UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA

Towards accreditation in metagenomics for clinical microbiology

Catarina Inês Marques de Sousa Mendes

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

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

2022

UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA

Towards accreditation in metagenomics for clinical microbiology

Catarina Inês Marques de Sousa Mendes

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

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

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

2022

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

-J. R. R. Tolkien, The Hobbit

Acknowledgements

Summary

Keywords: one, two, three, four, five

Resumo

Keywords: um, dois, três, quatro, cinco

Thesis Outline

The work described in the present thesis intended to evaluate the use of bioinformatics methods for the analysis of metagenomic data to allow the rapid identification, virulence analysis and antimicrobial susceptibility prediction of pathogens with clinical relevance. Ultimately, the applicability of metagenomic methods is to be evaluated in a clinical setting as an alternative to current golden standards. Given the dependence of these methodologies on bioinformatics post-processing of the raw data obtained, the major applications and pitfalls of metagenomics are to be identified.

The thesis comprises 10 chapters, organised as follows:

In **Chapter 1** the issues addressed throughout the thesis are put into context, highlighting the current impact of genomics in clinical microbiology, both as a diagnostic or a surveillance tool. The entire process in clinical microbiology for bacterial and viral infections is showcased through its different approaches over time: classical biochemical and molecular methods, whole-genome sequencing, and sequencing through metagenomics, both metataxonomics and shotgun, with a focus on the computational requirements necessary. This chapter elaborates on the evolution of whole-genome sequencing to metagenomic approaches, introducing the possibility of the identification and characterisation of a potential pathogen without the need for a priori knowledge of the causative agent of disease. The importance of bioinformatics analysis of these data was underlined, showcasing its complexity and the major pitfalls, such as reproducibility and transparency of the analysis methods.

Chapter 2 consists of the application of the shotgun metagenomics approach to nine body fluid samples and one tissue sample from patients at the University Medical Center Groningen (UMCG) as to compare against current golden standards practises in the diagnosis of disease. In this study, the accuracy and reliability of the bioinformatics analyses were evaluated and compared against the results obtained from traditional culture methods. Our aim was to evaluate the applicability of shotgun metagenomics in a routine diagnostic setting, and not only in cases where traditional methods fail to provide an answer. Most pathogens identified by culture were also identified by metagenomics. Substantial differences were noted between the taxonomic classification tools, highlighting the potential and limitations of shotgun metagenomics as a diagnostic tool. The fact that, when applying shotgun metagenomics to diagnostics, the results are highly dependent on the tools, and especially the database

that was chosen for the analysis greatly impacts its applicability in a clinical setting. This chapter is included in the following publication: *N. Couto, L. Schuele, E.C. Raangs, M. P. Machado, C. I. Mendes, T. F. Jesus, M. Chlebowicz, S. Rosema, M. Ramirez, J. A. Carriço, I. B. Autenrieth, A. W. Friedrich, S. Peter and J. W. Rossen. Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. Sci Rep 8, 13767 (2018). DOI: https://doi.org/10.1038/s41598-018-31873-w*

Chapter 3 describes the application of both second and third-generation sequencing technologies, also known as next-generation and long-read sequencing, to tap-water samples collected at the University Medical Center Groningen. Our aim was to evaluate the applicability of shotgun metagenomics, but this time in a surveillance setting. Building on the findings from Chapter 2, a hybrid assembly approach was used to increase resolution power. In this sample a new variant of a colistin resistance (mcr) determinants was detected, named mcr-5.4, and through hybrid assembly leveraging both short and long-read sequences, its context was determined, albeit with questionable success. This chapter is included in the following publication: *G. Fleres, N. Couto, L. Schuele, M. A. Chlebowicz, C. I. Mendes, L. W. M. van der Sluis, J. W. A. Rossen, A. W Friedrich, S. García-Cobos, Detection of a novel mcr-5.4 gene variant in hospital tap water by shotgun metagenomic sequencing, Journal of Antimicrobial Chemotherapy, Volume 74, Issue 12, December 2019, Pages 3626–3628. DOI: https://doi.org/10.1093/jac/dkz363.*

With the lessons learnt in Chapters 2 and 3, we developed in **Chapter 4** DEN-IM, a one-stop, user-friendly, containerised and reproducible workflow for the analysis of Dengue virus short-read sequencing data from both amplicon and shotgun metagenomics approaches. This takes into particular consideration the dependency on software and database versions used in the metagenomic bioinformatics downstream analysis in the results obtained. Dengue virus represents a public health threat and economic burden in affected countries, with the risk of exposure, increasing, not only driven by travel to endemic regions but also due to the broader dissemination of the mosquito vector, making the burden of dengue very significant. This makes it a particularly relevant target organism for the development of a straightforward workflow for both the identification and characterization of the virus. DEN-IM was designed to perform a comprehensive analysis in order to generate either de novo assemblies or consensus of full viral coding sequences and to identify their serotype and genotype, including the identification of co-infection cases whose prevalence is increasingly found in highly endemic areas. It was developed in Nextflow, a simple and scalable workflow management system. All tools and dependencies are provided in Docker containerised images. All these steps ensure reproducibility and transparency of the bioinformatic process. This chapter is included in the following publication: *C. I. Mendes*[*], *E. Lizarazo*[*] *, M. P. Machado, D. N. Silva, A. Tami, M. Ramirez, N. Couto, J. W. A. Rossen, J. A. Carriço, DEN-IM: dengue virus genotyping from amplicon and shotgun metagenomic sequencing. Microbial Genomics, Volume 6, Issue 3, March 2020. DOI: https://doi.org/10.1099/mgen.0.000328.*

[*]These authors contributed equally to this work.

A key process in metagenomic data analysis is the de novo assembly of raw sequence data since it allows recovering contigs representing the replicons present in the sample, be it genomes, plasmids, or bacteriophages, from a pool of mixed raw reads. **Chapter 5** employs the same core principles as in Chapter 4, describing a one-stop, user-friendly, containerised, and reproducible workflow, named LMAS, to assess the performance of de novo assembly algorithms for the assembly of second-generation metagenomic sequencing data. The LMAS workflow, which allows users to evaluate performance given a known standard community was implemented in Nextflow, ensuring the transparency and reproducibility of the results obtained. Similarly to Chapter 4, the use of Docker containers provides additional flexibility. The results are presented in an interactive HTML report where global and reference specific performance metrics can be explored. Currently, 12 de novo assemblers are implemented in LMAS, with the possibility of expansion as novel algorithms are developed.

Despite the advantages of reproducible, containerised workflow, Chapters 4 and 5 still do not guarantee the interoperability of results obtained from various sources. Chapter 5 highlighted the impact that the tool choice can have on downstream results when working with metagenomic data, therefore, and due to the lack of standardisation, it is pivotal that results from various tools can be compared for their applicability in the clinic. With a focus on antimicrobial resistance, **Chapter 6** presents a standardised output specification for the bioinformatic detection of antimicrobial resistance directly from genomes or metagenomes. This addresses the problem of combining the outputs of disparate antimicrobial resistance gene detection tools into a single unified format, implemented into a python package and command-line utility hAMRonization. As the detection of antimicrobial resistance directly from genomic or metagenomic data has become a standard procedure in public health, with hAMRonization allowing for the comparison of results within bioinformatics workflows, as these tools, although implementing similar principles, differ in supported inputs, search algorithms, parameterisation, and underlying reference databases.

Chapter 7 presents a direct application of a standardised specification, such as the one presented in Chapter 6. For this purpose, a SARS-CoV-2 contextual data specification package based on harmonisable, publicly available community standards was developed and implemented through a collection template, as well as a variety of protocols and tools to support both the harmonisation and submission of sequence data and contextual information to public biorepositories. In addition to the reproducibility and interoperability of data and software, transparency is also a keystone in the use of bioinformatics methods for the analysis of metagenomic data. This chapter is included in the following publication: *E. J. Griffiths, R. E. Timme, C. I. Mendes, A. J. Page, N. Alikhan, D. Fornika, F. Maguire, J. Campos, D. Park, I. B. Olawoye, P. E. Oluniyi, D. Anderson, A. Christoffels, A. G. da Silva, R. Cameron, D. Dooley, L. S. Katz, A. Black, I. Karsch-Mizrachi, T. Barrett, A. Johnston, T. R. Connor, S. M. Nicholls, A. A. Witney, G. H. Tyson, S. H. Tausch, A. R. Raphenya, B. Alcock, D. M. Aanensen, E. Hodcroft, W. W. L. Hsiao, A. T. R. Vasconcelos, D. R. MacCannell on behalf of the Public Health Alliance for Genomic Epidemiology (PHA4GE) consortium, Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-*

2 contextual data specification package. GigaScience, Volume 11, 2022, giac003. DOI: https://doi.org/10.1093/gigascience/giac003.

Chapter 8 showcases an effort to raise standards on the development and distribution of code for bioinformatic analysis. For this, seven recommendations are presented that help researchers implement software testing in microbial bioinformatics. We propose collaborative software testing as an opportunity to continuously engage software users, developers, and students to unify scientific work across domains. As automated software testing remains underused in scientific software, our set of recommendations not only ensures that appropriate effort can be invested in producing high quality and robust software, but also increases engagement in its sustainability. This chapter is included in the following publication: *B. C. L. van der Putten*[*]*, C. I. Mendes*[*]*, B. M. Talbot, J. de Korne-Elenbaas, R. Mamede, P. Vila-Cerqueira, L. P. Coelho, C. A. Gulvik, L. S. Katz, The Asm Ngs Hackathon Participants, Software testing in microbial bioinformatics: a call to action. Microbial Genomics, Volume 8, Issue 3. DOI: https://doi.org/10.1099/mgen.0.000790.*

Chapter 9 corresponds to the general discussion. This chapter provides a summary of the main results obtained in this thesis and its integrated discussion.

