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

2022

As opiniões expressas nesta publicação são da exclusiva responsabilidade do seu autor.

"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

List of Tables

List of Figures

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 (Vos et al., 2020). 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 (Ritchie et al., 2020), making it one of the deadliest pandemics in history. Coronavirus has been responsible for three of the eighteen major pandemics registered throughout modern history (Piret et al., 2021), 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 (C. J. Murray et al., 2022). In 2017, the WHO released The Global Priority Pathogens (GPP) list (Organization, 2017) 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.

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 (World Health Organization, 2021). 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 (Micah et al., 2021).

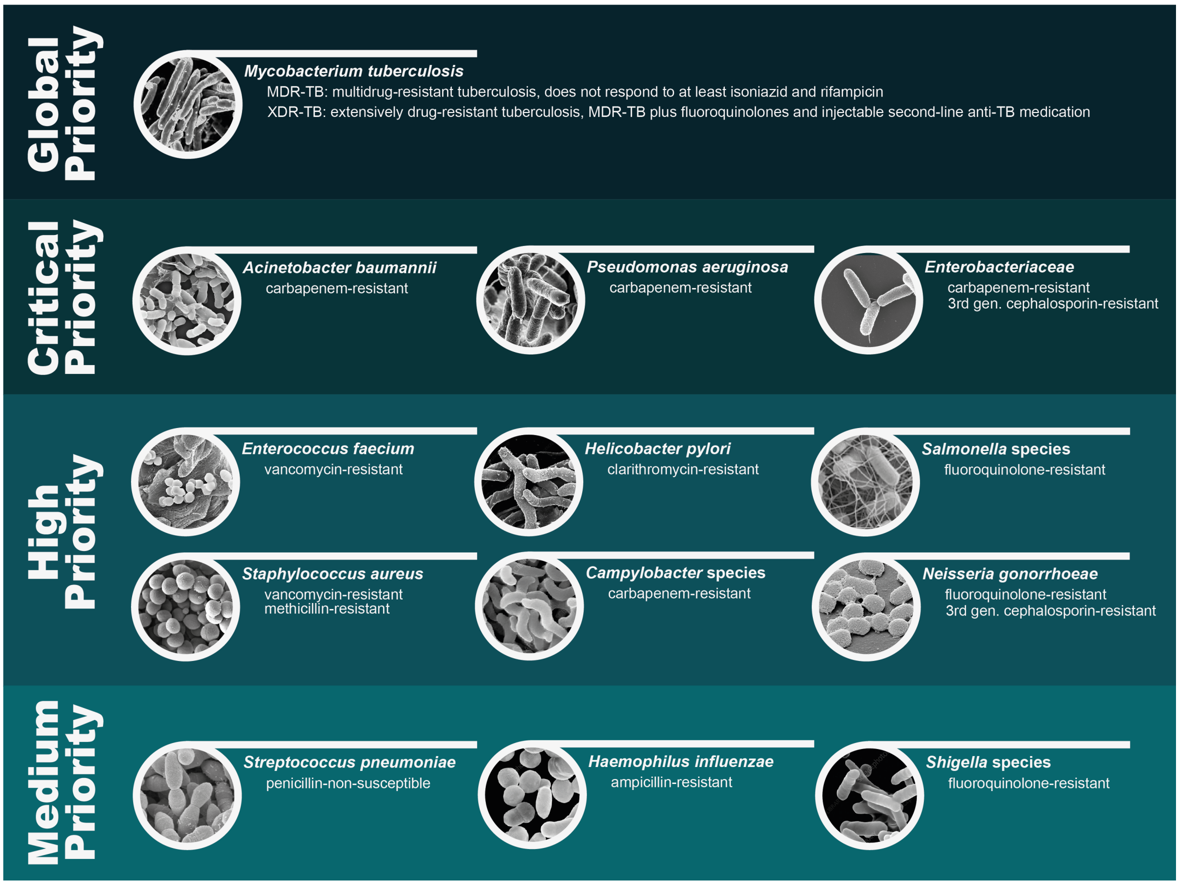


Figure 1.1: **World Health Organisation Global Priority Pathogens list.** This catalogue includes, besides *Mycobacterium tuberculosis* considered the number one global priority, a list of twelve microorganisms grouped under three priority tiers according to their antimicrobial resistance: critical (*Acinetobacter 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 Organization, 2017.

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 (Didelot et al., 2012). 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 (Foxman et al., 2005).

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 (Muhamad Rizal et al., 2020; Giuliano et al., 2019; Muhamad Rizal et al., 2020). 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 (Giuliano et al., 2019). 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 (Patel, 2015). 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 (Scerbo et al., 2016). 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 (Giuliano et al., 2019)

Following identification, antibiotic-susceptibility testing is essential for guiding clinicians 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 (Benkova et al., 2020). 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) (Allerberger, 2012; Foxman et al., 2005). DNA macrorestriction analysis by PFGE, which revolutionised precise separation of DNA fragments, became the most widely implemented DNA fingerprinting technique (Allerberger, 2012), becoming the golden standard for bacterial typing (Neoh et al., 2019).

In the early 2000s, Multilocus sequence typing (MLST) was proposed as a portable, universal, and definitive method for characterising bacteria (Maiden, 2006). 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 (M. V. Larsen et al., 2012). 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 (Maiden, 2006). So far, MLST schemes for 115 bacterial organisms have been published and made freely available[1], Jolley et al., 2018)

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 (Jauneikaite et al., 2015; Paton et al., 2019). 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 (Diep et al., 2019).

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) (Burrell et al., 2017). Viral

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

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 molecular methods with an agent which detects a virus-associated protein, such as an antibody (Cassedy et al., 2021).

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 (Cassedy et al., 2021). Syndromic testing (see subsubsection 1.1.1.1. Bacterial infections) is now fully integrated into the standard testing practices of many clinical laboratories (Dien Bard et al., 2020). 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 (Cassedy et al., 2021).

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 (Cassedy et al., 2021). 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 (Y. Chan et al., 2021; Bobrovitz et al., 2021). 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 (Koczula et al., 2016; Cassedy et al., 2021; Di Nardo et al., 2021).

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 (Groseclose et al., 2017; J. Murray et al., 2017). 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 (Teutsch, 2010).

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 (Organization, 2005). 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 (J. Murray et al., 2017; Melo-Cristino et al., 2006). 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 (Andrews et al., 2020; McWeeney, 1894; Baker et al., 2011; D. A. Larsen et al., 2020).

The emergence and re-emergence of infectious diseases are closely linked to the biology and ecology of infectious agents, their hosts, and their vectors (Destoumieux-Garzón et al., 2018). "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 (Mackenzie et al., 2019). 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 (Rugarabamu, 2021). 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 (Destoumieux-Garzón et al., 2018).

