UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA

Towards accreditation in metagenomics for clinical microbiology

Catarina Inês Marques de Sousa Mendes

Orientador: Doutor João André Nogueira Custódio Carriço
Co-orientador: Professor Doutor Mário Nuno Ramos de Almeida Ramirez

Documento provisório
Tese especialmente elaborada para obtenção do grau de Doutor em Ciências e Tecnologias da Saúde, especialidade em Biologia Computacional

2022

UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA

Towards accreditation in metagenomics for clinical microbiology

Catarina Inês Marques de Sousa Mendes

Orientador: Doutor João André Nogueira Custódio Carriço
Co-orientador: Professor Doutor Mário Nuno Ramos de Almeida Ramirez

Documento provisório
Tese especialmente elaborada para obtenção do grau de Doutor em Ciências e Tecnologias da Saúde, especialidade em Biologia Computacional

Fundação para a Ciência e Tecnologia
SFRH/BD/129483/2017 and COVID/BD/152583/2022

2022

"The greatest adventure is what lies ahead.
Today and tomorrow are yet to be said.
The chances, the changes are all yours to make.
The mould of your life is in your hands to break."

-J. R. R. Tolkien, The Hobbit

Acknowledgements

Summary

Keywords: one, two, three, four, five

Resumo

Keywords: um, dois, três, quatro, cinco

Thesis Outline

The work described in the present thesis intended to XXXXX.

The thesis comprises X chapters, organised as follows:

Chapter 1 corresponds to the general introduction that highlights the ...

Chapter 2 is dedicated to ...

Chapter 3 consists in ...

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8 corresponds to the general discussion. In this chapter is provided a summary of the main results obtained in this thesis and its integrated discussion.

Chapter 9 Contains the main conclusions driven from this work. It also includes perspectives for future work.

Abbreviation

Table of Contents

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1

General Introduction

1.1 The global impact of microbial pathogens

The Global Burden of Disease (GBD) 2019 study reported that microbial pathogens are responsible for more than 400 million years of life lost annually across the globe, a higher burden than either cancer or cardiovascular disease [1]. In particular, lower respiratory infections, diarrhoeal diseases, HIV/AIDS and tuberculosis were amongst the five leading causes of global total years of life lost. More recently, the COVID-19 pandemic, declared as such by the World Health Organization (WHO) on 11 March 2020 after the emergence and global spread of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and as of January 2022, has caused more than 5.63 million deaths worldwide [2], making it one of the deadliest pandemics in history. Coronavirus has been responsible for three of the eighteen major pandemics registered throughout modern history [3], all occurring after the year 2000. *Yersinia pestis*, responsible for three pandemics of plague, *Vibrio cholerae*, with seven cholera pandemics, and Influenza A virus, the causative agent of five flu pandemics, are responsible for the remaining, with Influenza being the only other pathogen with a pandemic registered after 2000. Recent decades have also witnessed the emergence of additional virulent pathogens, including the Ebola virus, West Nile virus, Dengue virus and Zika virus, particularly in lower-income countries.

In addition to the emergence of virulent pathogens, the rise of antimicrobial resistance (AMR) poses a major threat to human health around the world. In 2019 there were an estimated 4.95 million deaths associated with bacterial AMR [4]. In 2017, the WHO released The Global Priority Pathogens (GPP) list [5] to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections (see Figure 1.1). Besides tuberculosis,

the global priority due to being the most common and lethal airborne AMR disease worldwide today, responsible for 250 000 deaths each year, it includes 12 groups of pathogens in three priority categories.

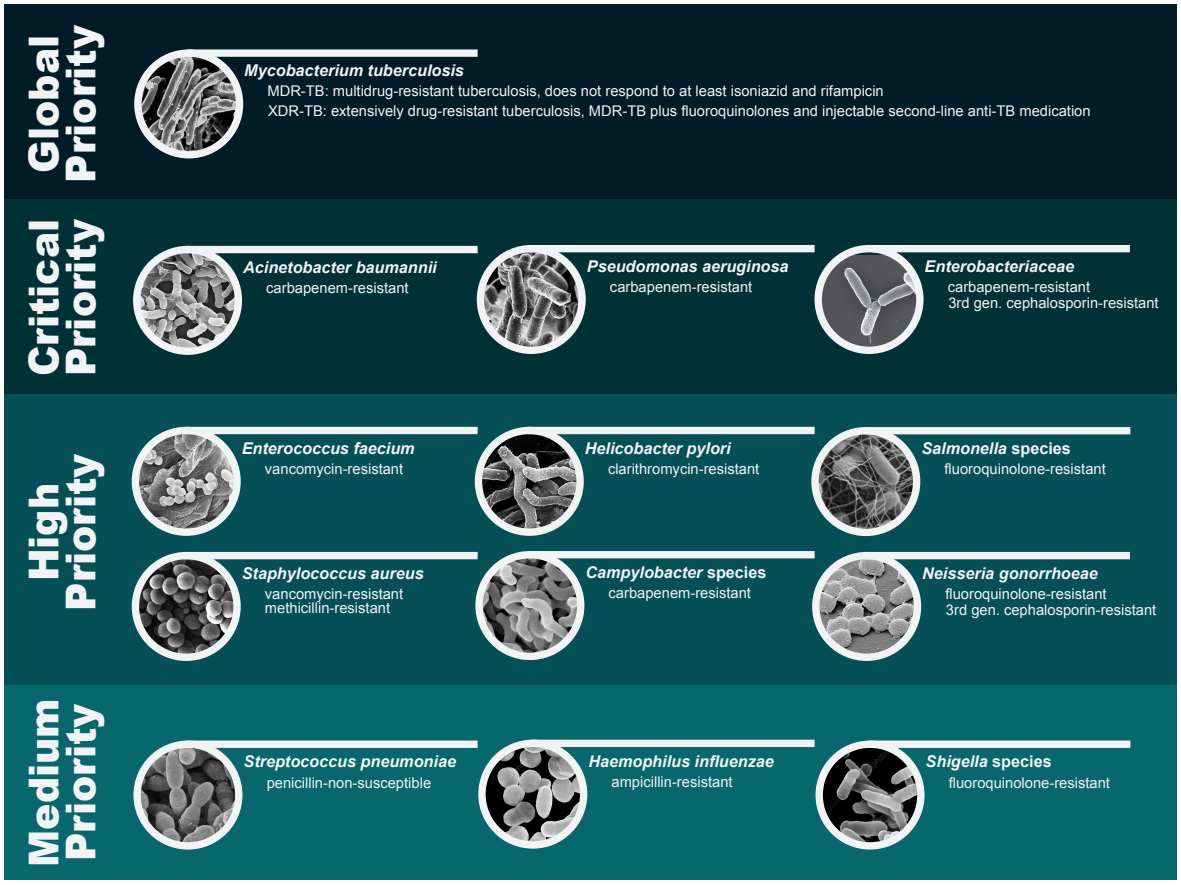

Figure 1.1: **World Health Organisation Global Priority Pathogens list.** This catalogue includes, besides *Mycobacterium tuberculosis* considered the number one global priority, a list of twelve microorganisms grouped under three priority tiers according to their antimicrobial resistance: critical (*Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacteriaceae*), high (*Enterococcus faecium*, *Helicobacter pylori*, *Salmonella* species, *Staphylococcus aureus*, *Campylobacter* species and *Neisseria gonorrhoeae*), and medium (*Streptococcus pneumoniae*, *Haemophilus influenzae* and *Shigella* species). The major objective was to encourage the prioritisation of funding and incentives, align research and development priorities of public health relevance, and garner global coordination in the fight against antimicrobial resistant bacteria. Adapted from [5].

Clinical microbiology is a discipline focused on rapidly characterising pathogen samples to direct the management of individual infected patients (diagnostic microbiology) and monitor the epidemiology of infectious disease (public health microbiology), including the detection of outbreaks and infection prevention. According to WHO's Global Expenditure on Health report from 2000 to 2019, of the 51 countries that reported health spending by disease and condition, an average of 37% of health spending went to infectious and parasitic diseases, corresponding to the largest share of health spending [6]. About 21% of total health spending went to three major infectious diseases — HIV/AIDS (9%), tuberculosis (1%) and malaria (11%) — and 16% went to other infectious and parasitic diseases. On average, 70% of external aid for health went to infectious and parasitic diseases in the 51 low and middle-income countries. Of the $54.8 billion estimated disbursed for health in 2020, $13.7 billion (25%) was targeted toward the COVID-19 health response [7].

1.1.1 Current standards for diagnostic in clinical microbiology

The past few decades have seen a major revolution in the operation of microbial laboratories, driven by the development of molecular technologies and ways to make these accessible, namely amplification-based polymerase chain reaction (PCR), matrix-assisted laser desorption/ionisation - time of flight (MALDI-TOF) and DNA-microarray-based hybridisation technology. These are used in conjunction with traditional techniques such as microscopy, culture and serology, not fully replacing them. Application of these methods differs by suspected infection type: bacterial, viral, fungal or parasitic. For the purpose of this dissertation work, we will be focusing on bacterial and viral infections.

1.1.1.1 Bacterial infections

For patients with bacterial infections, the crucial steps are (1) to grow an isolate from a specimen, (2) identify its species, and (3) determine its pathogenic potential and test its susceptibility to antimicrobial drugs [8]. Together this information facilitates the specific and rational treatment of patients. For public health purposes, knowledge also needs to be gained about (4) the relatedness of the pathogen to other strains of the same species to investigate transmission routes and enable the recognition of outbreaks [9] (see Figure 1.2).

The current gold standard for bacterial pathogen identification in diagnostic microbiology laboratories involves the isolation of the pathogen through culture followed by biochemical testing, a multi-step process that can take days to weeks before obtaining results, depending on the fastidiousness of the organism and if it can be cultured [10, 11]. Although culture allows the identification of a wide variety of organisms, some pathogens can escape routine investigation due to strict metabolic necessities for growth or the requirement for specific biochemical tests needed for their identification. Additionally, results will be obscured if a mixed culture is obtained, particularly if the cultures are obtained from sites with a microbiota, such as the gut and the skin, increasing the risk of contamination from normal flora, and leading to false results [11]. After successful growth in culture, Gram staining and MALDI-TOF mass spectrometry are often used for identification with good accuracy as long as the pathogen is presented in the the coexisting database [12]. An alternate rapid identification method is PCR where nucleic acid fragments are detected through specific primers, being highly sensitive and specific, to the point where PCR may detect bacteria that are not viable after a patient has been treated for an infection and it is limited to the primer used [13]. Syndromic panels, an extension of PCR by using multiple primers (multiplex PCR) to simultaneously amplify the nucleic acids of multiple targets in a single reaction, tried to address this issue by allowing for the identification of multiple bacteria and other important information such as the detection of antibiotic resistance or virulence genes [11]

Following identification, antibiotic-susceptibility testing is essential for guiding clini-

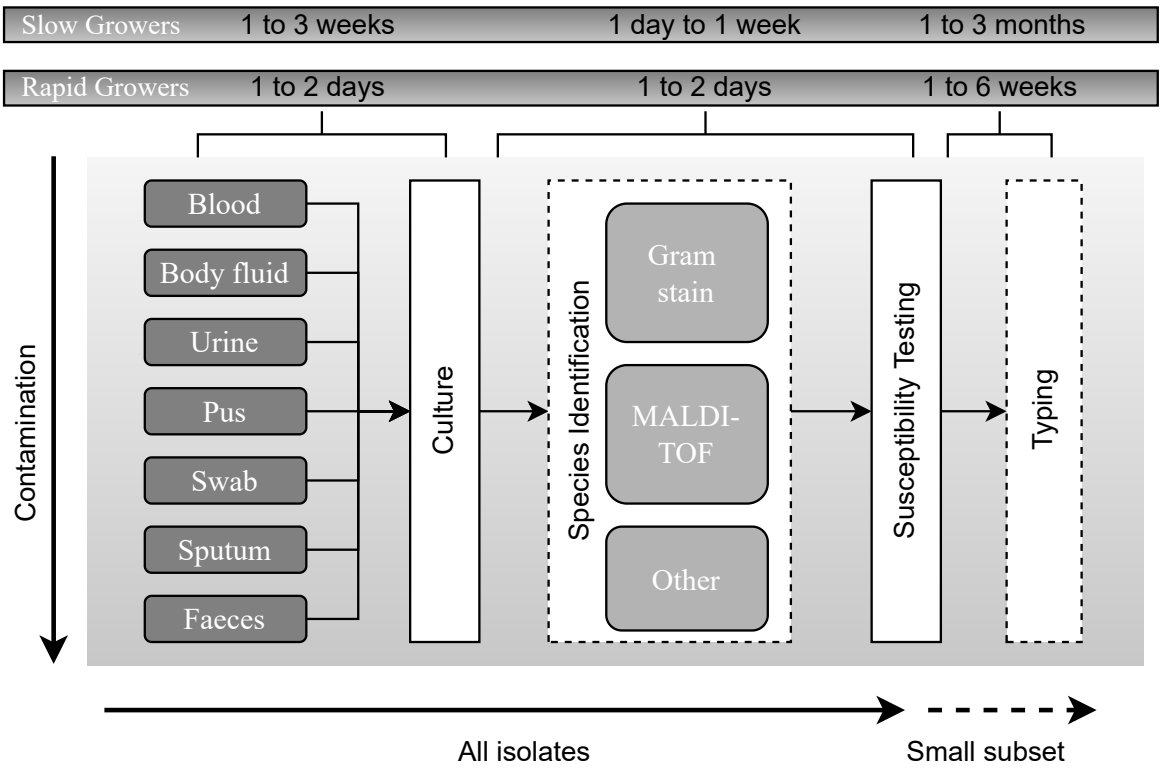

Figure 1.2: **Principles of current processing of bacterial pathogens.** Schematic representation of the current workflow for processing samples for bacterial pathogens is presented, with high complexity and a typical timescale of a few weeks to a few months. Samples that are likely to be normally sterile are often cultured on rich medium that will support the grown of any culturable organism. Samples contaminated with colonising flora present a challenge for growing the infecting pathogen. Many types of culture media (referred to as selective media) are used to favour the growth of the suspected pathogen. Once an organism is growing, the likely pathogens are then processed through a complex pathway that has many contingencies to determine species and antimicrobial susceptibility. Broadly, there are two approaches. One approach uses MALDI-TOF for species identification prior to setting up susceptibility testing. The other uses Gram staining followed by biochemical testing to determine species; susceptibility testing is often set up simultaneously with doing biochemical tests. Lastly, depending on the species and perceived likelihood of an outbreak, a small subset of isolates may be chosen for further investigation using a wide range of typing tests. Adapted from [8]

cians in selecting an appropriate treatment. Conventional detection methods of bacterial resistance, such as disc diffusion, antimicrobial gradient strip and broth microdilution, are widely used but results cannot be obtained earlier than 48 hours after receiving a sample, which may lead to prolonged use or overuse of broad-spectrum antibiotics [14]. Similarly to bacterial identification, MALDI-TOF and PCR have been increasingly adopted as solutions with lower turnaround times, although no phenotypic information is retrieved, nor information on the minimum inhibitory concentration (MIC) for a given antibiotic.

Choosing an appropriate bacterial typing technique for epidemiological studies depends on the resources available and the minimum intended resolution, ranging from DNA fingerprinting to multilocus sequence typing, Pulsed-field gel electrophoresis (PFGE) and sequence-based typing (see section 1.2. A genomic approach to clinical microbiology) [9, 15]. DNA macrorestriction analysis by PFGE, which revolutionised precise separation of DNA fragments, became the most widely implemented DNA fingerprinting technique [15],

becoming the golden standard for bacterial typing [16].

In the early 2000s, Multilocus sequence typing (MLST) was proposed as a portable, universal, and definitive method for characterising bacteria [17]. Instead of enzyme restriction of bacteria DNA, separation of the restricted DNA bands using a PFGE chamber, followed by clonal assignment of bacteria based on banding patterns, MLST relies on the amplification through PCR sequences of internal fragments of housekeeping genes (usually 5 to 7), approximately 450-500 basepairs (bp) in size, followed by its the sequence, usually my Sanger methods (see subsubsection 1.2.1.1. The first-generation of DNA sequencing). For each house-keeping gene, the different sequences present within a bacterial species are assigned as distinct alleles and, for each isolate, the alleles at each of the (usually) seven loci define the allelic profile or sequence type [18]. As with PFGE, different schemes, defining what house-keeping gene fragments are used, are available depending on the species. Unlike PFGE, the provision of freely accessible, curated databases of MLST nucleotide sequence data enables the direct comparison of bacterial isolates, providing the basis of a common language for bacterial typing [17]. So far, MLST schemes for 115 bacterial organisms have been published and made freely available[1], [19])

Depending on the organism identified, further and/or particular typing schemes can be applied. For *S. pneumoniae*, one of the pathogens listed in the WHO's GPP list, the typing of the polysaccharide capsule, usually through Quellung reaction, is paramount for disease surveillance and pre- and post-pneumococcal vaccine evaluation as the capsule, with over 90 serotypes reported, is the dominant surface structure of the organism and plays a critical role in virulence [20, 21]. For the *Salmonella* species, also in the GPP list, the serotype is usually determined by agglutination of the bacteria with specific antisera to identify variants of somatic (O) and flagella (H) antigens that, in various combinations, characterise more than 2600 reported serotypes [22].

1.1.1.2 Viral infections

The traditional approaches to the laboratory diagnosis of viral infections have been (1) direct detection in patient material of virions, viral antigens, or viral nucleic acids, (2) isolation of the virus in cultured cells, followed by identification of the isolate, and (3) detection and measurement of antibodies in the patient's serum (serology) [23]. Viral diagnostics is therefore generally organised into two primary categories, indirect and direct detection, depending on the method used.

Indirect detection methods involve the propagation of virus particles via their introduction to a suitable host cell line (virus isolation), as viruses rely on host organisms to replicate. This is a relatively slow diagnostic method, sometimes taking weeks for the virus to propagate, usually followed by microscopy for its identification, or more commonly, through

[1] `https://pubmlst.org/organisms`

molecular methods with an agent which detects a virus-associated protein, such as an antibody [24].

Direct detection methods negate the need for virus propagation, detecting the virus directly from the suspect source through nucleic acid and immunological methods. PCR and reverse transcription-PCR (RT-PCR) are widely applied methods for the detection of both DNA and RNA viruses, respectively, driven by increased awareness of the clinical value of, and demand for, prompt information about viral loads, viral sequence data, and potential antiviral resistance information [24]. Syndromic testing (see subsubsection 1.1.1.1. Bacterial infections) is now fully integrated into the standard testing practices of many clinical laboratories [25]. Limitations of these assays include no detection of off-target pathogens, a lack of full susceptibility information, cost, and false-positive results. Real-time quantitative PCR (qPCR) remains the front line tool in aetiological diagnosis, measuring the production of the target amplicon throughout the reaction and providing quantitative results with high specificity and sensibility, albeit with a significant cost due to sophisticated apparatus despite high-throughput systems being widely established [24].

Immunoassays employ singular-epitope specificity antibodies as the primary means to detect viruses within a sample and provide a much more cost-efficient alternative to nucleic acid detection [24]. One major application is seroprevalence assays, an essential technique for identifying patients who have been exposed to a virus (historical exposure), detecting asymptomatic infection or evaluating vaccine efficacy [26, 27]. Lateral flow immunoassays (LFIA) are extensively used for detecting virus-associated protein directly from the source through labelled antibodies binding to their cognate antigens, usually read by way of a colour change at a test line. Besides being very cost-effective, LFAIs have a turnaround time of minutes and the colour change can be observed with the naked eye, therefore facilitating rapid diagnosis but its results are limited to semi-quantitative and it does not typically achieve sensitivity comparable to nucleic-acid detection [24, 28, 29].

1.1.2 Surveillance and infection prevention in public health

Infectious disease surveillance is critical for improving population health, generating information that drives action not only in the management of infected patients but also in the prevention of new ones by identifying emerging health conditions that may have a significant impact by (1) describing the current burden and epidemiology of the disease, (2) to monitoring trends, and (3) identifying outbreaks and new pathogens [30, 31]. Public health surveillance systems (PHSS) are composed of the ongoing systematic collection, analysis, and interpretation of data, and its integration with the timely dissemination of results to those who can undertake effective prevention and control activities [32].

Traditional PHSS can have different approaches based on the epidemiology and clinical presentation of the disease and the goals of surveillance. In passive surveillance systems,

medical professionals in the community and at health facilities report cases to the public health agency, which conducts data management and analysis once the data are received and communicate with the responsible entities. Globally, the WHO as described in the International Health Regulations what is notifiable by every country to WHO, such as Severe acute respiratory syndrome (SARS) and Viral haemorrhagic fevers (Ebola, Lassa, Marburg), as well as guiding what public health measures should be implemented [33]. Active surveillance aims to detect every case, not relying on a reporting structure, and can have many approaches from sentinel sites or network of sites that capture cases of a given condition, such as respiratory tract infections, within a catchment population [31, 34]. The application of environmental surveillance methods, performed prospectively to detect pathogens prior to the recording of clinical cases or to monitor their abundance in the environment to assess the potential risk of disease, has been proven as a viable alternative, particularly in wastewater [35–38].

The emergence and re-emergence of infectious diseases are closely linked to the biology and ecology of infectious agents, their hosts, and their vectors [39]. "One Health" is a collaborative and multi-disciplinary approach to designing and implementing programmes, policies, legislation and research in which multiple sectors communicate and work together to achieve better public health outcomes [40]. It recognises that people's health is closely connected to animals' health and shared environment, focusing on zoonotic and vector-borne diseases, antimicrobial resistance, food safety, food security and environmental contamination [41]. This is crucial to (1) understanding the emergence and re-emergence of infectious and non-communicable chronic diseases and (2) in creating innovative control strategies. A better knowledge of causes and consequences of certain human activities, lifestyles, and behaviours in ecosystems is crucial for a rigorous interpretation of disease dynamics and to drive public policies, but it requires breaking down the interdisciplinary barriers that still separate human and veterinary medicine from ecological, evolutionary, and environmental sciences [39].

1.2 A genomic approach to clinical microbiology

Since the publication of the first complete microbial genome a quarter of a century ago, that of the bacterium *Haemophilus influenzae* [42], genomics has transformed the field of microbiology, and in particular its clinical application (see Figure 1.3).

The paper describing the DNA-sequencing method with chain-terminating inhibitors used in the sequencing of the first microbial genome [43], which earned the late Frederick Sanger his share of the 1980 Nobel Prize in Chemistry alongside Walter Gilbert, was, in 2014, the top fourth in the number of citations with 60335, highlighting its impact in the field of biological sciences, and by extension medicine [44]. Currently, this number has increased

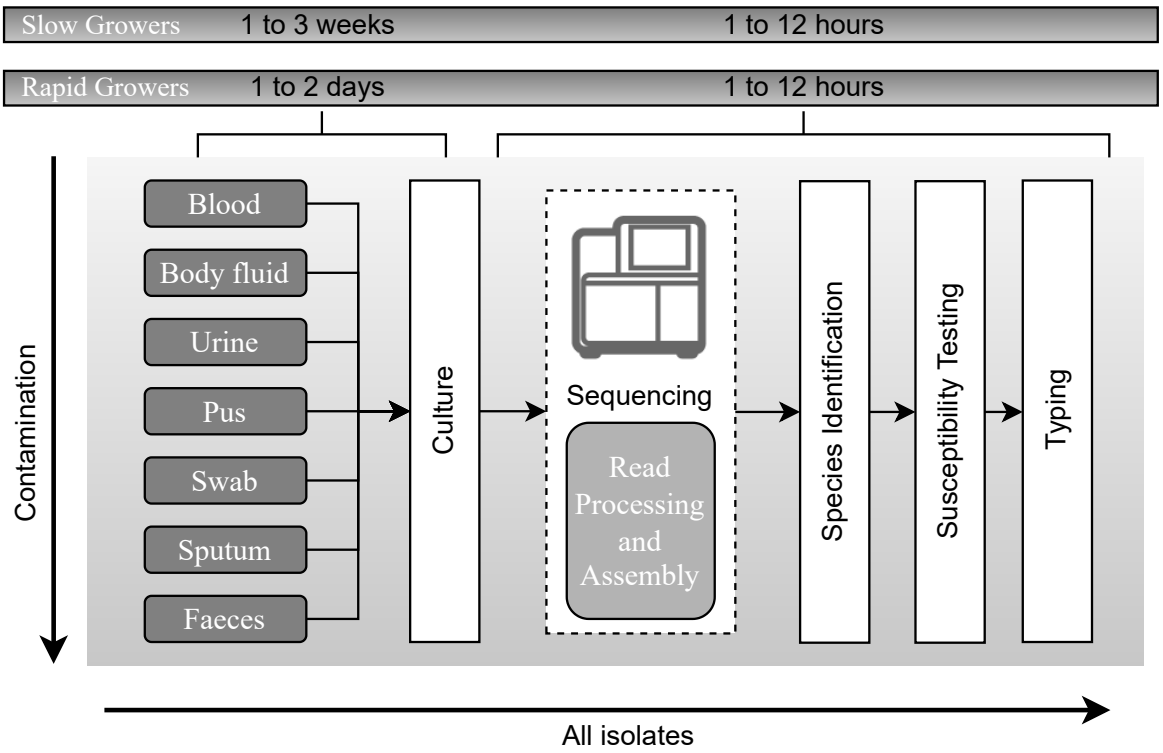

Figure 1.3: **Principles of current processing of bacterial pathogens based on whole genome sequencing.** Schematic representation of the workflow for processing samples for bacterial pathogens after adoption of whole genome sequencing, with an expected timescale that could fit within a single day. The culture steps would be the same as currently used in a routine microbiology laboratory. Once a likely pathogen is ready for sequencing, DNA will be extracted, taking as little as 2 hours to prepare the DNA for sequencing. After sequencing, the main processes for yielding information will be computational. Automated sequence assembly algorithms are necessary for processing the raw sequence data, from which species, relationship to other isolates of the same species, antimicrobial resistance profile and virulence gene content can be assessed. All the results will also be used for outbreak detection and infectious diseases surveillance. Adapted from [8]

to 84546 according to PubMed Central® (PMC)[23]. Since its emergence, reductions in cost, technical advances in sequencing technologies and new computational developments have made genomic sequencing one of the most influential tools in biomedical research, yielding unprecedented insights into microbial evolution and diversity, and the complexity of the genetic variation in both commensal and pathogenic microbes. The emerging application of genomic technologies in the clinic to combat infectious diseases is transforming clinical diagnostics and the detection and surveillance of outbreaks.

1.2.1 Twenty five years of microbial genome sequencing

Since the discovery of the structure of DNA [45], great strides have been made in understanding the complexity and diversity of genomes in health and disease. The development and commercialisation of high-throughput, massively parallel sequencing, has democratised

[2] `https://pubmed.ncbi.nlm.nih.gov/`
[3] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/`

sequencing by offering individual laboratories, either in research or in health, access to the technology. Over the last quarter of a century, three main revolutions can be considered in genomic sequencing (see Figure 1.4).