Chapter 10 Contains the main conclusions driven from this work. It also includes perspectives for future work.

[*]These authors contributed equally to this work.

Abbreviation

bp basepairs

cg/wg MLST core-genome/whole genome Multilocus Sequence Typing

dNTP deoxynucleotide triphosphate

ddNTP dideoxynucleotide triphosphate

qPCR Real-Time Quantitative PCR

rRNA ribossomal RNA

AMR Antimicrobial Resistance

CDS Coding Sequence

DENV Dengue virus

GBD Global Burden of Disease

GPP Global Priority Pathogens

HPC high-performance computing

HTS high-throughput sequencing

LFIA Lateral Flow Immunoassays

MERS Middle East Respiratory Syndrome

MLST Multilocus Sequence Typing

NCR non-coding region

NIST National Institute of Standards and Technology

OLC Overlap-Layout Consensus

OTUs Operational Taxonomic Units

PCR Polymerase Chain Reaction

PFGE Pulse Field Gel Electrophoresis

PHSS Public Health Surveillance Systems

PMC PubMed Central®

QC quality control

RT-PCR Reverse Transcription Polymerase Chain Reaction

SARS-CoV-2 Acute Respiratory Syndrome Coronavirus 2

SMg Shotgun Metagenomics

SMRT Single Molecule Real-Time Sequencing

SNP Single Nucleotide Polymorphism

STEC Shiga toxin-producing *Escherichia coli*

VCS Version control systems

WGS Whole Genome Sequencing

WHO World Health Organization

Table of Contents

List of Tables

List of Figures

Chapter 1

General Introduction

1.1 The global impact of microbial pathogens

The Global Burden of Disease (GBD) 2019 study reported that microbial pathogens are responsible for more than 400 million years of life lost annually across the globe, a higher burden than either cancer or cardiovascular disease [1]. In particular, lower respiratory infections, diarrhoeal diseases, HIV/AIDS and tuberculosis were amongst the five leading causes of global total years of life lost. More recently, the COVID-19 pandemic, declared as such by the World Health Organization (WHO) on 11 March 2020 after the emergence and global spread of the Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has caused more than 5 million deaths worldwide [2], making it one of the deadliest pandemics in history. Coronavirus has been responsible for three of the 18 major pandemics recorded throughout modern history [3], all occurring after the year 2000. *Yersinia pestis*, responsible for three plague pandemics, *Vibrio cholerae*, with seven cholera pandemics, and Influenza A virus, the causal agent of five flu pandemics, are responsible for the remaining, Influenza being the only other pathogen with a pandemic registered after the year 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. Besides tuberculosis, the global priority due to being the most common and lethal airborne AMR disease worldwide today, responsible for 250,000 deaths each year, it includes 12 groups of pathogens in three priority categories.

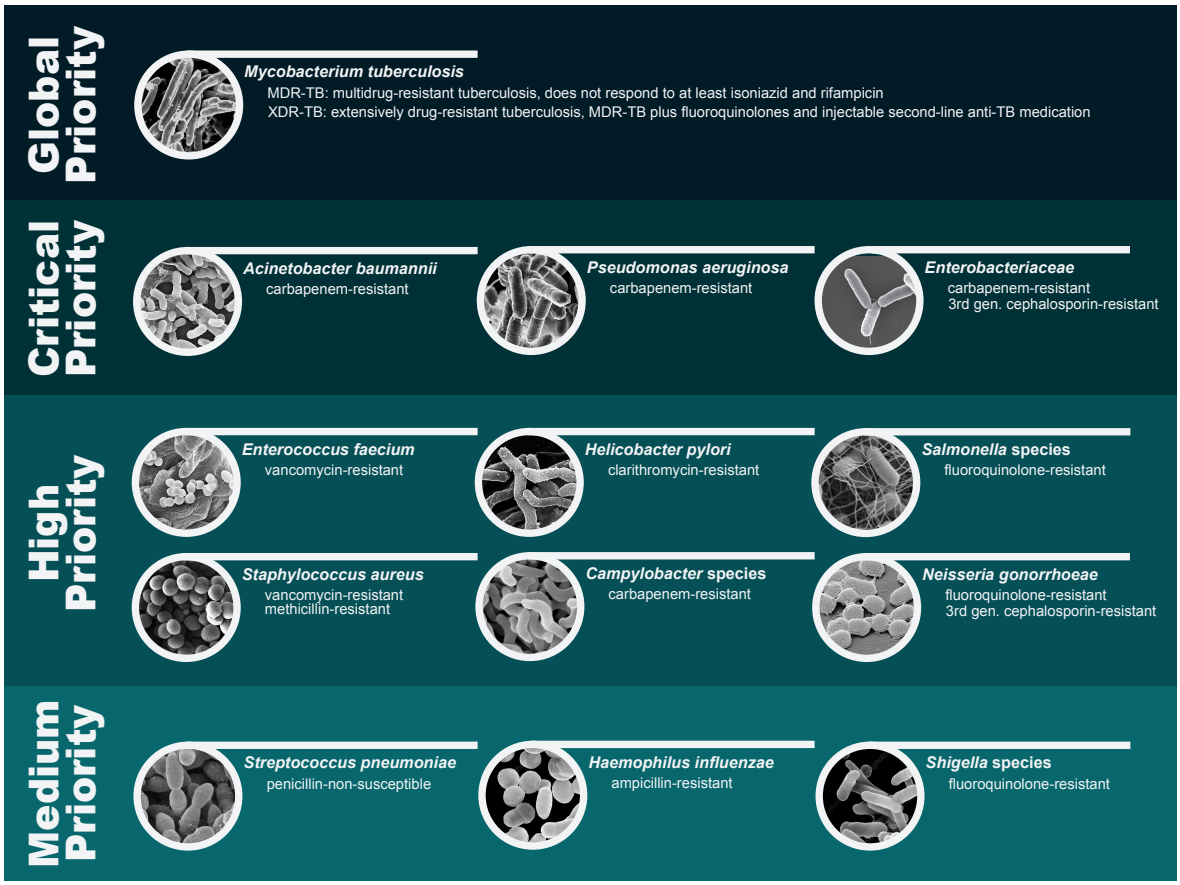


Figure 1.1: **World Health Organisation Global Priority Pathogens list.** This catalogue includes, besides *Mycobacterium tuberculosis* considered the number one global priority, a list of twelve microorganisms grouped under three priority tiers according to their antimicrobial resistance: critical (*Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacteriaceae*), high (*Enterococcus faecium*, *Helicobacter pylori*, *Salmonella* species, *Staphylococcus aureus*, *Campylobacter* species and *Neisseria gonorrhoeae*), and medium (*Streptococcus pneumoniae*, *Haemophilus influenzae* and *Shigella* species). The major objective was to encourage the prioritisation of funding and incentives, align research and development priorities of public health relevance, and garner global coordination in the fight against antimicrobial resistant bacteria. Adapted from [4].

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 the WHO Global Health Spending 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 [5]. 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 health aid went to infectious and parasitic diseases in 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 [6].

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. Application of these methods differs by suspected infection type: bacterial, viral, fungal or parasitic. For the purpose of this dissertation work, we will focus on bacterial (see Section 1.1.1.1 and viral infections (see Section 1.1.1.2.

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 [7]. 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 allow recognition of outbreaks [8] (see Figure 1.2).

The current gold standard for bacterial pathogen identification in diagnostic microbiology laboratories involves the isolation of the pathogen through culture followed by biochemical testing, a multi-step process that can take days to weeks before obtaining results, depending on the fastidiousness of the organism and if it can be cultured [9, 10]. 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. Furthermore, 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 by normal flora and leading to false results [10]. 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 [11]. 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 [12]. Syndromic panels, an extension of PCR by using multiple primers (multiplex PCR) to simultaneously amplify nucleic acids from 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 [10].