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* (Hood et al., 1996), genomics has transformed the field of microbiology, and in particular its clinical application. The paper describing the DNA-sequencing method with chain-terminating inhibitors used in the sequencing of the first microbial genome (Sanger et al., 1977), 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 (Van Noorden et al., 2014). Currently, this number has increased to 84546 according to PubMed Central® (PMC)[2][3]. 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

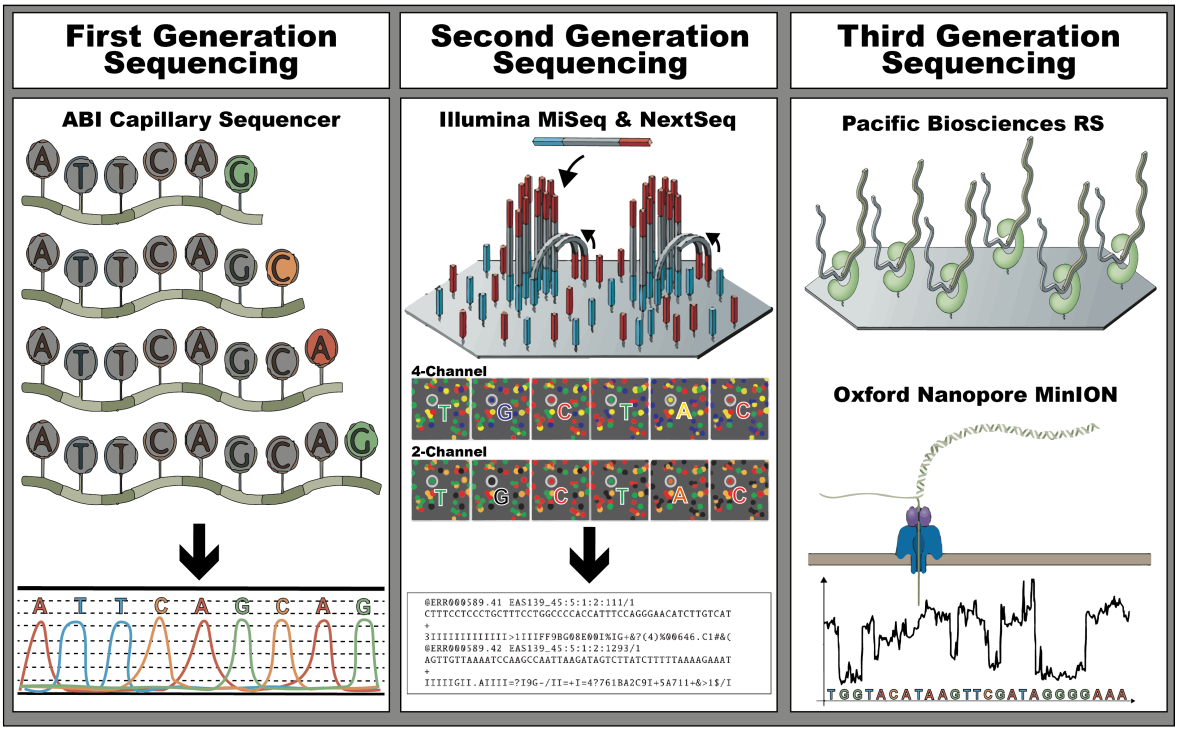
Since the discovery of the structure of DNA (Watson et al., 1953), great strides have been made in understanding the complexity and diversity of genomes in health and disease. The development and commercialisation of high-throughput, massively parallel sequencing, has democratised sequencing by offering 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 1.2).

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) (Koch et al., 2021; Collins et al., 1995). The first genomes of the pathogenic *Mycobacterium tuberculosis* (S. T. Cole et al., 1998), *Yersinia pestis* (Parkhill et al., 2001), *Escherichia coli* K-12 (Blattner et al., 1997) were sequenced using this technology, requiring years of effort and significant budgets but providing insights into the genomic complexity of these

