italy/Siena_7/2020	2020-03-20	Outpatient	29,838 (99.78%)	38.01	1-30; 29872-29903	B.1	62,353	1,160x	EPI_ISL_583961	MW134566	SRR12847862

Italy/Siena_8/2020	2020-03-20	Outpatient	29,825 (99.73%)	38.0	1-30; 29864-29903	B.1	7,020	170x	EPI_ISL_583962	MW134567	SRR12847861
Italy/Siena_3/2020	2020-03-23	ICU	27,664 (92.51%)	38.17	1-54; 1579-2321; 5359-6020; 21444-22135; 29852-29903	B.1	3,183	75x	EPI_ISL_583963	MW134568	SRR12847860
Italy/Siena_9/2020	2020-03-21	Emergency room	29,132 (97.42%)	38.08	1-30; 21444-22136; 29866-29903	B.1.1.35	4,466	100x	EPI_ISL_583964	MW134569	SRR12847872
Italy/Siena_15/2020	2020-05-14	COVID unit	29,133 (97.42%)	38.01	1-30; 21448-22106; 29866-29903	B.1	83,475	2,100x	EPI_ISL_583965	MW134570	SRR12847871
Italy/Siena_17/2020	2020-05-08	COVID unit	26,295 (87.93%)	37.84	1-54; 1579-2342; 10688-11285; 12181-12818; 21444-22133; 26857-26900; 28352-28952; 29866-29903	B.1	1,731	45x	EPI_ISL_583966	MW134571	SRR12847870
Italy/Siena_23/2020	2020-03-29	Outpatient	28,427 (95.06%)	38.12	1-30; 17460-18165; 21444-22106; 29866-29903	B.1	3,425	90x	EPI_ISL_583967	MW134572	SRR12847869

[a] With reference to the Wuhan Hu-1 genome. All samples miss regions 1-30 and 29866-29903, which are not covered by the primers