Following identification, antibiotic-susceptibility testing is essential to guide clinicians

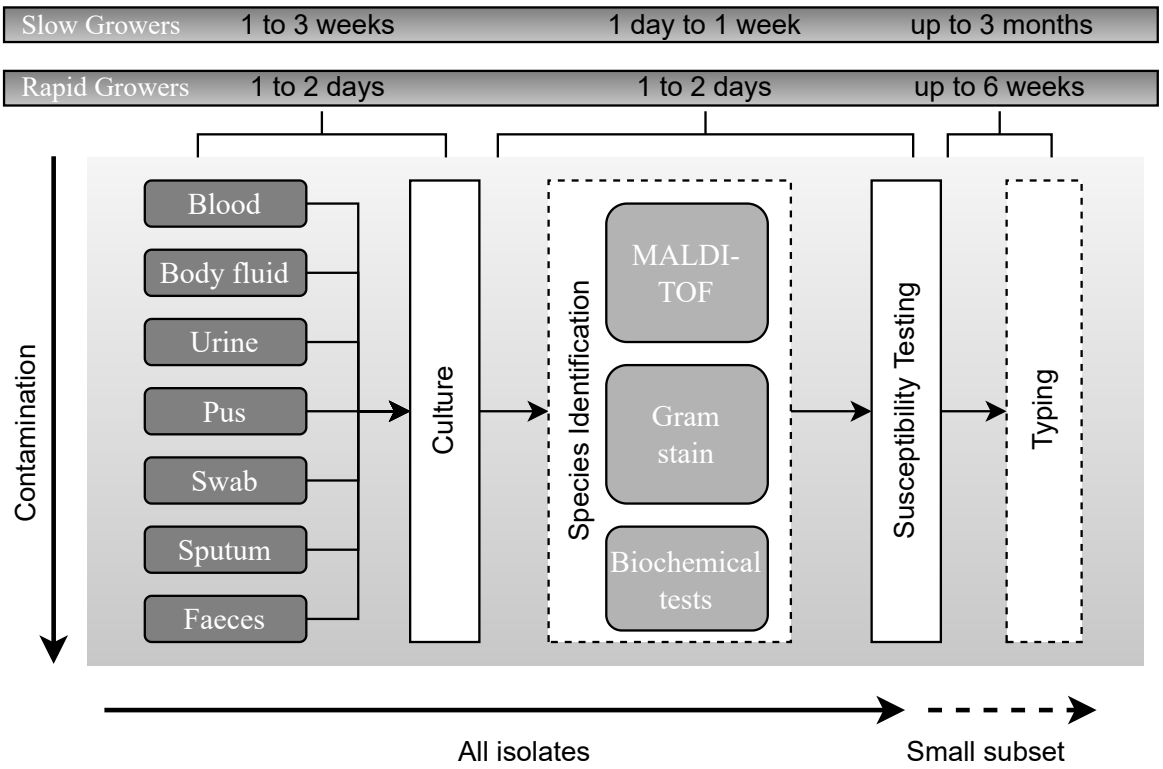


Figure 1.2: **Principles of current processing of bacterial pathogens.** Schematic representation of the current workflow for processing samples for bacterial pathogens is presented, with high complexity and a typical timescale of a few weeks to a few months. Samples that are likely to be normally sterile are often cultured on rich medium that will support the grown of any culturable organism. Samples contaminated with colonising flora present a challenge for growing the infecting pathogen. Many types of culture media (referred to as selective media) are used to favour the growth of the suspected pathogen. Once an organism is growing, the likely pathogens are then processed through a complex pathway that has many contingencies to determine species and antimicrobial susceptibility. Broadly, there are two approaches. One approach uses MALDI-TOF for species identification prior to setting up susceptibility testing. The other uses Gram staining followed by biochemical testing to determine species; susceptibility testing is often set up simultaneously with doing biochemical tests. Lastly, depending on the species and perceived likelihood of an outbreak, a small subset of isolates may be chosen for further investigation using a wide range of typing tests. Adapted from [7].

in selecting an appropriate treatment. Conventional methods of bacterial resistance detection, such as disc diffusion, antimicrobial gradient strip, and broth microdilution, are widely used, but results cannot be obtained before 48 hours after receiving a sample, which can lead to prolonged use or overuse of broad-spectrum antibiotics [13]. Similarly to bacterial identification, MALDI-TOF and PCR have been increasingly adopted as solutions with shorter turnaround times, although no phenotypic information is recovered, nor information on the minimum inhibitory concentration (MIC) for a given antibiotic.

Choosing an appropriate bacterial typing technique for epidemiological studies depends on the available resources and the minimum intended resolution, ranging from DNA fingerprinting to multilocus sequence typing, Pulse Field Gel Electrophoresis (PFGE), and sequence-based typing (see section 1.2. A genomic approach to clinical microbiology) [8, 14]. DNA macrorestriction analysis by PFGE, which revolutionised precise separation of DNA fragments, became the most widely implemented DNA fingerprinting technique [14],

becoming the golden standard for bacterial typing [15].

In the early 2000s, Multilocus Sequence Typing (MLST) was proposed as a portable, universal, and definitive method for characterising bacteria [16]. Instead of enzyme restriction of bacteria DNA, separation of 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 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 [17]. As with PFGE, different schemes, defining what housekeeping 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 [16]. So far, MLST schemes for more than 100 bacterial organisms have been published and made freely available[1], [18])

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 Global Priority Pathogens (GPP) list, the typing of the polysaccharide capsule, usually through Quellung reaction, is paramount for disease surveillance and evaluation of the pre- and post-pneumococcal vaccine, since the capsule, with over 90 serotypes reported, is the dominant surface structure of the organism and plays a critical role in virulence [19, 20]. 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 [21].

1.1.1.2 Viral infections

Traditional approaches to 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 patient serum (serology) [22]. Viral diagnostics is therefore generally organised into two primary categories, indirect and direct detection, depending on the method used 1.3.

Indirect detection methods involve the propagation of virus particles through their introduction to a suitable host cell line (virus isolation), since 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,

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

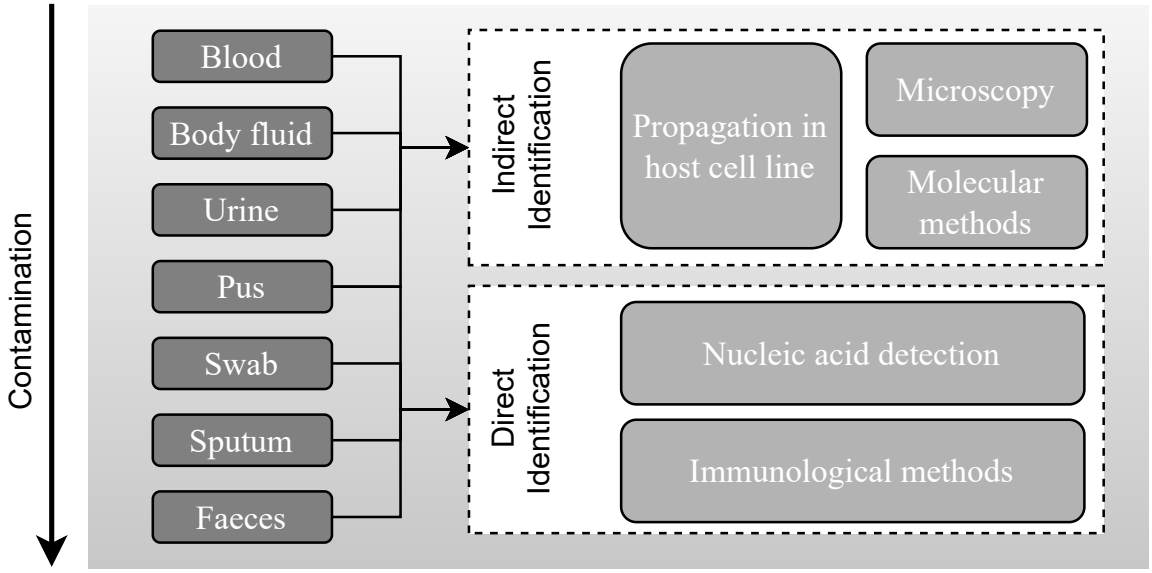


Figure 1.3: **Principles of current processing of viral pathogens.** Schematic representation of the current workflow for processing samples for viral pathogens is presented. Samples that are likely to be normally sterile are often cultured and isolated in suitable host cell lines (indirect identification). This supports the identification through microscopy of molecular methods, but the virus can take weeks to propagate. Direct identification is much faster, relying on nucleic acid detection of immunologic assays for the identification of the pathogen, without the need of virus propagation.

through molecular methods with an agent that detects a virus-associated protein, such as an antibody [23].

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 Polymerase Chain Reaction (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 [23]. Syndromic testing (see subsubsection 1.1.1.1. Bacterial infections) is now fully integrated into standard testing practises of many clinical laboratories [24]. The limitations of these assays include the absence of 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 [23].

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 [23]. A major application is seroprevalence assays, an essential technique to identify patients who have been exposed to a virus (historical exposure), detect asymptomatic infection, or evaluate the efficacy of vaccines [25, 26]. Lateral Flow Immunoassays (LFIA) are widely used to detect virus-associated proteins directly from the source through anti-

bodies labelled that binding to their cognate antigens, usually read by means of a colour change at a test line. In addition to being very cost-effective, LFIA have a turnaround time of minutes and the colour change can be observed with the naked eye, therefore facilitating rapid diagnosis, but their results are limited to semiquantitative and typically do not achieve sensitivity comparable to nucleic acid detection [23, 27, 28].

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 [29, 30]. Public Health Surveillance Systems (PHSS) consist of the ongoing systematic collection, analysis and interpretation of data, and its integration with the timely dissemination of results to those who can carry out effective prevention and control activities [31].

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 health facilities report cases to the public health agency, which conducts data management and analysis once the data is received and communicates with the responsible entities. Globally, the WHO as described in the International Health Regulations what is notifiable by all countries, such as severe acute respiratory syndrome (SARS) and viral hemorrhagic fevers (Ebola, Lassa, Marburg), as well as guiding which public health measures should be implemented [32]. 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 [30, 33]. 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 [34–37].

The emergence and reemergence of infectious diseases are closely linked to the biology and ecology of infectious agents, their hosts, and their vectors [38]. "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 [39]. It recognises that people's health is closely related to the health of animals and the shared environment, focussing on zoonotic and vector-borne diseases, antimicrobial resistance, food safety, food security, and environmental contamination [40]. This is crucial to (1) understanding the emergence and re-emergence of infectious and noncommunicable chronic diseases and (2) in creating innovative control

strategies. A better understanding of the 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 [38].