[2] https://pubmed.ncbi.nlm.nih.gov/

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



```
@ERR000589.41 EAS139_45:5:1:2:111/1
CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
+
3IIIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
@ERR000589.42 EAS139_45:5:1:2:1293/1
AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTAAAAGAAAT
+
IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

Figure 1.2: **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 Hagemann, 2015; Nicholas J. Loman and Pallen, 2015; Goodwin et al., 2016; Wang et al., 2021; Metzker, 2010; Xu et al., 2020

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 (Sanger et al., 1977; Hagemann, 2015). 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 (Hagemann, 2015).

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 (Slatko et al., 2018). 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) (Goodwin et al., 2016). 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 (Nicholas J. Loman and Pallen, 2015). 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 (hybridisation). 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 (Slatko et al., 2018; Detter et al., 2014). Although widely implemented via DNA chips or microarrays, has largely been displaced by other methods, including sequencing by synthesis (Goodwin et al., 2016).

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 (Giani et al., 2020) 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 (Slatko et al., 2018; Goodwin et al., 2016; Hagemann, 2015).

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 (Hagemann, 2015). They can be further classified as either single-nucleotide addition or cyclic reversible termination or as single-nucleotide addition (Goodwin et al., 2016).

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 (Goodwin et al., 2016). The ThermoFisher Ion 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 (Goodwin et al., 2016). 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) (Hagemann, 2015; Nicholas J Loman et al., 2012).

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

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

[5] https://www.thermofisher.com/pt/en/home/brands/ion-torrent.html

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 (Goodwin et al., 2016). 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 (Nicholas J Loman et al., 2012; Gupta et al., 2014; Hagemann, 2015).

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 (Hagemann, 2015). 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 (Hoang et al., 2022). 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 (Korlach et al., 2012). 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 sequencing (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 (Hoang et al., 2022; Nicholas J. Loman and Pallen, 2015). This technology has accuracy comparable with the Illumina systems but requires a higher initial investment cost, are much larger machines in comparison with

[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/

the benchtop counterparts, and have much lower throughput and longer library preparation protocols (Hoang et al., 2022; Wenger et al., 2019).

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 (Hoang et al., 2022; Nicholas J. Loman and Pallen, 2015). 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 (Didelot et al., 2012). 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. (Cirulli et al., 2010; Genetics, 2019; Goodwin et al., 2016).

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 (Goodwin et al., 2016). 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 (Cassedy et al., 2021).

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 (Tagini et al., 2017). 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 (Rossen et al., 2018b). 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 (Rossen et al., 2018b).

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 (Goodwin et al., 2016). 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 (Nicholas J. Loman and Pallen, 2015).

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 (Lo et al., 2020). In bacterial infections, initiatives such as Pathogenwatch[10]

[10] https://pathogen.watch/

offers a web-based platform for AMR analysis and phylogeny generation of *Campylobacter*, *Klebsiella*, *Neisseria gonorrhoeae*, *Staphylococcus aureus*, and *Salmonella Typhi* (Afolayan et al., 2021). 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 (Mamede et al., 2021). Enterobase[13] allows for the analysis and visualisation of genomic variation within enteric bacteria (Zhou et al., 2020). 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 (Argimón et al., 2022). 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 (Shu et al., 2017). 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 (Pickett et al., 2012). 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 (Borges, Pinheiro, et al., 2018). 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 (Hadfield et al., 2018)

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 (Beckett et al., 2021; Lancet, 2021). 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 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 (Smith et al., 2009; Team, 2009). By the time the pandemic was declared, on 11 of June of 2009, Smith et al., 2009 reported the origins and

[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/

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 (Piret et al., 2021), had its first publicly available sequence on 5 of July 2015 (GenBank accession number KT006149)(R. Lu et al., 2015), 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) (Kandeil et al., 2016; Al-Shomrani et al., 2020).

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) (Wu et al., 2020). 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 (Maxmen, 2021). 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 (Lancet, 2021; Beckett et al., 2021; Borges, Isidro, et al., 2022).

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 (Schuele, Cassidy, Peker, et al., 2021). 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 (Nicholas J. Loman and Pallen, 2015; Nicholas J. Loman, Constantinidou, J. Z. M. Chan, et al., 2012). 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 (Rossen et al., 2018a). 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 (Dunne et al., 2012). 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 or transcriptome in patient samples

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

[20] `https://www.gisaid.org/`

(Chiu et al., 2019). 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).

Albeit lacking consensus in the field, metagenomics can be classified into two variants as proposed by (Marchesi et al., 2015): (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%), (Srinivasan et al., 2015; Abellan-Schneyder et al., 2021). 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] (J. R. Cole et al., 2009; DeSantis et al., 2006; Pruesse et al., 2007). 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 (Schloss and Handelsman, 2005). 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 (Abellan-Schneyder et al., 2021). Furthermore, targeting of 16S variable regions with short-read sequencing platforms cannot achieve the taxonomic resolu-

[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/`

tion afforded by sequencing the entire (1500 bp) gene (Johnson et al., 2019). 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 (Quick et al., 2016). 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 (Schuele, Cassidy, Lizarazo, et al., 2020; Wylie et al., 2015).

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 (Chiu et al., 2019), 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 (Couto et al., 2018; Rossen et al., 2018b).

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) (Miller et al., 2009; Palmer et al., 2006; Chiu et al., 2019). 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 Wilson et al., 2014, prompting appropriate targeted antibiotic treatment and eventual recovery of the pa-

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

tient. 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 (Couto et al., 2018; Vijayvargiya et al., 2019; Sanabria et al., 2020; Hirakata et al., 2021).

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 (Nicholas J. Loman, Constantinidou, Christner, et al., 2013). 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 (Huang et al., 2017). 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 (Li et al., 2021).

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 (Carleton et al., 2019; Chiu et al., 2019). 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 (Salter et al., 2014; O'Neil et al., 2013; Feehery et al., 2013). 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 (Chiu et al., 2019; McIntyre et al., 2017). 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 how to deal with the increasing amount of data produced in a fit-for-purpose manner (Carriço et al., 2018).

Bioinformatics is an interdisciplinary research field that applies methodologies from computer science, applied mathematics and statistics to the study of biological phenomena(Carriço et al., 2018). 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 Angers-Loustau et al., 2018; Gruening et al., 2019; Sczyrba et al., 2017; Couto et al., 2018.

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

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

[27] https://www.nist.gov/

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 (Cock et al., 2010). Originally developed at the Wellcome Trust Sanger Institute, the FASTQ has emerged as a common file format for sharing sequencing read data (see 1.2). The FASTQ can be considered as an extension of the 'FASTA sequence file format', originally invented by Pearson et al., 1988, 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 (Ewing et al., 1998) and is defined by the following equation:

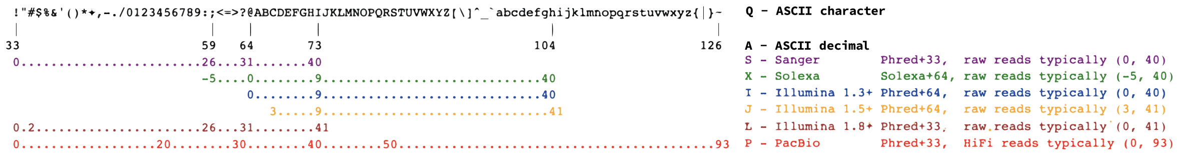
$$\mathbf{Q}\text{PHRED} = -10 \times \log \mathbf{P} \tag{1.1}$$

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

Since their introduction, PHRED scores have become the *de facto* standard for representing sequencing read base qualities (Cock et al., 2010). 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 **??**.

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%



```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~    Q - ASCII character
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126    A - ASCII decimal
 0........................26...31.......40                                                         S - Sanger        Phred+33,  raw reads typically (0, 40)
                          -5....0........9.............................40                          X - Solexa        Solexa+64, raw reads typically (-5, 40)
                                0........9.............................40                          I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
                                   3.....9..............................41                         J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
 0.2......................26...31........41                                                        L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
 0..................20........30........40........50..........................................93   P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

Figure 1.3: **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 structure 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. Escalona et al., 2016 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 1.4).