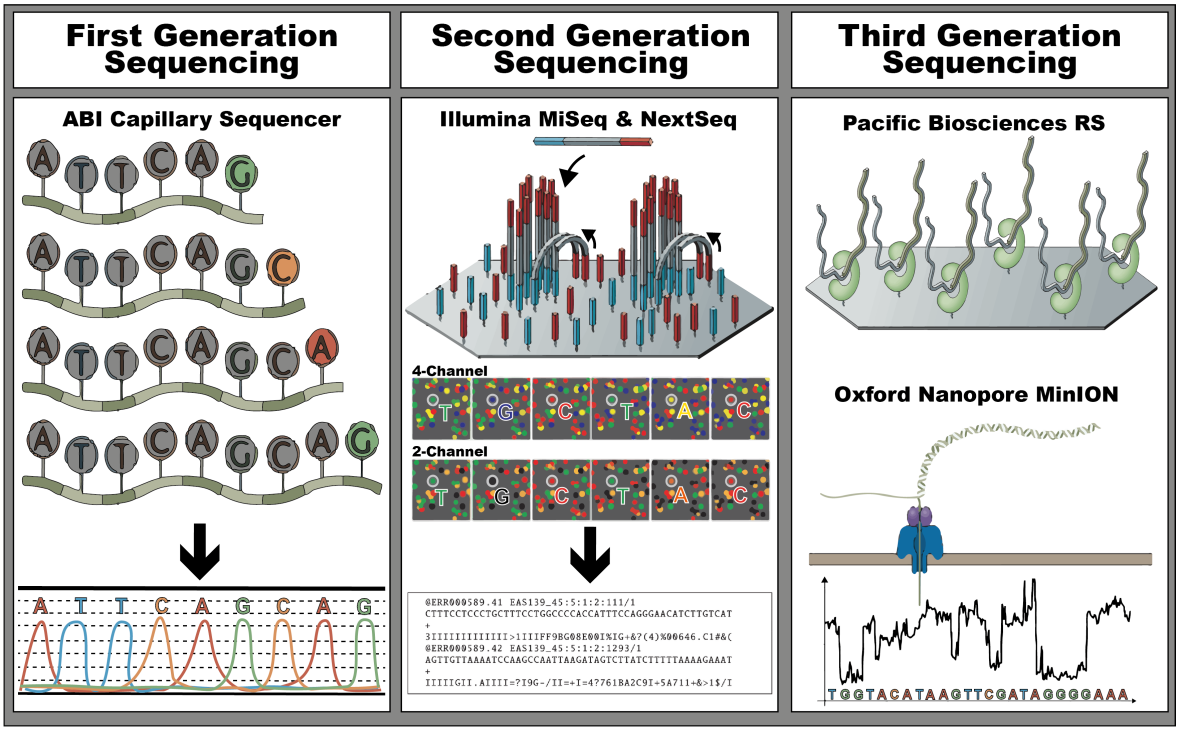

Figure 1.4: **The three revolutions in sequencing technology that have transformed the landscape of bacterial genome sequencing.** The first-generation, also known as Sanger sequencers, is represented by the ABI Capillary Sequencer (Applied Biosystems). During the sequencing reaction, at each nucleotide incorporation event a fluorescently labelled ddNTP is incorporated, terminating the elongation of the DNA molecule. The resulting electropherogram for sequencing reaction is below, and is read from left to right. The second-generation, also known as high-throughput sequencers, is represented by the MiSeq, a 4-channel sequencer, and the NextSeq, a 2-channel sequencer (Illumina), both sequencing by synthesis instruments. For both instruments, the loaded flowcell is sequenced in massive parallel reactions, with each nucleotide incorporation emitting a light signal that is captured and latter basecalled into a fastq file, with indication of the confidence of the call, presented bellow. In a 4-channel instrument each nucleotide has it's own marker (A: yellow, T: green, C: red, G: blue) but in a 2-channel instrument only 2 markers exist (A: green plus red, T: green, C: red, G: no marker). These instrument allow the sequencing of both ends of the DNA fragment. Lastly, the third-generation, also known as long-read sequencers, is represented by Pacific Bioscience BS sequencer and Oxford Nanopore MinION sequencer. In the first, immobilised polymerases in a SMRT Cell incorporating nucleotides with identifying fluorescent labels. In the latter, a nanopore embedded in a solid-state membrane causes a a change in an ionic current across the membrane each time a nucleotide is pushed though the pore. This difference in potential is then used for basecalling. Adapted from [46–51]

1.2.1.1 The first-generation of DNA sequencing

In the late 1980s, automated Sanger sequencing machines could sequence approximately 1,000 bases per day, having been applied in the 1990s to large bacterial genomes and the first unicellular and multicellular eukaryotic genomes, including the completion of a high-quality, reference sequence of the human genome under the Human Genome Project (HGP) [52, 53]. The first genomes of the pathogenic *Mycobacterium tuberculosis* [54], *Yersinia*

pestis [55], *Escherichia coli* K-12 [56] were sequenced using this technology, requiring years of effort and significant budgets but providing insights into the genomic complexity of these organisms. Some of the complete genome sequences produced during this era are still used today as high-quality references.

Simplistically, in Sanger sequencing, also known as "first-generation" DNA sequencing, a DNA polymerase is used to synthesize numerous copies of the sequence of interest using dideoxynucleotide triphosphates (ddNTPs) spiked into the reaction. At each nucleotide incorporation event, there is a chance that a ddNTP will be added and the growing DNA chain will be terminated, resulting in a collection of DNA molecules of varying lengths [43, 46]. Modern Sanger sequencing uses fluorescently labelled ddNTPs that allow the amplification step to be performed in a single reaction, resulting in a mixture of single-stranded DNA fragments of various lengths, each tagged at one end with a fluorophore indicating the identity of the 3' nucleotide that, after separation through capillary electrophoresis, the resulting electropherogram with four-colour fluorescence intensity can be interpreted by a base-calling software and producing 600–1000 bases of accurate sequence [46].

The Sanger sequencing technology remains very useful for applications where high-throughput is not required due to its cost-effectiveness, relatively low sample load and accuracy of sequencing even in repetitive genomic regions, although input DNA must consist of a relatively pure population of sequences [57]. One of the most common uses is thus individual sequencing reactions using a specific DNA primer on a specific template, such as MLST of bacterial genomes.

1.2.1.2 The second-generation of DNA sequencing

The release of the first truly high-throughput sequencing platform in the mid-2000s heralded a 50,000-fold drop in the cost of DNA sequencing in comparison with the first-generation technologies and led to the denomination of next-generation sequencing (NGS) [48]. This trend has continued throughout the next two decades of continued development and improvement, allied to the emergence of benchtop sequencing platforms with a high-throughput of sequencing data and turnaround times of days, making it a standard in any microbiology and public health laboratories [47]. Second-generation sequencing methods can be grouped into two major categories: (1) sequencing by hybridisation and (2) sequencing by synthesis.

1.2.1.2.1 Sequencing by hybridisation

Sequencing by hybridisation, also known as sequencing by ligation, originally developed in the 1980s, relies on the binding of one strand of DNA to its complementary strand (hy-

bridisation). By repeated hybridisation and washing cycles, it was possible to build larger contiguous sequence information, based upon overlapping information from the probe hybridisation spot, being sensitive to even single-base mismatches when the hybrid region is short or if specialised mismatch detection proteins are present [57, 58]. Although widely implemented via DNA chips or microarrays, has largely been displaced by other methods, including sequencing by synthesis [48].

1.2.1.2.2 Sequencing by synthesis

Sequencing by synthesis methods are a further development of Sanger sequencing, without the ddNTPs terminators, in combination with repeated cycles, run in parallel, of synthesis, imaging, and methods to incorporate additional nucleotides in the growing chain. All second-generation sequencing by synthesis approaches relies on a 'library' preparation using native or amplified DNA usually obtained through (1) DNA extraction, (2) DNA fragmentation and fragment size selection, and (3) ligation of adapters and optional barcodes to the ends of each fragment. This is generally followed by a step of DNA amplification. The resulting library is loaded on a flow cell and sequenced in massive parallel sequencing reactions [59] Besides having much shorter read lengths than first-generation methods, with reads ranging from 45 to 300 bases, and an intrinsically higher error rate, the massively parallel sequencing of millions to billions of short DNA sequence reads allows for the obtainment of millions of accurate sequences based upon the identification of a consensus (agreement) sequences [46, 48, 57].

Many of the currently available sequencing by synthesis methods approaches have been described as cyclic array sequencing platforms, as they involve dispersal of target sequences across the surface of a two-dimensional array, followed by sequencing of those targets [46]. They can be further classified as either single-nucleotide addition or cyclic reversible termination or as single-nucleotide addition [48].

The first relies on a single signal to mark the incorporation of a dNTP into an elongating strand, avoiding the use of terminators. As a consequence, each of the four nucleotides must be added iteratively to a sequencing reaction to ensure only one deoxynucleotide triphosphate (dNTP) is responsible for the signal. The Roche 454 Life Sciences pyrosequencing device [4], was the first and most popular instrument implementing this technology, but discontinued since 2013 with support to the platform ceasing since 2016. This system distributes template-bound beads into a PicoTiterPlate along with beads containing an enzyme cocktail. As a dNTP is incorporated into a strand, an enzymatic cascade occurs, resulting in a bio-luminescence signal which is captured by a camera, which can be attributed to the incorporation of one or more identical dNTPs at a particular bead [48]. The ThermoFisher Ion

[4] `https://web.archive.org/web/20161226040638/http://454.com/`, snapshot from 26 December 2016

Torrent system [5], released in 2010 and still available today, replaces the optical sensor, using instead H+ ions that are released as each dNTP is incorporated in the enzymatic cascade, and the consequential change in pH, to detect a signal [48]. Alongside the 454 pyrosequencing system, this system has difficulty in enumerating long repeats, additionally, the throughput of the method depends on the number of wells per chip, ranging from 10 megabases to 1000 megabases of 100 base reads in length, but with a very short run time (three hours) [46, 60].

The latter is defined by their use of terminator molecules that are similar to those used in the first-generation of sequencing, preventing elongation of the DNA molecule, but unlike the first methods, it is reversible. To begin the process, a DNA template is primed by a sequence that is complementary to an adapter region, which will initiate polymerase binding to this double-stranded DNA region. During each cycle, a mixture of all four individually labelled and 3'-blocked dNTPs are added. After the incorporation of a single dNTP to each elongating complementary strand, unbound dNTPs are removed and the surface is imaged to identify which dNTP was incorporated at each cluster by optical capture. The fluorophore and blocking group can then be removed and a new cycle can begin [48]. The Illumina systems, which use this technology, accounts for the largest market share for sequencing instruments compared to other platforms[6], allowing paired-end sequencing and having the highest throughput (from 25 million reads for a MiSeq instrument to 1.2 billion reads for a NextSeq instrument[7]), with read lengths ranging from 45 to 300 bases in length with high accuracy, albeit with long running times (4 to 55 hours), rendering this technology a good choice for many sequencing applications where large read length is not required [46, 60, 61].

1.2.1.3 The third-generation of DNA sequencing

Despite their wide adoption, second-generation methods require library preparation and an enrichment or amplification step. These steps are time-consuming, introduce biases related to preferential capture or amplification of certain regions, and produce reads with relatively small size, making transversing repetitive genomic regions impossible if they are larger than the read length [46]. Third-generation sequencing technologies, also known as long-read sequencing or single-molecule sequencing, are characterised by the generation of ultra-long-reads, albeit at a much lower throughput than the second-generation [62]. They also have the potential to go beyond four-base sequencing to reveal genome-wide patterns of methylation and other chemical modifications that control the biology of bacteria or the virulence of pathogens [63]. Currently, commercial long-read sequencing is supported by two companies: Pacific Biosciences[8] and Oxford Nanopore Technologies[9].

The basis of Pacific Biosciences sequencers is known as single-molecule real-time se-

[5] https://www.thermofisher.com/pt/en/home/brands/ion-torrent.html
[6] https://www.forbes.com/companies/illumina/?sh=774358a91aa6
[7] https://www.illumina.com/systems/sequencing-platforms.html
[8] https://www.pacb.com/
[9] https://nanoporetech.com/

quencing (SMRT), which takes place in single-use SMRT Cells. These contain multiple immobilised polymerases which, after binding to an adaptor sequence, begins replication incorporating nucleotides with identifying fluorescent labels. The sequence of fluorescence pulses is recorded into a movie which is then converted into a nucleotide sequence. After the polymerase completes replication of one DNA strand, it continues to sequence the opposite adapter and second strand. As a result, it is possible to generate multiple passes of the same template depending on the lifetime of the polymerase [47, 62]. This technology has accuracy comparable with the Illumina systems but requires a higher initial investment cost, are much larger machines in comparison with the benchtop counterparts, and have much lower throughput and longer library preparation protocols [62, 64].

Oxford Nanopore Technologies makes use of nanopores in small, portable single-molecule sequencing devices, capable of generating ultra-long sequences in real-time at a relatively low cost. Biological nanopores are embedded in solid-state membranes within disposable flow cells which, when a DNA strand passes through the pore driven by a motor protein, each nucleotide causes a change in an ionic current across the membrane, which is later base called [47, 62]. This process is free from fluorescence labels and amplification requirements, and after one strand is processed, the pore is available to sequence the next available strand. Sequence quality and length depend on the loaded library but are usually much lower than the alternative counterparts, and its throughput is dependent on the number and lifespan of the nanopore within the flowcell, but still much lower than the alternatives. Despite this, its portability, fast advances, and continued improvement of the flowcells make this a fast adopted technology for long-read sequencing.

1.2.2 DNA sequencing in clinical diagnosis and surveillance

Whole-genome sequencing (WGS) is becoming one of the most widely used applications of microbial genome sequencing. The major advantage of WGS is to yield all the available DNA information content on isolates in a single rapid step following culture (sequencing without culture will be discussed in the subsection 1.2.3. From genomics to metagenomics). In principle, after obtaining a pure culture, either bacterial (see subsubsection 1.1.1.1. Bacterial infections) or viral (see subsubsection 1.1.1.2. Viral infections), the data from sequencing contain all the information currently used for diagnostic and typing needs, and much more, thus opening the prospect for large-scale research into pathogen genotype-phenotype associations from routinely collected data [8]. The cost of producing massive amounts of information requires a new framework with expert handling and processing of computer-driven genomic information, as well as capable computational infrastructures, but through this technology, researchers and clinicians can obtain the most comprehensive view of genomic information and associated biological implications, transforming clinical diagnostics and the detection and surveillance of outbreaks. [48, 65, 66].

Targeted sequencing is also proving invaluable to clinical microbial and research, not only by allowing more individual samples to be sequenced within a single run, significantly reducing costs and the amount of data generated, but also, due to the smaller target size, obtaining results with very high confidence due to the high coverage obtained [48]. This has been particularly useful in viral genomics where sections, such as the capsid, or the complete viral genome can be selectively targeted directly from the suspected sample, offering a more time-effective method to achieve the same output as traditional nucleic acid amplification methods [24].

1.2.2.1 Sequencing in the routine laboratory workflow

WGS has been used in the routine laboratory workflow when typing of pathogens by a method having the highest possible discriminatory power is required either through single nucleotide polymorphism (SNP) or core-genome/whole genome MLST (cg/wg MLST) analysis, for example during hospital outbreaks [67]. Additionally, in bacterial diagnostics, WGS can be used to reveal the presence of AMR genes, or genes associated with virulence and pathogenicity, as well as to discover new genetic mechanisms for the three previously defined important clinical features of a bacterium [68]. The implementation of WGS in routine diagnostics requires several adaptations in the laboratory workflow, from the 'wet' laboratory part (extraction, library preparation, sequencing), to the 'dry' bioinformatics part where genomic data is analysed and its results interpreted by specialised personnel [68].

Currently, sequencing technologies are used in a case-by-case approach, with its adoption being much more present in a research setting than in a diagnostic one. Sequencing is mostly used after a diagnostic through the identification of the causative agent has already been performed. Although substantial advances have been made in reducing response time, most of the current systems do not yet generate enough data fast enough for a truly rapid response for it to be used in the clinical setting [48]. High-throughput DNA sequencing has found additional new applications in drug discovery and in functional genomics with, for example, SNP-based analysis to identify new drug targets [47].

Although the second-generation DNA sequencing methods have shed light on fundamental aspects of microbial ecology and function, they suffer from issues associated with short read length (see 1.2.1.2) and cannot reliably reconstruct long repeats because of uncertainties in mapping read, even when paired-end sequencing is used. Third-generation sequencing methods (see 1.2.1.3) have become increasingly used in microbiology, although their accuracy and low throughput make it challenging to implement in a clinical diagnostic setting.

1.2.2.2 Sequencing and genomic surveillance

Most notably, WGS has become a common tool in surveillance and infection prevention, allowing for pathogen identification and tracking, establishing transmission routes and outbreak control [69]. In bacterial infections, initiatives such as Pathogenwatch[10] offers a web-based platform for AMR analysis and phylogeny generation of *Campylobacter*, *Klebsiella*, *Neisseria gonorrhoeae*, *Staphylococcus aureus*, and *Salmonella Typhi* [70]. The Center for Genomic Epidemiology website[11] offers services for phylogenetic tree building and AMR prediction. Chewie Nomenclature Server[12] allows users to share genome-based gene-by-gene typing schemas and to maintain a common nomenclature, simplifying the comparison of results [71]. Enterobase[13] allows for the analysis and visualisation of genomic variation within enteric bacteria [72]. Microreact[14], from the same developers as Pathogenwatch, combines clustering, geographical and temporal data into an interactive visualisation with trees, maps, timelines and tables for a multitude of microorganisms, both bacterial and viral [73]. Particularly for viruses, GISAID[15] promotes the rapid sharing of data from all influenza viruses and the coronavirus causing COVID-19, including the genetic sequences and related clinical and epidemiological data [74]. ViPR[16] provides access to sequence records, gene and protein annotations, immune epitopes, 3D structures, host factor data, and other data types for over 14 viral families, including *Coronaviridae*, from which SARS-CoV-2 belongs to, and *Faviviridae*, the family of Dengue and Zika virus [75]. INSaFLU[17] supplies public health laboratories and influenza researchers with a web-based suite for effective and timely influenza and SARS-CoV-2 laboratory surveillance, identifying the type and subtype/lineage, detection of putative mixed infections and intra-host minor variants [76]. Nextrain[18] provide a continually-updated view of publicly available data alongside powerful analytic and visualisation tools o aid epidemiological understanding and improve outbreak response for 10 pathogens: Influenza, SARS-CoV-2, West Nile virus, Mumps, Zika, West African Ebola, Dengue, Measles, Enterovirus D68 and Tuberculosis [77]

In outbreak detection and surveillance, genetic sequencing techniques combined with epidemiological data have undoubtedly provided immeasurable insights regarding evolutionary relationships and transmission pathways in various environments [78, 79]. In a pandemic setting, this approach, although not novel, has been revolutionary, particularly in the COVID-19 setting.

In the 2009 swine-origin Influenza A H1N1 pandemic, the first complete genome was

[10] `https://pathogen.watch/`
[11] `https://www.genomicepidemiology.org/`
[12] `https://chewbbaca.online/`
[13] `https://enterobase.warwick.ac.uk/`
[14] `https://microreact.org/`
[15] `https://www.gisaid.org/`
[16] `https://www.viprbrc.org/`
[17] `https://insaflu.insa.pt/`
[18] `https://nextstrain.org/`

publicly available on the 25 of April of 2009 (GenBank accession number FJ966079), about a month after records of increased flu activity in Mexico and 10 days after the first confirmed cases in California, United States of America [80, 81]. By the time the pandemic was declared, on 11 of June of 2009, [80] reported the origins and evolutionary genomics of the pandemic influenza A variant with a collection of 813 complete influenza genome sets, 17 of which belonging to the newly swine influenza viruses (GenBank accessions numbers GQ229259–GQ229378). The MERS pandemic, declared as such in 2015 [3], had its first publicly available sequence on 5 of July 2015 (GenBank accession number KT006149)[82], with a sequence from a camel, thought to be an intermediate host for the virus, available as early as 7 of March 2016 (GenBank accession number KU740200) [83, 84].

The SARS-CoV-2 has brought a new meaning to genomic surveillance, with the first sequence from a COVID-19 patient being made publicly available as early as 12 January 2020 from a case of respiratory disease from the Wuhan outbreak (GenBank accession number MN908947) [85]. At the date of the pandemic declaration by WHO, at 11 March 2020, over 400 complete SARS-CoV-2 sequences were deposited on GISAID[19], hitting over one million sequences in April 2021 [86]. Currently, over 8 million complete viral sequences are available at GISAID[20], being one of the most highly sequenced genomes of any organism on the planet. This richness in genomic information has been basal to identifying new variants of risk and new variants of concern with a myriad of different origins, identifying routes of transmission across borders, including the identification of "super-spreaders" events, and informing infection control measures [78, 79, 87].

1.2.3 From genomics to metagenomics

Despite the increasing adoption of DNA sequencing methods in clinical microbiology, the sequencing of genetic material from a pure culture requires *a priori* knowledge of what to expect from a particular clinical sample or patient [88]. In most cases, this knowledge is enough to request the most appropriate test, such as multiplexed panels or specific culture media, but this is not always the case. In recent years, there has been a growing interest in using metagenomics to deliver culture-independent approaches to microbial ecology, surveillance and diagnosis (see Figure 1.5)[47, 89]. Metagenomic DNA sequence allows detailed characterisation of pathogens in all kinds of samples originating from humans, animals, food and the environment, ligating the diagnostics to surveillance in a true "one health" fashion [90]. Unlike PCR or microarrays, it usually does not require primer or probe design, it can be easily multiplexed, and the specificity and selectivity of the sequencing can be adjusted computationally after acquiring the data [91]. While most molecular assays target only a limited number of pathogens, metagenomic approaches characterise all DNA or RNA present in a sample, enabling analysis of the entire microbiome as well as the human host genome

[19] `http://web.archive.org/web/20200311053731/https://www.gisaid.org/`
[20] `https://www.gisaid.org/`

or transcriptome in patient samples [92]. Whether or not it can entirely replace routine microbiology depends on several conditions and future developments, both technological and computational (see section 1.3. The role of bioinformatics).

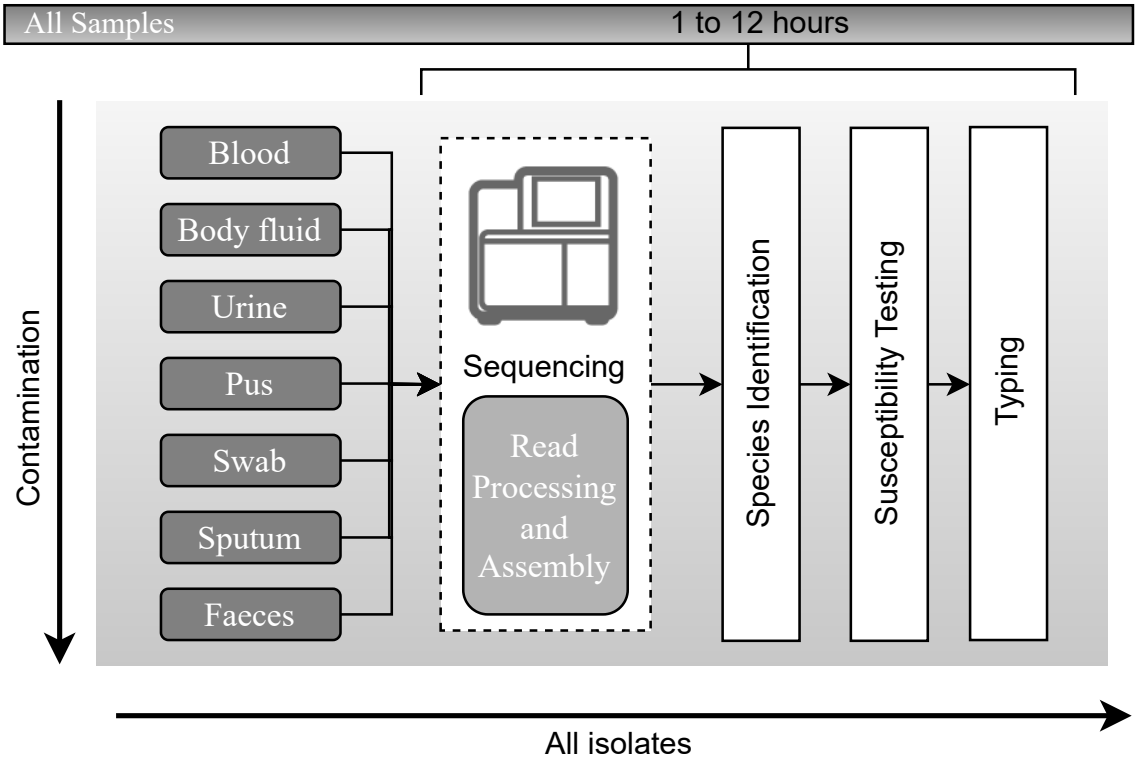

Figure 1.5: **Hypothetical workflow based on metagenomic sequencing.** Schematic representation of the hypothetical workflow for the direct processing of samples from suspected sources of pathogens after adoption of metagenomic sequencing, with an expected timescale that could fit within a single day. Adapted from [8]

Albeit lacking consensus in the field, metagenomics can be classified into two variants as proposed by [93]: (1) metaxonomics where marker genes ubiquitous in many taxa are targeted and sequenced, and (2) the untargeted "shotgun" sequencing of all microbial genomes present in a sample.

1.2.3.1 Metataxonomics and Targeted Metagenomics

Molecular barcoding approaches can be combined with second-generation high-throughput sequencing to achieve unprecedented depths of coverage in microbial community profiling, being defined as metataxonomics. For profiling bacterial species, the most popular approach is 16S ribosomal RNA (rRNA) gene sequencing, an 1500 base pair gene coding for a catalytic RNA that is part of the 30S ribosomal subunit. Traditionally, the variable regions of the 16S rRNA gene (V-regions) are targeted, or ranges thereof (V1-V2, V1-V3, V3-V4, V4, V4-V5, V6-V8, and V7-V9), and are specific to bacterial genus (96%) and for some, even species (87.5%), [94, 95]. Moreover, dedicated 16S databases that include near full length sequences for a large number of strains and their taxonomic placements exist, such

as RDP[21], Greengenes[22], silva[23] and NCBI's 16S ribosomal RNA project[24] [96–98]. The sequence from an unknown strain can be compared against the sequences in these databases, after very closely related sequences are grouped into Operational Taxonomic Units (OTUs), and infer likely taxonomy, with the assumption that sequences of >95% identity represent the same genus, whereas sequences of >97% identity represent the same species [99]. Additionally, NCBI also provides the 23S ribosomal RNA project for Bacteria and Archaea metataxonomics.

Because this approach is PCR-based, it suffers from the same issues described previously for conventional PCR, requiring primer design. Additionally, it must necessarily account for intragenomic variation between 16S gene copies. Microbial profiles generated using different primer pairs need independent validation of performance, and the comparison of data sets across V-regions using different databases might be misleading due to differences in nomenclature and varying precisions in classification, and specific but important taxa are not picked up by certain primer pairs (e.g., *Bacteroidetes* is missed using primers 515F-944R) or due to the database used [95]. Furthermore, targeting of 16S variable regions with short-read sequencing platforms cannot achieve the taxonomic resolution afforded by sequencing the entire (1500 bp) gene [100]. The emergence of third generating sequencing technologies (see subsubsection 1.2.1.3. The third-generation of DNA sequencing) allows for this limitation to be overcome but currently, only a fraction of the databases includes complete 16S rRNA sequences.