Figure 1 Phylogenetic analysis of Italy/Siena SARS-CoV-2 genomes


Tree scale: 0.0001
China Wuhan-Hu-1
hCoV-19 Scotland GCVR-17228A 2020 EPI ISL 469993 2020-03-30
hCoV-19 Italy Siena 11 2020-03-21
hCoV-19 Italy Siena 21 2020-04-27
hCoV-19 England 20136087004 2020 EPI ISL 420637 2020-03-26
hCoV-19 Italy Siena 3 2020-03-23
hCoV-19 Spain CM-ISCIII-202018188 2020 EPI ISL 539526 2020-03-21
hCoV-19 Italy Siena 6 2020-03-21
hCoV-19 Spain MD-HLP-101 2020 EPI ISL 530027 2020-04-21
hCoV-19 Spain CL-IBV-97006296 2020 EPI ISL 537666 2020-04-21
hCoV-19 Iceland 466 2020 EPI ISL 424489 2020-03-20
hCoV-19 England LIVE-9EA09 2020 EPI ISL 453356 2020-04-11
hCoV-19 France ARA24023 2020 EPI ISL 508931 2020-04-11
hCoV-19 France IHU-0230 2020 EPI ISL 569393 2020-03-27
hCoV-19 Italy Siena 18 2020-06-07
hCoV-19 Italy Siena 19 2020-06-07
hCoV-19 Belgium ULG-10054 2020 EPI ISL 424656 2020-04-05
hCoV-19 Spain PV-IBV-000913 2020 EPI ISL 452654 2020-03-27
hCoV-19 England BRIS-12D312 2020 EPI ISL 481925 2020-05-04
hCoV-19 Scotland EDB4367 2020 EPI ISL 449300 2020-04-24
hCoV-19 Italy Siena 7 2020-03-20
hCoV-19 Italy Siena 17 2020-05-08
hCoV-19 Italy Siena 15 2020-05-14
hCoV-19 Italy Siena2 2020-03
hCoV-19 Italy Siena 23 2020-03-29
hCoV-19 Switzerland BL-UHB-42195622 2020 EPI ISL 528180 2020-03-17
hCoV-19 France ARA12915 2020 EPI ISL 420618 2020-03-23
hCoV-19 England 20179044504 2020 EPI ISL 528433 2020-04-25
hCoV-19 Scotland EDB4024 2020 EPI ISL 453106 2020-04-15
hCoV-19 England CAMB-7B71A 2020 EPI ISL 442314 2020-04-03
hCoV-19 Scotland CVR2694 2020 EPI ISL 484642 2020-04-18
hCoV-19 England CAMB-7E146 2020 EPI ISL 434044 2020-04-16
hCoV-19 Italy Siena 8 2020-03-20
hCoV-19 Switzerland GE-HUG-8102 2020 EPI ISL 415458 2020-03-01
hCoV-19 Spain CT-IBV-004163 2020 EPI ISL 510073 2020-04-21
hCoV-19 England CAMB-1B2F1E 2020 EPI ISL 456736 2020-05-20
hCoV-19 England OXON-B160C 2020 EPI ISL 479030 2020-04-29
hCoV-19 England OXON-AC6AA 2020 EPI ISL 448475 2020-04-07
hCoV-19 Italy Siena1 2020-03
hCoV-19 England PORT-2D3D2B 2020 EPI ISL 499324 2020-04-21
hCoV-19 Wales PHWC-163A39 2020 EPI ISL 474030 2020-05-10
hCoV-19 Wales PHWC-2807B 2020 EPI ISL 431926 2020-03-25
hCoV-19 Wales PHWC-1652C6 2020 EPI ISL 479446 2020-05-27
hCoV-19 Wales PHWC-1649A4 2020 EPI ISL 479331 2020-04-22
hCoV-19 Italy Siena 4 2020-03-23
hCoV-19 Italy Siena 9 2020-03-21
hCoV-19 England CAMB-1ABCF3 2020 EPI ISL 470248 2020-04-21
hCoV-19 England NORT-294AA7 2020 EPI ISL 488576 2020-03-27
hCoV-19 England BRIS-123446 2020 EPI ISL 442760 2020-04-06
hCoV-19 England CAMB-7716B 2020 EPI ISL 439371 2020-04-01
hCoV-19 Italy Siena 10 2020-03-21
hCoV-19 Portugal PT1308 2020 EPI ISL 511267 2020-03-17
hCoV-19 England LIVE-9DB52 2020 EPI ISL 449619 2020-04-29
hCoV-19 England BRIS-12658C 2020 EPI ISL 469917 2020-04-10
hCoV-19 Netherlands un-EMC-740 2020 EPI ISL 523495 2020-04-29
hCoV-19 England BRIS-12962C 2020 EPI ISL 458810 2020-04-20
hCoV-19 England LEED-2A8633 2020 EPI ISL 539131 2020-03-20
hCoV-19 England QEUH-54DFC9 2020 EPI ISL 532260 2020-05-12
hCoV-19 Belgium ULG-10068 2020 EPI ISL 424636 2020-04-03
hCoV-19 England SHEF-C8532 2020 EPI ISL 432867 2020-03-30
hCoV-19 England CAMB-1AB30D 2020 EPI ISL 470256 2020-04-19
hCoV-19 England SHEF-D29A2 2020 EPI ISL 484518 2020-04-25
hCoV-19 Scotland EDB016 2020 EPI ISL 425822 2020-03-11
hCoV-19 Portugal PT1035 2020 EPI ISL 511582 2020-03-26
hCoV-19 Portugal PT1421 2020 EPI ISL 511380 2020-03-26
hCoV-19 England CAMB-7B501 2020 EPI ISL 433969 2020-04-12

A total of 67 viral genomes are displayed, including (i) 16 genomes from Siena, (ii) 50 representative B.1 genomes randomly selected from the GISAID database, and (iv) the SARS-CoV-2 Wuhan Hu-1 reference genome. Outpatients samples are represented in green, while samples collected from COVID units in blue.

Data Availability

The genomic sequences and the raw sequencing reads of all the 15 isolates were deposited in GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) databases. Genomic data were also deposited in the GISAID EpiCov (https://www.gisaid.org/) database contributing to global surveillance.