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 se-

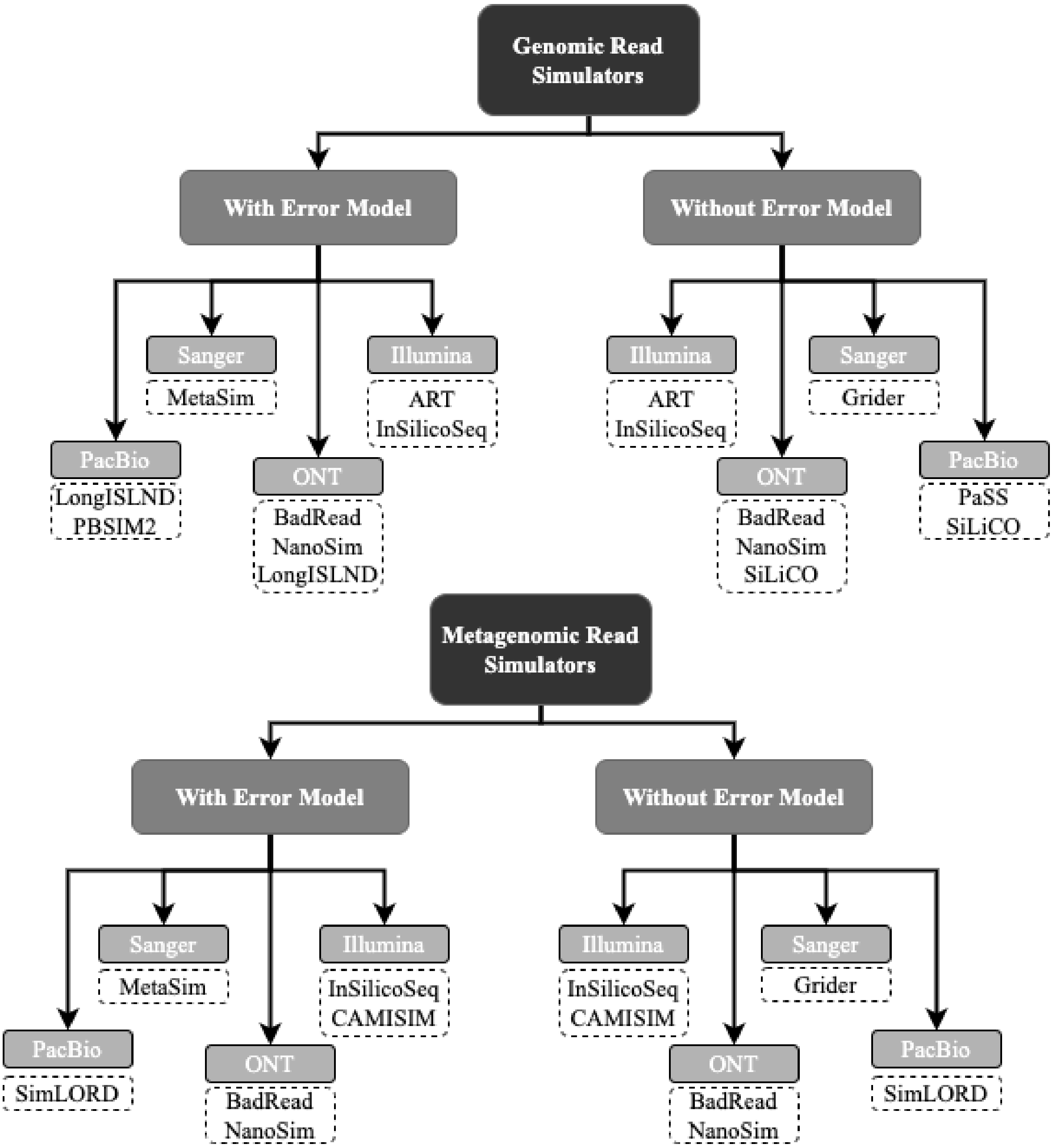


Figure 1.4: **Sequence simulators for genomic and metagenomic data.** For first generation sequencing, Metasim (`https://github.com/gwcbi/metagenomics_simulation`) and Grider (`https://sourceforge.net/projects/biogrinder/`) can generate mock genomic and metagenomic data, with and without error models 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 BioSciences (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 Escalona et al., 2016.

quences 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 (Bolger et al., 2014), Cutadapt (Martin, 2011) or fastp (Chen et al., 2018) to trim and/or filter adaptors, low quality and low complexity sequences. For long-read sequencing, tools like NanoPlot and NanoStats (De Coster et al., 2018), 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 (Derrick E Wood et al., 2014; Derrick E. Wood et al., 2019) and Braken (J. Lu et al., 2017). 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.

These tools are also extremely useful to assess the contents of a metagenomic sample. Alternatives such as Midas (Nayfach et al., 2016), Kaiju, (Menzel et al., 2016), and MetaPhlAn2 (Truong et al., 2015) 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.

1.3.2.1 Metataxonomics and Operational Taxonomic Units

Metataxonomics (see subsubsection 1.2.3.1. Metataxonomics and Targeted Metagenomics) is the most widely used technique for microbial diversity analysis (Hilton et al.,

[28] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[29] https://github.com/rrwick/Filtlong/

[30] https://benlangmead.github.io/aws-indexes/k2

2016), 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)[31] (Bolyen et al., 2019) 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 (Schloss, Westcott, et al., 2009) and UPARSE (Edgar, 2013) are also a popular alternative although resulting outputs differing significantly between pipelines despite using the same inputs having been reported by Marizzoni et al., 2020, 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 (Westcott et al., 2015). 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 (Hao et al., 2011). 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 (Westcott et al., 2015).

[31] `https://qiime2.org/`

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 (Carriço et al., 2018). 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 (Wick et al., 2021).

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 (Ayling et al., 2020). 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 (Rausch et al., 2009; Lischer et al., 2017). 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 (Bayat et al., 2020). 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 (Carriço et al., 2018). 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 1.5). 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.

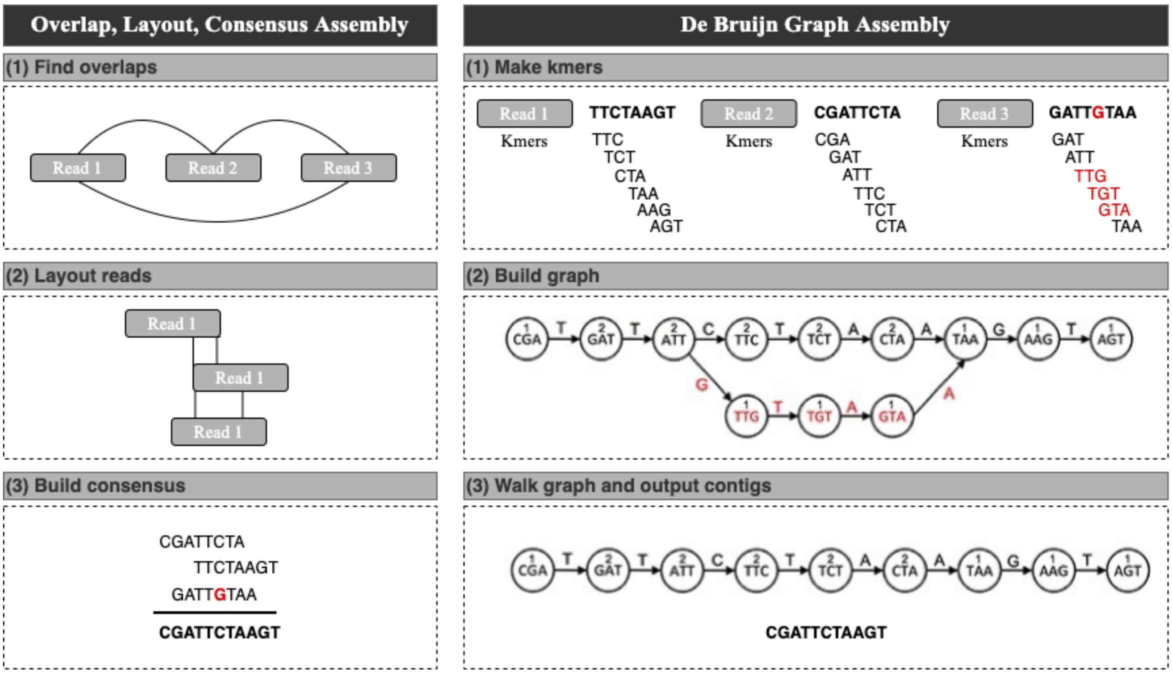


Figure 1.5: **Approaches to *de novo* genome assemble.** In Overlap, Layout, Consensus assembly, (1) overlaps are found between reads and an overlap graph constructed (edges indicate overlapping reads). (2) Reads are laid out into contigs based on the overlaps (lines indicate overlapping portions). (3) 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 Ayling et al., 2020

.

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 (see Figure 1.5), 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.

These type of assemblers were very popular in the early 2010s, with assemblers such as Celera[32], Genovo[33], xGenovo[34] and BBAP[35] having been widely used (Myers et al., 2000; Laserson et al., 2010; Afiahayati et al., 2013; Lin et al., 2017). 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[36], raven[37] and canu[38], the latter being a a fork of the Celera Assembler, have become staples in the community, showing good reliability and amassing over 3000 citations (Vaser et al., 2019; Koren et al., 2017; Wick et al., 2021).