While viruses are an integral part of the microbiota, no universal viral marker genes are available to perform such taxonomic assignments. Amplification of whole viral genomes is possible and, in 2015, RNA extracted from whole blood, serum, re-suspended swabs and urine, after targeted amplification of the whole viral genome, proved invaluable in the track of the Ebola virus disease epidemic in West Africa, responsible for >11 thousand deaths, allowing for the characterisation of the infectious agent the determination of its evolutionary rate, signatures of host adaptation, identification and monitoring of diagnostic targets and responses to vaccines and treatments [101]. As an alternative, broad scope viral targeted sequence capture (TSC) panels offer depletion of background nucleic acids and improve the recovery of viral reads by targeting coding sequence from a multitude viral genera, such as VirCapSeq-VERT Capture Panel[25] but do not guarantee the full recovery of the viral genome, and can present biases towards certain genera [102, 103].

[21] `http://rdp.cme.msu.edu/`

[22] `https://greengenes.secondgenome.com/`

[23] `https://www.arb-silva.de/`

[24] `https://www.ncbi.nlm.nih.gov/refseq/targetedloci/`

[25] `https://sequencing.roche.com/content/dam/rochesequence/worldwide/resources/brochure-vircapseq-vert-capture-panel-SEQ1000117.pdf`

1.2.3.2 Shotgun Metagenomics

Shotgun metagenomics can offer relatively unbiased pathogen detection and characterisation. The capacity to detect all potential pathogens — bacteria, viruses, fungi and parasites — in a sample has great potential utility in the diagnosis of infectious disease [92], potentially able to provide genotyping, antimicrobial resistance and virulence profiling in a single methodological step. This comes with the cost of producing massive amounts of information that require expert handling and processing, as well as capable computational infrastructures [68, 104].

Clinical applications of shotgun metagenomics derive its roots from the use of microarrays (see subsection 1.1.1. Current standards for diagnostic in clinical microbiology), where it was successfully applied in in-depth microbiome analysis of different sites in the human body, it was the emergence of second-generation sequencing technology and its high throughput of genomic data at a competitive price that made the sequencing of all genomic content, DNA and/or RNA) if a clinical sample a viable possibility for diagnostics (see subsubsection 1.2.1.2. The second-generation of DNA sequencing) [92, 105, 106]. The first reported case that demonstrated the utility of shotgun metagenomics was in 2014 with the clinical diagnosis of neuroleptospirosis in a 14-year-old immunodeficient and critically ill boy with meningoencephalitis by [107], prompting appropriate targeted antibiotic treatment and eventual recovery of the patient. In this case, traditional methods, including an invasive brain biopsy, failed to provide answers, until the shotgun sequencing of cerebrospinal fluid identified 475 of 3,063,784 sequence reads (0.016%) corresponding to leptospira, for which clinical assays were negative due to its very low abundance. Ever since many other reports of successful application of shotgun metagenomics in clinical metagenomics have been reported. but all in edge cases where traditional diagnostic methods have failed or as proof-of-concept [104, 108–110].

In public health microbiology, shotgun metagenomics combined with transmission network analysis allowed the investigation and quick action on the food supply of the 2013 outbreak of Shiga toxin-producing *Escherichia coli* (STEC) strain O104:H4 from faecal specimens obtained from patients [111]. A similar approach was followed in the detection of *Salmonella enterica* subsp. *enterica* serovar Heidelberg from faecal samples in two though to be unrelated outbreaks in the United States of America, as well as the *in situ* abundance and level of intrapopulation diversity of the pathogen, and the possibility of co-infections with *Staphylococcus aureus*, overgrowth of commensal *Escherichia coli*, and significant shifts in the gut microbiome during infection relative to reference healthy samples [112]. More recently, shotgun metagenomic sequencing has evidenced alterations in the gut microbiota of a subset of COVID-19 patients that present the uncommon gastrointestinal (GI) symptoms, shedding a higher understanding of gut–lung axis affecting the progression of COVID-19 [113].

Clinical diagnostic applications have lagged behind research advances. A significant

challenge with shotgun metagenomic approach is the large variation in the pathogen load between patient samples, as evidenced in the studies presented. A low pathogen load and high contamination of host DNA or even the present microbiome may result in enough data to produce the high-resolution subtype needed to distinguish and cluster the cases that were caused by the same outbreak pathogen source, or, extremely, the undetection of the causative agent [92, 114]. Differential lysis of human host cells followed by degradation of background DNA has proven an effective method to reduce host contamination, but limitations include potential decreased sensitivity for microorganisms without cell walls, such as *Mycoplasma* spp. or parasites; a possible paradoxical increase in exogenous background contamination by use of additional reagent [115–117]. Additionally, it is often unclear whether a detected microorganism is a contaminant, coloniser or *bona fide* pathogen, and the lack of golden standards remains one of the biggest challenges when applying these methods in clinical microbiology for diagnosis.

In addition to negative controls, already a common practice in any sequencing assay and in particular in metataxonomics (see subsubsection 1.2.3.1. Metataxonomics and Targeted Metagenomics), positive controls can be a way to circumvent the lack of golden standards, either through the spike of the samples with a known amount of a specific DNA/RNA or though the sequencing of samples with known composition and abundance. Well-characterised reference standards and controls are needed to ensure shotgun metagenomics assay quality and stability over time [92, 118]. Most available metagenomic reference materials are highly tailored to a specific application. For example, the ZymoBIOMICS Microbial Community Standard[26] is the first commercially available standard for microbiomics and metagenomics studies, providing mock a mock community with defined composition and abundance consisting of Gram-positive, Gram-negative and yeast. It is useful to determine the limit of detection of an assay, and the effectiveness and biases of a given protocol. Standards with a more limited spectrum of organisms are also available, such as the National Institute of Standards and Technology (NIST)[27] reference materials for mixed microbial DNA detection, which contain only bacteria. Thus, these materials may not apply to untargeted shotgun metagenomics analyses.

1.3 The role of bioinformatics

As stated previously (see section 1.2. A genomic approach to clinical microbiology and subsection 1.2.3. From genomics to metagenomics), one of the biggest challenges when dealing with genomic, and in particular metagenomic, data is the lack of golden standards. This is also applicable to the bioinformatic analysis, required due to the amount of data produced by genomic sequencing technologies. This is currently one of the bottlenecks in the deployment of sequencing technology in clinical microbiology as there's no standard in

[26] `https://www.zymoresearch.com/collections/zymobiomics-microbial-community-standards`
[27] `https://www.nist.gov/`

how to deal with the increasing amount of data produced in a fit-for-purpose manner [119].

Bioinformatics is an interdisciplinary research field that applies methodologies from computer science, applied mathematics and statistics to the study of biological phenomena[119]. With the widespread use and continuous development of sequencing technologies, bioinformatics has become a cornerstone in modern clinical microbiology.

Major efforts are being made on the standardisation and assessment of software for the analysis of genomic data, both commercial and open-source [104, 120–122].

1.3.1 The FASTQ file

In all sequencing technologies (see subsection 1.2.1. Twenty five years of microbial genome sequencing), many copies of the source DNA are randomly fragmented and sequenced. To these sequences, we refer to as reads. In the case of second-generation sequencing (see subsubsection 1.2.1.2. The second-generation of DNA sequencing), one or both ends of the fragment can be sequenced. If a fragment is sequenced from one end, we refer to it as single-end sequencing. If a fragment is sequenced on both ends, spanning the entire fragment, it is called paired-end sequencing.

All sequencing technologies, regardless of generation, produce data in the same standard file format: the FASTQ, a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores [123]. Originally developed at the Wellcome Trust Sanger Institute, the FASTQ has emerged as a common file format for sharing sequencing read data (see 1.3). The FASTQ can be considered as an extension of the 'FASTA sequence file format', originally invented by [124], which includes just the sequence information. A FASTQ file normally uses four lines per sequence:

- **Line 1** begins with a '@' character and is followed by a sequence identifier and an optional description;
- **Line 2** is the raw sequence letters;
- **Line 3** begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again;
- **Line 4** encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

In FASTQ both the sequence letter and quality score are each encoded with a single ASCII character for brevity. The quality of a sequence in a FASTQ file is represented by a quality value Q is an integer mapping of p, where p is the probability that the corresponding

base call is incorrect (see Table 1.1). This is called the PHRED score [125] and is defined by the following equation:

$$\mathbf{Q}_{\text{PHRED}} = -10 \times \log \mathbf{P} \tag{1.1}$$

The PHRED quality scores **Q** is defined as a property which is logarithmically related to the base-calling error probability **P**.

Table 1.1: **PHRED quality scores are logarithmically linked to error probabilities.** A PHRED Score of 20 indicates the likelihood of finding 1 incorrect base call among 100 bases. In other words, the precision of the base call is 99%. **Q** scores are classified as a property that is associated logarithmically with the probabilities of base calling error **P**.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.90%
40	1 in 10,000	99.99%
50	1 in 100,000	100.00%
60	1 in 1,000,000	100.00%

Since their introduction, PHRED scores have become the *de facto* standard for representing sequencing read base qualities [123]. Despite this convention, the encoding of the Phread score can vary when it is translated to its ASCII representation in the FASTQ file format. For example, the Sanger FASTQ files use ASCII 33–126 to encode PHRED qualities from 0 to 93 (i.e. PHRED scores with an ASCII offset of 33). A full list of available encoding is available in 1.6.

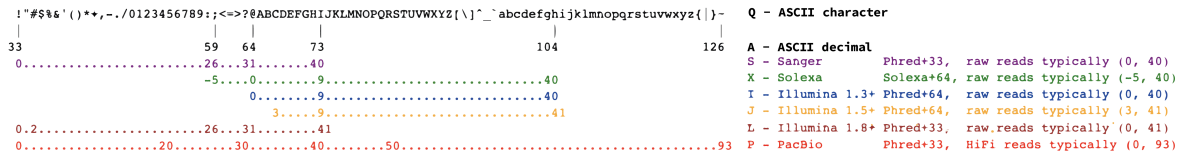

Figure 1.6: **Range of FASTQ quality scores andd their corresponding ASCII encoding.** For raw reads, the range of scores will depend on the technology and the base caller used. Starting in Illumina 1.8, the quality scores have returned to the use of the Sanger format (PHRED+33). For processed reads and long accurate reads, scores may be even higher with, For example, quality values of up to 93 observed in reads from PacBio HiFi reads.

1.3.1.1 FASTQ file simulation

With the lack of golden standards for metagenomic analysis, the use of simulated mock communities, with known composition, abundance and genomic information, provides a ground truth against which evaluations of success can be made. Given their standard struc-

ture and adoption, the generation of simulated FASTQ files from a reference, or a set of references, is very straightforward.

Multiple computational tools for the simulation of sequencing data, particularly for second and third-generation sequencing technologies, have been developed in recent years, which could be used to compare existing and new bioinformatic analytical pipelines. [126] provides a comprehensive assessment of 23 different read-simulation tools, highlighting their distinct functionality, requirements and potential applications, as well as providing a selection of suggestions for different simulation tools depending on their purpose. For *in silico* genomic and metagenomic sequence generation, a pletora of tools are available for first, second and third-generation reads (see Figure 1.7).

1.3.1.2 FASTQ quality assessment and quality control

Quality assessment and control is a basal step to any analysis, and aims to (1) remove and/or filter low quality and low complexity reads, (2) trim adapters, and (3) remove host sequences from the samples' raw data. There are many tools available but the most commonly used are FastQC[28] (Babraham Bioinformatics) for quality control, followed by Trimmomatic [127], Cutadapt [128] or fastp [129] to trim and/or filter adaptors, low quality and low complexity sequences. For long-read sequencing, tools like NanoPlot and NanoStats [130], and Filtlong[29] can perform the equivalent quality assessment and control, adapter trimming and low quality trimming, respectively.

1.3.2 Direct taxonomic assignment and characterisation

A piece of important information that can be retrieved directly from the quality-controlled read data: (1) the identification and characterisation of the microbes present in a sample and (2) their relative abundance. Taxonomic classification methods can vary depending on the sequencing methodology used: pure culture, metataxonomics and amplicon metagenomics, and shotgun metagenomics.

From pure culture, taxonomic identification of the read content of a sample is useful to assess contamination. Tools like Kraken2 [131, 132] and Braken [133]. These tools, relying on a database, assign taxonomic labels to reads and are therefore biased to the contents of the database used. Various databases are available[30], varying in size and content (archaea, bacteria, viral, plasmid, human and eukaryotic pathogens), and therefore in sensitivity depending on the resources available and the purpose intended. Alternatively, there are options to create custom databases.

[28] `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`
[29] `https://github.com/rrwick/Filtlong/`
[30] `https://benlangmead.github.io/aws-indexes/k2`

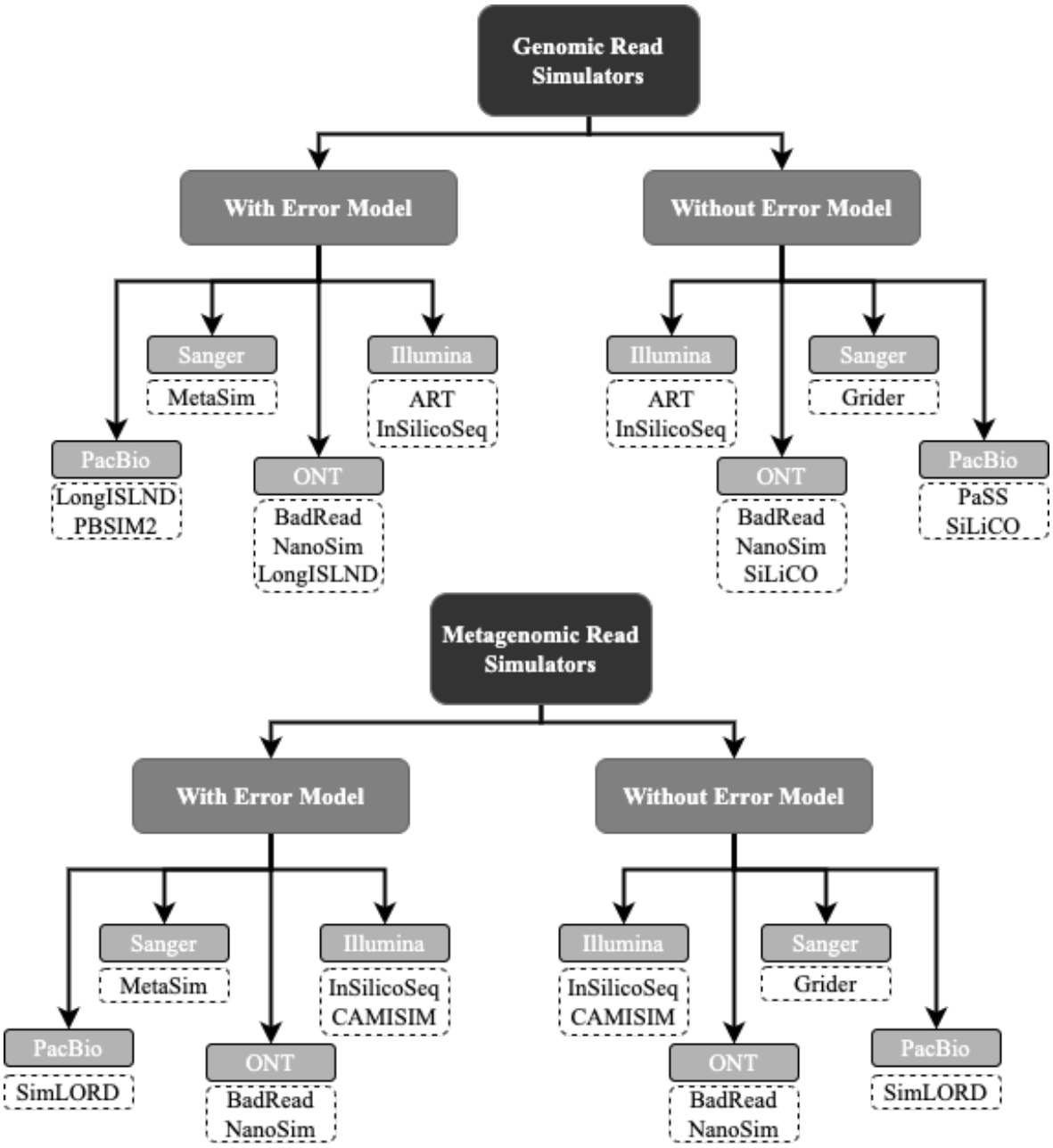

Figure 1.7: **Sequence simulators for genomic and metagenomic data.** For first generation sequencing, Metasim (`https://github.com/gwcbi/metagenomics_simulation`) and Grider (`https://sourceforge.net/projects/biogrinder/`) can generate mock genomic and metagenomic data, with and without error models respectively. For Illumina data, ART (`https://www.niehs.nih.gov/research/resources/software/biostatistics/art/index.cfm`), InSilicoSeq (`https://github.com/HadrienG/InSilicoSeq`) and CAMISIM (`https://github.com/CAMI-challenge/CAMISIM`) represent options for in silico data generation. Due to their differences, the third generation Pacific Bio-Sciences (PacBio) and Oxford Nanopore (ONT) have distict software for in silico data generation. The first can be accomplished by LongISLND (`https://bioinform.github.io/longislnd/`) and PBSIM2 (`https://github.com/yukiteruono/pbsim2`) got genomic data, and SimLORD (`https://bitbucket.org/genomeinformatics/simlord/src`) fot metagenomic data, with and without error model. The latter BadRead (`https://github.com/rrwick/Badread`) and NanoSim (`https://github.com/bcgsc/NanoSim`) can genenrate genomic and metagenomic *in silico* data, with and withouth error model. Additionally, for genomic data, LongISLND and SiLiCO (`https://github.com/ethanagb/SiLiCO`) generate data with and without error, respectively. Adapted from [126].

These tools are also extremely useful to assess the contents of a metagenomic sample. Alternatives such as Midas [134], Kaiju, [135], and MetaPhlAn2 [136] offer the same analysis as Kraken and Bracken using different algorithms, and with the disadvantage that they come prepackaged with their own databases, without the option to create a tailored database, limiting their applicability. Kaiju differs from the other tools by using a protein reference database, instead of nucleotide, but no pre-built version is available, requiring significant resources to build and index the database pre-use. The long-read data of third-generation sequencing technologies (see subsubsection 1.2.1.3. The third-generation of DNA sequencing) can be treated as single-end reads, and all tools mentioned accommodate the classification of single-end files.

1.3.3 From reads to genomes

Due to the limitations of current sequencing technologies (see subsection 1.2.1. Twenty five years of microbial genome sequencing), the order of the reads produced by these machines cannot be preserved. Therefore, to obtain the true original genomic sequence the process of "genome assembly" has to occur. The term "draft genome" is commonly used because these sequencing technologies do not generate a single closed genome, particularly short-read such as in second generation sequencing (see subsubsection 1.2.1.2. The second-generation of DNA sequencing) which need to be assembled into usually a series of sequences (contigs) that may cover up to 95% to 99% of the strain genome [119]. Long-read technologies (see subsubsection 1.2.1.3. The third-generation of DNA sequencing) allow for this value to reach 100%, effectively producing closed, complete genomes, notwithstanding that this value can sometimes overcome the 100% due to overlap [137].

Assembling reads into contigs has many advantages, namely that longer sequences are more informative, allowing the consideration of whole genes or even gene clusters within a genome and to understand larger genetic variants and repeats. Additionally, it has the effect of removing most sequencing errors, though this can be at the expense of new assembly errors [138]. Two methods are used to obtain draft genomes: (1) through reference-guided sequence assembly, or (2), through *de novo* sequence assembly.

1.3.3.1 Genomes through reference-guided sequence assembly

A reference-guided genome assembly uses an already sequenced reference genome to assemble a new genome, making use of the similarity between target and reference species to gain additional information, which often lead to a more complete and improved genome [139, 140]. This process is usually done through the mapping of the reads to a closely related reference sequence, and as more and more species get sequenced, the chance that a genome of the same or related species is already available, in which a significant proportion of the

reads can be mapped, increase greatly. This process usually includes the following steps: (1) the reference genome has to be indexed, allowing compression of the input text while still permitting fast sub-string queries, (2) for each short-read several sub sequences (seeds) are taken and searched to find their exact matches in the reference (candidate regions), (3) each short-read is then aligned to all corresponding candidate regions, and (4) the consensus sequence is computed in which the reference sequence is corrected when there is enough evidence of an difference based on the mapped reads, identifying the differences between it and the newly generated consensus sequence [141]. Besides variants, the new consensus genome might have insertions or deletions with respect to the reference genome.

Besides the generation of a consensus sequence, the mapping of the reads to the reference sequence can be used to estimate sequence depth and breadth of coverage. Depth of coverage, often referred to simply as coverage, refers to the average number of times each nucleotide position in the strain's genome has a read that aligns to that position. Depending on the study goals, bacterial species and the intended analyses, the optimal depth of coverage varies. In public repositories, most submissions have a depth of coverage ranging from 15 to 500 times [119]. Breadth of coverage is defined as the ratio of covered sequence on the reference by the aligned reads.

1.3.3.2 Genomes through *de novo* sequence assembly

De novo assembly refers to the bioinformatics process whereby reads are assembled into a draft genome using only the sequence information of the reads. Two methods are used to obtain draft genomes without the need of a reference genome: (1) through Overlap, Layout and Consensus, or (2) De Bruijn graph assembly (see Figure 1.8). The *de novo* assembly methods provide longer sequences that are more informative than shorter sequencing data and can provide a more complete picture of the microbial community in a given sample.

1.3.3.2.1 Overlap, Layout and Consensus assembly

First generation sequencing technology (see subsubsection 1.2.1.1. The first-generation of DNA sequencing) produces far fewer reads than second generation sequencing technology (see subsubsection 1.2.1.2. The second-generation of DNA sequencing, but individual reads are longer (500–1000 bp). Assembly of Sanger data usually uses overlap-layout consensus (OLC) approaches, in which:

- Overlaps are computed by comparing all reads to all other reads;
- Overlaps are grouped together to form contigs;
- A consensus contiguous sequence, or contig, is determined by picking the most likely nucleotides from the overlapping reads.

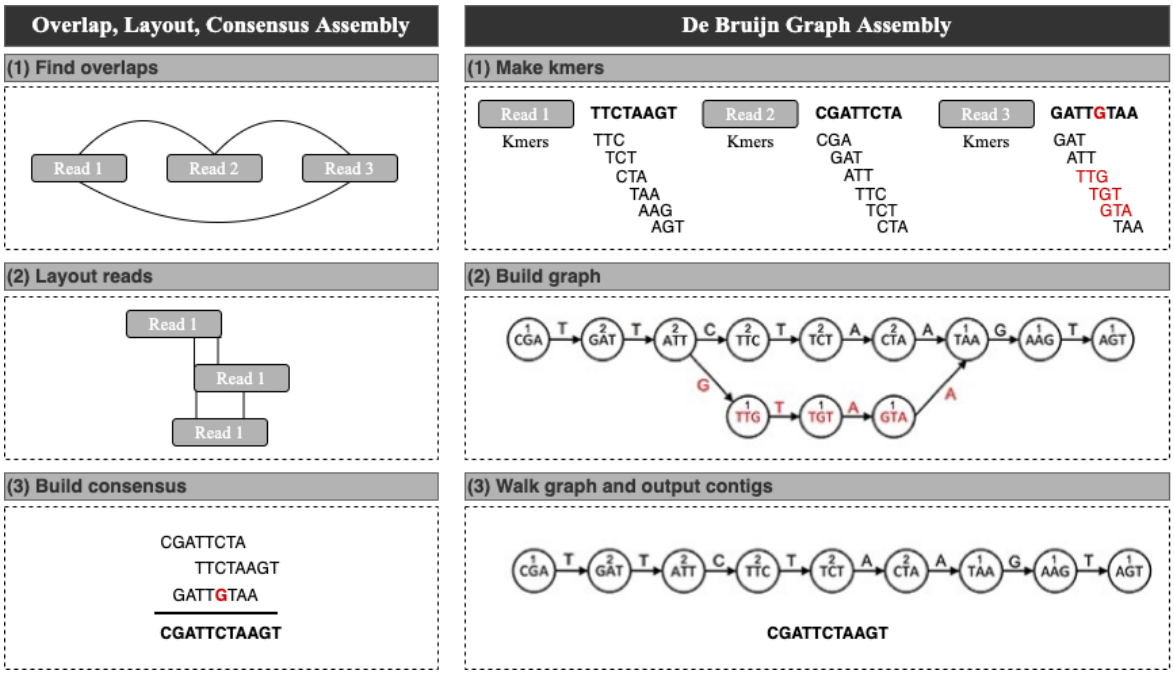

Figure 1.8: **Approaches to *de novo* genome assemble.** In Overlap, Layout, Consensus assembly, (1) overlaps are found between reads and an overlap graph constructed (edges indicate overlapping reads). (2) Reads are laid out into contigs based on the overlaps (lines indicate overlapping portions). (3) The most likely sequence is chosen to construct consensus sequence. In De Bruijn graph assembly, (1) reads are decomposed into kmers of a determined size by sliding a window of size k (in here of k=3) across the reads. (2) The kmers become vertices in the De Bruijn graph, with edges connecting overlapping kmers. Polymorphisms (red) form branches in the graph. A count is kept of how many times a kmer is seen, shown here as numbers above kmers. (3) Contigs are built by walking the graph from edge nodes. A variety of heuristics handle branches in the graphs—for example, low coverage paths, as shown here, may be ignored. Adapted from [138]

.

These type of assemblers were very popular in the early 2010s, with assemblers such as Celera[31], Genovo[32], xGenovo[33] and BBAP[34] having been widely used [142–145]. With the emergence of third-generation sequencing (see subsubsection 1.2.1.3. The third-generation of DNA sequencing), OLC assemblers have been increasingly developed and adopted by the community to assembly long-read data. In the latest years, ra[35], raven[36] and canu[37], the latter being a a fork of the Celera Assembler, have become staples in the community, showing good reliability and amassing over 3000 citations [137, 146, 147].

1.3.3.2.2 De Bruijn graph assembly

In the De Bruijn assembly graph, reads are split into overlapping k-mers where nodes of

[31] `https://www.cbcb.umd.edu/software/celera-assembler`
[32] `https://cs.stanford.edu/genovo`
[33] `http://xgenovo.dna.bio.keio.ac.jp/`
[34] `http://homepage.ntu.edu.tw/~youylin/BBAP.html`
[35] `https://github.com/lbcb-sci/ra`
[36] `https://github.com/lbcb-sci/raven`
[37] `https://github.com/marbl/canu`

the graph represent k-mers where:

- A directed edge from node N_a to node N_b indicates that N_b is next to N_a in a read;
- The number of nodes in the De Bruijn graph is theoretically the total number of identical k-mers in the genome;
- The weight on the edge indicates the number of times N_b is observed next to N_a in all reads.

Thus, the weight of an edge indicates the possibility that two k-mers appear after each other in the DNA sequence. A path in the graph where all edges have the highest weight is the most likely to be a part of the genome [141].

Most second-generation sequencing (see subsubsection 1.2.1.2. The second-generation of DNA sequencing assemblers, such as SPAdes[38], SKESA[39] and MEGAHIT[40], use a multiple k-mer De Bruijn graph, starting with the lowest size and iteratively adding k-mers of increasing length to connect the graph [148–150]. Older assemblers, such as Velvet[41], Ray[42] and SoapDeNovo2[43] use a single k-mer strategy for the De Bruijn graph construction [151–153].

1.3.3.3 Assembly quality assessment and quality control

The success of an assembly is evaluated in two steps: (1) globally, through intrinsic characteristics of the assembly itself, and (2) relative to a reference genome. The computation of the global metrics is performed through statistics inherent to the complete set of contigs assembled per sample, independent of the species/sample of origin. Commonly, these statistics include information on contig number, its median size and number ambiguous bases. The comparison with a reference sequence allows for statistics such as number of misassemblies, meaning contigs that do not reflect the structural organisation in the reference sequence, to be computed.