1.2 A genomic approach to clinical microbiology

Since the publication of the first complete microbial genome, a quarter of a century ago, of the bacterium *Haemophilus influenzae* [41], genomics has transformed the field of microbiology, and in particular its clinical application (see Figure 1.4).

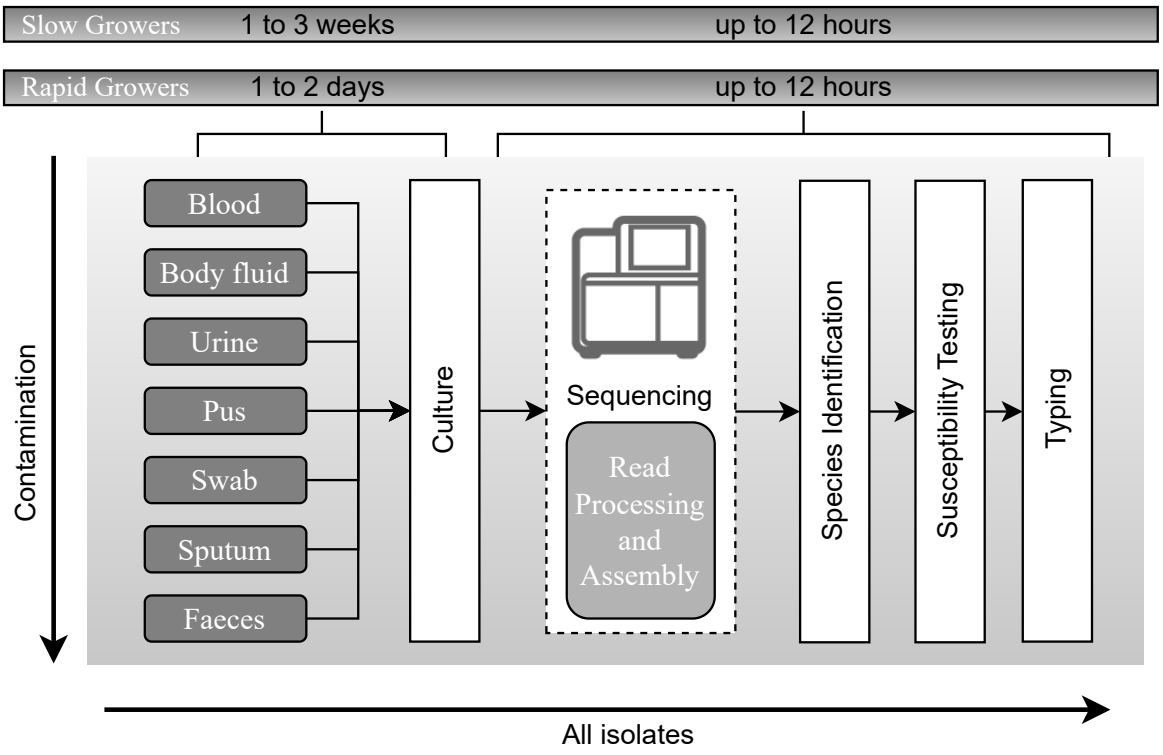


Figure 1.4: **Principles of current processing of bacterial pathogens based on whole genome sequencing.** Schematic representation of the workflow for processing samples for bacterial pathogens after the adoption of whole genome sequencing, with an expected timescale that could fit within a single day. The culture steps would be the same as currently used in a routine microbiology laboratory (see Figure 1.2). Once a likely pathogen is ready for sequencing, DNA is extracted, taking as little as 2 hours to prepare the DNA for sequencing. After sequencing, the main processes for yielding information is computational. Automated sequence assembly algorithms are necessary for processing the raw sequence data, from which species, relationship to other isolates of the same species, antimicrobial resistance profile and virulence gene content can be assessed. All the results can also be used for outbreak detection and infectious diseases surveillance. Adapted from [7]

The paper describing the DNA-sequencing method with chain-terminating inhibitors used in the sequencing of the first microbial genome [42], 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 over 60000, highlighting its impact in the field of biological sciences, and by extension medicine [43]. Currently, this number has increased to over 84000 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 [44], 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: the first, the second and the third generations of sequencing (see Figure 1.5).

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 [51]. The first genomes of pathogenic *Mycobacterium tuberculosis* [52], *Yersinia pestis* [53], *Escherichia coli* K-12 [54] were sequenced using this technology, requiring years of effort and significant budgets, but providing insights into the genomic complexity of these organisms. Some of the complete genome sequences produced during this era are still used today as high-quality references.

Simplistically, in first-generation DNA sequencing, also known as Sanger sequencing, a DNA polymerase is used to synthesise numerous copies of the sequence of interest using ddNTP) in the reaction. At each nucleotide incorporation event, there is a chance that a ddNTP will be added and the growing DNA chain will terminate, resulting in a collection of DNA molecules of varying lengths [42, 45]. Modern Sanger sequencing uses fluorescently labelled ddNTP 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

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

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

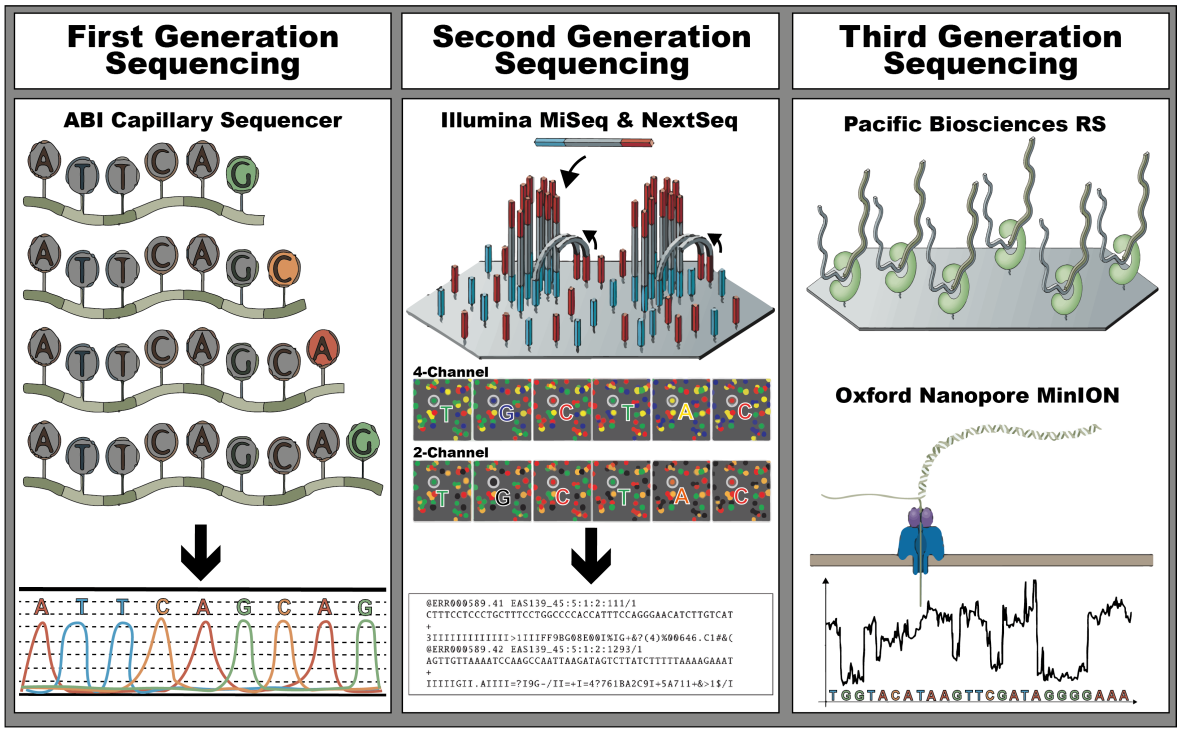


Figure 1.5: **The three revolutions in sequencing technology that have transformed the landscape of bacterial genome sequencing.** The first-generation, also known as Sanger sequencers, is represented by the ABI Capillary Sequencer (Applied Biosystems). During the sequencing reaction, at each nucleotide 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 MiSeq, a 4-channel sequencer, and 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 below. In a 4-channel instrument each nucleotide has its 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 instruments allow the sequencing of both ends of the DNA fragment. Lastly, the third-generation, also known as long-read sequencers, is represented by the 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 change in an ionic current across the membrane each time a nucleotide is pushed through the pore. This difference in potential is then used for basecalling. Adapted from [45–50]

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 per reaction[45].

The first generation 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 [55]. 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 the second generation as next-generation sequencing (NGS) [47]. This trend has continued over the next two decades of continued development and improvement, allied to the emergence of benchtop sequencing platforms with a high-throughput sequencing data and turnaround times of days, making it a standard in any microbiology and public health laboratories [46]. 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 on 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 [55, 56]. Although widely implemented via DNA chips or microarrays, has largely been displaced by other methods, including sequencing by synthesis [47].

1.2.1.2.2 Sequencing by synthesis

Sequencing by synthesis methods is a further development of Sanger sequencing, without the ddNTP 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 by (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 (4) loaded into a flow cell and (5) sequenced in massive parallel sequencing reactions [57]. Besides having much shorter read lengths than first-generation methods, with reads ranging from 45 to 300 bases,. These have an intrinsically higher error rate, the massively parallel sequencing of millions to billions of short DNA sequence reads allows the obtainment of millions of accurate sequences based on the identification of consensus (agreement) sequences [45, 47, 55].