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 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.

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

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 (Bayat et al., 2020).

1.3.3.3 Assembly quality assessment and quality control

1.4 Bioinformatic Analysis for Shotgun 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

For virulence gene detection and antimicrobial resistance characterization a mapping approach, with an adequate database, is usually followed, using as reference the Virulence Factors Database (Chen et al., 2016), and ResFinder (Zankari et al., 2012) or CARD (Jia et al., 2017) for antimicrobial resistance gene detection. Besides mapping, other strategies have been applied, like Mash Screen (Ondov et al., 2016), that offer similar results in a faster way. Similar strategies can be applied to plasmid detection by using the PlasmidFinder (Carattoli et al., 2014) or RefSeq plasmid (O'Leary et al., 2016) databases. The minimap 2 tool (Li, 2018) is a good alternative to map long-read data to any of the resistance and virulence databases mentioned. It is possible to genotype the bacterial population in a metagenomic sample, but only for short-read sequencing data. MetaMLST (Zolfo et al., 2017) reconstructs the MultiLocus Sequence Typing (MLST) loci directly from the sequencing data and provides a pre-built database for the analysis.

A plethora of open-source tools are available specifically for metagenomic data, both short and long-read data, and several combinations of these tools can be used to characterize the causative agent in a patient's infection in a fraction of the time required by the traditional methods. Alongside, there are several commercial alternatives, such as CLC Genomics Workbench (QIAGEN Bioinformatics), Taxonomer (Flygare et al., 2016) and BaseSpace (Illumina), that offer ready to use complete workflows at the cost of lack of transparency, reproducibility and control in the analysis. Several steps that can be implemented to ensure the transparency and reproducibility of the chosen 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 containerizing all the software tools with one of the many solutions available, like Docker (https://www.docker.com/) or Singularity (Kurtzer et al., 2017). The use of workflow managers, like nextflow (Tommaso et al., 2017) or the Galaxy Project

(Afgan et al., 2016), will push reproducibility to the next level by taking advantage of the containerization and scalability, enabling the workflow to be executed with the same parameters in the same conditions in a multitude of different environments. The FlowCraft project (https://github.com/assemblerflow/flowcraft) leverages the combination of Nextflow and docker/singularity containers to assemble, monitor and report scientific pipelines created from the combination of pre-built components, many of them supporting metagenomic analysis. Additional difficulties of metagenomic data are the overpowering quantities of host DNA that are often sequenced (Couto et al., 2018), making the microbial community close to undetectable, the presence of contaminants, from the bench process to the biota, and the cost associated with this methodology. They account for major caveats and must be made aware of when analysing the data. The basic strategies for analysing metagenomic data can be simplified in the scheme in Figure 1. One of the biggest challenges when doing metagenomic analysis is differentiating between colonization and infection and to successfully discriminate between a potential pathogen and background microbiota. In the latter, when analysing samples from presumably sterile sites, like CSF and blood, it is safe to assume that all organisms found are of interest. In locations with a microbiota, the use of spiked metagenomic samples as positive control might guide the detection of the possible pathogens by comparing relative abundance between the samples. 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. These controls should be processed similarly to the samples and the taxonomic results should be filtered out from the final reports.

1.4.1 Metagenome Assembly

Several limitations arise when using just the sequencing data. Although relatively fast and providing quantitative information, it's strictly dependent on the content of the databases used. In addition it lacks context information, as linking the characterizing information to a given identified organism isn't possible. Longer sequences are more informative than shorter sequencing data 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 (Nurk et al., 2017) and MegaHIT (Li et al., 2015). These tools, in comparison to single-cell data assemblers, are better at dealing with the combination of intragenomic and intergenomic repeats and uneven sequencing coverage (Olson et al., 2017). Assembler using multiple k-mers, like the ones suggested, substantially outperform single k-mer assemblers, and smaller k-mers improve the recovery of low-abundance genomes, larger k-mer lead to a better recovery of highly abundant ones (Sczyrba et al., 2017). For long-read data, no dedicated metagenomic assembler is not yet available, but several assembler for long-read data as available, including Canu (Koren et al., 2017) and Unicycler (Wick et al., 2017). The latter allows for hybrid assemblies to be constructed, combining short and long-read information to produce the best assembly possible. 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 (Teeling Glockner, 2012) as traditional assemblers follow the basic principle that the coverage in a sample is constant. 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 softwares can work with single metagenomic samples, most make use of differential coverage of multiple samples to improve the binning process (Sedlar et al., 2016). 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 binning softwares and four taxonomic binning softwares (Sczyrba et al., 2017) revealed that, for taxonomic independent approaches, MaxBin 2.0 (Wu et al., 2016) had the highest completeness and purity in the bins obtained, with 20% better results in comparison with the second best ranked tool. For taxonomic binning, working similarly to the direct taxonomic assignment of the sequencing data, PhyloPythiaS+ (Gregor et al., 2016) obtained better results in accuracy, completeness and purity, followed by Kraken (Wood Salzberg, 2014) 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. 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 (Parks et al., 2015) assesses the quality of the recovered genomes, estimating completeness and contamination by evaluating ubiquitous single-copy genes. 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 (Mikheenko et al., 2016) 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. VALET (https://github.com/marbl/VALET) is a de novo pipeline for detecting mis-assemblies without the need for references, relying instead on coverage and length to do the assignment, as well as providing severa visual representations of assembly quality. After the mis-assemblies have been detected, they can be visualized in Icarus (Mikheenko et al., 2016) for metaQUAST, or IGV (interactive genome viewer) for VALET. Anvi'o (Eren et al., 2015) is an analysis and visualization platform that empowers binning refinement and genome completeness and contamination evaluation with interactive interface. All downstream processes used in single cell genomes can be applied to each of the resulting binned genomes, that now represent a taxonomic unit recovered from the original metagenomic sample. The typical workflow usually involves antimicrobial resistance and ccvirulence detection. Similar approaches can be used as described in the Direct Taxonomic Assignment and Characterization section by using alignment methods, such as BLAST (Altschul et al., 1990) or DIAMOND (Buchfink et al., 2015), to compare against the Virulence Factors Database, and the ResFinder or CARD databases. Genotyping can be done through the mlst software (https://github.com/tseemann/mlst). The reconstructed genomes allow for the use comparative genomics against other references by using, for example, cgMLST or SNP analysis, and playing a major role in early outbreak detection.