4.4 References

1. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses., Gorbalenya, A.E., Baker, S.C. et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 5, 536–544 (2020). https://doi.org/10.1038/s41564-020-0695-z.

2. Wang H, Li X, Li T, Zhang S, Wang L, Wu X, Liu J. The genetic sequence, origin, and diagnosis of SARS-CoV-2. Eur J Clin Microbiol Infect Dis. 2020 Sep;39(9):1629-1635. Doi: 10.1007/s10096-020-03899-4.

3. Ahmad A T N, Kisa F, Taj M, Urooj F, Indrakant K S, Archana S, Shaikh M A, Gururao H, Gulam M H, Imtaiyaz H. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, Volume 1866, Issue 10, 2020.doi: 10.1016/j.bbadis.2020.165878.

4. Candido D S, Claro I M, de Jesus J G, Souza W M, Moreira F R R, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020 Sep 4;369(6508):1255-1260. doi: 10.1126/science.abd2161.

5. Cusi MG, Pinzauti D, Gandolfo C, Anichini G, Pozzi G, Santoro F. 2020. Whole genome sequence of SARS-CoV-2 isolate Siena-1/2020. Microbiol Resour Announc9:e00944-20. https://doi.org/10.1128/MRA.00944-20.

6. Rambaut A, Holmes EC, Hill V, O'Toole Á, McCrone JT, Ruis C, Du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv https://doi.org/10.1101/2020.04.17.046086.

7. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.

8. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066. https://doi.org/10.1093/nar/gkf436.

9. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35:518–522. https://doi.org/10.1093/molbev/msx281.

10. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.

Chapter 5. Conclusions

Oxford Nanopore sequencing is rapidly renewing the entire microbiology field. Its ease of use and cost effective applications, coupled with long-read sequencing, represent an interesting solution in both research and diagnostic fields. In the present thesis, I have proposed the installation and implementation of an Oxford Nanopore-based sequencing platform, a useful and flexible system providing benefits in bacterial genomic and bacterial profiling applications.

The platform has been proven to be extremely efficient in complete bacterial *de novo* genome assemblies, upon high molecular weight DNA isolation. Moreover coupling accurate Illumina reads with long Nanopore reads further improves the quality of the final assembly. The possibility of performing population studies as well as antimicrobial resistance profiling, represents an extremely important application of such platform, providing rapid and real-time results helping clinicians to define efficient treatment. Its latest involvement monitoring SARS-CoV-2 circulating lineages, has also pointed out the importance of relying on rapid and effective tools enabling pathogen surveillance and tracking pathogen spread. The GridION x5 platform, which is able to run 5 flow cells simultaneously, can improve the number of tested samples, decreasing overall costs and saving time. Further improvements are required in order to simplify and automate those approaches, even for non-expert users and make nanopore sequencing suitable for routine-like diagnostic applications.

Appendix 1 - Scientific curriculum vitae of David Pinzauti

Work experience

January 2021 – Present

Research fellowship at the Department of Medical Biotechnologies, University of Siena, for the "Characterization of antibiotic resistance determinants in bacterial genomes by sequencing and bioinformatic analysis"

Education

September 2017 - Present

PhD programme in Medical Biotechnologies – Department of Medical Biotechnologies, University of Siena, Italy.

Main areas of interest: BACTERIAL GENOMICS, WHOLE GENOME SEQUENCING, NGS, DATA ANALYSIS

October 2015 - September 2017

Master's Degree in Medical Biotechnologies, Department of Medical Biotechnologies, University of Siena, Italy.

Mark: 100/110 cum laude

Thesis: "Bacterial Genome Sequencing using the MinION Nanopore Sequencer"

October 2012 - October 2015

Bachelor's Degree in Medical-Pharmaceutical Biotechnologies, University of Florence, Italy.

Mark: 103/110

Thesis: "*In vitro* screening of novel antimicrobial compounds"

2011 - 2012

Scientific High School Diploma – Liceo Scientifico Piero Gobetti, Florence

Training courses

- 2020, May. "**Metagenomics applied to surveillance of pathogens and antimicrobial resistance**" on-line course Coursera platform, organized by Technical University of Denmark – DTU, Denmark.
- 2020, April. "**Whole genome sequencing of bacterial genomes – tool and application**" on-line course Coursera platform, organized by Technical University of Denmark – DTU,

Denmark.