Many of the approaches currently available for sequencing by synthesis methods 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 [45]. They can also be classified as single nucleotide addition or cyclic reversible termination or as single nucleotide addition [47].

The first relies on a single signal to mark the incorporation of a deoxynucleotide triphosphate (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 that only one 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 in 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 [47]. 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 [47]. 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) [45, 58].

The latter is defined by their use of terminator molecules that are similar to those used in the first-generation of sequencing, preventing elongation of the DNA molecule, but unlike the first methods, it is reversible. To begin the process, a DNA template is primed by a sequence that is complementary to an adapter region, which will initiate polymerase binding to this double-stranded DNA region. During each cycle, a mixture of all four individually labelled and 3'-blocked dNTPs are added. After incorporation of a single dNTP into each elongating complementary strand, the 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 [47]. 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 [45,

[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

[6] https://www.forbes.com/companies/illumina/?sh=774358a91aa6

[7] https://www.illumina.com/systems/sequencing-platforms.html

58, 59].

1.2.1.3 The third-generation of DNA sequencing

Despite their wide adoption, second-generation methods require in the library preparation 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 [45]. 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 [60]. 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 or the virulence of pathogens [61]. 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 SMRT, which takes place in single-use SMRT Cells. These contain multiple immobilised polymerases, which, after binding to an adaptor sequence, begin 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, multiple passes of the same template can be generated depending on the lifetime of the polymerase [46, 60]. This technology has accuracy comparable with the Illumina systems but requires a higher initial investment cost, are much larger machines in comparison with the benchtop counterparts, and have much lower throughput and longer library preparation protocols [60, 62].

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 [46, 60]. 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.

[8] https://www.pacb.com/

[9] https://nanoporetech.com/

1.2.2 DNA sequencing in clinical diagnosis and surveillance

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 [7]. 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 (see Section 1.3), 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. [47, 63, 64].

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 [47]. 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 [23].

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 Multilocus Sequence Typing (cg/wg MLST) analysis, for example during hospital outbreaks [65].

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 [66].

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 [47]. 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 [46].

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 The third-generation of DNA sequencing) have become increasingly used in microbiology, although their accuracy and low throughput make it difficult to implement in a clinical diagnostic setting.

1.2.2.2 Sequencing and genomic surveillance

Most notably, WGS has become a common tool in infection surveillance and prevention, allowing identification and tracking of pathogens, establishing transmission routes and outbreak control [67]. In bacterial infections, initiatives such as Pathogenwatch[10] offers a web-based platform for AMR analysis and phylogeny generation of *Campylobacter*, *Klebsiella*, *Neisseria gonorrhoeae*, *Staphylococcus aureus*, and *Salmonella Typhi* [68]. 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 [69]. Enterobase[13] allows for the analysis and visualisation of genomic variation within enteric bacteria [70]. 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 [71]. 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 [72]. 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 [73]. 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

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

subtype/lineage, detection of putative mixed infections and intra-host minor variants [74]. Nextrain[18] provide a continually-updated view of publicly available data alongside powerful analytic and visualisation tools to 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 [75].

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 [76, 77]. 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 [78, 79]. By the time the pandemic was declared, on 11 of June of 2009, [78] reported the origins and evolutionary genomics of the pandemic influenza A variant with a collection of 813 complete influenza genome sets, 17 of which belonging to the newly swine influenza viruses (GenBank accessions numbers GQ229259–GQ229378). The Middle East Respiratory Syndrome (MERS) pandemic, declared as such in 2015 [3], had its first publicly available sequence on 5 of July 2015 (GenBank accession number KT006149)[80], 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) [81, 82].

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) [83]. 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 [84]. 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 [76, 77, 85].

[18] https://nextstrain.org/

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

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

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 [86]. In most cases, this knowledge is enough to request the most appropriate test, such as syndromic panels or specific culture media, but this is not always the case. In recent years, there has been a growing interest in using metagenomics to deliver culture-independent approaches to microbial ecology, surveillance and diagnosis (see Figure 1.6)[46, 87]. 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 [88]. 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 [89] (see 1.3). 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 [90]. Whether or not it can entirely replace routine microbiology depends on several conditions and future developments, both technological and computational.

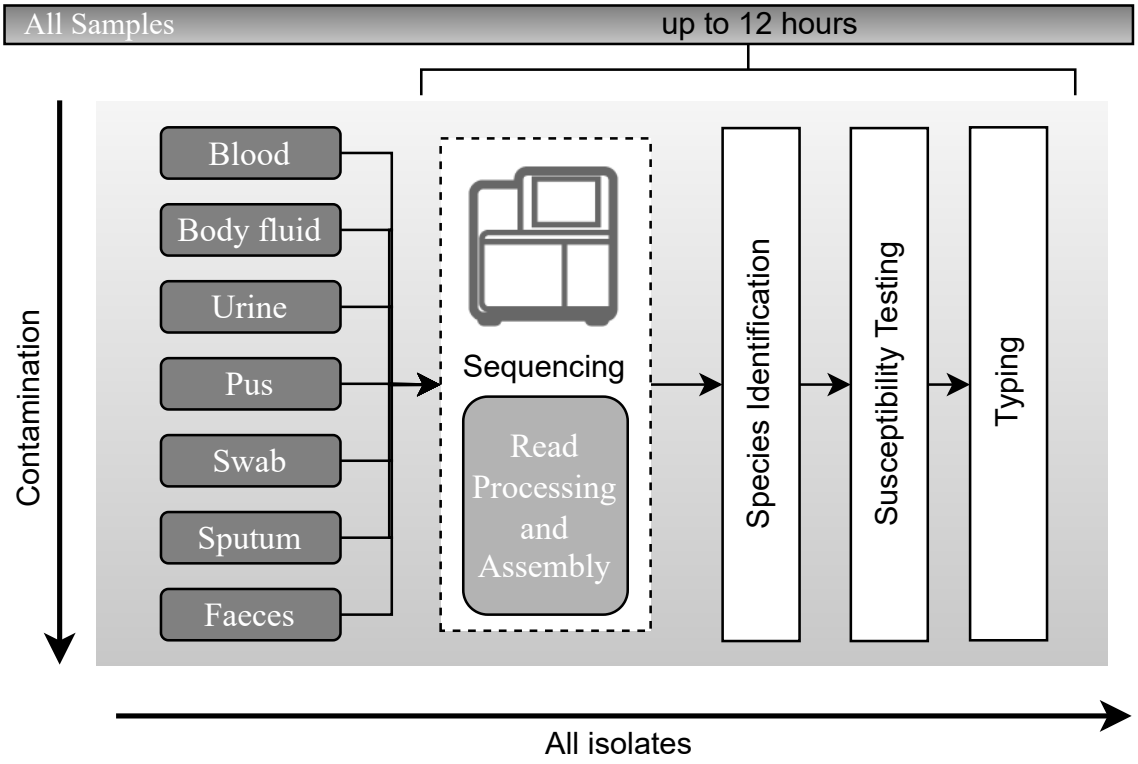


Figure 1.6: **Hypothetical workflow based on metagenomic sequencing.** Schematic representation of the hypothetical workflow for the direct processing of samples from suspected pathogen sources after adoption of metagenomic sequencing, with an expected timescale that could be accommodated in a single day. Adapted from [7].

Albeit lacking consensus in the field, metagenomics can be classified into two variants as proposed by [91]: (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 ribossomal RNA (rRNA) gene sequencing, an 1500 base pair gene that encodes 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%), [92, 93]. 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] [94–96]. The sequence of an unknown strain can be compared with the sequences in these databases, after very closely related sequences are grouped into Operational Taxonomic Units (OTUs), an operational definition used to classify groups of closely related individuals. This allows the deduction of probable taxonomy, with the assumption that sequences of $>$95% identity represent the same genus, whereas sequences of $>$97% identity represent the same species [97].

Furthermore, it must necessarily account for intragenomic variation between 16S gene copies. Furthermore, targeting 16S variable regions with short-read sequencing platforms cannot achieve the taxonomic resolution afforded by sequencing the entire gene and is limited by the database chosen [98]. 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

[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/

rate, signatures of host adaptation, identification and monitoring of diagnostic targets and responses to vaccines and treatments [99]. 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 [100, 101].

1.2.3.2 Shotgun Metagenomics

SMg 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 [90], 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 [66, 102].

Clinical applications of SMg 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 sequencing of all genomic content, DNA and/or RNA, in a clinical sample a viable possibility for diagnostics (see subsubsection 1.2.1.2. The second-generation of DNA sequencing The second-generation of DNA sequencing) [90, 103, 104]. The first reported case that demonstrated the utility of SMg 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 [105], prompting appropriate targeted antibiotic treatment and eventual recovery of the patient. In this case, traditional methods, including an invasive brain biopsy, failed to provide answers, until the shotgun sequencing of cerebrospinal fluid identified 475 of 3,063,784 sequence reads (0.016%) corresponding to leptospira, for which clinical assays were negative due to its very low abundance. Ever since many other reports of successful application of SMg in clinical metagenomics have been reported. but all in edge cases where traditional diagnostic methods have failed or as proof-of-concept [102, 106–108].

In public health microbiology, SMg 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 [109]. 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

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

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 [110]. 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 [111].