1.4.2 Virus in Metagenomic Analysis

One of the biggest advantages of using metagenomic methods is the detection of not only bacterial organisms, but also viral and eukaryotic pathogens. Besides the limitations inherent to the metagenomic process, the retrieval of viral genomes from clinical samples has added difficulties. The fragments of viral genomes are typically orders of magnitude less abundant, the viral genomes often deviate considerably from reference genomes, and the high intrapopulation viral diversity can lead to ambiguous sequence reconstruction or broken assemblies (Rose et al., 2016). Adding to this, the relatively few viral reference genomes can render classification problematic. For fungi, there's an underrepresentation of the diversity of this group in databases as it remains understudied compared to bacterial microbiomes (Donovan et al., 2018). Of the tools mentioned for read classification, Kraken's MiniKraken and MetaPhlAn 2 databases is the most inclusive, including information of virus, bacteria, human and fungi. None other method of the ones described in this review provide a database as inclusive but many, such as Midas, allow the user to build custom databases although requiring very high computational power. Alternatively, the assembly based method can be implemented followed by an alignment search to a database that includes fungi and viral genomes, such as NCBI's RefSeq (O'Leary et al., 2016) or GenBank (Benson et al., 2005).

1. GENERAL INTRODUCTION

Bibliography

Abellan-Schneyder, Isabel et al. (Feb. 2021). "Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing". EN. In: *mSphere*. 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).

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

Afolayan, Ayorinde O. et al. (Dec. 2021). "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, S267–S274. ISSN: 1537-6591. DOI: 10.1093/cid/ciab785.

Allerberger, Franz (Jan. 2012). "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, 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).

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

Angers-Loustau, Alexandre et al. (Dec. 2018). "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, p. 459. ISSN: 2046-1402. DOI: 10.12688/f1000research.14509.2. URL: https://f1000research.com/articles/7-459/v2 (visited on 03/25/2021).

Argimón, Silvia et al. (2022). "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).

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

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

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

Beckett, Angela H., Kate F. Cook, and Samuel C. Robson (2021). “A pandemic in the age of next-generation sequencing”. In: *The Biochemist* 43.6, 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).

Benkova, M., O. Soukup, and J. Marek (2020). “Antimicrobial susceptibility testing: currently used methods and devices and the near future in clinical practice”. en. In: *Journal of Applied Microbiology* 129.4. _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).

Blattner, Frederick R. et al. (Sept. 1997). “The Complete Genome Sequence of *Escherichia coli* K-12”. en. In: *Science* 277.5331, 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).

Bobrovitz, Niklas et al. (June 2021). “Global seroprevalence of SARS-CoV-2 antibodies: A systematic review and meta-analysis”. In: *PLoS ONE* 16.6, 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).

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

Bolyen, Evan et al. (Aug. 2019). “Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2”. en. In: *Nature Biotechnology* 37.8. 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).

Borges, Vítor, Joana Isidro, et al. (Jan. 2022). “SARS-CoV-2 introductions and early dynamics of the epidemic in Portugal”. en. In: *Communications Medicine* 2.1. 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).

Borges, Vítor, Miguel Pinheiro, et al. (June 2018). “INSaFLU: an automated open web-based bioinformatics suite “from-reads” for influenza whole-genome-sequencing-based surveillance”. In: *Genome Medicine* 10.1, 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).

Burrell, Christopher J., Colin R. Howard, and Frederick A. Murphy (2017). “Laboratory Diagnosis of Virus Diseases”. In: *Fenner and White’s Medical Virology*, 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).

Carleton, Heather A. et al. (July 2019). “Metagenomic Approaches for Public Health Surveillance of Foodborne Infections: Opportunities and Challenges”. In: *Foodborne Pathogens and Disease* 16.7. 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).

Carriço, J. A. et al. (2018). “A primer on microbial bioinformatics for nonbioinformaticians”. en. In: *Clinical Microbiology and Infection* 24.4, 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).

Cassedy, A., A. Parle-McDermott, and R. O’Kennedy (Apr. 2021). “Virus Detection: A Review of the Current and Emerging Molecular and Immunological Methods”. In: *Frontiers in Molecular Biosciences* 8, 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).

Chan, YuYen et al. (June 2021). “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. 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).

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

Chiu, Charles Y. and Steven A. Miller (June 2019). “Clinical metagenomics”. en. In: *Nature Reviews Genetics* 20.6. 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).

Cirulli, Elizabeth T. and David B. Goldstein (June 2010). “Uncovering the roles of rare variants in common disease through whole-genome sequencing”. en. In: *Nature Reviews Genetics* 11.6. 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).

Cock, Peter J. A. et al. (Apr. 2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". In: *Nucleic Acids Research* 38.6, 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).

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

Cole, S. T. et al. (June 1998). "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence". en. In: *Nature* 393.6685. 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).

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

Couto, Natacha et al. (Dec. 2018). "Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens". en. In: *Scientific Reports* 8.1, 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).

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

DeSantis, T. Z. et al. (July 2006). "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB". In: *Applied and Environmental Microbiology* 72.7, 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).

Destoumieux-Garzón, Delphine et al. (Feb. 2018). "The One Health Concept: 10Years Old and a Long Road Ahead". In: *Frontiers in Veterinary Science* 5, 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).

Detter, J.C. et al. (2014). "Nucleic acid sequencing for characterizing infectious and/or novel agents in complex samples". en. In: *Biological Identification*. Elsevier, 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).

Di Nardo, Fabio et al. (Jan. 2021). "Ten Years of Lateral Flow Immunoassay Technique Applications: Trends, Challenges and Future Perspectives". en. In: *Sensors* 21.15. 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).

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

Dien Bard, Jennifer and Erin McElvania (Dec. 2020). “Panels and Syndromic Testing in Clinical Microbiology”. In: *Clinics in Laboratory Medicine* 40.4, 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).

Diep, Benjamin et al. (2019). “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. ISSN: 1664-302X. URL: https://www.frontiersin.org/article/10.3389/fmicb.2019.02554 (visited on 01/31/2022).

Dunne, W. M., L. F. Westblade, and B. Ford (Aug. 2012). “Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory”. en. In: *European Journal of Clinical Microbiology & Infectious Diseases* 31.8, 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).

Edgar, Robert C. (Oct. 2013). “UPARSE: highly accurate OTU sequences from microbial amplicon reads”. en. In: *Nature Methods* 10.10. 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).

Escalona, Merly, Sara Rocha, and David Posada (Aug. 2016). “A comparison of tools for the simulation of genomic next-generation sequencing data”. In: *Nature reviews. Genetics* 17.8, 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).

Ewing, Brent and Phil Green (Mar. 1998). “Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities”. en. In: *Genome Research* 8.3. 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).

Feehery, George R. et al. (2013). “A Method for Selectively Enriching Microbial DNA from Contaminating Vertebrate Host DNA”. en. In: *PLoS ONE* 8.10. 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).

Foxman, Betsy et al. (Nov. 2005). “Choosing an appropriate bacterial typing technique for epidemiologic studies”. In: *Epidemiologic perspectives & innovations : EP+I* 2, 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).

Genetics, Nature Reviews (June 2019). “A genomic approach to microbiology”. en. In: *Nature Reviews Genetics* 20.6, 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).

Giani, Alice Maria et al. (Jan. 2020). “Long walk to genomics: History and current approaches to genome sequencing and assembly”. en. In: *Computational and Structural Biotechnology Journal* 18, 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).

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

Goodwin, Sara, John D. McPherson, and W. Richard McCombie (June 2016). “Coming of age: ten years of next-generation sequencing technologies”. en. In: *Nature Reviews Genetics* 17.6. 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).