The assessment and evaluation of genome assemblies has been a relevant field ever since the emergence of the assembly process itself. The most widely adopted is QUAST[44], can evaluate assemblies both with a reference genome, as well as without a reference, producing many reports, summary tables and plots to help compare and assess assembly success [154].

[38] https://github.com/ablab/spades/
[39] https://github.com/ncbi/SKESA/
[40] https://github.com/voutcn/megahit/
[41] https://www.ebi.ac.uk/~zerbino/velvet/
[42] https://sourceforge.net/projects/denovoassembler/f
[43] https://sourceforge.net/projects/soapdenovo2/
[44] http://quast.sourceforge.net/quast

1.3.4 Reproducibility and transparency

Several steps that can be implemented to ensure the transparency and reproducibility of the chosen bioinformatic workflow. Favouring open-source tools, with clear documentation describing the methodology implemented, and stating the version of the software used and which parameters were used enables the comparison of results. This can be simplified by containerising all the software tools with one of the many solutions available, like Docker[45] or Singularity[46] [155].

The use of workflow managers, like nextflow[47], snakemake[48] or the Galaxy Project[49], will push reproducibility to the next level by taking advantage of the containerisation and scalability, enabling the workflow to be executed with the same parameters in the same conditions in a multitude of different environments [156–158].

1.4 Bioinformatic Analysis for Metagenomics

As mentioned previously (see subsection 1.2.3. From genomics to metagenomics, Metagenomic shotgun sequencing circumvents the need for cultivation and, compared with metataxonomics, avoids biases from primer choice, enables the detection of organisms across all domains of life and *de novo* assembly of genomes and functional genome analyses. However, highly uneven sequencing depth of different organisms and low depth of coverage per species are drawbacks that limit taxa

1.4.0.1 Metataxonomics

Metataxonomics (see subsubsection 1.2.3.1. Metataxonomics and Targeted Metagenomics) is the most widely used technique for microbial diversity analysis [159], and due to its particularities, the analysis of this data is also very particular. Data analyses are mostly carried out through specialised pipelines that wrap and combine several tools, offering the possibility to follow a simple protocol with default configurations or choose between a plethora of different configurations to adjust for any particular needs. Quantitative Insights Into Microbial Ecology 2 (QUIIME2)[50] [160] has become the *de facto* tool for metataxonomic analysis as a framework with an ever-growing suite of plugins and intuitive data visualisation tools for the assessment of results. Mothur [161] and UPARSE [162] are

[45] `https://www.docker.com/`
[46] `https://sylabs.io/`
[47] `https://www.nextflow.io/`
[48] `https://snakemake.github.io/`
[49] `https://galaxyproject.org/`
[50] `https://qiime2.org/`

also a popular alternative although resulting outputs differing significantly between pipelines despite using the same inputs having been reported by [163], with a magnitude that is comparable to differences in upstream sample treatment and sequencing procedures. A typical workflow starts with quality filtering, error correction and removal of chimeric sequences. These quality control steps are followed by either taxonomic assignment of reads or a clustering step where reads are gathered into OTUs given their sequence identity, followed by statistical analysis to assess differences between given groups. Taxonomic assignment methods classify query sequences based on the best hit found in reference databases of annotated sequences, being heavily dependent on the completeness of the reference databases (see subsubsection 1.2.3.1. Metataxonomics and Targeted Metagenomics). Classification is further limited by lack of species annotation in most reference databases [164]. Alternatively, the same approach of direct taxonomic classification, without OTU clustering, can be followed as with genomic and shotgun metagenomic data, given that the databases include rRNA sequences.

OTU clustering methods can be categorised into: (1) computationally expensive hierarchical methods that cluster sequences based on a distance matrix measuring the difference between each pair of sequences, (2) less expensive heuristic methods cluster sequences into OTUs based on a pre-defined threshold, generally, with a sequence being selected as a seed and the rest of the sequences being analysed sequentially and added to existing or new clusters according to the defined threshold, and (3) model based clustering methods that do not rely on a pre-defined and fixed threshold, defining OTUs based on a soft threshold and carrying out the clustering process based on methods such as an unsupervised probabilistic Bayesian clustering algorithm [165]. These methods offer the possibility to cluster sequences based on criteria that do not depend on reference databases and are especially useful in less characterised microbial communities or with a high representation of uncultured microbes. Due to the assumptions made with this strategy, it is sensitive to under or overestimation of the number of OTUs in a sample as defining a threshold to accurately cluster sequences is difficult [164].

1.4.0.2 Shotgun metagenomics

A plethora of open-source tools are available specifically for shotgun metagenomic data, and several combinations of these tools can be used to characterise the causative agent in a patient's infection in a fraction of the time required by the traditional methods.

A major additional difficulty of shotgun metagenomic data is the overpowering quantities of host DNA that are often sequenced, making the microbial community sometimes close to undetectable [104]. The presence of contaminants, from the bench process to the pre-existing biota, and the cost associated with this methodology, are also major hinders in its applicability in the clinic. They account for major caveats and must be made aware of when analysing the data.

The basic strategies for analysing shotgun metagenomic data can be simplified in the scheme in Figure **??**. One of the biggest challenges when doing metagenomic analysis is differentiating between colonisation and infection by successfully discriminating between a potential pathogen and background microbiota. In the latter, when analysing samples from presumably sterile sites, like Cerebrospinal fluid (CSF) and blood, it is safe to assume that all organisms found are of interest. In locations with a microbiota, the inclusion of negative controls is essential for the correct identification of contaminants in the taxonomic results, whether originated from the sample collection, handling or sequencing process. The use of spiked metagenomic samples as positive control might guide the detection of the possible pathogens by comparing relative abundance between the samples. These controls should be processed similarly to the samples and the taxonomic results should be filtered out from the final report.

As explored in subsection 1.3.3. From reads to genomes, longer sequences are more informative than shorter sequencing data, as the one obtained from second-generation sequencing (see subsubsection 1.2.1.2. The second-generation of DNA sequencing), and can provide a more complete picture of the microbial community in a given samples. Several dedicated metagenomic assembly tools are available, such as metaSPAdes[51] and MegaHIT[52] [150, 166]. These tools, in comparison to single-cell data assemblers, are better at dealing with the combination of intra and intergenomic repeats and uneven sequencing coverage [167]. For third-generation sequencing, dedicated metagenomic assemblers have recently emerged, such as meta-flye[53] which expands on the original flye assembler by overcoming a k-mer selection limitation on low abundance species [168]. Nevertheless, the use of non dedicated assemblers for metagenomics may come with the cost of wrongly interpret variation as error, especially in samples that contained closely related species and the construction of chimeric sequences as traditional assemblers follow the basic principle that the coverage in a sample is constant [169].

The assembly-based approach requires the grouping of the different contigs into bins, ideally each collecting the sequences that belong to a microorganism present in the sample. The binning process can be taxonomy dependent, relying on a database to aggregate the sequences, or independent. The independent approach has the benefit of not relying on a database, but instead it uses the composition of each sequence and coverage profiles to cluster together sequences that might belong to the same organism. These algorithms don't require prior knowledge about the genomes in a given sample, instead relying on features inherent to the sequences in the sample. Although most binning software can work with single metagenomic samples, most make use of differential coverage of multiple samples to improve the binning process [170]. It allows the handling of complex ecosystems and might be crucial when analysing samples recovered from sites with a complex microbiota. A comparison of five taxonomic independent and four taxonomic binning software by [122]

[51] `https://github.com/ablab/spades/`
[52] `https://github.com/voutcn/megahit/`
[53] `https://github.com/fenderglass/Flye/`

revealed that, for taxonomic independent approaches, MaxBin2[54] had the highest completeness and purity in the bins obtained []. For taxonomic binning, working similarly to the direct taxonomic assignment of the sequencing data, PhyloPythiaS+[55] obtained better results in accuracy, completeness and purity, followed by Kraken[56] that still obtained decent results with the added benefit of very high speed of analysis, ease of use and inclusion of the pre-built databases [131, 171].

The last step on the assembly methodology is the evaluation of the completeness and contamination of the bins. When using a taxonomic binner, the effects of contamination are mitigated as the sequence clustering is performed based on matches with reference database. The contaminants, if present in the database, will be separated into different bins or just added to the bin of unclassified sequences. When using a taxonomic independent binning software, the composition and abundance might not be enough to discriminate between all the organisms, with the possible result of having bins with contaminating sequences of other organisms present in the sample. CheckM[57] assesses the quality of the recovered genomes, estimating completeness and contamination by evaluating ubiquitous single-copy genes [172].

Another problem with metagenomic assembly is the high number of ambiguities that fail to being resolved, mostly due to the possible presence of several strains of the same species or species that are closely related. When faced with this ambiguities the assembler usually breaks the sequence, leading to fragmented reconstructions of genomes. MetaQUAST[58] that besides computing several metrics to evaluate assembly quality like number of contigs, maximum contig length, etc, also uses reference-based method, either provided by the user or by identifying the appropriate reference sequences by 16S ribosomal RNA identification, to identify mis-assemblies and structural variants [173]. All downstream processes used in whole genome sequencing draft genomes can be applied to each of the resulting binned genomes, that now represent a taxonomic unit recovered from the original metagenomic sample.

[54] https://sourceforge.net/projects/maxbin2/
[55] https://github.com/algbioi/ppsp
[56] https://github.com/DerrickWood/kraken2/
[57] https://ecogenomics.github.io/CheckM/
[58] https://github.com/ablab/quast

Bibliography

[1] Theo Vos et al. "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019". English. In: *The Lancet* 396.10258 (Oct. 2020). Publisher: Elsevier, pp. 1204–1222. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(20)30925-9. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30925-9/fulltext (visited on 01/28/2022).

[2] Hannah Ritchie et al. *Coronavirus Pandemic (COVID-19)*. 2020. URL: https://ourworldindata.org/coronavirus (visited on 01/28/2022).

[3] Jocelyne Piret and Guy Boivin. "Pandemics Throughout History". In: *Frontiers in Microbiology* 11 (Jan. 2021), p. 631736. ISSN: 1664-302X. DOI: 10.3389/fmicb.2020.631736. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7874133/ (visited on 01/28/2022).

[4] Christopher JL Murray et al. "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis". English. In: *The Lancet* 0.0 (Jan. 2022). Publisher: Elsevier. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(21)02724-0. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02724-0/fulltext (visited on 01/28/2022).

[5] World Health Organization. *Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis*. Technical documents. World Health Organization, 2017, 87 p.

[6] World Health Organization. *Global expenditure on health: public spending on the rise?* en. Section: xi, 74 p. Geneva: World Health Organization, 2021. ISBN: 978-92-4-004121-9. URL: https://apps.who.int/iris/handle/10665/350560 (visited on 02/01/2022).

[7] Angela E. Micah et al. "Tracking development assistance for health and for COVID-19: a review of development assistance, government, out-of-pocket, and other private spending on health for 204 countries and territories, 1990–2050". English. In: *The Lancet* 398.10308 (Oct. 2021). Publisher: Elsevier, pp. 1317–1343. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(21)01258-7. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)01258-7/fulltext (visited on 02/01/2022).

[8] Xavier Didelot et al. “Transforming clinical microbiology with bacterial genome sequencing”. en. In: *Nature Reviews Genetics* 13.9 (Sept. 2012), pp. 601–612. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3226. URL: http://www.nature.com/articles/nrg3226 (visited on 01/28/2022).

[9] Betsy Foxman et al. “Choosing an appropriate bacterial typing technique for epidemiologic studies”. In: *Epidemiologic perspectives & innovations : EP+I* 2 (Nov. 2005), p. 10. ISSN: 1742-5573. DOI: 10.1186/1742-5573-2-10. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1308839/ (visited on 01/31/2022).

[10] Nurnabila Syafiqah Muhamad Rizal et al. “Advantages and Limitations of 16S rRNA Next-Generation Sequencing for Pathogen Identification in the Diagnostic Microbiology Laboratory: Perspectives from a Middle-Income Country”. en. In: *Diagnostics* 10.10 (Oct. 2020). Number: 10 Publisher: Multidisciplinary Digital Publishing Institute, p. 816. ISSN: 2075-4418. DOI: 10.3390/diagnostics10100816. URL: https://www.mdpi.com/2075-4418/10/10/816 (visited on 02/04/2022).

[11] Christopher Giuliano, Chandni R. Patel, and Pramodini B. Kale-Pradhan. “A Guide to Bacterial Culture Identification And Results Interpretation”. In: *Pharmacy and Therapeutics* 44.4 (Apr. 2019), pp. 192–200. ISSN: 1052-1372. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6428495/ (visited on 02/04/2022).

[12] Robin Patel. “MALDI-TOF MS for the Diagnosis of Infectious Diseases”. In: *Clinical Chemistry* 61.1 (Jan. 2015), pp. 100–111. ISSN: 0009-9147. DOI: 10.1373/clinchem.2014.221770. URL: https://doi.org/10.1373/clinchem.2014.221770 (visited on 02/04/2022).

[13] Michelle H. Scerbo et al. “Beyond Blood Culture and Gram Stain Analysis: A Review of Molecular Techniques for the Early Detection of Bacteremia in Surgical Patients”. eng. In: *Surgical Infections* 17.3 (June 2016), pp. 294–302. ISSN: 1557-8674. DOI: 10.1089/sur.2015.099.

[14] M. Benkova, O. Soukup, and J. Marek. “Antimicrobial susceptibility testing: currently used methods and devices and the near future in clinical practice”. en. In: *Journal of Applied Microbiology* 129.4 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jam.14704, pp. 806–822. ISSN: 1365-2672. DOI: 10.1111/jam.14704. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jam.14704 (visited on 02/04/2022).

[15] Franz Allerberger. “Molecular Typing in Public Health Laboratories: From an Academic Indulgence to an Infection Control Imperative”. In: *Journal of Preventive Medicine and Public Health* 45.1 (Jan. 2012), pp. 1–7. ISSN: 1975-8375. DOI: 10.3961/jpmph.2012.45.1.1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3278599/ (visited on 01/31/2022).

[16] Hui-min Neoh et al. "Pulsed-field gel electrophoresis (PFGE): A review of the "gold standard" for bacteria typing and current alternatives". en. In: *Infection, Genetics and Evolution* 74 (2019), p. 103935. ISSN: 1567-1348. DOI: 10.1016/j.meegid.2019.103935. URL: https://www.sciencedirect.com/science/article/pii/S156713481930156X (visited on 01/31/2022).

[17] Martin C.J. Maiden. "Multilocus Sequence Typing of Bacteria". en. In: *Annual Review of Microbiology* 60.1 (Oct. 2006), pp. 561–588. ISSN: 0066-4227, 1545-3251. DOI: 10.1146/annurev.micro.59.030804.121325. URL: https://www.annualreviews.org/doi/10.1146/annurev.micro.59.030804.121325 (visited on 01/31/2022).

[18] Mette V. Larsen et al. "Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria". en. In: *Journal of Clinical Microbiology* 50.4 (Apr. 2012), pp. 1355–1361. ISSN: 0095-1137, 1098-660X. DOI: 10.1128/JCM.06094-11. URL: https://journals.asm.org/doi/10.1128/JCM.06094-11 (visited on 01/31/2022).

[19] Keith A. Jolley, James E. Bray, and Martin C. J. Maiden. "Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications". In: *Wellcome Open Research* 3 (Sept. 2018), p. 124. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.14826.1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6192448/ (visited on 01/31/2022).

[20] Elita Jauneikaite et al. "Current methods for capsular typing of Streptococcus pneumoniae". en. In: *Journal of Microbiological Methods* 113 (June 2015), pp. 41–49. ISSN: 0167-7012. DOI: 10.1016/j.mimet.2015.03.006. URL: https://www.sciencedirect.com/science/article/pii/S0167701215000858 (visited on 01/31/2022).

[21] James C. Paton and Claudia Trappetti. "Streptococcus pneumoniae Capsular Polysaccharide". EN. In: *Microbiology Spectrum* (Apr. 2019). Publisher: ASM PressWashington, DC. DOI: 10.1128/microbiolspec.GPP3-0019-2018. URL: https://journals.asm.org/doi/abs/10.1128/microbiolspec.GPP3-0019-2018 (visited on 01/31/2022).

[22] Benjamin Diep et al. "Salmonella Serotyping; Comparison of the Traditional Method to a Microarray-Based Method and an in silico Platform Using Whole Genome Sequencing Data". In: *Frontiers in Microbiology* 10 (2019). ISSN: 1664-302X. URL: https://www.frontiersin.org/article/10.3389/fmicb.2019.02554 (visited on 01/31/2022).

[23] Christopher J. Burrell, Colin R. Howard, and Frederick A. Murphy. "Laboratory Diagnosis of Virus Diseases". In: *Fenner and White's Medical Virology* (2017), pp. 135–154. DOI: 10.1016/B978-0-12-375156-0.00010-2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149825/ (visited on 01/28/2022).

[24] A. Cassedy, A. Parle-McDermott, and R. O'Kennedy. "Virus Detection: A Review of the Current and Emerging Molecular and Immunological Methods". In: *Frontiers in Molecular Biosciences* 8 (Apr. 2021), p. 637559. ISSN: 2296-889X. DOI: 10.3389/fmolb.2021.637559. URL: https://www.frontiersin.org/articles/10.3389/fmolb.2021.637559/full (visited on 02/01/2022).

[25] Jennifer Dien Bard and Erin McElvania. "Panels and Syndromic Testing in Clinical Microbiology". In: *Clinics in Laboratory Medicine* 40.4 (Dec. 2020), pp. 393–420. ISSN: 0272-2712. DOI: 10.1016/j.cll.2020.08.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7528880/ (visited on 02/04/2022).

[26] YuYen Chan et al. "Determining seropositivity—A review of approaches to define population seroprevalence when using multiplex bead assays to assess burden of tropical diseases". en. In: *PLOS Neglected Tropical Diseases* 15.6 (June 2021). Publisher: Public Library of Science, e0009457. ISSN: 1935-2735. DOI: 10.1371/journal.pntd.0009457. URL: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0009457 (visited on 02/01/2022).

[27] Niklas Bobrovitz et al. "Global seroprevalence of SARS-CoV-2 antibodies: A systematic review and meta-analysis". In: *PLoS ONE* 16.6 (June 2021), e0252617. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0252617. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8221784/ (visited on 02/01/2022).

[28] Katarzyna M. Koczula and Andrea Gallotta. "Lateral flow assays". en. In: *Essays in Biochemistry* 60.1 (June 2016). Ed. by Pedro Estrela, pp. 111–120. ISSN: 0071-1365, 1744-1358. DOI: 10.1042/EBC20150012. URL: https://portlandpress.com/essaysbiochem/article/60/1/111/78237/Lateral-flow-assays (visited on 02/01/2022).

[29] Fabio Di Nardo et al. "Ten Years of Lateral Flow Immunoassay Technique Applications: Trends, Challenges and Future Perspectives". en. In: *Sensors* 21.15 (Jan. 2021). Number: 15 Publisher: Multidisciplinary Digital Publishing Institute, p. 5185. ISSN: 1424-8220. DOI: 10.3390/s21155185. URL: https://www.mdpi.com/1424-8220/21/15/5185 (visited on 02/01/2022).

[30] Samuel L. Groseclose and David L. Buckeridge. "Public Health Surveillance Systems: Recent Advances in Their Use and Evaluation". In: *Annual Review of Public Health* 38.1 (2017). _eprint: https://doi.org/10.1146/annurev-publhealth-031816-044348, pp. 57–79. DOI: 10.1146/annurev-publhealth-031816-044348. URL: https://doi.org/10.1146/annurev-publhealth-031816-044348 (visited on 02/07/2022).

[31] Jillian Murray and Adam L. Cohen. "Infectious Disease Surveillance". In: *International Encyclopedia of Public Health* (2017), pp. 222–229. DOI: 10.1016/B978-0-12-803678-5.00517-8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149515/ (visited on 02/07/2022).

[32] Steven M. Teutsch. "Considerations in Planning a Surveillance System". eng. In: *Principles & Practice of Public Health Surveillance*. 3rd ed. Oxford University Press, 2010. ISBN: 978-0-19-537292-2. DOI: 10.1093/acprof:oso/9780195372922.003.0002. URL: https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780195372922.001.0001/acprof-9780195372922-chapter-2 (visited on 02/07/2022).

[33] World Health Organization. *International Health Regulations (2005)*. en. Second edition. The Ukrainian version is published by the Center for Implementation of International Health Regulations, Ukraine. Geneva: World Health Organization, 2005. ISBN: 978-92-4-158041-0. URL: https://www.who.int/publications-detail-redirect/9789241580410 (visited on 02/07/2022).

[34] J. Melo-Cristino, Letícia Santos, and Mário Ramirez. "Estudo Viriato: Actualização de dados de susceptibilidade aos antimicrobianos de bactérias responsáveis por infecções respiratórias adquiridas na comunidade em Portugal em 2003 e 2004". pt. In: *Revista Portuguesa de Pneumologia* 12.1 (Jan. 2006), pp. 9–30. ISSN: 0873-2159. DOI: 10.1016/S0873-2159(15)30419-0. URL: https://www.sciencedirect.com/science/article/pii/S0873215915304190 (visited on 02/07/2022).

[35] Jason R Andrews et al. "Environmental Surveillance as a Tool for Identifying High-risk Settings for Typhoid Transmission". In: *Clinical Infectious Diseases* 71.Supplement_2 (July 2020), S71–S78. ISSN: 1058-4838. DOI: 10.1093/cid/ciaa513. URL: https://doi.org/10.1093/cid/ciaa513 (visited on 02/07/2022).

[36] E. J. McWeeney. "Demonstration of the Typhoid Bacillus in Suspected Drinking Water by Parietti's Method". eng. In: *British Medical Journal* 1.1740 (May 1894), pp. 961–962. ISSN: 0007-1447. DOI: 10.1136/bmj.1.1740.961.

[37] Stephen Baker et al. "Combined high-resolution genotyping and geospatial analysis reveals modes of endemic urban typhoid fever transmission". eng. In: *Open Biology* 1.2 (Oct. 2011), p. 110008. ISSN: 2046-2441. DOI: 10.1098/rsob.110008.

[38] David A. Larsen and Krista R. Wigginton. "Tracking COVID-19 with wastewater". en. In: *Nature Biotechnology* 38.10 (Oct. 2020). Number: 10 Publisher: Nature Publishing Group, pp. 1151–1153. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0690-1. URL: https://www.nature.com/articles/s41587-020-0690-1 (visited on 02/07/2022).

[39] Delphine Destoumieux-Garzón et al. "The One Health Concept: 10Years Old and a Long Road Ahead". In: *Frontiers in Veterinary Science* 5 (Feb. 2018), p. 14. ISSN: 2297-1769. DOI: 10.3389/fvets.2018.00014. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5816263/ (visited on 03/08/2022).

[40] John S Mackenzie and Martyn Jeggo. "The One Health Approach—Why Is It So Important?" In: *Tropical Medicine and Infectious Disease* 4.2 (May 2019), p. 88. ISSN: 2414-6366. DOI: 10.3390/tropicalmed4020088. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6630404/ (visited on 02/07/2022).

[41] Sima Ernest Rugarabamu. *The One-Health Approach to Infectious Disease Outbreaks Control*. en. Publication Title: Current Perspectives on Viral Disease Outbreaks - Epidemiology, Detection and Control. IntechOpen, Sept. 2021. ISBN: 978-1-83881-911-8. DOI: 10.5772/intechopen.95759. URL: https://www.intechopen.com/chapters/75084 (visited on 02/07/2022).

[42] D. W. Hood et al. "DNA repeats identify novel virulence genes in Haemophilus influenzae." en. In: *Proceedings of the National Academy of Sciences* 93.20 (Oct. 1996), pp. 11121–11125. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.93.20.11121. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.93.20.11121 (visited on 01/28/2022).

[43] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (Dec. 1977), pp. 5463–5467. ISSN: 0027-8424. DOI: 10.1073/pnas.74.12.5463.

[44] Richard Van Noorden, Brendan Maher, and Regina Nuzzo. "The top 100 papers". en. In: *Nature News* 514.7524 (Oct. 2014). Cg_type: Nature News Section: News Feature, p. 550. DOI: 10.1038/514550a. URL: http://www.nature.com/news/the-top-100-papers-1.16224 (visited on 02/07/2022).

[45] J. D. Watson and F. H. C. Crick. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". en. In: *Nature* 171.4356 (Apr. 1953). Number: 4356 Publisher: Nature Publishing Group, pp. 737–738. ISSN: 1476-4687. DOI: 10.1038/171737a0. URL: https://www.nature.com/articles/171737a0 (visited on 02/08/2022).

[46] Ian S. Hagemann. "Overview of Technical Aspects and Chemistries of Next-Generation Sequencing". en. In: *Clinical Genomics*. Elsevier, 2015, pp. 3–19. ISBN: 978-0-12-404748-8. DOI: 10.1016/B978-0-12-404748-8.00001-0. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780124047488000010 (visited on 02/08/2022).

[47] Nicholas J. Loman and Mark J. Pallen. "Twenty years of bacterial genome sequencing". en. In: *Nature Reviews Microbiology* 13.12 (Dec. 2015). Number: 12 Publisher: Nature Publishing Group, pp. 787–794. ISSN: 1740-1534. DOI: 10.1038/nrmicro3565. URL: https://www.nature.com/articles/nrmicro3565 (visited on 02/08/2022).

[48] Sara Goodwin, John D. McPherson, and W. Richard McCombie. "Coming of age: ten years of next-generation sequencing technologies". en. In: *Nature Reviews Genetics* 17.6 (June 2016). Number: 6 Publisher: Nature Publishing Group, pp. 333–351. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.49. URL: https://www.nature.com/articles/nrg.2016.49 (visited on 02/08/2022).

[49] Yunhao Wang et al. “Nanopore sequencing technology, bioinformatics and applications”. en. In: *Nature Biotechnology* 39.11 (Nov. 2021). Number: 11 Publisher: Nature Publishing Group, pp. 1348–1365. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01108-x. URL: https://www.nature.com/articles/s41587-021-01108-x (visited on 03/01/2022).

[50] Michael L. Metzker. “Sequencing technologies — the next generation”. en. In: *Nature Reviews Genetics* 11.1 (Jan. 2010), pp. 31–46. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg2626. URL: http://www.nature.com/articles/nrg2626 (visited on 03/01/2022).

[51] Liu Xu and Masahide Seki. “Recent advances in the detection of base modifications using the Nanopore sequencer”. en. In: *Journal of Human Genetics* 65.1 (Jan. 2020). Number: 1 Publisher: Nature Publishing Group, pp. 25–33. ISSN: 1435-232X. DOI: 10.1038/s10038-019-0679-0. URL: https://www.nature.com/articles/s10038-019-0679-0 (visited on 03/01/2022).

[52] Linda Koch, Catherine Potenski, and Michelle Trenkmann. “Sequencing moves to the twenty-first century”. en. In: *Nature Research* (Feb. 2021). Bandiera_abtest: a Cg_type: Milestones Publisher: Nature Publishing Group. DOI: 10.1038/d42859-020-00100-w. URL: https://www.nature.com/articles/d42859-020-00100-w (visited on 02/08/2022).