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 [90, 112]. 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 [113–115]. 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 the quality and stability of the SMg assay over time [90, 116]. 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 SMg analyses.

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

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 is no standard in how to deal with the increasing amount of data produced in a fit-for-purpose manner [117].

Bioinformatics is an interdisciplinary research field that applies methodologies from computer science, applied mathematics and statistics to the study of biological phenomena[117]. 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 [102, 118–120].

1.3.1 From molecules to reads

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 The second-generation of DNA sequencing), one or both ends of the fragment can be sequenced. If a fragment is sequenced from one end, we refer to it as single-end sequencing. If a fragment is sequenced on both ends, spanning the entire fragment, it is called paired-end sequencing.

1.3.1.1 The FASTQ file

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 [121]. Originally developed at the Wellcome Trust Sanger Institute, the FASTQ has emerged as a common file format for sharing sequencing read data (see 1.4). The FASTQ can be considered as an extension of the 'FASTA sequence file format', originally invented by [122], 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 [123] and is defined by the following equation:

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

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

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

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

Since their introduction, PHRED scores have become the *de facto* standard for representing sequencing read base qualities [121]. Despite this convention, the encoding of the Phread score can vary when translated to its ASCII representation in the FASTQ file format. For example, Sanger FASTQ files use ASCII 33–126 to encode PHRED qualities from 0 to 93 (that is, PHRED scores with an ASCII offset of 33). A full list of encoding is available in Figure 1.7.

1.3.1.2 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

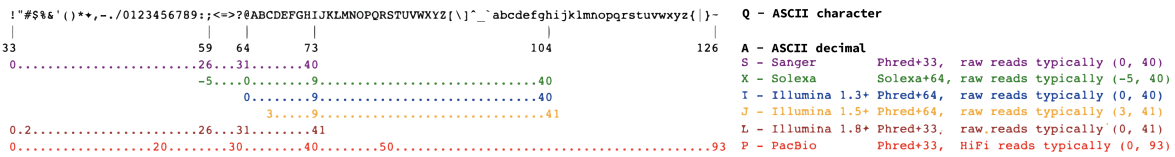


Figure 1.7: **Range of FASTQ quality scores and their corresponding ASCII encoding.** For raw reads, the range of scores will depend on the technology and the base caller used. Starting in Illumina 1.8, the quality scores have returned to the use of the Sanger format (PHRED+33). For processed reads and long accurate reads, scores may be even higher with, for example, quality values of up to 93 observed in reads from PacBio HiFi reads.

ground truth against which success evaluations 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 have been developed in recent years for the simulation of sequencing data, particularly for second and third-generation sequencing technologies, which could be used to compare existing and new bioinformatic analytical pipelines. [124] 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 plethora of tools are available for first, second and third-generation reads (see Figure 1.8).

1.3.1.3 FASTQ quality assessment and quality control

Quality assessment and control is a basal step to any analysis, and aims to (1) remove and/or filter low quality and low complexity reads, (2) trim adapters, and (3) remove host sequences from the samples' raw data. There are many tools available but the most commonly used are FastQC[28] (Babraham Bioinformatics) for quality control, followed by Trimmomatic [125], Cutadapt [126] or fastp [127] to trim and/or filter adaptors, low quality and low complexity sequences. For long-read sequencing, tools like NanoPlot and NanoStats [128], and Filtlong[29] can perform the equivalent quality assessment and control, adapter trimming and low quality trimming, respectively.

1.3.1.4 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

[28] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[29] https://github.com/rrwick/Filtlong/

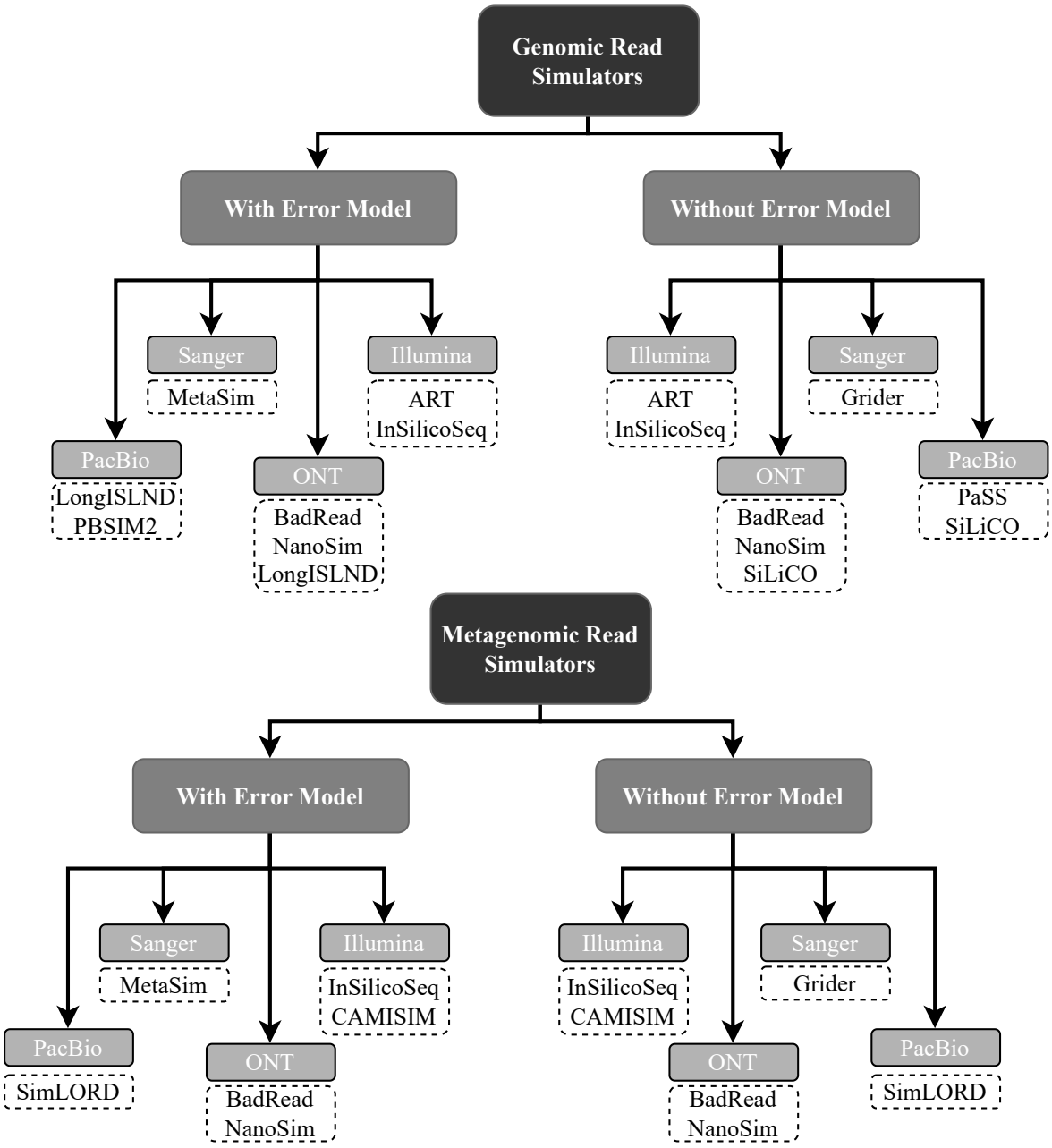


Figure 1.8: **Sequence simulators for genomic and metagenomic data.** For first generation sequencing, Metasim (`https://github.com/gwcbi/metagenomics_simulation`) and Grider (`https://sourceforge.net/projects/biogrinder/`) can generate mock genomic and metagenomic data, with and without error models, 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 [124].

metagenomics, and shotgun metagenomics.

From pure culture, taxonomic identification of the read content of a sample is useful to assess contamination. Tools like Kraken2 [129, 130] and Braken [131]. 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 [132], Kaiju, [133], and MetaPhlAn2 [134] 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. Long-read data from 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 mentioned tools accommodate the classification of single-end files.

1.3.2 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, where FASTQ files, containing the sequencing information, are converted into FASTA files with informative genomic sequences. Information can be inferred through genome annotation software, such as Prokka[31], identifying and labelling all relevant features in a genome sequence, such as predicted coding regions and their putative products, noncoding RNAs, signal peptides, and so on [135].

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 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 [117]. 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 [136].

[30] `https://benlangmead.github.io/aws-indexes/k2`
[31] `https://github.com/tseemann/prokka`

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 [137]. Two methods are used to obtain draft genomes: (1) through reference-guided sequence assembly, or (2), through *de novo* sequence assembly.

1.3.2.1 The FASTA file

In bioinformatics, the FASTA format is a text-based format to represent nucleotide or amino acid sequences using single-letter codes, preceded by a sequence name or any other information relative to the sequence. Similarly to FASTQ (see **??**), it was developed by the Wellcome Trust Sanger Institute, the FASTQ has emerged as a common file format for sharing sequence data [122]. The FASTA file follows the following conformation:

- The **first line** of a FASTA file starts with a ">" (greater-than) symbol, signifying the comment portion;
- The **subsequent lines** containing the actual sequence itself represented in the standard IUB/IUPAC amino acid and nucleic acid codes [138], usually 80 characters in length.