Groseclose, Samuel L. and David L. Buckeridge (2017). “Public Health Surveillance Systems: Recent Advances in Their Use and Evaluation”. In: *Annual Review of Public Health* 38.1. _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).

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

Gupta, Anuj Kumar and U. D. Gupta (Jan. 2014). “Chapter 19 - Next Generation Sequencing and Its Applications”. en. In: *Animal Biotechnology*. Ed. by Ashish S. Verma and Anchal Singh. San Diego: Academic Press, 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).

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

Hagemann, Ian S. (2015). “Overview of Technical Aspects and Chemistries of Next-Generation Sequencing”. en. In: *Clinical Genomics*. Elsevier, 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).

Hao, Xiaolin, Rui Jiang, and Ting Chen (Mar. 2011). "Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering". en. In: *Bioinformatics* 27.5, 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).

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

Hirakata, Shota et al. (2021). "The application of shotgun metagenomics to the diagnosis of granulomatous amoebic encephalitis due to Balamuthia mandrillaris: a case report". In: *BMC Neurology* 21.1, 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).

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

Hood, D. W. et al. (Oct. 1996). "DNA repeats identify novel virulence genes in Haemophilus influenzae." en. In: *Proceedings of the National Academy of Sciences* 93.20, 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).

Huang, Andrew D. et al. (Jan. 2017). "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, 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).

Jauneikaite, Elita et al. (June 2015). "Current methods for capsular typing of Streptococcus pneumoniae". en. In: *Journal of Microbiological Methods* 113, 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).

Johnson, Jethro S. et al. (Nov. 2019). "Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis". en. In: *Nature Communications* 10.1. 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).

Jolley, Keith A., James E. Bray, and Martin C. J. Maiden (Sept. 2018). "Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications". In: *Wellcome Open Research* 3, 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).

Kandeil, Ahmed et al. (Apr. 2016). "Complete Genome Sequence of Middle East Respiratory Syndrome Coronavirus Isolated from a Dromedary Camel in Egypt". In: *Genome*

Announcements 4.2, 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).

Koch, Linda, Catherine Potenski, and Michelle Trenkmann (Feb. 2021). “Sequencing moves to the twenty-first century”. en. In: *Nature Research*. 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).

Koczula, Katarzyna M. and Andrea Gallotta (June 2016). “Lateral flow assays”. en. In: *Essays in Biochemistry* 60.1. 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).

Koren, Sergey et al. (May 2017). “Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation”. en. In: *Genome Research* 27.5. 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).

Korlach, Jonas and Stephen W Turner (June 2012). “Going beyond five bases in DNA sequencing”. en. In: *Current Opinion in Structural Biology*. Nucleic acids/Sequences and topology 22.3, 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).

Lancet, The (Feb. 2021). “Genomic sequencing in pandemics”. English. In: *The Lancet* 397.10273. 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).

Larsen, David A. and Krista R. Wigginton (Oct. 2020). “Tracking COVID-19 with wastewater”. en. In: *Nature Biotechnology* 38.10. 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).

Larsen, Mette V. et al. (Apr. 2012). “Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria”. en. In: *Journal of Clinical Microbiology* 50.4, 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).

Laserson, Jonathan, Vladimir Jojic, and Daphne Koller (2010). “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, 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).

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

Lin, You-Yu et al. (2017). “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, 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).

Lischer, Heidi E. L. and Kentaro K. Shimizu (Nov. 2017). “Reference-guided de novo assembly approach improves genome reconstruction for related species”. In: *BMC Bioinformatics* 18.1, 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).

Lo, Stephanie W. and Dorota Jamrozy (Sept. 2020). “Genomics and epidemiological surveillance”. en. In: *Nature Reviews Microbiology* 18.9. 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).

Loman, Nicholas J et al. (May 2012). “Performance comparison of benchtop high-throughput sequencing platforms”. en. In: *Nature Biotechnology* 30.5, 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).

Loman, Nicholas J., Chrystala Constantinidou, Jacqueline Z. M. Chan, et al. (Sept. 2012). “High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity”. en. In: *Nature Reviews Microbiology* 10.9. 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).

Loman, Nicholas J., Chrystala Constantinidou, Martin Christner, et al. (2013). “A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4”. In: *JAMA* 309.14, 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).

Loman, Nicholas J. and Mark J. Pallen (Dec. 2015). “Twenty years of bacterial genome sequencing”. en. In: *Nature Reviews Microbiology* 13.12. 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).

Lu, Jennifer et al. (Jan. 2017). “Bracken: estimating species abundance in metagenomics data”. en. In: *PeerJ Computer Science* 3. 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).

Lu, Roujian et al. (Aug. 2015). "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, 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).

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

Maiden, Martin C.J. (Oct. 2006). "Multilocus Sequence Typing of Bacteria". en. In: *Annual Review of Microbiology* 60.1, 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).

Mamede, Rafael et al. (Jan. 2021). "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, pp. D660–D666. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa889. URL: https://doi.org/10.1093/nar/gkaa889 (visited on 02/18/2022).

Marchesi, Julian R. and Jacques Ravel (July 2015). "The vocabulary of microbiome research: a proposal". In: *Microbiome* 3.1, 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).

Marizzoni, Moira et al. (2020). "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. ISSN: 1664-302X. URL: https://www.frontiersin.org/article/10.3389/fmicb.2020.01262 (visited on 03/04/2022).

Martin, Marcel (May 2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1, 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).

Maxmen, Amy (Apr. 2021). "One million coronavirus sequences: popular genome site hits mega milestone". en. In: *Nature* 593.7857. 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).

McIntyre, Alexa B. R. et al. (2017). "Comprehensive benchmarking and ensemble approaches for metagenomic classifiers". In: *Genome Biology* 18.1, 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).

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

Melo-Cristino, J., Letícia Santos, and Mário Ramirez (Jan. 2006). "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, 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).

Menzel, Peter, Kim Lee Ng, and Anders Krogh (Apr. 2016). "Fast and sensitive taxonomic classification for metagenomics with Kaiju". en. In: *Nature Communications* 7.1. 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).

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

Micah, Angela E. et al. (Oct. 2021). "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. 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).

Miller, Melissa B. and Yi-Wei Tang (Oct. 2009). "Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology". In: *Clinical Microbiology Reviews* 22.4, 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).

Muhamad Rizal, Nurnabila Syafiqah et al. (Oct. 2020). "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. 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).

Murray, Christopher JL et al. (Jan. 2022). "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis". English. In: *The Lancet* 0.0. 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).

Murray, Jillian and Adam L. Cohen (2017). "Infectious Disease Surveillance". In: *International Encyclopedia of Public Health*, 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).

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

Nayfach, Stephen et al. (Nov. 2016). "An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography". en. In: *Genome Research* 26.11, pp. 1612–1625. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.201863.115. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.201863.115 (visited on 03/03/2022).