- 2019, May. "**Introduction to Machine Learning algorithms**" ALMALE (2018DU0092), organized by the University of Siena, project "Tuscan Start-Up Academy 4.0", Regione Toscana funds.

Languages

ITALIAN: native
ENGLISH: very good knowledge of English language (written and spoken)
2017: **B2 English qualification**, Centro Linguistico Ateneo (CLA),
University of Siena, Italy

Technical skills

Very good knowledge of standard and molecular microbiology techniques and excellent knowledge of DNA extraction protocols, focusing on the recovery of High Molecular Weight genomic DNA. Excellent knowledge of Next Generation Sequencing (NGS) technologies, focusing on bacterial Whole Genome Sequencing, Targeted Sequencing and Metagenomics approaches. I have developed excellent skills and knowledge of Oxford Nanopore sequencing technology and I am very proficient in sequencing library preparation starting from genomic DNA, amplicons or viral samples. Excellent knowledge of the data analysis methods for raw NGS data: genome assembly, identification of virulence and antimicrobial resistance genes, identification of genomic variations and mutations, comparative genomics, and phylogenetics.

Computer skills

Excellent knowledge of Windows operating system, Microsoft Office and LibreOffice packages. Good knowledge of softwares for image manipulation such as GIMP or Inkscape. Excellent knowledge of Linux operating system focusing on Debian and Ubuntu distributions. I am very proficient in the definition of virtual environments for data analysis and in the installation of bioinformatic tools. I have acquired a basic knowledge of the **python** programming language and good knowledge of its related tools and repositories (**GitHub** and **Conda**). Basic knowledge of **R software**, mainly as regards graphical representation packages (**ggplot2**). Excellent knowledge of databases (e.g. NCBI, CARD, VFDB, Argannot, Transposon Registry) and websites for data analysis.

List of Publications

- Genome Sequences of 10 SARS-CoV-2 Viral Strains Obtained by Nanopore Sequencing of Nasopharyngeal Swabs in Maltahole. Manuele Biazzo, Silvia Madeddu, Elfath Elnifro, Tessabella Sultana, Josie Muscat, Christian A. Scerri, Francesco Santoro, **David Pinzauti**. *Microbiol Resour Announc*. 2021 January 28. doi: 0.1128/MRA.01375-20
- Whole-Genome Sequence of SARS-CoV-2 Isolate Siena-1/2020. Maria Grazia Cusi, **David Pinzauti**, Claudia Gandolfo, Gabriele Anichini, Gianni Pozzi, Francesco Santoro. *Microbiol Resour Announc*. 2020 September 24. doi: 10.1128/MRA.00944-20.

Conferences

- 2020, 1-3 December. Nanopore Community Meeting, virtual conference organized by Oxford Nanopore Technologies, UK
- 2020, 21-22 September. 48° National Congress of Italian Microbiology Society (SIM), virtual conference, Italy. "Whole genome sequence of the SARS-CoV-2 Siena-1/2020 viral isolate". Maria Grazia Cusi, David Pinzauti, Claudia Gandolfo, Gabriele Anichini, Gianni Pozzi, Francesco Santoro. **Poster**.
- 2020, 17-18 June. London Calling 2020, virtual conference organized by Oxford Nanopore Technologies, UK.
- 2019, 19-22 June. XXXIII SIMGBM Congress, Microbiology 2019, Florence, Italy. "DNA isolation methods for nanopore sequencing of *Streptococcus mitis* genome". David Pinzauti, Francesco Iannelli, Gianni Pozzi, Francesco Santoro. **Poster**.
- 2019, 19-22 June. XXXIII SIMGBM Congress, Microbiology 2019, Florence, Italy. "Complete Genome Sequence of *Lactobacillus crispatus* M247 strain and its derivative Mu5 lacking the auto-aggregation phenotype". Lorenzo Colombini, Francesco Santoro, Anna Maria Cuppone, David Pinzauti, Gianni Pozzi, Francesco Iannelli. **Poster**.
- 2019, 22-24 June. London Calling 2019, organized by Oxford Nanopore Technologies, London, UK. "DNA isolation methods for nanopore sequencing of *Streptococcus mitis* genome". David Pinzauti, Francesco Iannelli, Gianni Pozzi, Francesco Santoro. **Poster**.
- 2018, 26-27 September. 46° National Congress of Italian Microbiology Society (SIM), Palermo, Italy. "Complete genome sequence of *Streptococcus mitis* integrating Nanopore and Illumina data". David Pinzauti, Francesco Iannelli, Gianni Pozzi, Francesco Santoro. **Poster**.