[53] Francis S. Collins and Leslie Fink. “The Human Genome Project”. In: *Alcohol Health and Research World* 19.3 (1995), pp. 190–195. ISSN: 0090-838X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875757/ (visited on 02/08/2022).

[54] S. T. Cole et al. “Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence”. en. In: *Nature* 393.6685 (June 1998). Number: 6685 Publisher: Nature Publishing Group, pp. 537–544. ISSN: 1476-4687. DOI: 10.1038/31159. URL: https://www.nature.com/articles/31159 (visited on 02/07/2022).

[55] J. Parkhill et al. “Genome sequence of Yersinia pestis, the causative agent of plague”. en. In: *Nature* 413.6855 (Oct. 2001). Number: 6855 Publisher: Nature Publishing Group, pp. 523–527. ISSN: 1476-4687. DOI: 10.1038/35097083. URL: https://www.nature.com/articles/35097083 (visited on 02/08/2022).

[56] Frederick R. Blattner et al. “The Complete Genome Sequence of *Escherichia coli* K-12”. en. In: *Science* 277.5331 (Sept. 1997), pp. 1453–1462. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.277.5331.1453. URL: https://www.science.org/doi/10.1126/science.277.5331.1453 (visited on 02/08/2022).

[57] Barton E. Slatko, Andrew F. Gardner, and Frederick M. Ausubel. “Overview of Next Generation Sequencing Technologies”. In: *Current protocols in molecular biology* 122.1 (Apr. 2018), e59. ISSN: 1934-3639. DOI: 10.1002/cpmb.59. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6020069/ (visited on 02/08/2022).

[58] J.C. Detter et al. “Nucleic acid sequencing for characterizing infectious and/or novel agents in complex samples”. en. In: *Biological Identification*. Elsevier, 2014, pp. 3–53. ISBN: 978-0-85709-501-5. DOI: 10.1533/9780857099167.1.3. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780857095015500015 (visited on 02/08/2022).

[59] Alice Maria Giani et al. “Long walk to genomics: History and current approaches to genome sequencing and assembly”. en. In: *Computational and Structural Biotechnology Journal* 18 (Jan. 2020), pp. 9–19. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2019.11.002. URL: https://www.sciencedirect.com/science/article/pii/S2001037019303277 (visited on 02/08/2022).

[60] Nicholas J Loman et al. “Performance comparison of benchtop high-throughput sequencing platforms”. en. In: *Nature Biotechnology* 30.5 (May 2012), pp. 434–439. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2198. URL: http://www.nature.com/articles/nbt.2198 (visited on 02/14/2022).

[61] Anuj Kumar Gupta and U. D. Gupta. “Chapter 19 - Next Generation Sequencing and Its Applications”. en. In: *Animal Biotechnology*. Ed. by Ashish S. Verma and Anchal Singh. San Diego: Academic Press, Jan. 2014, pp. 345–367. ISBN: 978-0-12-416002-6. DOI: 10.1016/B978-0-12-416002-6.00019-5. URL: https://www.sciencedirect.com/science/article/pii/B9780124160026000195 (visited on 02/14/2022).

[62] Minh Thuy Vi Hoang et al. “Long-Reads-Based Metagenomics in Clinical Diagnosis With a Special Focus on Fungal Infections”. In: *Frontiers in Microbiology* 12 (2022). ISSN: 1664-302X. URL: https://www.frontiersin.org/article/10.3389/fmicb.2021.708550 (visited on 02/14/2022).

[63] Jonas Korlach and Stephen W Turner. “Going beyond five bases in DNA sequencing”. en. In: *Current Opinion in Structural Biology*. Nucleic acids/Sequences and topology 22.3 (June 2012), pp. 251–261. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2012.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0959440X12000681 (visited on 02/14/2022).

[64] Aaron M. Wenger et al. “Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome”. en. In: *Nature Biotechnology* 37.10 (Oct. 2019), pp. 1155–1162. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0217-9. URL: http://www.nature.com/articles/s41587-019-0217-9 (visited on 02/14/2022).

[65] Elizabeth T. Cirulli and David B. Goldstein. “Uncovering the roles of rare variants in common disease through whole-genome sequencing”. en. In: *Nature Reviews Genetics* 11.6 (June 2010). Number: 6 Publisher: Nature Publishing Group, pp. 415–425. ISSN: 1471-0064. DOI: 10.1038/nrg2779. URL: https://www.nature.com/articles/nrg2779 (visited on 02/18/2022).

[66] Nature Reviews Genetics. “A genomic approach to microbiology”. en. In: *Nature Reviews Genetics* 20.6 (June 2019), pp. 311–311. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-019-0131-5. URL: http://www.nature.com/articles/s41576-019-0131-5 (visited on 01/26/2022).

[67] F. Tagini and G. Greub. “Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review”. In: *European Journal of Clinical Microbiology & Infectious Diseases* 36.11 (2017), pp. 2007–2020. ISSN: 0934-9723. DOI: 10.1007/s10096-017-3024-6. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5653721/ (visited on 02/08/2022).

[68] J. W. A. Rossen et al. “Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology”. eng. In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 24.4 (Apr. 2018), pp. 355–360. ISSN: 1469-0691. DOI: 10.1016/j.cmi.2017.11.001.

[69] Stephanie W. Lo and Dorota Jamrozy. “Genomics and epidemiological surveillance”. en. In: *Nature Reviews Microbiology* 18.9 (Sept. 2020). Number: 9 Publisher: Nature Publishing Group, pp. 478–478. ISSN: 1740-1534. DOI: 10.1038/s41579-020-0421-0. URL: https://www.nature.com/articles/s41579-020-0421-0 (visited on 02/18/2022).

[70] Ayorinde O. Afolayan et al. “Overcoming Data Bottlenecks in Genomic Pathogen Surveillance”. eng. In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 73.Supplement_4 (Dec. 2021), S267–S274. ISSN: 1537-6591. DOI: 10.1093/cid/ciab785.

[71] Rafael Mamede et al. “Chewie Nomenclature Server (chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas”. In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D660–D666. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa889. URL: https://doi.org/10.1093/nar/gkaa889 (visited on 02/18/2022).

[72] Zhemin Zhou et al. “The EnteroBase user’s guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity”. en. In: *Genome Research* 30.1 (Jan. 2020). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 138–152. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.251678.119. URL: https://genome.cshlp.org/content/30/1/138 (visited on 02/18/2022).

[73] Silvia Argimón et al. “Microreact: visualizing and sharing data for genomic epidemiology and phylogeography”. In: *Microbial Genomics* 2.11 (). Publisher: Microbiology Society, e000093. ISSN: 2057-5858, DOI: 10.1099/mgen.0.000093. URL:

https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000093 (visited on 02/18/2022).

[74] Yuelong Shu and John McCauley. "GISAID: Global initiative on sharing all influenza data – from vision to reality". In: *Eurosurveillance* 22.13 (Mar. 2017), p. 30494. ISSN: 1025-496X. DOI: 10.2807/1560-7917.ES.2017.22.13.30494. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5388101/ (visited on 02/18/2022).

[75] Brett E. Pickett et al. "Virus Pathogen Database and Analysis Resource (ViPR): A Comprehensive Bioinformatics Database and Analysis Resource for the Coronavirus Research Community". en. In: *Viruses* 4.11 (Nov. 2012). Number: 11 Publisher: Molecular Diversity Preservation International, pp. 3209–3226. ISSN: 1999-4915. DOI: 10.3390/v4113209. URL: https://www.mdpi.com/1999-4915/4/11/3209 (visited on 02/18/2022).

[76] Vítor Borges et al. "INSaFLU: an automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance". In: *Genome Medicine* 10.1 (June 2018), p. 46. ISSN: 1756-994X. DOI: 10.1186/s13073-018-0555-0. URL: https://doi.org/10.1186/s13073-018-0555-0 (visited on 02/18/2022).

[77] James Hadfield et al. "Nextstrain: real-time tracking of pathogen evolution". In: *Bioinformatics* 34.23 (2018), pp. 4121–4123. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty407. URL: https://doi.org/10.1093/bioinformatics/bty407 (visited on 02/18/2022).

[78] Angela H. Beckett, Kate F. Cook, and Samuel C. Robson. "A pandemic in the age of next-generation sequencing". In: *The Biochemist* 43.6 (2021), pp. 10–15. ISSN: 0954-982X. DOI: 10.1042/bio_2021_187. URL: https://doi.org/10.1042/bio_2021_187 (visited on 02/23/2022).

[79] The Lancet. "Genomic sequencing in pandemics". English. In: *The Lancet* 397.10273 (Feb. 2021). Publisher: Elsevier, p. 445. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(21)00257-9. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00257-9/fulltext (visited on 02/23/2022).

[80] Gavin J. D. Smith et al. "Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic". en. In: *Nature* 459.7250 (June 2009). Number: 7250 Publisher: Nature Publishing Group, pp. 1122–1125. ISSN: 1476-4687. DOI: 10.1038/nature08182. URL: https://www.nature.com/articles/nature08182 (visited on 02/23/2022).

[81] Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. "Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans". In: *New England Journal of Medicine* 360.25 (June 2009). Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa0903810, pp. 2605–2615. ISSN: 0028-

4793. DOI: 10.1056/NEJMoa0903810. URL: https://doi.org/10.1056/NEJMoa0903810 (visited on 02/23/2022).

[82] Roujian Lu et al. "Complete Genome Sequence of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) from the First Imported MERS-CoV Case in China". In: *Genome Announcements* 3.4 (Aug. 2015), e00818–15. ISSN: 2169-8287. DOI: 10.1128/genomeA.00818-15. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4536671/ (visited on 02/23/2022).

[83] Ahmed Kandeil et al. "Complete Genome Sequence of Middle East Respiratory Syndrome Coronavirus Isolated from a Dromedary Camel in Egypt". In: *Genome Announcements* 4.2 (Apr. 2016), e00309–16. ISSN: 2169-8287. DOI: 10.1128/genomeA.00309-16. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4850855/ (visited on 02/23/2022).

[84] Badr M. Al-Shomrani et al. "Genomic Sequencing and Analysis of Eight Camel-Derived Middle East Respiratory Syndrome Coronavirus (MERS-CoV) Isolates in Saudi Arabia". In: *Viruses* 12.6 (June 2020), p. 611. ISSN: 1999-4915. DOI: 10.3390/v12060611. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7354450/ (visited on 02/23/2022).

[85] Fan Wu et al. "A new coronavirus associated with human respiratory disease in China". en. In: *Nature* 579.7798 (Mar. 2020), pp. 265–269. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2008-3. URL: http://www.nature.com/articles/s41586-020-2008-3 (visited on 02/23/2022).

[86] Amy Maxmen. "One million coronavirus sequences: popular genome site hits mega milestone". en. In: *Nature* 593.7857 (Apr. 2021). Bandiera_abtest: a Cg_type: News Number: 7857 Publisher: Nature Publishing Group Subject_term: SARS-CoV-2, Databases, Epidemiology, pp. 21–21. DOI: 10.1038/d41586-021-01069-w. URL: https://www.nature.com/articles/d41586-021-01069-w (visited on 02/23/2022).

[87] Vítor Borges et al. "SARS-CoV-2 introductions and early dynamics of the epidemic in Portugal". en. In: *Communications Medicine* 2.1 (Jan. 2022). Number: 1 Publisher: Nature Publishing Group, pp. 1–11. ISSN: 2730-664X. DOI: 10.1038/s43856-022-00072-0. URL: https://www.nature.com/articles/s43856-022-00072-0 (visited on 02/23/2022).

[88] Leonard Schuele et al. "Future potential of metagenomics in microbiology laboratories". In: *Expert Review of Molecular Diagnostics* 21.12 (2021). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14737159.2021.2001329, pp. 1273–1285. ISSN: 1473-7159. DOI: 10.1080/14737159.2021.2001329. URL: https://doi.org/10.1080/14737159.2021.2001329 (visited on 01/31/2022).

[89] Nicholas J. Loman et al. “High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity”. en. In: *Nature Reviews Microbiology* 10.9 (Sept. 2012). Number: 9 Publisher: Nature Publishing Group, pp. 599–606. ISSN: 1740-1534. DOI: 10.1038/nrmicro2850. URL: https://www.nature.com/articles/nrmicro2850 (visited on 02/08/2022).

[90] J. W. A. Rossen, A. W. Friedrich, and J. Moran-Gilad. “& ESCMID Study Group for Genomic and Molecular Diagnostics (ESGMD). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology”. In: *Clin. Microbiol. Infect.* 24 (2018). DOI: 10.1016/j.cmi.2017.11.001. URL: https://doi.org/10.1016/j.cmi.2017.11.001.

[91] W. M. Dunne, L. F. Westblade, and B. Ford. “Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory”. en. In: *European Journal of Clinical Microbiology & Infectious Diseases* 31.8 (Aug. 2012), pp. 1719–1726. ISSN: 1435-4373. DOI: 10.1007/s10096-012-1641-7. URL: https://doi.org/10.1007/s10096-012-1641-7 (visited on 02/24/2022).

[92] Charles Y. Chiu and Steven A. Miller. “Clinical metagenomics”. en. In: *Nature Reviews Genetics* 20.6 (June 2019). Number: 6 Publisher: Nature Publishing Group, pp. 341–355. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0113-7. URL: https://www.nature.com/articles/s41576-019-0113-7 (visited on 02/08/2022).

[93] Julian R. Marchesi and Jacques Ravel. “The vocabulary of microbiome research: a proposal”. In: *Microbiome* 3.1 (July 2015), p. 31. ISSN: 2049-2618. DOI: 10.1186/s40168-015-0094-5. URL: https://doi.org/10.1186/s40168-015-0094-5 (visited on 02/24/2022).

[94] Ramya Srinivasan et al. “Use of 16S rRNA Gene for Identification of a Broad Range of Clinically Relevant Bacterial Pathogens”. en. In: *PLOS ONE* 10.2 (June 2015). Publisher: Public Library of Science, e0117617. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0117617. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117617 (visited on 02/24/2022).

[95] Isabel Abellan-Schneyder et al. “Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing”. EN. In: *mSphere* (Feb. 2021). Publisher: American Society for Microbiology 1752 N St., N.W., Washington, DC. DOI: 10.1128/mSphere.01202-20. URL: https://journals.asm.org/doi/abs/10.1128/mSphere.01202-20 (visited on 02/24/2022).

[96] J. R. Cole et al. “The Ribosomal Database Project: improved alignments and new tools for rRNA analysis”. In: *Nucleic Acids Research* 37.suppl_1 (Jan. 2009), pp. D141–D145. ISSN: 0305-1048. DOI: 10.1093/nar/gkn879. URL: https://doi.org/10.1093/nar/gkn879 (visited on 02/24/2022).

[97] T. Z. DeSantis et al. “Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB”. In: *Applied and Environmental Microbiology* 72.7 (July 2006), pp. 5069–5072. ISSN: 0099-2240. DOI: 10.1128/AEM.03006-05. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1489311/ (visited on 02/24/2022).

[98] Elmar Pruesse et al. “SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB”. In: *Nucleic Acids Research* 35.21 (Dec. 2007), pp. 7188–7196. ISSN: 0305-1048. DOI: 10.1093/nar/gkm864. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2175337/ (visited on 02/24/2022).

[99] Patrick D. Schloss and Jo Handelsman. “Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness”. In: *Applied and Environmental Microbiology* 71.3 (Mar. 2005). Publisher: American Society for Microbiology, pp. 1501–1506. DOI: 10.1128/AEM.71.3.1501-1506.2005. URL: https://journals.asm.org/doi/10.1128/AEM.71.3.1501-1506.2005 (visited on 02/24/2022).

[100] Jethro S. Johnson et al. “Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis”. en. In: *Nature Communications* 10.1 (Nov. 2019). Number: 1 Publisher: Nature Publishing Group, p. 5029. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13036-1. URL: https://www.nature.com/articles/s41467-019-13036-1 (visited on 02/24/2022).

[101] Joshua Quick et al. “Real-time, portable genome sequencing for Ebola surveillance”. In: *Nature* 530.7589 (Feb. 2016), pp. 228–232. ISSN: 0028-0836. DOI: 10.1038/nature16996. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817224/ (visited on 02/28/2022).

[102] Leonard Schuele et al. “Assessment of Viral Targeted Sequence Capture Using Nanopore Sequencing Directly from Clinical Samples”. en. In: *Viruses* 12.12 (Dec. 2020). Number: 12 Publisher: Multidisciplinary Digital Publishing Institute, p. 1358. ISSN: 1999-4915. DOI: 10.3390/v12121358. URL: https://www.mdpi.com/1999-4915/12/12/1358 (visited on 02/24/2022).

[103] Todd N. Wylie et al. “Enhanced virome sequencing using targeted sequence capture”. In: *Genome Research* 25.12 (Dec. 2015), pp. 1910–1920. ISSN: 1088-9051. DOI: 10.1101/gr.191049.115. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4665012/ (visited on 02/24/2022).

[104] Natacha Couto et al. “Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens”. en. In: *Scientific Reports* 8.1 (Dec. 2018), p. 13767. ISSN: 2045-2322. DOI: 10.1038/s41598-018-31873-w. URL: http://www.nature.com/articles/s41598-018-31873-w (visited on 03/25/2021).

[105] Melissa B. Miller and Yi-Wei Tang. "Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology". In: *Clinical Microbiology Reviews* 22.4 (Oct. 2009), pp. 611–633. ISSN: 0893-8512. DOI: 10.1128/CMR.00019-09. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2772365/ (visited on 02/28/2022).

[106] Chana Palmer et al. "Rapid quantitative profiling of complex microbial populations". In: *Nucleic Acids Research* 34.1 (2006), e5. ISSN: 0305-1048. DOI: 10.1093/nar/gnj007. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1326253/ (visited on 02/28/2022).

[107] Michael R. Wilson et al. "Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing". In: *New England Journal of Medicine* 370.25 (June 2014). Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1401268, pp. 2408–2417. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1401268. URL: https://doi.org/10.1056/NEJMoa1401268 (visited on 02/28/2022).

[108] Prakhar Vijayvargiya et al. "Application of metagenomic shotgun sequencing to detect vector-borne pathogens in clinical blood samples". en. In: *PLOS ONE* 14.10 (Feb. 2019). Publisher: Public Library of Science, e0222915. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0222915. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0222915 (visited on 02/28/2022).

[109] Adriana Sanabria et al. "Shotgun-Metagenomics on Positive Blood Culture Bottles Inoculated With Prosthetic Joint Tissue: A Proof of Concept Study". In: *Frontiers in Microbiology* 11 (2020). ISSN: 1664-302X. URL: https://www.frontiersin.org/article/10.3389/fmicb.2020.01687 (visited on 02/28/2022).

[110] Shota Hirakata et al. "The application of shotgun metagenomics to the diagnosis of granulomatous amoebic encephalitis due to Balamuthia mandrillaris: a case report". In: *BMC Neurology* 21.1 (2021), p. 392. ISSN: 1471-2377. DOI: 10.1186/s12883-021-02418-y. URL: https://doi.org/10.1186/s12883-021-02418-y (visited on 02/28/2022).

[111] Nicholas J. Loman et al. "A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4". In: *JAMA* 309.14 (2013), pp. 1502–1510. ISSN: 0098-7484. DOI: 10.1001/jama.2013.3231. URL: https://doi.org/10.1001/jama.2013.3231 (visited on 02/28/2022).

[112] Andrew D. Huang et al. "Metagenomics of Two Severe Foodborne Outbreaks Provides Diagnostic Signatures and Signs of Coinfection Not Attainable by Traditional Methods". In: *Applied and Environmental Microbiology* 83.3 (Jan. 2017), e02577–16. ISSN: 0099-2240. DOI: 10.1128/AEM.02577-16. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5244306/ (visited on 02/28/2022).

[113] Sijia Li et al. “Microbiome Profiling Using Shotgun Metagenomic Sequencing Identified Unique Microorganisms in COVID-19 Patients With Altered Gut Microbiota”. In: *Frontiers in Microbiology* 12 (2021). ISSN: 1664-302X. URL: https://www.frontiersin.org/article/10.3389/fmicb.2021.712081 (visited on 02/28/2022).

[114] Heather A. Carleton et al. “Metagenomic Approaches for Public Health Surveillance of Foodborne Infections: Opportunities and Challenges”. In: *Foodborne Pathogens and Disease* 16.7 (July 2019). Publisher: Mary Ann Liebert, Inc., publishers, pp. 474–479. ISSN: 1535-3141. DOI: 10.1089/fpd.2019.2636. URL: https://www.liebertpub.com/doi/10.1089/fpd.2019.2636 (visited on 02/28/2022).

[115] Susannah J. Salter et al. “Reagent and laboratory contamination can critically impact sequence-based microbiome analyses”. In: *BMC Biology* 12.1 (Nov. 2014), p. 87. ISSN: 1741-7007. DOI: 10.1186/s12915-014-0087-z. URL: https://doi.org/10.1186/s12915-014-0087-z (visited on 02/28/2022).

[116] Dominic O’Neil, Heike Glowatz, and Martin Schlumpberger. “Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity”. en. In: *Current Protocols in Molecular Biology* 103.1 (2013). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471142727.mb0419s103, pp. 4.19.1–4.19.8. ISSN: 1934-3647. DOI: 10.1002/0471142727.mb0419s103. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb0419s103 (visited on 02/28/2022).

[117] George R. Feehery et al. “A Method for Selectively Enriching Microbial DNA from Contaminating Vertebrate Host DNA”. en. In: *PLoS ONE* 8.10 (2013). Publisher: Public Library of Science. DOI: 10.1371/journal.pone.0076096. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810253/ (visited on 02/28/2022).

[118] Alexa B. R. McIntyre et al. “Comprehensive benchmarking and ensemble approaches for metagenomic classifiers”. In: *Genome Biology* 18.1 (2017), p. 182. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1299-7. URL: https://doi.org/10.1186/s13059-017-1299-7 (visited on 02/28/2022).

[119] J. A. Carriço et al. “A primer on microbial bioinformatics for nonbioinformaticians”. en. In: *Clinical Microbiology and Infection* 24.4 (2018), pp. 342–349. ISSN: 1198-743X. DOI: 10.1016/j.cmi.2017.12.015. URL: https://www.sciencedirect.com/science/article/pii/S1198743X17307097 (visited on 02/18/2022).

[120] Alexandre Angers-Loustau et al. “The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies”. en. In: *F1000Research* 7 (Dec. 2018), p. 459. ISSN: 2046-1402. DOI: 10.12688/f1000research.14509.2. URL: https://f1000research.com/articles/7-459/v2 (visited on 03/25/2021).

[121] Bjorn Gruening et al. “Recommendations for the packaging and containerizing of bioinformatics software”. en. In: *F1000Research* 7 (Mar. 2019), p. 742. ISSN: 2046-1402. DOI: 10.12688/f1000research.15140.2. URL: https://f1000research.com/articles/7-742/v2 (visited on 03/25/2021).

[122] Alexander Sczyrba et al. “Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software”. en. In: *Nature Methods* 14.11 (Nov. 2017). Number: 11 Publisher: Nature Publishing Group, pp. 1063–1071. ISSN: 1548-7105. DOI: 10.1038/nmeth.4458. URL: https://www.nature.com/articles/nmeth.4458 (visited on 03/20/2022).

[123] Peter J. A. Cock et al. “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants”. In: *Nucleic Acids Research* 38.6 (Apr. 2010), pp. 1767–1771. ISSN: 0305-1048. DOI: 10.1093/nar/gkp1137. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/ (visited on 03/02/2022).

[124] W R Pearson and D J Lipman. “Improved tools for biological sequence comparison.” In: *Proceedings of the National Academy of Sciences of the United States of America* 85.8 (Apr. 1988), pp. 2444–2448. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC280013/ (visited on 03/02/2022).

[125] Brent Ewing and Phil Green. “Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities”. en. In: *Genome Research* 8.3 (Mar. 1998). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 186–194. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.8.3.186. URL: https://genome.cshlp.org/content/8/3/186 (visited on 03/02/2022).

[126] Merly Escalona, Sara Rocha, and David Posada. “A comparison of tools for the simulation of genomic next-generation sequencing data”. In: *Nature reviews. Genetics* 17.8 (Aug. 2016), pp. 459–469. ISSN: 1471-0056. DOI: 10.1038/nrg.2016.57. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5224698/ (visited on 03/03/2022).

[127] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15 (2014), pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170. URL: https://doi.org/10.1093/bioinformatics/btu170 (visited on 03/02/2022).

[128] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1 (May 2011), p. 10. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: http://journal.embnet.org/index.php/embnetjournal/article/view/200 (visited on 03/02/2022).

[129] Shifu Chen et al. “fastp: an ultra-fast all-in-one FASTQ preprocessor”. In: *Bioinformatics* 34.17 (2018), pp. i884–i890. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty560. URL: https://doi.org/10.1093/bioinformatics/bty560 (visited on 03/02/2022).

[130] Wouter De Coster et al. “NanoPack: visualizing and processing long-read sequencing data”. In: *Bioinformatics* 34.15 (2018), pp. 2666–2669. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty149. URL: https://doi.org/10.1093/bioinformatics/bty149 (visited on 03/02/2022).

[131] Derrick E Wood and Steven L Salzberg. “Kraken: ultrafast metagenomic sequence classification using exact alignments”. en. In: *Genome Biology* 15.3 (2014), R46. ISSN: 1465-6906. DOI: 10.1186/gb-2014-15-3-r46. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46 (visited on 03/18/2022).

[132] Derrick E. Wood, Jennifer Lu, and Ben Langmead. “Improved metagenomic analysis with Kraken 2”. In: *Genome Biology* 20.1 (Nov. 2019), p. 257. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1891-0. URL: https://doi.org/10.1186/s13059-019-1891-0 (visited on 03/03/2022).

[133] Jennifer Lu et al. “Bracken: estimating species abundance in metagenomics data”. en. In: *PeerJ Computer Science* 3 (Jan. 2017). Publisher: PeerJ Inc., e104. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.104. URL: https://peerj.com/articles/cs-104 (visited on 03/03/2022).

[134] Stephen Nayfach et al. “An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography”. eng. In: *Genome Research* 26.11 (Nov. 2016), pp. 1612–1625. ISSN: 1549-5469. DOI: 10.1101/gr.201863.115.

[135] Peter Menzel, Kim Lee Ng, and Anders Krogh. “Fast and sensitive taxonomic classification for metagenomics with Kaiju”. en. In: *Nature Communications* 7.1 (Apr. 2016). Number: 1 Publisher: Nature Publishing Group, p. 11257. ISSN: 2041-1723. DOI: 10.1038/ncomms11257. URL: https://www.nature.com/articles/ncomms11257 (visited on 03/03/2022).

[136] Duy Tin Truong et al. “MetaPhlAn2 for enhanced metagenomic taxonomic profiling”. en. In: *Nature Methods* 12.10 (Oct. 2015). Number: 10 Publisher: Nature Publishing Group, pp. 902–903. ISSN: 1548-7105. DOI: 10.1038/nmeth.3589. URL: https://www.nature.com/articles/nmeth.3589 (visited on 03/03/2022).

[137] Ryan R. Wick and Kathryn E. Holt. “Benchmarking of long-read assemblers for prokaryote whole genome sequencing”. en. In: *F1000Research* 8 (Feb. 2021), p. 2138. ISSN: 2046-1402. DOI: 10.12688/f1000research.21782.4. URL: https://f1000research.com/articles/8-2138/v4 (visited on 03/25/2021).