Neoh, Hui-min et al. (2019). "Pulsed-field gel electrophoresis (PFGE): A review of the "gold standard" for bacteria typing and current alternatives". en. In: *Infection, Genetics and Evolution* 74, 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).

O'Neil, Dominic, Heike Glowatz, and Martin Schlumpberger (2013). "Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity". en. In: *Current Protocols in Molecular Biology* 103.1. _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).

Organization, World Health (2005). *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. ISBN: 978-92-4-158041-0. URL: https://www.who.int/publications-detail-redirect/9789241580410 (visited on 02/07/2022).

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

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

Parkhill, J. et al. (Oct. 2001). "Genome sequence of Yersinia pestis, the causative agent of plague". en. In: *Nature* 413.6855. 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).

Patel, Robin (Jan. 2015). "MALDI-TOF MS for the Diagnosis of Infectious Diseases". In: *Clinical Chemistry* 61.1, 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).

Paton, James C. and Claudia Trappetti (Apr. 2019). "Streptococcus pneumoniae Capsular Polysaccharide". EN. In: *Microbiology Spectrum*. 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).

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

Pickett, Brett E. et al. (Nov. 2012). "Virus Pathogen Database and Analysis Resource (ViPR): A Comprehensive Bioinformatics Database and Analysis Resource for the Coronavirus Research Community". en. In: *Viruses* 4.11. 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).

Piret, Jocelyne and Guy Boivin (Jan. 2021). "Pandemics Throughout History". In: *Frontiers in Microbiology* 11, 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).

Pruesse, Elmar et al. (Dec. 2007). "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB". In: *Nucleic Acids Research* 35.21, 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).

Quick, Joshua et al. (Feb. 2016). "Real-time, portable genome sequencing for Ebola surveillance". In: *Nature* 530.7589, 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).

Rausch, Tobias et al. (May 2009). "A consistency-based consensus algorithm for de novo and reference-guided sequence assembly of short reads". In: *Bioinformatics* 25.9, 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).

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

Rossen, J. W. A., A. W. Friedrich, and J. Moran-Gilad (2018a). "& 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. DOI: 10.1016/j.cmi.2017.11.001. URL: https://doi.org/10.1016/j.cmi.2017.11.001.

— (Apr. 2018b). "Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology". English. In: *Clinical Microbiology and Infection* 24.4. Publisher: Elsevier, pp. 355–360. ISSN: 1198-743X. DOI: 10.1016/j.cmi.2017.11.001. URL: https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X(17)30630-4/fulltext (visited on 02/18/2022).

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

Salter, Susannah J. et al. (Nov. 2014). “Reagent and laboratory contamination can critically impact sequence-based microbiome analyses”. In: *BMC Biology* 12.1, 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).

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

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

Scerbo, Michelle H. et al. (June 2016). “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, pp. 294–302. ISSN: 1557-8674. DOI: 10.1089/sur.2015.099.

Schloss, Patrick D. and Jo Handelsman (Mar. 2005). “Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness”. In: *Applied and Environmental Microbiology* 71.3. 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).

Schloss, Patrick D., Sarah L. Westcott, et al. (Dec. 2009). “Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities”. In: *Applied and Environmental Microbiology* 75.23. 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).

Schuele, Leonard, Hayley Cassidy, Erley Lizarazo, et al. (Dec. 2020). “Assessment of Viral Targeted Sequence Capture Using Nanopore Sequencing Directly from Clinical Samples”. en. In: *Viruses* 12.12. 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).

Schuele, Leonard, Hayley Cassidy, Nilay Peker, et al. (2021). “Future potential of metagenomics in microbiology laboratories”. In: *Expert Review of Molecular Diagnostics* 21.12. 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).

Sczyrba, Alexander et al. (Nov. 2017). "Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software". en. In: *Nature Methods* 14.11, pp. 1063–1071. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4458. URL: http://www.nature.com/articles/nmeth.4458 (visited on 03/25/2021).

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

Shu, Yuelong and John McCauley (Mar. 2017). "GISAID: Global initiative on sharing all influenza data – from vision to reality". In: *Eurosurveillance* 22.13, 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).

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

Smith, Gavin J. D. et al. (June 2009). "Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic". en. In: *Nature* 459.7250. 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).

Srinivasan, Ramya et al. (June 2015). "Use of 16S rRNA Gene for Identification of a Broad Range of Clinically Relevant Bacterial Pathogens". en. In: *PLOS ONE* 10.2. 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).

Tagini, F. and G. Greub (2017). "Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review". In: *European Journal of Clinical Microbiology & Infectious Diseases* 36.11, 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).

Team, Novel Swine-Origin Influenza A (H1N1) Virus Investigation (June 2009). "Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans". In: *New England Journal of Medicine* 360.25. 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).

Teutsch, Steven M. (2010). "Considerations in Planning a Surveillance System". eng. In: *Principles & Practice of Public Health Surveillance*. 3rd ed. Oxford University Press.

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).

Truong, Duy Tin et al. (Oct. 2015). "MetaPhlAn2 for enhanced metagenomic taxonomic profiling". en. In: *Nature Methods* 12.10. 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).

Van Noorden, Richard, Brendan Maher, and Regina Nuzzo (Oct. 2014). "The top 100 papers". en. In: *Nature News* 514.7524. 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).

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

Vijayvargiya, Prakhar et al. (Feb. 2019). "Application of metagenomic shotgun sequencing to detect vector-borne pathogens in clinical blood samples". en. In: *PLOS ONE* 14.10. 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).

Vos, Theo et al. (Oct. 2020). "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. 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).

Wang, Yunhao et al. (Nov. 2021). "Nanopore sequencing technology, bioinformatics and applications". en. In: *Nature Biotechnology* 39.11. 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).

Watson, J. D. and F. H. C. Crick (Apr. 1953). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". en. In: *Nature* 171.4356. 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).

Wenger, Aaron M. et al. (Oct. 2019). "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome". en. In: *Nature Biotechnology* 37.10, 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).

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

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

Wilson, Michael R. et al. (June 2014). “Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing”. In: *New England Journal of Medicine* 370.25. 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).

Wood, Derrick E and Steven L Salzberg (2014). “Kraken: ultrafast metagenomic sequence classification using exact alignments”. en. In: *Genome Biology* 15.3, 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 03/03/2022).

Wood, Derrick E., Jennifer Lu, and Ben Langmead (Nov. 2019). “Improved metagenomic analysis with Kraken 2”. In: *Genome Biology* 20.1, 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).

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

Wu, Fan et al. (Mar. 2020). “A new coronavirus associated with human respiratory disease in China”. en. In: *Nature* 579.7798, 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).

Wylie, Todd N. et al. (Dec. 2015). “Enhanced virome sequencing using targeted sequence capture”. In: *Genome Research* 25.12, 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).

Xu, Liu and Masahide Seki (Jan. 2020). “Recent advances in the detection of base modifications using the Nanopore sequencer”. en. In: *Journal of Human Genetics* 65.1. 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).

Zhou, Zhemin et al. (Jan. 2020). “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. 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).

Chapter 2

Conclusion

2. CONCLUSION

Appendix A

Appendix