A multiple sequence FASTA format can be obtained by concatenating several single sequence FASTA files in a common file (also known as multi-FASTA format). The extension of the file indicates the type of sequence (nucleotide or amino acid) present (see Table 1.2). For genomic data, the ".fasta", ".fa", ".fna", and ".ffn" are the most used, with the first two being generic and the last two specific for nucleic acid and coding regions of a genome.

Table 1.2: The standard filename extension for a text file containing FASTA formatted sequences.

Extension	Sequence	Definition
fasta, fa	generic FASTA	Any generic fasta file. See below for other common FASTA file extensions
fna	FASTA nucleic acid	Used generically to specify nucleic acids.
ffn	FASTA nucleotide of gene regions	Contains coding regions for a genome.
faa	FASTA amino acid	Contains amino acid sequences. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA

1.3.2.2 Genomes through reference-guided sequence assembly

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

of the same or related species is already available, in which a significant proportion of the reads can be mapped, increase greatly. This process usually includes the following steps: (1) the reference genome has to be indexed, allowing compression of the input text while still permitting fast sub-string queries, (2) for each short-read several sub sequences (seeds) are taken and searched to find their exact matches in the reference (candidate regions), (3) each short-read is then aligned to all corresponding candidate regions, and (4) the consensus sequence is computed in which the reference sequence is corrected when there is enough evidence of an difference based on the mapped reads, identifying the differences between it and the newly generated consensus sequence [141]. In addition to 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 [117]. The breadth of coverage is defined as the ratio of covered sequence on the reference by aligned reads.

1.3.2.3 Genomes through *de novo* sequence assembly

The *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 1.9). The *de novo* assembly methods provide longer sequences that are more informative than shorter sequencing data and can provide a more complete picture of the microbial community in a given sample.

1.3.2.3.1 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 The second-generation of DNA sequencing), but individual reads are longer (500 to 1000 bp). Assembly of Sanger data usually uses Overlap-Layout Consensus (OLC) approaches, in which:

- Overlaps are computed by comparing all reads to all other reads;
- Overlaps are grouped together to form contigs;

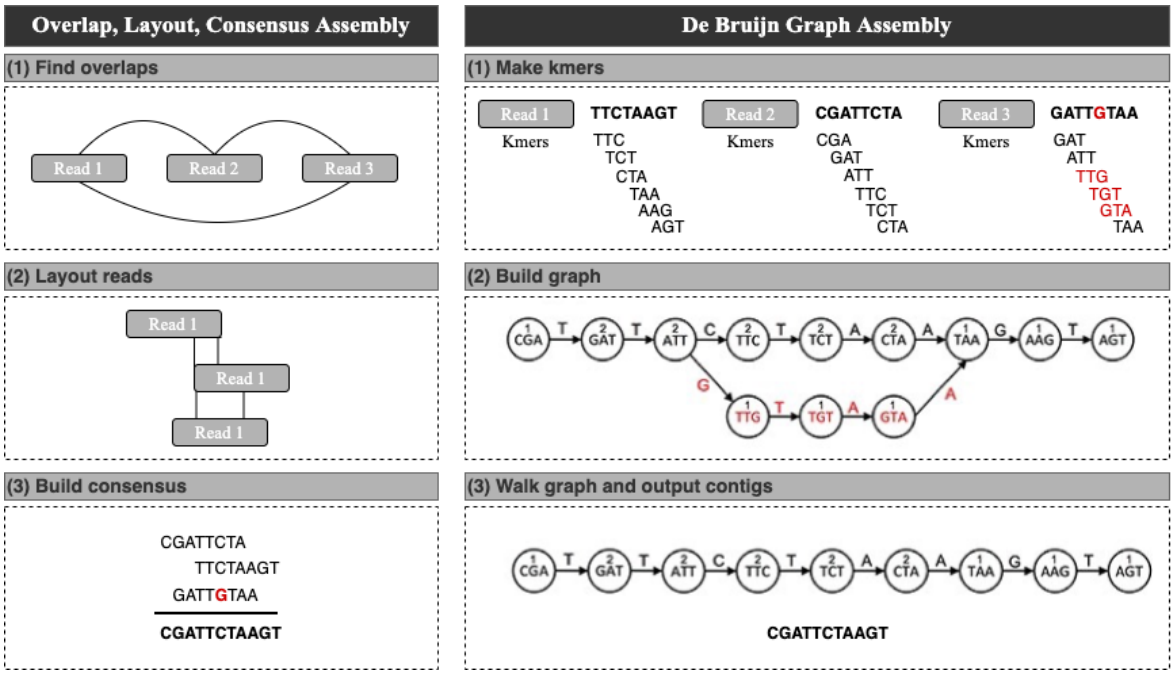


Figure 1.9: **Approaches to *de novo* genome assemble.** In Overlap, Layout, Consensus assembly, (1) overlaps are found between reads and an overlap graph constructed (edges indicate overlapping reads). (2) Reads are laid out into contigs based on the overlaps (lines indicate overlapping portions). (3) The most likely sequence is chosen to construct consensus sequence. In the 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 the numbers above kmers. (3) Contigs are built by walking the graph from the edge nodes. A variety of heuristics handle branches in the graphs—for example, low coverage paths, as shown here, may be ignored. Adapted from [137]

.

- A consensus contiguous sequence, or contig, is determined by picking the most likely nucleotides from the overlapping reads.

These types 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 [142–145]. With the emergence of third-generation sequencing (see subsubsection 1.2.1.3. The third-generation of DNA sequencing 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 [136, 146, 147].

[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`

1.3.2.3.2 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.

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

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

1.3.2.4 Assembly quality assessment and quality control

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

Assessment and evaluation of genome assemblies has been a relevant field ever since the emergence of the assembly process itself. The most widely adopted is QUAST[44], can

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

evaluate assemblies both with a reference genome, as well as without a reference, producing many reports, summary tables and plots to help compare and assess assembly success [154], but alternatives, such as GenomeQC[45] exist [155].

1.3.3 Reproducibility, replicability and transparency

Computational algorithms have become an essential component of microbiome research, with great efforts by the scientific community to raise standards on the development and distribution of code. A lack of reproducibility in computational biology research can be attributed to many factors such as an incomplete or erroneous descriptions of the software used, incomplete documentation on how to run an analysis, or failing to make available the relevant computer code needed [156]. As early as 1990, movements for reproducible research, with special focus on computation-intensive scientific work, have arisen, brought on by the growing use of computational workflows for analysing data across a range of disciplines [157]

Despite the presented efforts, the effectiveness in computational reproducibility is still questionable. Stodden et al [158] reported that, in 22 randomly selected publications who's results relied on the use of computational and data-enabled methods and deemed to be reproducible (i.e. provided data and/or code), only 14% were straightforward to reproduce with minimal effort. Similar results have been observed in comparable studies [159–161]

Several steps can be implemented to ensure the transparency and reproducibility of the chosen bioinformatic workflow. Despite these efforts, sustainability and reproducibility are still major issues. In the field of microbial bioinformatics are not yet widely adopted. The FAIR Principles, standing for Findability, Accessibility, Interoperability, and Reusability, put specific emphasis on enhancing the ability to find and reuse not only data but also the algorithms, tools, and workflows that led to that data [162]. The FAIR guiding principles can be summarised as follows:

- **To be Findable:**
  - F1. (meta)data are assigned a globally unique and persistent identifier
  - F2. data are described with rich metadata (defined by R1 below)
  - F3. metadata clearly and explicitly include the identifier of the data it describes
  - F4. (meta)data are registered or indexed in a searchable resource
- **To be Accessible:**
  - A1. (meta)data are retrievable by their identifier using a standardised communications protocol

[45] `https://github.com/HuffordLab/GenomeQC`

        - A1.1 the protocol is open, free, and universally implementable
        - A1.2 the protocol allows for an authentication and authorisation procedure, where necessary
    - A2. metadata are accessible, even when the data are no longer available

- **To be Interoperable:**
    - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
    - I2. (meta)data use vocabularies that follow FAIR principles
    - I3. (meta)data include qualified references to other (meta)data

- **To be Reusable:**
    - R1. meta(data) are richly described with a plurality of accurate and relevant attributes
    - R1.1. (meta)data are released with a clear and accessible data usage license
    - R1.2. (meta)data are associated with detailed provenance
    - R1.3. (meta)data meet domain-relevant community standards

Several steps have been recommended by experts to ensure the FAIR'ness of both software and data [163–166]. Favouring open-source tools, with clear documentation describing the methodology implemented and stating the version of the software used and which parameters were used, enables the comparison of results. This can be simplified by containerising all software tools with one of the many available solutions, such as Docker[46] or Singularity[47] [167]. The use of workflow managers, like nextflow[48], snakemake[49] or the Galaxy Project[50], will push reproducibility to the next level by taking advantage of the containerisation and scalability, enabling the workflow to be executed with the same parameters in the same conditions in a multitude of different environments [168–170].

When developing software, in the field of microbial bioinformatics, good software engineering practices are not yet widely adopted. An example of such is the widespread use of **VSC!** (**VSC!**), which have long been used to maintain code repositories in the software industry, are now finding new applications in science [163]. Git[51], a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency, provides a powerful way to track and compare versions, retrace errors, explore new approaches in a structured manner, while maintaining a full audit

[46] `https://www.docker.com/`
[47] `https://sylabs.io/`
[48] `https://www.nextflow.io/`
[49] `https://snakemake.github.io/`
[50] `https://galaxyproject.org/`
[51] `https://git-scm.com/`

trail. Remote **VSC!** hosting services, such as GitHub[52], allows for this functionality to be expanded by placing the software in a central location so that it can be accessed by multiple developers and users, facilitating collaboration and auditability. Another example is the use of continued validation through software testing. Modern software engineering advocates reliable software testing standards and best practices. Different approaches are employed: from unit testing to system testing, going from testing every individual component to testing a tool as a whole, verifying and demonstrating that that the published code and data are working properly [171].