[138] Martin Ayling, Matthew D Clark, and Richard M Leggett. "New approaches for metagenome assembly with short reads". In: *Briefings in Bioinformatics* 21.2 (Mar. 2020), pp. 584–594. ISSN: 1477-4054. DOI: 10.1093/bib/bbz020. URL: https://doi.org/10.1093/bib/bbz020 (visited on 03/08/2022).

[139] Tobias Rausch et al. "A consistency-based consensus algorithm for de novo and reference-guided sequence assembly of short reads". In: *Bioinformatics* 25.9 (May 2009), pp. 1118–1124. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp131. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732307/ (visited on 03/08/2022).

[140] Heidi E. L. Lischer and Kentaro K. Shimizu. "Reference-guided de novo assembly approach improves genome reconstruction for related species". In: *BMC Bioinformatics* 18.1 (Nov. 2017), p. 474. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1911-6. URL: https://doi.org/10.1186/s12859-017-1911-6 (visited on 03/08/2022).

[141] Arash Bayat et al. *Methods for De-novo Genome Assembly*. preprint. LIFE SCIENCES, June 2020. DOI: 10.20944/preprints202006.0324.v1. URL: https://www.preprints.org/manuscript/202006.0324/v1 (visited on 03/08/2022).

[142] E. W. Myers et al. "A whole-genome assembly of Drosophila". eng. In: *Science (New York, N.Y.)* 287.5461 (Mar. 2000), pp. 2196–2204. ISSN: 0036-8075. DOI: 10.1126/science.287.5461.2196.

[143] Jonathan Laserson, Vladimir Jojic, and Daphne Koller. "Genovo: De Novo Assembly for Metagenomes". In: *Research in Computational Molecular Biology*. Ed. by David Hutchison et al. Vol. 6044. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 341–356. ISBN: 978-3-642-12682-6 978-3-642-12683-3. DOI: 10.1007/978-3-642-12683-3_22. URL: http://link.springer.com/10.1007/978-3-642-12683-3_22 (visited on 03/09/2022).

[144] Afiahayati, Kengo Sato, and Yasubumi Sakakibara. "An extended genovo metagenomic assembler by incorporating paired-end information". en. In: *PeerJ* 1 (Oct. 2013). Publisher: PeerJ Inc., e196. ISSN: 2167-8359. DOI: 10.7717/peerj.196. URL: https://peerj.com/articles/196 (visited on 03/09/2022).

[145] You-Yu Lin et al. "De novo assembly of highly polymorphic metagenomic data using in situ generated reference sequences and a novel BLAST-based assembly pipeline". In: *BMC Bioinformatics* 18.1 (2017), p. 223. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1630-z. URL: https://doi.org/10.1186/s12859-017-1630-z (visited on 03/09/2022).

[146] Robert Vaser and Mile Šikić. *Yet another de novo genome assembler*. en. preprint. Bioinformatics, May 2019. DOI: 10.1101/656306. URL: http://biorxiv.org/lookup/doi/10.1101/656306 (visited on 03/09/2022).

[147] Sergey Koren et al. “Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation”. en. In: *Genome Research* 27.5 (May 2017). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 722–736. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.215087.116. URL: https://genome.cshlp.org/content/27/5/722 (visited on 03/09/2022).

[148] Anton Bankevich et al. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing”. en. In: *Journal of Computational Biology* 19.5 (May 2012), pp. 455–477. ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2012.0021. URL: http://www.liebertpub.com/doi/10.1089/cmb.2012.0021 (visited on 03/18/2022).

[149] Alexandre Souvorov, Richa Agarwala, and David J. Lipman. “SKESA: strategic k-mer extension for scrupulous assemblies”. In: *Genome Biology* 19.1 (Oct. 2018), p. 153. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1540-z. URL: https://doi.org/10.1186/s13059-018-1540-z (visited on 03/14/2022).

[150] Dinghua Li et al. “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph”. In: *Bioinformatics* 31.10 (May 2015), pp. 1674–1676. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv033. URL: https://doi.org/10.1093/bioinformatics/btv033 (visited on 03/14/2022).

[151] Daniel R. Zerbino and Ewan Birney. “Velvet: Algorithms for de novo short read assembly using de Bruijn graphs”. en. In: *Genome Research* 18.5 (May 2008). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 821–829. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.074492.107. URL: https://genome.cshlp.org/content/18/5/821 (visited on 03/14/2022).

[152] Sébastien Boisvert, François Laviolette, and Jacques Corbeil. “Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies”. en. In: *Journal of Computational Biology* 17.11 (Nov. 2010), pp. 1519–1533. ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2009.0238. URL: http://www.liebertpub.com/doi/10.1089/cmb.2009.0238 (visited on 03/14/2022).

[153] Ruibang Luo et al. “SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler”. In: *GigaScience* 1.1 (Dec. 2012), pp. 2047–217X–1–18. ISSN: 2047-217X. DOI: 10.1186/2047-217X-1-18. URL: https://doi.org/10.1186/2047-217X-1-18 (visited on 03/14/2022).

[154] Alexey Gurevich et al. “QUAST: quality assessment tool for genome assemblies”. eng. In: *Bioinformatics (Oxford, England)* 29.8 (Apr. 2013), pp. 1072–1075. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt086.

[155] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. "Singularity: Scientific containers for mobility of compute". en. In: *PLOS ONE* 12.5 (Nov. 2017). Publisher: Public Library of Science, e0177459. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0177459. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177459 (visited on 03/17/2022).

[156] Paolo Di Tommaso et al. "Nextflow enables reproducible computational workflows". en. In: *Nature Biotechnology* 35.4 (Apr. 2017), pp. 316–319. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3820. URL: http://www.nature.com/articles/nbt.3820 (visited on 03/24/2021).

[157] Felix Mölder et al. "Sustainable data analysis with Snakemake". In: *F1000Research* 10 (Apr. 2021), p. 33. ISSN: 2046-1402. DOI: 10.12688/f1000research.29032.2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8114187/ (visited on 03/17/2022).

[158] Enis Afgan et al. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update". In: *Nucleic Acids Research* 46.W1 (July 2018), W537–W544. ISSN: 0305-1048. DOI: 10.1093/nar/gky379. URL: https://doi.org/10.1093/nar/gky379 (visited on 03/17/2022).

[159] Sarah K. Hilton et al. "Metataxonomic and Metagenomic Approaches vs. Culture-Based Techniques for Clinical Pathology". In: *Frontiers in Microbiology* 7 (2016). ISSN: 1664-302X. URL: https://www.frontiersin.org/article/10.3389/fmicb.2016.00484 (visited on 03/03/2022).

[160] Evan Bolyen et al. "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2". en. In: *Nature Biotechnology* 37.8 (Aug. 2019). Number: 8 Publisher: Nature Publishing Group, pp. 852–857. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0209-9. URL: https://www.nature.com/articles/s41587-019-0209-9 (visited on 03/03/2022).

[161] Patrick D. Schloss et al. "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities". In: *Applied and Environmental Microbiology* 75.23 (Dec. 2009). Publisher: American Society for Microbiology, pp. 7537–7541. DOI: 10.1128/AEM.01541-09. URL: https://journals.asm.org/doi/10.1128/AEM.01541-09 (visited on 03/04/2022).

[162] Robert C. Edgar. "UPARSE: highly accurate OTU sequences from microbial amplicon reads". en. In: *Nature Methods* 10.10 (Oct. 2013). Number: 10 Publisher: Nature Publishing Group, pp. 996–998. ISSN: 1548-7105. DOI: 10.1038/nmeth.2604. URL: https://www.nature.com/articles/nmeth.2604 (visited on 03/04/2022).

[163] Moira Marizzoni et al. "Comparison of Bioinformatics Pipelines and Operating Systems for the Analyses of 16S rRNA Gene Amplicon Sequences in Human Fecal Samples". In: *Frontiers in Microbiology* 11 (2020). ISSN: 1664-302X. URL: https:

//www.frontiersin.org/article/10.3389/fmicb.2020.01262 (visited on 03/04/2022).

[164] Sarah L. Westcott and Patrick D. Schloss. "De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units". In: *PeerJ* 3 (Dec. 2015), e1487. ISSN: 2167-8359. DOI: 10.7717/peerj.1487. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4675110/ (visited on 03/04/2022).

[165] Xiaolin Hao, Rui Jiang, and Ting Chen. "Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering". en. In: *Bioinformatics* 27.5 (Mar. 2011), pp. 611–618. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btq725. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq725 (visited on 03/04/2022).

[166] Sergey Nurk et al. "metaSPAdes: a new versatile metagenomic assembler". en. In: *Genome Research* 27.5 (May 2017), pp. 824–834. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.213959.116. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.213959.116 (visited on 03/25/2021).

[167] Nathan D Olson et al. "Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes". In: *Briefings in Bioinformatics* 20.4 (Aug. 2017), pp. 1140–1150. ISSN: 1467-5463. DOI: 10.1093/bib/bbx098. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6781575/ (visited on 03/17/2022).

[168] Mikhail Kolmogorov et al. "metaFlye: scalable long-read metagenome assembly using repeat graphs". en. In: *Nature Methods* 17.11 (Nov. 2020). Number: 11 Publisher: Nature Publishing Group, pp. 1103–1110. ISSN: 1548-7105. DOI: 10.1038/s41592-020-00971-x. URL: https://www.nature.com/articles/s41592-020-00971-x (visited on 03/20/2022).

[169] Hanno Teeling and Frank Oliver Glöckner. "Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective". eng. In: *Briefings in Bioinformatics* 13.6 (Nov. 2012), pp. 728–742. ISSN: 1477-4054. DOI: 10.1093/bib/bbs039.

[170] Karel Sedlar, Kristyna Kupkova, and Ivo Provaznik. "Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics". eng. In: *Computational and Structural Biotechnology Journal* 15 (2017), pp. 48–55. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2016.11.005.

[171] Ivan Gregor et al. "PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes". en. In: *PeerJ* 4 (Feb. 2016). Publisher: PeerJ Inc., e1603. ISSN: 2167-8359. DOI: 10.7717/peerj.1603. URL: https://peerj.com/articles/1603 (visited on 03/20/2022).

[172] Donovan H. Parks et al. “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes”. en. In: *Genome Research* 25.7 (July 2015). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1043–1055. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.186072.114. URL: https://genome.cshlp.org/content/25/7/1043 (visited on 03/20/2022).

[173] Alla Mikheenko, Vladislav Saveliev, and Alexey Gurevich. “MetaQUAST: evaluation of metagenome assemblies”. eng. In: *Bioinformatics (Oxford, England)* 32.7 (Apr. 2016), pp. 1088–1090. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv697.

Chapter 2

Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens

This chapter is a reproduction of the following publication:

N. Couto, L. Schuele, E.C. Raangs, M. P. Machado, C. I. Mendes, T. F. Jesus, M. Chlebowicz, S. Rosema, M. Ramirez, J. A. Carriço, I. B. Autenrieth, A. W. Friedrich, S. Peter and J. W. Rossen. Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. Sci Rep 8, 13767 (2018). `https://doi.org/10.1038/s41598-018-31873-w`

The supplementary information referred throughout the text can be consulted in this chapter before the section of references.

As mentioned in Chapter 1.2.3.2, shotgun metagenomic (SMg) approaches have been a growing interest to deliver clinically relevant results without *a priori* knowledge of what to expect from a particular clinical sample or patient. The capacity to detect all potential pathogens in a sample has great potential utility in the diagnosis of infectious disease. However, it is unclear how the variety of available methods impacts the end results.

In this publication SMg was applied to nine body fluid samples and one tissue sample from patients at the University Medical Center Groningen (UMCG), included one sample from peritoneal fluid, five from pus (3 abscesses and 2 empyema), two from synovial fluid of knees with prosthesis, one from sputum and one from a bone biopsy. The results of microbial identification through whole genome sequencing (WGS) and SMg were compared to standard culture-based microbiological methods.

In order to evaluate and compare the accuracy and reliability of the bioinformatics analyses in providing the closest results to culture and WGS of any cultured isolates, three different bioinformatic pipelines (two commercially and one freely available) were used. Most

pathogens identified by culture were also identified through metagenomics, but substantial differences were noted between the taxonomic classification tools.

My contribution to this publication included the bioinformatics analysis of all the samples using a unix-based approach. I performed quality assessment and quality control of the WGS and SMg data, the removal of host sequencing from the samples, and the taxonomic identification of the remaining reads in each sample though 3 different methods: MetaPhlan2, Kraken and MiDAS. Gene detection directly from the reads for bacterial typing was also performed using metaMLST, ReMatCh, and Bowtie2 and Samtools. Finally, the reads were assembled using the SPAdes genome assembler, with and without metagenomic mode according to the sample being processed.

Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens

Natacha Couto[1], Leonard Schuele[1,2], Erwin C. Raangs[1], Miguel P. Machado[3], Catarina I. Mendes[1,3], Tiago F. Jesus[3], Monika Chlebowicz[1], Sigrid Rosema[1], Mário Ramirez[3], João A. Carriço[3], Ingo B. Autenrieth[2], Alex W. Friedrich[1], Silke Peter[2], John W. Rossen[1]

[1] University of Groningen, University Medical Center Groningen, Department of Medical Microbiology, Groningen, The Netherlands;

[2] Institute of Medical Microbiology and Hygiene, University of Tübingen, Germany;

[3] Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Portugal.

2.1 Abstract

High throughput sequencing has been proposed as a one-stop solution for diagnostics and molecular typing directly from patient samples, allowing timely and appropriate implementation of measures for treatment, infection prevention and control. However, it is unclear how the variety of available methods impacts the end results. We applied shotgun metagenomics on diverse types of patient samples using three different methods to deplete human DNA prior to DNA extraction. Libraries were prepared and sequenced with Illumina chemistry. Data was analysed using methods likely to be available in clinical microbiology laboratories using genomics. The results of microbial identification were compared to standard culture-based microbiological methods. On average, 75% of the reads were corresponded to human DNA, being a major determinant in the analysis outcome. None of the kits was clearly superior suggesting that the initial ratio between host and microbial DNA or other sample characteristics were the major determinants of the proportion of microbial reads. Most pathogens identified by culture were also identified through metagenomics, but substantial differences were noted between the taxonomic classification tools. In two cases the high number of human reads resulted in insufficient sequencing depth of bacterial DNA for identification. In three samples, we could infer the probable multilocus sequence type of the most abundant species. The tools and databases used for taxonomic classification and antimicrobial resistance identification had a key impact on the results, recommending that efforts need to be aimed at standardisation of the analysis methods if metagenomics is to be used routinely in clinical microbiology.

2.2 Introduction

Classical microbial culture is still considered the gold standard in medical microbiology. Several molecular detection techniques have been implemented but these are generally geared towards specific pathogens (e.g. specific RT-PCR or microarrays). Even when unbiased molecular approaches are used, such as 16S/18S rRNA gene sequencing, these do not provide all the information that can be obtained by culturing, e.g., antimicrobial susceptibility and molecular typing information. However, microbial culture is laborious and time-consuming and new methods are needed to replace it. Ideally, a single method should provide rapid identification and characterisation of clinically relevant pathogens directly from a sample in order to guide therapy, predict potential treatment failures and to reveal possible transmission events.

Shotgun metagenomics (SMg) is a culture-independent technique that provides valuable information not only at the identification level, but also at the level of molecular characterisation. Studies have shown that it has added value in terms of detection sensitivity and personalised treatment in clinical microbiology, when identifying bacteria [1, 2] or viruses [3]. Indeed Gyarmati et al., 2016 [4], used a sequence-based metagenomics approach directly from blood to detect non-culturable, difficult-to-culture and non-bacterial pathogens. The authors were able, through SMg, to detect viral and fungal pathogens together with bacteria, which had not been detected through classical microbiology. Additionally, SMg can be used for infection prevention, having the potential to identify transmission events directly from clinical samples [5]. For example, SMg was proven valuable for the identification of inter-host nucleotide variations occurring after direct transmission of noroviruses causing gastroenteritis [5]. Hasman and colleagues (2014) [1] were able to identify urinary pathogens directly from urine, as well as antimicrobial resistant genes compatible with the resistant phenotype determined through antimicrobial susceptibility testing. They also identified almost perfect phylogenetic matches between whole-genome sequence (WGS) data obtained by metagenomics and WGS of pure isolates.

Despite the promise of SMg of becoming a one-stop solution in clinical microbiology, SMg still has several challenges to overcome. One of the greatest challenges is the choice of the extraction and sequencing protocols, as well of the type of controls [6]. The extraction protocol should efficiently and specifically isolate microbial DNA/RNA, while removing the host DNA/RNA [7]. However, the variety of clinical samples used in the diagnosis of distinct types of infection (e.g. tissues versus fluids), poses a serious challenge for standardisation, an essential step if these methods are to be used by routine diagnostic laboratories. The sequencing protocol is also dependent on the pathogens of interest (e.g. bacteria versus viruses), sequencing strategy (DNA and/or RNA), required turnaround time, sequencing depth and error tolerance [6]. The use of defined controls is necessary for validation of each experiment and these should be adapted for every type of infection and sample type and should consist of a combination of known positive specimens, pathogen-negative patient

specimens and pathogen-negative patient specimens spiked with live microorganisms or pure DNA [6].

Another potential challenge are the metagenomics analysis tools. Recent studies have evaluated the different SMg sequence classification methods [8]. These use different methodologies for classification: sequence similarity-based methods, sequence composition-based methods and hybrid methods [8]. They differ not only in the algorithms for detecting the microorganisms present, but also in the databases used. This high variability leads to different results, not only at the microorganism classification level but also when evaluating the relative abundance of these pathogens [8]. A recent study evaluated the accuracy of 38 bioinformatics methods using both *in silico* and *in vitro* generated mock bacterial communities. Dozens to hundreds of species were falsely predicted by the most popular software, and no software clearly outperformed the others [8]. In the absence of studies comparing the outputs of different analysis methods in clinical samples, users may decide which methods to use based on personal experience with a given tool, availability of the tool in the laboratory or its ease of use. This poses a great challenge when providing reproducible results and creates uncertainty regarding the reliability of the information derived. This is a major barrier to the implementation of SMg approaches in routine clinical microbiology laboratories.

In this study, the aim was to identify the critical steps when using SMg for the identification and characterization of microbial pathogens directly from clinical specimens using methods that are likely to be available in clinical microbiology laboratories wanting to implement genomics for pathogen identification or molecular epidemiology studies. For this purpose, we used three human-DNA depletion kits and evaluated a diverse set of bioinformatics tools (commercial and non-commercial) in order to investigate how well they performed and what would the differences be in terms of taxonomic classification, antimicrobial resistance gene detection and typing directly from patient samples, bypassing culture.

2.3 Methods

2.3.1 Sample collection

Nine body fluid samples and one tissue sample entering the Medical Microbiology laboratory were selected for metagenomics sequencing. These included one sample from peritoneal fluid, five from pus (3 abscesses and 2 empyema), two from synovial fluid of knees with prosthesis, one from sputum and one from a bone biopsy (Table **??**). All samples were stored at 4ºC for a variable period (2-10 days). The samples used for the present analyses were collected during routine diagnostics and infection prevention and control investigations. All procedures were carried out according to guidelines and regulations of University Medical Centre Groningen (UMCG) concerning the use of patient materials for the validation of

Table 2.1: Characteristics of the samples and mapping of trimmed reads against a human genome hg19 (%) using CLC Genomics Workbench v10.0.1.

Sample	Sample type	DNA extraction method	Total number of reads	Mapped reads against hg19	**Unmapped reads**
Sample 1	Peritoneal fluid	Ultra-Deep Microbiome Prep (Molzym)	5892978	5,249,063 (89.2%)	632,951 (10.8%)
Sample 2	Pus (abscess)	Ultra-Deep Microbiome Prep (Molzym)	9603346	7,828.746 (81.6%)	1,770,558 (18.4%)
Sample 3	Synovial fluid	Ultra-Deep Microbiome Prep (Molzym)	8615810	8,254,594 (95.9%)	355,200 (4.1%)
Sample 4	Synovial fluid	Ultra-Deep Microbiome Prep (Molzym)	6078166	6,015,945 (99.0%)	61,099 (1.0%)
Sample 5	Pus (abscess)	Ultra-Deep Microbiome Prep (Molzym)	8368930	309,588 (3.7%)	8,052,272 (96.3%)
Sample 6	Pus (empyema)	QIAamp DNA Microbiome Kit (Qiagen)	2912802	2,877,066 (98.8%)	34,506 (1.1%)
Sample 7	Pus (empyema)	QIAamp DNA Microbiome Kit (Qiagen)	1486700	922,932 (62.2%)	561,772 (37.8%)
Sample 8	Bone biopsy	Micro-DXTM (Molzym)	6534866	229,149 (3.5%)	6,303,803 (96.5%)
Sample 9	Pus (abscess)	Micro-DXTM (Molzym)	6173132	6,081,612 (98.5%)	89,922 (1.5%)
Sample 10	Sputum	Micro-DXTM (Molzym)	7596836	7,337,832 (96.7%)	235,520 (3.3%)
Negative control	Water	QIAamp DNA Microbiome Kit (Qiagen)	1730738	1,706,861 (98.9%)	19,805 (1.2%)

clinical methods, which are in compliance with the guidelines of the Federation of Dutch Medical Scientific Societies (FDMSS). Every patient entering the UMCG is informed that samples taken may be used for research and publication purposes, unless they indicate that they do not agree to it. This procedure has been approved by the Medical Ethical Committee of the UMCG. Informed consent was obtained from all individuals or their guardians prior to study participation. All samples were used after performing and completing a conventional microbiological diagnosis and were coded to protect patients' confidentiality. All experiments were performed in accordance with the guidelines of the Declaration of Helsinki and the institutional regulations.

2.3.2 Classic culturing and susceptibility testing

The samples were cultured following methods routinely used in our institution. Briefly, samples were streaked onto five plates (Mediaproducts BV, Groningen, The Netherlands) - blood agar (aerobic), chocolate agar (aerobic), McConkey agar (aerobic), Brucella agar (anaerobic) and Sabouraud Dextrose +AV (aerobic) - and incubated overnight under aerobic and anaerobic atmosphere at 37ºC. The two pus samples were also plated onto Phenylethyl alcohol sheep blood agar (PEA), Kanamycin vancomycin laked blood (KVLB) agar and Bacteroides bile esculin (BBE) agar and incubated under anaerobic conditions overnight. The isolates recovered were subjected to susceptibility testing by Vitek 2 using either the AST-P559 (Gram-positive bacteria) or the AST-N344 (Gram-negative bacteria) card (bioMérieux, Marcy-l'Étoile, France) and identified by MALDI-TOF MS (Bruker Daltonik, Gmbh, Germany) using standard protocols.

2.3.3 DNA extraction, library preparation and sequencing

The DNA for metagenomic sequencing was isolated using the Ultra-Deep Microbiome Prep (Molzym Life Science, Bremen, Germany), Micro-Dx™kit (Molzym Life Science) or QIAamp DNA Microbiome Kit (Qiagen, Hilden, Germany) directly from the clinical samples and a negative control consisting of a mock sample of DNA and RNA free water

(Table **??**). These kits include human DNA depletion steps. The QIAamp DNA Microbiome Kit was used according to the manufacturer's protocol with an additional 5 min air-dry step before elution. For microbial lysis, a Precellys 24 homogeniser (Bertin, Montigny-le-Bretonneux, France) set to 3 times 30 seconds at 5000 rpm separated by 30 seconds was used. After extraction, DNA was quantified with the Qubit 2.0 (Life Technologies, ThermoFisher Scientific, Waltham, Massachusetts, EUA) and NanoDrop 2000 (ThermoFisher Scientific). The DNA quality was assessed using the Genomic DNA ScreenTape and Agilent 2200 TapeStation System (Agilent Technologies, California, United States of America). Isolated DNA was purified using Agencourt AMPure XP beads (Beckman Coulter, California, United States of America) according to the manufacturer's instructions, to eliminate small DNA fragments and chemical contaminants (e.g. benzonase). The DNA was then diluted to 0.2 ng/μl and 1 ng was used for the library preparation, using the Nextera XT Library Preparation kit (Illumina, California, United States of America), according to the manufacturer's protocol. Cluster generation and sequencing were performed with the MiSeq Reagent Kit v2 500-cycles Paired-End in a MiSeq instrument (Illumina). Samples were sequenced in batches of 5 samples on a single flow cell.

For the DNA extraction of bacterial isolates (when an isolate was recovered from culture), we used the UltraClean Microbial DNA Isolation Kit (Mo Bio), with some modifications. We started with solid cultures and resuspended a 10 μl-loopfull of culture directly into the tube with the microbeads and microbead solution. The library preparation, cluster generation and sequencing was performed as described above. Strains were sequenced in batches of 12 to 16 on a single flow cell.

2.3.4 Bioinformatics analyses

In order to evaluate and compare the accuracy and reliability of the bioinformatics analyses in providing the closest results to culture and WGS of any cultured isolates, three different pipelines (two commercially and one freely available) were used (Figure 1). Different tools to perform raw read quality control, filtering and trimming were used and reads were mapped against the human genome (hg19) before performing taxonomic classification. Reads mapping to hg19 where removed from the analysis to increase the efficiency of the bioinformatics tools. Typing (MLST), phylogenetic analysis, plasmid analysis, detection of antimicrobial resistance and virulence genes was performed. To determine the appropriateness of SMg as predictor of the WGS (chromosome and plasmids), SMg results obtained were compared with the results of WGS of any bacterial isolates obtained from culturing the sample.

All the parameters used in each approach are available in Supplementary Table 1.

2.3.4.1 Unix-based approach

For the metagenomics data, read quality control and cleaning was performed using FastQC v0.11.5 and Trimmomatic v0.36, respectively, through the INNUca v2.6 pipeline[1], excluding assembly and polishing. Using a reference mapping approach against the human genome (UCSC hg19), human reads were discarded using Bowtie 2 v2.3.2 [9] and SAMtools v1.3.1 [10]. Those paired reads that did not map against the human genome were used in subsequent analyses. The bacterial species were identified through Kraken v0.10.5-beta [11] using the miniKraken database (pre-built 4 GB database constructed from complete bacterial, archaeal and viral genomes in RefSeq, as of Dec. 8, 2014), MIDAS [12] using the midas_db_v1.2 database (>30,000 bacterial reference genomes, as of May 9, 2018) and MetaPhlAn2 v2.0 [13] using the database provided by the tool ($\sim$13,500 bacterial and archaeal, $\sim$3,500 viral, and $\sim$110 eukaryotic reference genomes, as of May 9, 2018). The sequence type (ST) was obtained through metaMLST v1.1 [14] based on the metamlstDB_2017. Antimicrobial resistance genes were detected using ReMatCh v3.2[2], a read mapping tool that uses Bowtie 2 v2.3.2 [9] and the following rules for gene presence/absence: genes were considered present when $\geq$ 80% of the reference sequence was covered and the sample sequence was $\geq$ 70% identical to the one used as reference. For that, ResFinder database (2231 genes, downloaded on 29-06-2017) was used as reference and, due to the low coverage of microbial metagenomics samples, a minimal coverage depth of 1 read was set to consider a reference sequence position as covered (and therefore present in the sample data), as well as to perform base call (used for sequence identity determination). Finally, the assembly was accomplished through SPAdes v3.10.1 [15].