Deposited Sequences

GenBank:

- *Streptococcus mitis* strain S022-V3-A4, complete genome. GenBank Accession no. CP047883, BioProject ID PRJNA208927. Pinzauti D, Iannelli F, Pozzi G, and Santoro F. 2021
- *Streptococcus mitis* strain S022-V7-A4, complete genome. GenBank Accession no. CP067992, BioProject ID PRJNA690734.. Pinzauti D, Iannelli F, Pozzi G, and Santoro F. 2020
- *Staphylococcus lugdunensis* strain MBAZ2, complete genome. GenBank Accession no. CP060160, BioProject ID PRJNA656409. Biazzo M, Taddei A R, Zollo A, Santoro F, and Pinzauti D. 2020
- *Mammaliicoccus fleurettii* strain ssch2, complete genome. GenBank Accession no. CP064058, BioProject ID PRJNA674009. Bidossi A, Pinzauti D, and Santoro F. 2020
- *Mammaliicoccus fleurettii* strain ssch3, complete genome. GenBank Accession no. CP064059, BioProject ID PRJNA674010. Bidossi A, Pinzauti D, and Santoro F. 2020
- Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ITA/Siena/2020 15 viral isolates, complete genomes. GenBank Accession no. MW134558-MW134572, Bioproject ID PRJNA669459. Cusi M G, Pinzauti D, Gandolfo C, Anichini G, Pozzi G, and Santoro F. 2020
- Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/MLT/Sliema/2020 10 viral isolates, complete genome. GenBank Accession no. MW079418-MW079427, Bioproject ID PRJNA667434. Biazzo M, Madeddu S, Santoro F, and Pinzauti D. 2020
- Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ITA/Siena-1/2020, complete genome. GenBank Accession no. MT531537. Cusi M G, Pinzauti D, Gandolfo C, Anichini G, Pozzi G, and Santoro F. 2020

Sequence Read Archive:

- *Streptococcus mitis* strain S022-V3-A4 whole genome sequencing. BioProject ID PRJNA690894. Pinzauti D, Iannelli F, Pozzi G, and Santoro F. 2021
- *Streptococcus mitis* strain S022-V7-A4 whole genome sequencing. BioProject ID PRJNA690894. Pinzauti D, Iannelli F, Pozzi G, and Santoro F. 2021
- *Mammaliicoccus fleurettii* strain ssch2 Genome sequencing and assembly. BioProject ID PRJNA674009. Bidossi A, Pinzauti D, and Santoro F. 2020
- *Mammaliicoccus fleurettii* strain ssch3 Genome sequencing and assembly. BioProject ID

PRJNA674010. Bidossi A, Pinzauti D, and Santoro F. 2020

- PCR-tiling and sequencing of SARS-CoV-2 genomes directly from clinical samples collected in Siena, Italy. BioProject ID PRJNA669459. Cusi M G, Pinzauti D, Gandolfo C, Anichini G, Pozzi G, and Santoro F. 2020
- PCR-tiling and sequencing of SARS-CoV-2 genomes directly from clinical samples collected in Sliema, Malta. BioProject ID PRJNA667434. Biazzo M, Madeddu S, Santoro F, and Pinzauti D. 2020
- Genome sequence of viral SARS-CoV-2 Siena-1 viral isolate. BioProject ID PRJNA658490. Cusi M G, Pinzauti D, Gandolfo C, Anichini G, Pozzi G, and Santoro F. 2020