Plasmid detection was achieved by running the script PlasmidCoverage[3], using the plasmid sequences downloaded from NCBI RefSeq (`ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plasmid/`, as of May 11, 2017). The script uses Bowtie 2 v2.2.9 [9], to map the pre-processed input reads against the plasmid database (Bowtie2 index for all plasmid sequences). For Bowtie 2 we used the '-k' option, allowing each read to map to as many plasmid sequences as present in the NCBI RefSeq plasmid database (since plasmid sequences are modular) [**barcia_identification_2011**, 16]. Then, this pipeline used SAMtools v1.3.1 [10] to estimate the coverage for each position, and reported the length of plasmid sequence covered (in percentage) and average depth (mean number of reads mapped against a given position in each plasmid). Plasmids with less than 80% of its length covered were excluded from the final results in line with what has described elsewhere [17]. The pATLAS tool[4] was used to visualise which plasmids were present.

For the WGS reads of the bacterial isolates, the whole INNUca v2.6 pipeline was run, including SPAdes assembly and polishing. Plasmids were detected as mentioned previously.

[1] `https://github.com/B-UMMI/INNUca/`
[2] `https://github.com/B-UMMI/ReMatCh/`
[3] `https://github.com/tiagofilipe12/PlasmidCoverage`
[4] `http://www.patlas.site/`

2.3.4.2 Commercial-based approach

The fastq files containing the reads were uploaded into CLC Genomics Workbench v10.1.1, using the following options: Illumina import, paired-reads, paired-end (forward-reverse) and minimum distance of 1 and a maximum distance of 1000 (default). The trimming was performed using the default settings, except the quality trimming score limit was set to 0.01 and we added a Trim adapter list containing Illumina adapters. The mapping was performed with the Map Reads to Reference tool, using the hg19 genome as reference. The default settings were used with the addition of the collect un-mapped reads option. The *de novo* assembly tool was used for the assembly (even for the metagenomics reads) and, apart from the word size, which was changed to 29, all the settings were default. Two tools were used for the microbial identification, Taxonomic Profiling and Find Best Matches using K-mer Spectra (Microbial Genomics Module). In both, the bacterial and fungal databases were downloaded from NCBI RefSeq (with the Only Complete Genomes option turned off; minimum length 500,000 nucleotides) on 08-07-2017 (bacterial, 70,868 sequences) and 25-05-2017 (fungal, 377 sequences). The antimicrobial resistance genes were detected, based on the assembled contigs, using the Find Resistance tool (Microbial Genomics Module) and were initially only considered present when they were $\geq$ 70% identical to the reference and $\geq$ 80% of the sequence was covered. The analysis was also repeated using $\geq$ 40% and $\geq$ 20% of sequence coverage for comparison purposes. The database containing the antimicrobial resistance genes was downloaded directly to the software from ResFinder[5] (downloaded on 05-07-2017, 2156 sequences). The MLST was determined through the Identify MLST tool (Microbial Genomics Module), using all MLST schemes available at PubMLST (04-03-2017). The same database used for plasmid detection in Unix, was used for mapping the reads in CLC Genomics Workbench. Again, plasmids with less than 80% of its length covered were excluded from the final results. For WGS reads we used the Trim Sequences tool and the assembly, antimicrobial resistance genes detection, and MLST determination were performed as before.

2.3.4.3 Web-based approaches

The fastq files containing the reads were uploaded into the BaseSpace[6] website. First, the raw forward and reverse fastq reads were subjected to FASTQ Toolkit for adapter/quality trimming and length filtering with standard settings and length filtering adjusted to a minimum of 100 and a maximum of 500. The trimmed reads were then used as input for all the following processes. The available microorganism identification apps Kraken v1.0.0, MetaPhlAn v1.0.0 and GENIUS v.1.1.0 were used with the standard settings/parameters. SEAR was used to detect antimicrobial resistance genes, maintaining the standard settings except for the clustering stringency which was set to 0.98 and the annotation stringency was

[5] `https://cge.cbs.dtu.dk/services/data.php`
[6] `https://basespace.illumina.com`

set to 40. The SPAdes Genome Assembler v3.9.0 app was run with the standard parameters for multi cell data type. For metagenomic datatype settings, the running mode was set to only assembly and careful mode was disabled.

The reads were uploaded into CosmosID[7] and Taxonomer[8] [18] directly without any quality trimming. We used the Full Analysis mode in Taxonomer.

2.3.4.4 wgMLST analyses

Typing was done by MLST and wgMLST analyses obtained using Ridom SeqSphere+ v4.0.1. The genomic data (assembled contigs) obtained from SMg was compared to the data obtained through WGS. Since no cg/wgMLST scheme was available for *Escherichia coli*, *Enterococcus faecalis*, *Ochrobactrum intermedium* and *Staphylococcus haemolyticus*, cgMLST and accessory genome schemes were constructed, using Ridom SeqSphere+ cgMLST Target Definer with the following parameters: a minimum length filter that removes all genes smaller than 50 bp; a start codon filter that discards all genes that contain no start codon at the beginning of the gene; a stop codon filter that discards all genes that contain no stop codon or more than one stop codon or that do not have the stop codon at the end of the gene; a homologous gene filter that discards all genes with fragments that occur in multiple copies within a genome (with identity of 90% and >100 bp overlap); and a gene overlap filter that discards the shorter gene from the cgMLST scheme if the two genes affected overlap $>$4 bp. The remaining genes were then used in a pairwise comparison using BLAST version 2.2.12 (parameters used were word size 11, mismatch penalty -1, match reward 1, gap open costs 5, and gap extension costs 2). All genes of the reference genome that were common in all query genomes with a sequence identity of $\geq$ 90% and 100% overlap and, with the default parameter stop codon percentage filter turned on, formed the final cgMLST scheme. The combination of all alleles in each strain formed an allelic profile that was used to generate minimum spanning trees using the parameter "pairwise ignore missing values" during distance calculation [19].

2.3.4.5 Statistical analysis

The sensitivity and positive predictive value of each taxonomic classification method were determined. Classical culture and MALDI-TOF identifications were considered as the gold standard. The true positives were considered when the same bacterial species were identified by culture/MALDI-TOF and the taxonomic classification method. The false positives were detected when bacterial species different from those identified by culture/MALDI-TOF, were identified by the taxonomic classification method. The false negatives were determined

[7] `https://app.cosmosid.com/login`
[8] `https://www.taxonomer.com/`

when the bacterial species identified by culture/MALDI-TOF were not identified by the taxonomic classification method.

2.4 Results

2.4.1 Classical identification

Nine body fluid samples and one tissue sample from 9 different patients were sequenced, including one sample from peritoneal fluid, five from pus (3 abscesses and 2 empyemas), two from synovial fluid of knees with prosthesis, one from sputum and one from a bone biopsy (Table 2.1). In total 15 different isolates obtained from the 10 samples were considered of possible clinical significance and were selected for species identification and antimicrobial susceptibility testing during routine work up of the samples (Table 2.2,**??** and **??**). In samples 2 and 3, only one colony-forming unit (CFU) of *Escherichia coli* and *Staphylococcus epidermidis*, respectively, was detected after 48 hours of incubation. In samples 2 and 5, the anaerobic cultures were mixed to such an extent, that no further characterization of the colonies was performed, and the results were reported as anaerobic mixed culture.

Antimicrobial susceptibility testing, revealed three isolates to be fully susceptible, while the others were resistant to at least one antimicrobial. Two isolates, one *Staphylococcus haemolyticus* and one *S. epidermidis* were oxacillin-resistant and positive in the cefoxitin test (Vitek 2).

There was fungal growth in 2 samples (1 and 5) that included two Candida species (one *Candida glabrata* and one *Candida albicans*). The different bacterial and fungal species identified in each sample are shown in Tables 2.2, **??** and **??**.

2.4.2 Comparison of standard procedures and shotgun metagenomics for the identification of clinically relevant pathogens

The tools used for taxonomic classification are shown in Figure **??**. The total number of reads and the total number of reads mapped against the human genome (hg19) varied between samples, ranging from 3.5% to 98.9% (Table 2.1). The abundance of human reads was not determined by the type of sample but was probably influenced by individual characteristics of each sample and the success of the methods used in depleting the human DNA. We identified the microorganisms present using different taxonomical methods, including three Unix-based tools (Kraken, Metaphlan2 and MIDAS), web-based tools including both commercial and freely available solutions (BaseSpace, Taxonomer and CosmosID) and one commercial approach having a graphical interface (CLC Genomics Workbench v10.0.1).

2. CRITICAL STEPS IN CLINICAL SHOTGUN METAGENOMICS FOR THE CONCOMITANT DETECTION AND TYPING OF MICROBIAL PATHOGENS

Table 2.2: Microorganisms identified by conventional methods, WGS and using shotgun metagenomics and the taxonomic classification methods in Unix.

Sample number	Culture result (CFU)[a]	Conventional identification (MALDI-TOF)	WGS-based identification	Shotgun metagenomics		
				Kraken[b]	MIDAS[c]	MetaPhlAn[c]
1	10^3 10^3 10	*E. faecium* *S. haemolyticus* *C. glabrata*	*E. faecium* *S. haemolyticus* -	*E. faecium* (34.6%) *S. haemolyticus* (10.1%) -	*E. faecium* (62.0%) *S. haemolyticus* (28.0%) -	*E. faecium* (66.6%) *S. haemolyticys* (27.7%) -
2	10^3 1 Not determined	*E. avium* *E. coli* Anaerobes	-# -# -#	Not identified* Not identified* Several species (29.5%)	Not identified* Not identified* Several species (100.0%)	Not identified* Not identified* Several species (100.0%)
3	1	*S. epidermidis*	-#	S. aureus (0.2%)	Not identified*	Not identified*
4	10^3	*S. aureus*	*S. aureus*	*S. aureus* (0.73%)	*S. aureus* (100%)	*S. aureus* (100%)
5	$\geq 10^5$ $\geq 10^5$ 10^3 10^3 Not determined 10	*E. coli* *K. oxytoca* *S. anginosus* *E. faecalis* Anaerobes *C. albicans*	*E. coli* *K. oxytoca* -# *E. faecalis* -# -#	*E. coli* (9.7%) *K. oxytoca* (0.5%) *S. anginosus* (0.07%) *E. faecalis* (0.3%) Several species (12.7%) -	*E. coli* (6.5%) *K. oxytoca* (0.3%) *S. anginosus* (0.01%) *E. faecalis* (0.9%) Several species (96.7%) -	*E. coli* (8.5%) *K. oxytoca* (0.3%) *Streptococcus spp.* (0.09%) *E. faecalis* (0.7%) Several species (90.4%) -
6	10^3	*E. faecium*	*E. faecium*	*E. faecium* (0.77%)	Not identified*	Not identified*
7	10^2	*S. aureus*	-#	*S. aureus* (82.9%)	*S. aureus* (100%)	*S. aureus* (100%)
8	10^3	*O. intermedium*	*O. intermedium*	*O. anthropi* (21.3%)	*O. intermedium* (99.4%)	*O. intermedium* (99.1%)
9	10^3	*S. aureus*	*S. aureus*	*S. aureus* (22.9%)	*S. aureus* (100%)	*S. aureus* (100%)
10	10^3	*S. marcescens*	-#	*S. marcescens* (64.7%)	*S. marcescens* (99.1%)	*S. marcescens* (100%)

[a]The number of colonies of a given species was estimated from the number of colonies with the same morphology on the same plate

[b]The relative abundance is calculated using total number of reads as denominator

[c]The relative abundance is calculated with the total number of classified reads as denominator

[d]miniKraken database was used

[#]Although there was a laboratory identification, no isolates were available for WGS

[*]No reads matched that specific pathogen, not even at the genus level

Table 2.3: Microorganisms identified by conventional methods, WGS and using shotgun metagenomics and the taxonomic classification methods in CLC Genomics Workbench.

Sample number	Culture result (CFU)[a]	Conventional identification (MALDI-TOF)	WGS-based identification	Shotgun metagenomics	
				Taxonomic Profiling (CLC)[b]	Best match with K-mer spectra (CLC)[c]
1	103 10^3 10	*E. faecium* *S. haemolyticus* *C. glabrata*	*E. faecium* *S. haemolyticus* -	*E. faecium* (71%) *S. haemolyticus* (24%) *C. glabrata* (100%)	*E. faecium* (41.4%) *S. haemolyticus* (13.8%) *C. glabrata* (0.5%)
2	10^3 1 Not determined	*E. avium* *E. coli* Anaerobes	-# -# -#	Not identified* Not identified* Several species (97%)	Not identified* Not identified* Several species (13.2%)
3	1	*S. epidermidis*	-#	Not identified*	*S. aureus* (4%)
4	10^3	*S. aureus*	*S. aureus*	Not identified*	*S. aureus* (9.7%)
5	$\geq 10^5$ $\geq 10^5$ 10^3 10^3 Not determined 10	*E. coli* *K. oxytoca* *S. anginosus* *E. faecalis* Anaerobes *C. albicans*	*E. coli* *K. oxytoca* -# E. faecalis -# -#	*E. coli* (25%) *K. michiganensis* (0.3%) Not identified* *E. faecalis* (2%) Several species (70.0%) Not identified*	*E. coli* (11.5%) Not identified* Not identified* *E. faecalis* (0.6%) Not identified* *C. albicans* (<0.05%)
6	10^3	*E. faecium*	*E. faecium*	Not identified*	*E. faecium* (4.0%)
7	10^2	*S. aureus*	-#	*S. aureus* (100%)	*S. aureus* (95.5%)
8	10^3	*O. intermedium*	*O. intermedium*	*O. intermedium* (86.0%)	*O. intermedium* (91.2%)
9	10^3	*S. aureus*	*S. aureus*	*S. aureus* (100%)	*S. aureus* (81.2%)
10	10^3	*S. marcescens*	-#	*S. marscescens* (100%)	*S. marcescens* (79.7%)

[a]The number of colonies of a given species was estimated from the number of colonies with the same morphology on the same plate

[b]The relative abundance is calculated using total number of reads as denominator

[c]The relative abundance is calculated with the total number of classified reads as denominator

[#]Although there was a laboratory identification, no isolates were available for WGS

[*]No reads matched that specific pathogen, not even at the genus level

Table 2.4: Microorganisms identified by conventional methods, WGS and using shotgun metagenomics and the taxonomic classification methods in webpages (BaseSpace, Taxonomer and CosmosID).

Sample number	Culture result (CFU)[a]	Conventional identification (MALDI-TOF)	WGS-based identification	Shotgun metagenomics				
				Genius (Basespace)[c]	Kraken (Basespace)[c, d]	MetaPhlAn (Basespace)[c]	Taxonomer (Utah)[b, e]	Cosmos ID[a]
1	10^3 10^3 10	*E. faecium* *S. haemolyticus* *C. glabrata*	*E. faecium* *S. haemolyticus* -	*E. faecium* (14.4%) *S. haemolyticus* (55.8%) -	*E. faecium* (25.0%) *S. haemolyticus* (20.1%) -	*E. faecium* (65.1%) *S. haemolyticys* (30.4%) -	*E. faecium* (22.9%) *S. haemolyticus* (20.1%) Not identified*	*E. faecium* (50.3%) *S. haemolyticus* (22.1%) *C. glabrata* (88.6%)
2	103 1 Not determined	*E. avium* *E. coli* Anaerobes	-# -# -#	Not identified* Not identified* Several species (94.0%)	Not identified* Not identified* Several species (27.0%)	Not identified* Not identified* Several species (54.2%)	Not identified* Not identified* Several species (14.2%)	Not identified* Not identified* Several species (100%)
3	1	*S. epidermidis*	-#	*S. aureus* (100%)	*S. aureus* (0.1%)	Not identified*	*S. pseudintermedius* (3.4%)	Not identified*
4	10^3	*S. aureus*	*S. aureus*	*S. aureus* (100%)	*S. aureus* (0.3%)	*S. aureus* (100%)	*S. aureus* (8.3%)	*S. aureus* (100%)
5	$\geq 10^5$ $\geq 10^5$ 10^3 10^3 Not determined 10	*E. coli* *K. oxytoca* *S. anginosus* *E. faecalis* Anaerobes *C. albicans*	*E. coli* *K. oxytoca* -# *E. faecalis* -# -#	*E. coli* (0.4%) Not identified* *S. anginosus* (0.03%) *E. faecalis* (0.8%) Several species (45.0%) -	*E. coli* (10.2%) *K. oxytoca* (0.5%) *S. anginosus* (0.4%) *E. faecalis* (0.3%) Several species (8.0%) -	*E. coli* (7.0%) *K. pneumoniae* (0.01%) *S. anginosus* (0.3%) *E. faecalis* (0.7%) Several species (89.1%) -	*E. coli* (3.6%) *K. michiganensis* (0.1%) *S. anginosus* (0.1%) *E. faecalis* (0.1%) Several species (60.3%) -	*E. coli* (7.6%) *K. oxytoca* (1.7%) *S. anginosus* (0.09%) *E. faecalis* (3.7%) Several species (86.2%) Not identified*
6	10^3	*E. faecium*	*E. faecium*	*E. faecium* (4.2%)	*E. faecium* (14.8%)	*E. faecium* (5.5%)	*E. faecium* (1.4%)	*E. faecium* (4.1%)
7	10^2	*S. aureus*	-#	*S. aureus* (100%)	*S. aureus* (93.8%)	*S. aureus* (100%)	*S. aureus* (14.2%)	*S. aureus* (100%)
8	10^3	*O. intermedium*	*O. intermedium*	*O. intermedium* (100%)	*O. nthropic* (88.9%)	*O. intermedium* (99.8%)	*O. intermedium* (13.1%)	*O. intermedium* (49.5%)
9	10^3	*S. aureus*	*S. aureus*	*S. aureus* (100%)	*S. aureus* (99.5%)	*S. aureus* (100%)	*S. aureus* (12.7%)	*S. aureus* (100%)
10	10^3	*S. marcescens*	-#	*S. marcescens* (32.5%)	*S. marcescens* (94.8%)	*Serratia* spp. (100%)	*S. marcescens* (1.4%)	S. *marscescens* (38.4%)

[a] The number of colonies of a given species was estimated from the number of colonies with the same morphology on the same plate

[b] The relative abundance is calculated using total number of reads as denominator

[c] The relative abundance is calculated with the total number of classified reads as denominator

[d] miniKraken database was used [e] Full Analysis mode was used

[#] Although there was a laboratory identification, no isolates were available for WGS

[*] No reads matched that specific pathogen, not even at the genus level

Table 2.5: Performance of the different taxonomic classification methods for each sample. Sensitivity and positive predictive value were calculated using culture/MALDI-TOF as standards

Method	Total number of bacteria identified[a]	True positives[a]	False positives	False negatives	Sensitivity (%)	PPV (%)
Culture/MALDI-TOF	9	9	0	0	1	1
MetaPhlAn (BaseSpace)	16	7	9	2	0.78	0.44
Genius (BaseSpace)	35	8	27	1	0.89	0.23
Kraken (BaseSpace)	959	7	952	2	0.78	0.01
Taxonomer (Full Analysis)	4649	8	4641	1	0.89	0
CosmosID	35	8	27	1	0.89	0.23
Taxonomic Profiling (CLC Genomics Workbench v10.0.1)	17	6	11	3	0.67	0.35
Best match K-mer spectra (CLC Genomics Workbench v10.0.1)	12	8	4	1	0.89	0.67
Kraken (Unix)	198	7	191	2	0.78	0.04
MetaPhlAn2 (Unix)	15	7	6	4	0.75	0.75
MIDAS (Unix)	34	7	26	2	0.88	0.5

[a] aExcluding the samples with non-identified anaerobic bacteria (Samples 2 and 5)

Abbreviations: PPV – positive predictive value

The taxonomic classification results for each sample are presented in Tables 2.2, 2.3 and 2.4. In 8 samples, all the microorganisms identified by classical culture were also identified through metagenomics (using at least one method). In sample 2, two of the bacterial species identified by classical culture, i.e., *E. coli* and one *Enterococcus avium* were not identified through shotgun metagenomics and in sample 3 there was no concordance between the results of MALDI-TOF and the taxonomical classification methods at the species level (Tables 2.2, 2.3 and 2.4). We identified *Ochrobactrum intermedium* in the negative control, but in low amounts (1.0% of the reads mapped to the reference genome with the accession number NZ_ACQA01000002 and only 1.4% of the reference genome was covered). The sensitivity and positive predictive value of each classification method is shown in Table **??**.

2.4.3 Determination of antimicrobial resistance

Metagenomics provides other sequence information in addition to pathogen detection. We determined the presence of antimicrobial-resistance genes in the SMg sequence data and

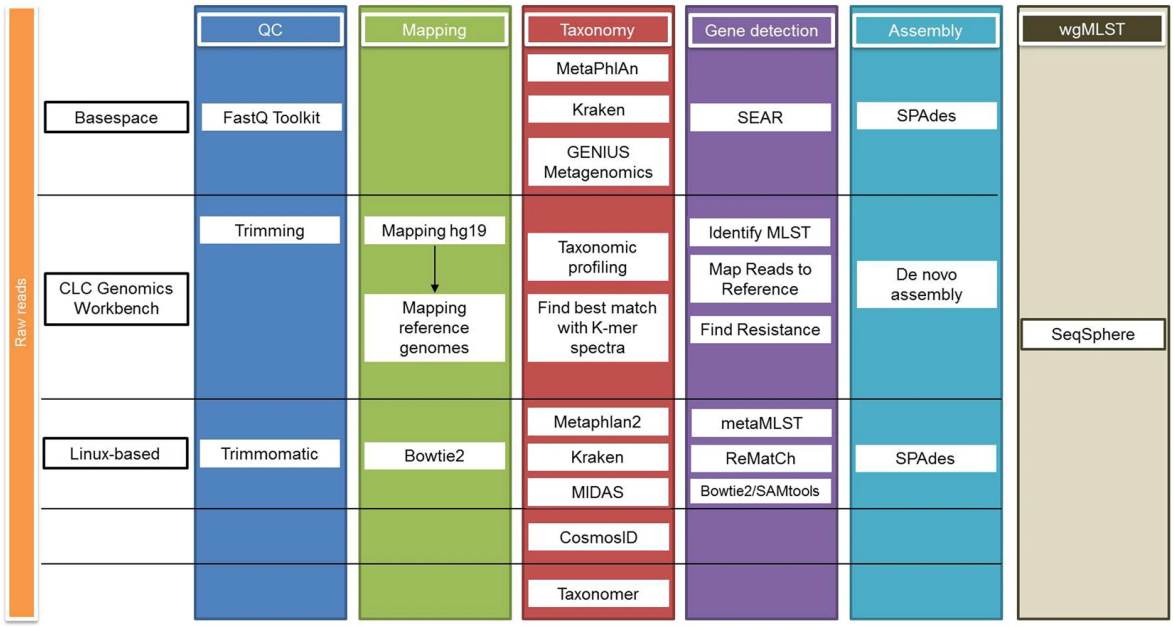

Figure 2.1: Scheme of the bioinformatic analysis of the metagenomics samples.

compared the results with those obtained from WGS and phenotypic resistance testing (Table **??**).

AMR genes found with CLC Genomics Workbench and ReMatCh in samples 1, 7 and 9 correlated well with phenotypic results. However, in the other 7 samples, not all antimicrobial resistance genes that could explain the phenotypic profile were identified. In addition, in samples 2, 5, 7 and 10, ReMatCh detected different resistance genes compared to those reported by CLC Genomics Workbench (Table **??**). Some of these differences (genes *norA*, *blaSST*-1, *fusA*) were due to slight differences in the databases used, however, the other resistance genes were present in both databases. Interestingly, in two samples (samples 2 and 5), we were able to identify several antimicrobial resistance genes usually found in anaerobic bacteria. These were not reported by classical microbiology methods, probably because they were not considered relevant pathogens worthy of subsequent susceptibility study (mixed anaerobic culture).

The SEAR app in BaseSpace (the only one available for antimicrobial resistance gene detection) crashed several times, although we performed the analysis repeatedly, using different parameters. We were only able to get results in 3 samples, with no resistance genes detected.

Table 2.6: Antimicrobial resistance phenotypes and antimicrobial resistance genes detected using different approaches.

Sample number	Conventional identification (MALDI-TOF)	Conventional susceptibility testing (VITEK 2)b	WGS	Shotgun metagenomics	
			CLC Genomics Workbench	ReMatCh (Unix)	CLC Genomics Workbencha
1	E. faecium S. haemolyticus	LEV, ERY, CLI OXA, GEN, CIP, FOS, ERY, CLI	erm(B), msr(C), ant(6')-Ia, aph(3')-III, dfrG blaZ, mecA, ant(6')-Ia, aph(3')-III, aac(6')-aph(2''), erm(C), mph(C), msr(A), dfrG	erm(B), msr(C), ant(6')-Ia, aph(3')-III, aac(6')-aph(2''), blaZ, mecA, erm(C), mph(C), msr(A), dfrG	erm(B), msr(C), ant(6')-Ia, aph(3')-III, aac(6')-aph(2''), blaZ, mecA, erm(C), mph(C), msr(A), dfrG
2	E. avium E. coli Anaerobes	DOX, CLI susceptible -	-# -# -#	Not detected Not detected catS, lnu(D), lsa(C), cepA-44, tet(Q)	Not detected Not detected catS, lnu(D), lsa(C), cepA-44, tet(Q), fusA
3	S. epidermidis	OXA, GEN, TEC, FUS, CIP, ERY, CLI	-#	Not detected	Not detected
4	S. aureus	PEN, ERY	blaZ, spc, erm(A)	Not detected	Not detected
5	E. coli K. oxytoca S. anginosus E. faecalis Anaerobes	susceptible AMX susceptible DOX, CLI -	-# blaOXY-1-3 -# tet(M), lsa(A) -#	- Not detected - tet(M) cfxA4, tet(Q)	- Not detected - tet(O) cfxA4, tet(Q)
6	E. faecium	PEN, AMX, CFX, IMP, GENhl, STRhl, LEV, ERY, CLI, AMP/SUL	erm(B), msr(C), ant(6')-Ia, aph(3')-III, aac(6')-aph(2''), dfrG	Not detected	Not detected
7	S. aureus	PEN	blaZ	blaZ, norA	blaZ
8	O. intermedium	AMX, PIP/TAZ, CFX, CPT, CTZ, IMP, FOX, TOB, FOS, NIT, TMP	blaOCH-2	blaOCH-5	blaOCH-2
9	S. aureus	PEN	-#	blaZ	blaZ
10	S. marcescens	AMX, AMC, CFX, FOX, NIT, POL	-#	blaSST-1, tet(41), oqxB, aac(6')-Ic	tet(41), oqxB, aac(6')-Ic

[a]The analysis aborted when the script tried to connect to NCBI

[b]Only non-susceptibility is indicated.

Abbreviations: AMP/SUL, ampicillin/sulbactam; AMX, amoxicillin; AMC, amoxicillin/clavulanate; CFX, cefuroxime; FOS, fosfomycin; FOX, cefoxitin; CIP, ciprofloxacin; CLI, clindamycin; DOX, doxycycline; ERY, erythromycin; FUS, fusidic acid; GEN, gentamicin; GENhl, gentamicin high-level; LEV, levofloxacin; NIT, nitrofurantoin; PEN, penicillin; POL, polymyxin B; STRhl, streptomycin high-level; TEC, teicoplanin.

Table 2.7: Results of MLST using by whole genome sequencing and shotgun metagenomics

Sample number	Conventional identification (MALDI-TOF)	WGS	Shotgun metagenomics	
		CLC Genomics Workbench v10.1.1	CLC Genomics Workbench v10.1.1	metaMLST (Unix-based)
1	*E. faecium* *S. haemolyticus*	ST117 ST25	Not detected (6 alleles identified correctly) Not detected (3 alleles identified correctly)	ST117 Not detected
2	*E. avium* *E. coli* Anaerobes	-# -# -#	- Not detected -	- Not detected -
3	*S. epidermidis*	-#	Not detected	Not detected
4	*S. aureus*	ST30	Not detected	Not detected
5	*E. coli* *K. oxytoca* *S. anginosus* *E. faecalis* Anaerobes	ST141 ST40 -# ST179 -#	ST141 Not detected - Not detected -	ST4508 Not detected - Not detected -#
6	*E. faecium*	ST117	Not detected	Not detected
7	*S. aureus*	ST30	ST30	ST667
8	*O. intermedium*	-	-	-
9	*S. aureus*	-#	Not detected	Not detected
10	*S. marcescens*	-#	-	-

Abbreviations: ST, sequence type

2.4.4 MLST and wgMLST analysis

In three cases when SMg data covered $\geq$ 93% of the genome we were able to identify the ST, which corresponded to the one found using WGS of the isolated bacteria using CLC Genomics Workbench (n=2) and metaMLST (n=1). These results are summarized in Table 2.7. Assembled genomes and metagenomes, were compared by wgMLST analysis using Ridom SeqSphere+. Figure 2.2 shows examples of the allele difference between the genomes obtained through WGS versus the genomes obtained through shotgun metagenomics.

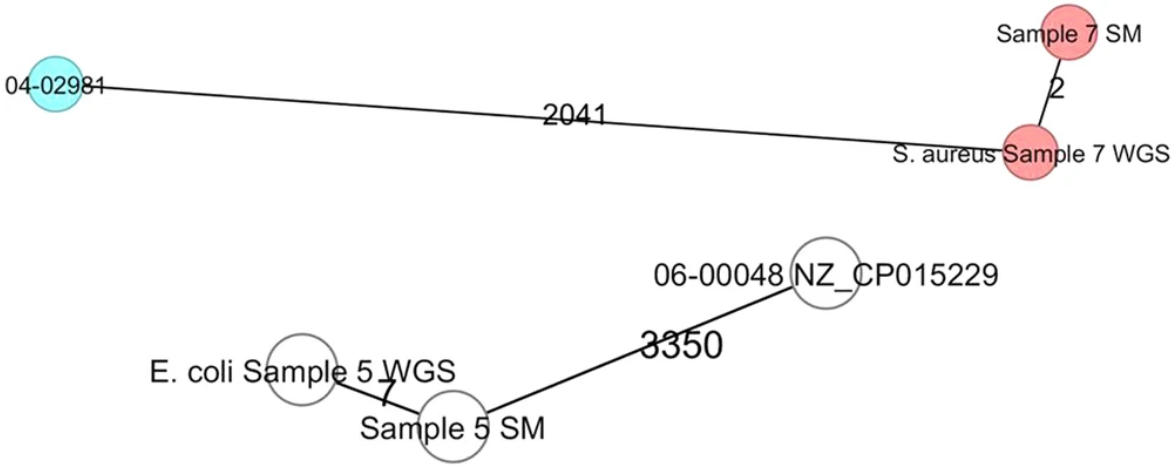

Figure 2.2: Minimum-spanning tree based on wgMLST allelic profiles of 2 S. aureus genomes and 2 E. coli genomes obtained through SM and WGS in comparison to reference strains 04-02981 (GenBank accession number NC_017340) and 06-00048 (NZ_CP015229), respectively. Each circle represents an allelic profile based on sequence analysis. The numbers on the connecting lines illustrate the numbers of target genes with differing alleles.

2.4.5 Characterisation of mobile genetic elements

Two different approaches, i.e. CLC Genomics Workbench and Bowtie2 were used to identify plasmids present in the sequence data. Both approaches used mapping of sequences against the same plasmid database. Since some plasmids present in the database are very similar and sequence reads may be mapped to more than one plasmid, we used the pAT-LAS tool, which provides an overview of the nodes (representing plasmid sequences) and links between plasmids (which connect similar plasmids), to enable the visualisation of the plasmids identified (Figure 2.3). A colour gradient indicates the sequence coverage of the plasmids. In most cases, the same plasmids were identified by both approaches, with some small differences in sequence coverage. When comparing the plasmids identified in the SMg dataset versus the WGS data, most of the plasmids were also detected in the isolates (an example is shown in Figure 2.4). However, some plasmids were not identified in any of the isolated bacteria and were probably residing in low-abundant species.

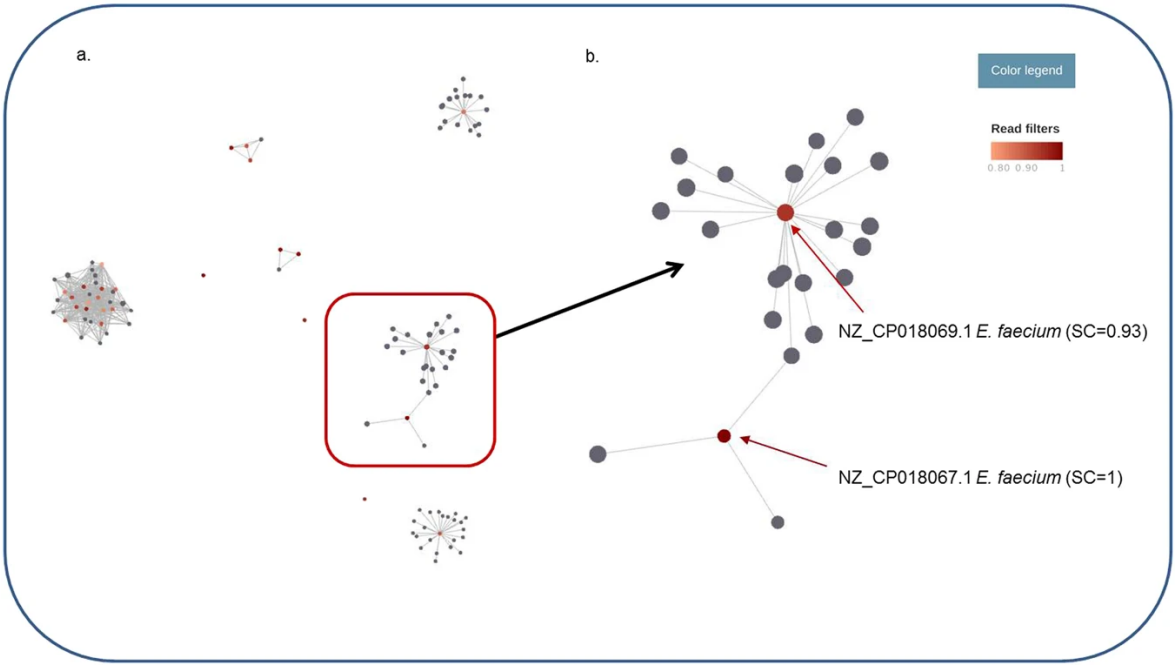

Figure 2.3: (a) Overview of the nodes (representing plasmid sequences) and links between plasmids (connecting similar plasmids) found in Sample 1 (SMg) using the pATLAS tool. (b) A closer look at one of the cloud of plasmids. The colour gradient in each cloud of plasmids represents the plasmid sequence coverage (SC), varying between 0-0.79 (grey) and 0.80-1 (red gradient).

2.5 Discussion

This study evaluated the suitability of SMg for the microbiological diagnosis and (patho- and epi-) typing of microorganisms directly from real patient samples. The whole procedure took between 48-54 hours to complete, which is shorter than culture-based methods if one includes typing. However, the amount of information derived from SMg in most cases, did not overcome the necessity for pathogen isolation and subsequent (phenotypic and genotypic) typing, which can take up to 1-2 weeks (particularly in slow-growing organisms). Nevertheless, SMg can help guide antimicrobial therapy and be helpful in cases where there is a suspicion of transmission and there is a need to quickly determine the genetic relationship between pathogens, although the success of SMg in individual patient samples can be highly variable, as reported here.

Different bioinformatics pipelines were evaluated to identify potential differences between them and identify those which could provide the clinical microbiologist with the maximum of relevant and accurate information. In terms of microbial identification, in both Unix and web-based approaches we would recommend MetaPhlAn, since it has good sensitivity and a good positive predictive value (PPV). The find best match K-mer spectra tool should

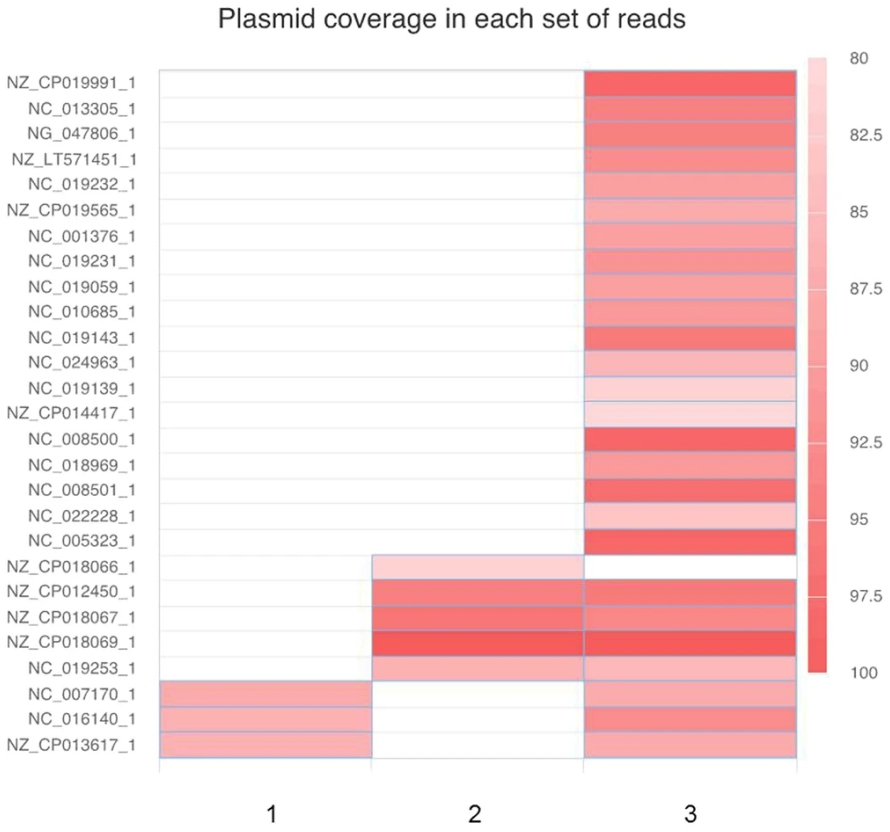

Figure 2.4: A heatmap comparing the identified plasmids using bowtie2 in *S. haemolyticus* WGS (1), *E. faecium* WGS (2) and in the SMg dataset (3) isolated from sample 1.

be used in the context of the CLC Genomics Workbench, since it had a higher sensitivity and PPV compared to the Taxonomic Profiling tool.

In a clinical setting, a combination of high sensitivity and high PPV of any new method is key. Popular software designed for bacterial identification, can predict dozens to hundreds of species in in vitro generated bacterial communities of known composition [8]. We observed the same when using Kraken and Taxonomer when comparing to culture-based methods. For both Kraken and Taxonomer, relative abundance cut-off values may be required to limit the number of species identified. However, which cut-off values should be used are a matter of debate, since in some cases, even if applying a cut-off value as low as 1.0% (comparable to what was found in the negative control) would have resulted in decreased sensitivity (e.g. the *Streptococcus anginosus* identified by culture in Sample 5 would have been disregarded). The methods that employ several parameters to infer microbial identification are superior, because they not only rely on the relative abundance of bacterial species, but also on the genome coverage and on the proportion of the genome that was covered. On the other hand, in some cases SMg may be more sensitive than culture in identifying pathogens, reflecting the higher sensitivity or the capacity to detect bacterial species which are non-culturable in the conditions used or that are no longer culturable, such as due to prior antimicrobial therapy. In such cases, other methods like 16S rDNA sequencing or the recently described 16S-23S rDNA sequencing method [20] may be used for discrepancy analyses. However,

here we decided to use culture-based methods as the gold standard, since this is still the method of choice in clinical microbiology.

One limitation of this study was the exclusion of culture-negative samples and thus their inclusion would have affected the calculation of the specificity values. However, as mentioned above, culture-negative samples do not necessarily mean that the samples are pathogen-free, but it might only reflect the low sensitivity or capacity of culture-based methods to detect non-culturable bacterial species. As with other (molecular) methods, several controls should be included to validate the obtained results, including a negative control. In our negative control, we detected an *O. intermedium* strain, although with only 1.0% of the reads mapping to the reference genome and covering only 1.4% of the reference genome (accession number NZ_ACQA01000002). These results may be due to contamination during library preparation (e.g. sample-to-sample contamination prior to indexing), the result of sequencing artefacts (e.g. demultiplexing errors), or to incorrect classification during data analysis (e.g. highly similar regions) [3]. Our samples and sequencing libraries were handled in laminar flow cabinets; however, we cannot also exclude the possibility of contamination. Furthermore, the reagents used may also be or become contaminated with DNA leading the detection of these contaminating species, something that has been described previously [7]. This poses a challenge for interpretation, because some positive samples also had very low numbers of reads for some pathogens (< 1%). When approaching this limit of detection, small numbers of pathogen reads will be difficult to interpret, as they can represent true-positives with low abundance in the sample, or artefacts such as contamination during library preparation[3].

In terms of antimicrobial resistance gene detection, ReMatCh (Unix) and the CLC Genomics Workbench Find Resistance tool gave comparable results. Since ReMatCh (Unix) performs the analysis at the read level, while CLC Genomics Workbench performs it at the contig level, we suggest that both strategies should be employed in parallel when looking for antimicrobial resistance genes. It is also important to emphasise that the contig-level approach employed by CLC Genomics Workbench may give negative results if the sequence coverage is set to a high percentage (e.g. above 80%). This is due to the assembly method, which may split the antimicrobial resistance genes into different contigs, when the number of reads is too low. This phenomenon was observed in Sample 1, for the *aac(6')-aph(2")* gene, which was split into 3 different contigs, each part corresponding to less than 40% of the gene. Only when applying a cut-off value of *geq* 20% for sequence coverage could we identify all three parts of the gene, which in total corresponded to 89% of the entire sequence. Finally, it is important to point out that the ResFinder database (used here), and other databases, focus on acquired genes, not including chromosomal point mutations resulting in antimicrobial resistance. However, a recently developed tool, PointFinder, was added to ResFinder for the detection of chromosomal point mutations associated with antimicrobial resistance [21] and an updated database will be available soon.

Another challenge is to infer where these antimicrobial resistance genes are located

(chromosome or plasmid). The study of mobile genetic elements, including plasmids, carrying antimicrobial resistance genes present in clinical samples is important to predict possible treatment failures and the spread of resistance within and across bacterial species. When performing bacterial isolation followed by WGS, information on polymicrobial infections may be lost. This is mainly driven by a bottleneck in culture, where some bacterial species are not isolated with standard work up protocols (frequently anaerobes and slow-growing organisms). The presence of antimicrobial resistance genes in plasmids of bacteria other than those isolated through culture poses a risk since they are not identified by conventional methods but could potentially be horizontally transmitted to pathogenic bacteria under the antimicrobial selective pressure of treatment. Antimicrobial administration may also select minority populations where these resistance determinants are found. Furthermore, the understanding of how plasmids are shared by different bacteria in a bacterial community (e.g. within an infection site or in the gut) can improve our understanding of how these elements disseminate across species and from patient to patient11. The SMg approach is clearly more efficient than culture in identifying the "cloud" of plasmids present in a given sample (Figure 4) and which can be potentially transferred to more pathogenic species generating problems of resistance, as was the case with the emerge of vancomycin resistance *S. aureus* [22].

Whole-genome sequencing has been used extensively for several purposes [23] and is considered to have the potential of playing an important role in clinical microbiology [24]. It is the ongoing goal of medical molecular microbiology to develop faster typing methods that can be used for outbreak surveillance. For this purpose, we assembled the metagenomics data and compared it with the assemblies given by WGS. Surprisingly, the assemblies provided by SPAdes in BaseSpace were closer to the assemblies provided by WGS. When comparing the genomes obtained through WGS and SMg, we could see that in 4 out of 8 bacterial isolates the number of different alleles was leq 7. This showed the potential of SMg to draw phylogenetic relationships from uncultured bacterial genomes, although more potentially limited than those obtained using WGS data from axenic cultures. As for the detection of resistance genes, a key limiting factor may be the number of bacterial reads, reflected in a lower genome coverage (e.g. samples 4 and 6). In these cases, we would have to either improve the human-DNA depletion step, improve the microbial enrichment or perform sequencing at a higher sequencing depth to have enough microbial reads to be able to get a more appropriate genome coverage. Yet, this last step will severely raise the sequencing costs, which might render the methodology unfeasible for routine application.

In this study, we evaluated the results of metagenomics pipelines using three different methods. CLC Genomics Workbench has advantages over the other methods. It does not require previous knowledge of Unix-based tools, it is arguably the most user-friendly and delivered reliable results for microbial identification and antimicrobial resistance gene detection. The downside was the assembly approaches, which provided lower wgMLST allele detection, when compared to the assemblies using SPAdes (BaseSpace and Unix). BaseSpace, the other commercial solution, on the other hand, provided only a few tools that can be used for metagenomics data. Furthermore, since Illumina did not develop the apps them-

selves, they offered no direct support. Contacting the developers (via email and posting on their forum) does not guarantee a solution to the issues in a time frame compatible with a routine clinical microbiology laboratory work. The dependence and no direct control over a third party to resolve software bugs and provide a stable platform illustrates a disadvantage of a cloud-based system like BaseSpace. Finally, the Unix-based pipeline complemented the data on antimicrobial resistance genes but did not offer better results in terms of microbial identification and MLST typing. However, many more freely available tools for this last purpose could have been used, potentially improving on the results obtained. Reference-guided assembly approaches, taking advantage of the species information derived in the first steps of our analysis pipelines, will deserve further study in the future since these may provide higher quality assemblies from metagenomics data. The main advantage of an open-source approach is its flexibility since it allows the user to choose the most adequate method for each desired outcome. There were several limitations to this study. First, the number of samples included was low and some of the bacterial isolates were not available for further WGS analysis. However, the extended data analyses performed in each sample limited the number of samples to be included. It is our intention to move forward with the most adequate pipelines for each purpose and apply them to additional patients' samples. Second, the samples differed greatly from each other. However, in our point of view, this was beneficial to the study, since it did not bias the analyses as it could have happened if only one type of sample had been used. Finally, we used three different extraction methods that could have influenced the final results. Yet, as can be seen in Table 1, the number of human reads differed between samples, even when using the same extraction kit. This suggests none of the kits is clearly superior to the others and that the ratio between host and microbial DNA or other individual sample characteristics will be the major determinants of the proportion of microbial reads.

In conclusion, this study showed the potential but also highlighted the problems of implementing shotgun metagenomics for the identification and typing of pathogens directly from clinical samples. Based on the results obtained here we can conclude that the tools and databases used for taxonomic classification and antimicrobial resistance will have a key impact on the results, cautioning about the comparison between studies using different methods and suggesting that efforts need to be directed towards standardisation of the analysis methods if SMg is to be used routinely in clinical microbiology.

2.6 Acknowledgements

We thank Peter Posma, Yvette Bisselink and Brigitte Dijkhuizen for excellent technical assistance. We thank Dr. Michael Lustig and colleagues from Molzym Life Science for helping with extraction protocols.

This project has received funding from the European Union's Horizon 2020 research and

innovation program under the Marie Skłodowska-Curie grant agreement 713660.This work was partly supported by the INTERREG VA (202085) funded project EurHealth-1Health, part of a Dutch-German cross-border network supported by the European Commission, the Dutch Ministry of Health, Welfare and Sport (VWS), the Ministry of Economy, Innovation, Digitalisation and Energy of the German Federal State of North Rhine-Westphalia and the German Federal State of Lower Saxony.

2.7 Author contributions statement

N.C., J.A.C., M.R., S.P., I.A., A.W.F. and J.W.A conceived the experiment(s), N.C., L.S. and E.C.R. conducted the experiment(s), N.C., L.S., M.M., C.I.M., T.F.J., S.R., M.C., J.A.C. and M.R. analysed the results, N.C. and L.S. wrote the manuscript. All authors reviewed the manuscript.

2.8 Additional information

2.8.1 Accession codes

The paired-trimmed-un-mapped reads (hg19) generated for each sample have been submitted to SRA under project number SRP126380. The cgMLST schemes are deposited in figshare under the DOI:10.6084/m9.figshare.5679376

2.8.2 Competing financial interests

The authors declare that they have no conflict of interest.

Bibliography

[1] Henrik Hasman et al. “Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples”. en. In: *Journal of Clinical Microbiology* 52.1 (Jan. 2014). Ed. by Y.-W. Tang, pp. 139–146. ISSN: 0095-1137, 1098-660X. DOI: 10.1128/JCM.02452-13. URL: https://journals.asm.org/doi/10.1128/JCM.02452-13 (visited on 03/18/2022).

[2] Matthias Willmann et al. “Antibiotic Selection Pressure Determination through Sequence-Based Metagenomics”. eng. In: *Antimicrobial Agents and Chemotherapy* 59.12 (Dec. 2015), pp. 7335–7345. ISSN: 1098-6596. DOI: 10.1128/AAC.01504-15.

[3] Erin H. Graf et al. “Unbiased Detection of Respiratory Viruses by Use of RNA Sequencing-Based Metagenomics: a Systematic Comparison to a Commercial PCR Panel”. eng. In: *Journal of Clinical Microbiology* 54.4 (Apr. 2016), pp. 1000–1007. ISSN: 1098-660X. DOI: 10.1128/JCM.03060-15.

[4] P. Gyarmati et al. “Metagenomic analysis of bloodstream infections in patients with acute leukemia and therapy-induced neutropenia”. eng. In: *Scientific Reports* 6 (Mar. 2016), p. 23532. ISSN: 2045-2322. DOI: 10.1038/srep23532.

[5] Nathan D Olson et al. “Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes”. In: *Briefings in Bioinformatics* 20.4 (Aug. 2017), pp. 1140–1150. ISSN: 1467-5463. DOI: 10.1093/bib/bbx098. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6781575/ (visited on 03/17/2022).

[6] Robert Schlaberg et al. “Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection”. eng. In: *Archives of Pathology & Laboratory Medicine* 141.6 (June 2017), pp. 776–786. ISSN: 1543-2165. DOI: 10.5858/arpa.2016-0539-RA.

[7] Teresa L. Street et al. “Molecular Diagnosis of Orthopedic-Device-Related Infection Directly from Sonication Fluid by Metagenomic Sequencing”. en. In: *Journal of Clinical Microbiology* 55.8 (Aug. 2017). Ed. by Nathan A. Ledeboer, pp. 2334–2347. ISSN: 0095-1137, 1098-660X. DOI: 10.1128/JCM.00462-17. URL: https://journals.asm.org/doi/10.1128/JCM.00462-17 (visited on 03/18/2022).

[8] Michael A. Peabody et al. “Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities”. en. In: *BMC Bioinformatics* 16.1 (Dec. 2015), p. 362. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0788-5. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0788-5 (visited on 03/18/2022).

[9] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. en. In: *Nature Methods* 9.4 (Apr. 2012), pp. 357–359. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1923. URL: http://www.nature.com/articles/nmeth.1923 (visited on 03/18/2022).

[10] Heng Li et al. “The Sequence Alignment/Map format and SAMtools”. eng. In: *Bioinformatics (Oxford, England)* 25.16 (Aug. 2009), pp. 2078–2079. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352.

[11] Derrick E Wood and Steven L Salzberg. “Kraken: ultrafast metagenomic sequence classification using exact alignments”. en. In: *Genome Biology* 15.3 (2014), R46. ISSN: 1465-6906. DOI: 10.1186/gb-2014-15-3-r46. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46 (visited on 03/18/2022).

[12] Stephen Nayfach et al. “An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography”. eng. In: *Genome Research* 26.11 (Nov. 2016), pp. 1612–1625. ISSN: 1549-5469. DOI: 10.1101/gr.201863.115.

[13] Nicola Segata et al. “Metagenomic microbial community profiling using unique clade-specific marker genes”. en. In: *Nature Methods* 9.8 (Aug. 2012), pp. 811–814. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2066. URL: http://www.nature.com/articles/nmeth.2066 (visited on 03/18/2022).

[14] Moreno Zolfo et al. “MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples”. In: *Nucleic Acids Research* 45.2 (Jan. 2017), e7. ISSN: 0305-1048. DOI: 10.1093/nar/gkw837. URL: https://doi.org/10.1093/nar/gkw837 (visited on 03/18/2022).

[15] Anton Bankevich et al. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing”. en. In: *Journal of Computational Biology* 19.5 (May 2012), pp. 455–477. ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2012.0021. URL: http://www.liebertpub.com/doi/10.1089/cmb.2012.0021 (visited on 03/18/2022).

[16] Chris Smillie et al. “Mobility of plasmids”. eng. In: *Microbiology and molecular biology reviews: MMBR* 74.3 (Sept. 2010), pp. 434–452. ISSN: 1098-5557. DOI: 10.1128/MMBR.00020-10.

[17] Tossawan Jitwasinkul et al. “Plasmid metagenomics reveals multiple antibiotic resistance gene classes among the gut microbiomes of hospitalised patients”. en. In: *Journal of Global Antimicrobial Resistance* 6 (Sept. 2016), pp. 57–66. ISSN: 2213-7165. DOI: 10.1016/j.jgar.2016.03.001. URL: https://www.sciencedirect.com/science/article/pii/S2213716516300261 (visited on 03/18/2022).

[18] Steven Flygare et al. “Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling”. In: *Genome Biology* 17.1 (May 2016), p. 111. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0969-1. URL: https://doi.org/10.1186/s13059-016-0969-1 (visited on 03/19/2022).

[19] Werner Ruppitsch et al. “Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes”. eng. In: *Journal of Clinical Microbiology* 53.9 (Sept. 2015), pp. 2869–2876. ISSN: 1098-660X. DOI: 10.1128/JCM.01193-15.

[20] Artur J. Sabat et al. “Targeted next-generation sequencing of the 16S-23S rRNA region for culture-independent bacterial identification - increased discrimination of closely related species”. eng. In: *Scientific Reports* 7.1 (June 2017), p. 3434. ISSN: 2045-2322. DOI: 10.1038/s41598-017-03458-6.

[21] Ea Zankari et al. “PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens”. eng. In: *The Journal of Antimicrobial Chemotherapy* 72.10 (Oct. 2017), pp. 2764–2768. ISSN: 1460-2091. DOI: 10.1093/jac/dkx217.

[22] José Melo-Cristino et al. “First case of infection with vancomycin-resistant Staphylococcus aureus in Europe”. eng. In: *Lancet (London, England)* 382.9888 (July 2013), p. 205. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(13)61219-2.

[23] Ruud H. Deurenberg et al. “Application of next generation sequencing in clinical microbiology and infection prevention”. eng. In: *Journal of Biotechnology* 243 (Feb. 2017), pp. 16–24. ISSN: 1873-4863. DOI: 10.1016/j.jbiotec.2016.12.022.

[24] J. W. A. Rossen et al. “Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology”. eng. In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 24.4 (Apr. 2018), pp. 355–360. ISSN: 1469-0691. DOI: 10.1016/j.cmi.2017.11.001.

BIBLIOGRAPHY

Chapter 3

Conclusion

3. CONCLUSION

Appendix A

Appendix