1.4 Bioinformatic Analysis for Metagenomics

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

1.4.0.1 Metataxonomics

Metataxonomics (see subsubsection 1.2.3.1. Metataxonomics and Targeted Metagenomics) is the most widely used technique for microbial diversity analysis [172], 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)[53] [173] 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 [174] and UPARSE [175] are also a popular alternative although resulting outputs differing significantly between pipelines despite using the same inputs having been reported by [176], 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

[52] `https://github.com/`
[53] `https://qiime2.org/`

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 [177]. Alternatively, the same approach of direct taxonomic classification, without OTUs clustering, can be followed as with genomic and shotgun metagenomic data, given that the databases include rRNA sequences.

OTUs 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 [178]. 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 [177].

1.4.0.2 Shotgun metagenomics

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

A major additional difficulty of shotgun metagenomic data is the overpowering quantities of host DNA that are often sequenced, making the microbial community sometimes close to undetectable [102]. The presence of contaminants, from the bench process to the pre-existing biota, and the cost associated with this methodology, are also major hinders in its applicability in the clinic. They account for major caveats and must be made aware of when analysing the data.

The basic strategies for analysing shotgun metagenomic data can be simplified in the scheme in Figure 1.10. One of the biggest challenges when doing metagenomic analysis is differentiating between colonisation and infection by successfully discriminating between a potential pathogen and background microbiota. In the latter, when analysing samples from presumably sterile sites, like cerebrospinal fluid and blood, it is safe to assume that all organisms found are of interest. In locations with a microbiota, the inclusion of negative controls is essential for the correct identification of contaminants in the taxonomic results, whether

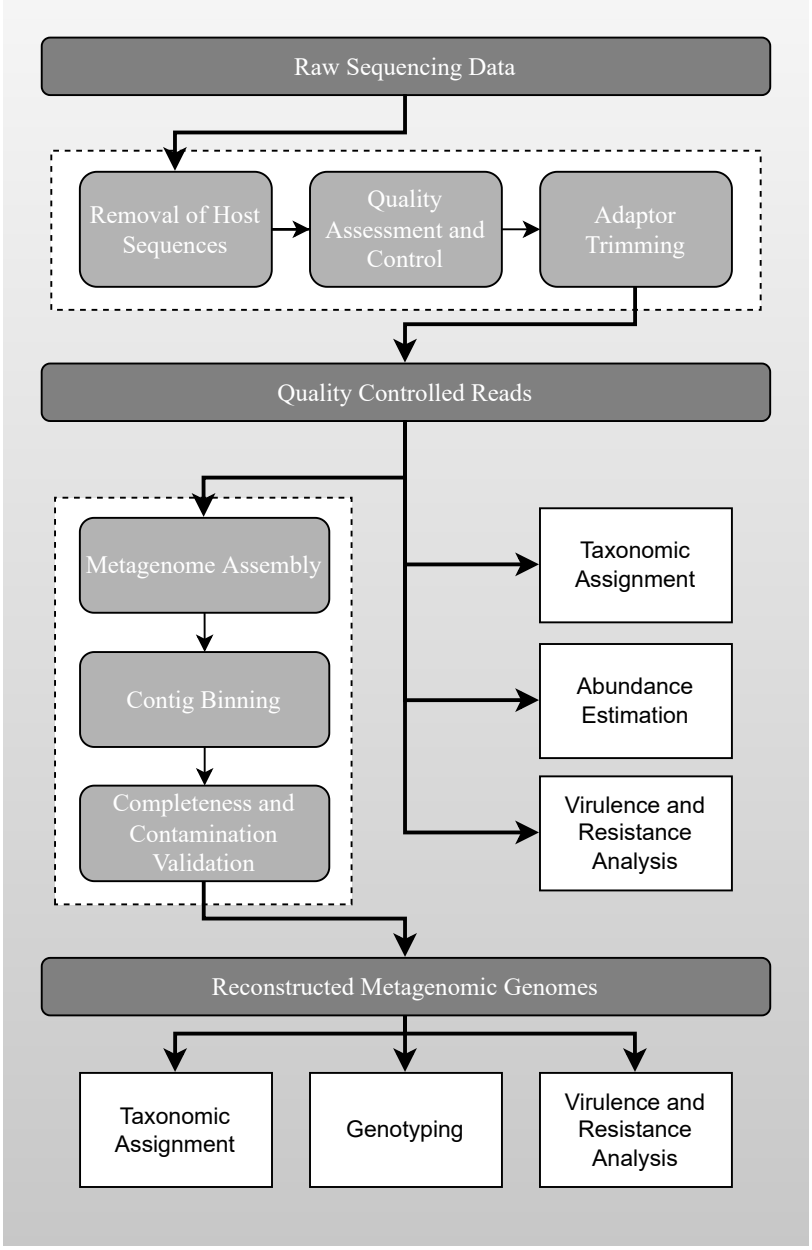


Figure 1.10: Typical bioinformatic analysis procedure for metagenomic data

originated from the sample collection, handling or sequencing process. The use of spiked metagenomic samples as positive control might guide the detection of the possible pathogens by comparing relative abundance between the samples. These controls should be processed similarly to the samples and the taxonomic results should be filtered out from the final report.

As explored in subsection 1.3.2. From reads to genomes, longer sequences are more informative than shorter sequencing data, as the one obtained from second-generation sequencing (see subsubsection 1.2.1.2. The second-generation of DNA sequencing The second-generation of DNA sequencing), and can provide a more complete picture of the microbial community in a given samples. Several dedicated metagenomic assembly tools are available,

such as metaSPAdes[54] and MegaHIT[55] [150, 179]. These tools, in comparison to single-cell data assemblers, are better at dealing with the combination of intra and intergenomic repeats and uneven sequencing coverage [180]. For third-generation sequencing, dedicated metagenomic assemblers have recently emerged, such as meta-flye[56] which expands on the original flye assembler by overcoming a k-mer selection limitation on low abundance species [181]. Nevertheless, the use of non dedicated assemblers for metagenomics may come with the cost of wrongly interpret variation as error, especially in samples that contained closely related species and the construction of chimeric sequences as traditional assemblers follow the basic principle that the coverage in a sample is constant [182].

The assembly-based approach requires the grouping of the different contigs into bins, ideally each collecting the sequences that belong to a microorganism present in the sample. The binning process can be taxonomy dependent, relying on a database to aggregate the sequences, or independent. The independent approach has the benefit of not relying on a database, but instead it uses the composition of each sequence and coverage profiles to cluster together sequences that might belong to the same organism. These algorithms don't require prior knowledge about the genomes in a given sample, instead relying on features inherent to the sequences in the sample. Although most binning software can work with single metagenomic samples, most make use of differential coverage of multiple samples to improve the binning process [183]. It allows the handling of complex ecosystems and might be crucial when analysing samples recovered from sites with a complex microbiota. A comparison of five taxonomic independent and four taxonomic binning software by [120] revealed that, for taxonomic independent approaches, MaxBin2[57] had the highest completeness and purity in the bins obtained [184]. For taxonomic binning, working similarly to the direct taxonomic assignment of the sequencing data, PhyloPythiaS+[58] obtained better results in accuracy, completeness and purity, followed by Kraken[59] 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 [129, 185].

1.5 Aims of the Thesis

Shotgun metagenomic approaches, defined by the sequencing of random DNA fragments of microbial organisms directly from the biological sample, is a promising methodology to obtain very fast results for the identification of pathogens and their virulence and resistance properties directly from samples, without the need for culture. Standardisation of the method and validation of the statistical metrics used to analyse and report the data are of major

[54] https://github.com/ablab/spades/
[55] https://github.com/voutcn/megahit/
[56] https://github.com/fenderglass/Flye/
[57] https://sourceforge.net/projects/maxbin2/
[58] https://github.com/algbioi/ppsp
[59] https://github.com/DerrickWood/kraken2/

importance to get this approach to be accredited and used in clinical settings.

The main objective of this work is to evaluate the use of bioinformatics methods for the analysis of metagenomic data to allow the rapid identification, virulence analysis and antimicrobial susceptibility prediction of pathogens with clinical relevance. The main goals are:

- Evaluate the current impact and applicability of metagenomics genomics in medical microbiology, both in a clinical and in surveillance and infection prevention settings;
- Develop novel methods and metrics to accurately identify and estimate relative abundance of pathogens of interest through a hybrid approach of read mapping and de novo assembly methods;
- Standardise the process of metagenomic analysis, allowing the comparison of results obtained across domains and stakeholders
- Develop computationally efficient and robust frameworks that allows scientists and/or medical experts with limited programming experience to rapidly and easily query the abundance of specific taxa and genes across the samples of interest, obtaining simple and intuitive reports.

As proof-of-concept, greater focus was given to clinically relevant taxa, such as Dengue virus. All methodologies and tools developed were tested and validated on both real and simulated data.

1.6 References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[32] World Health Organization. *International Health Regulations (2005)*. en. Second edition. The Ukrainian version is published by the Center for Implementation of International Health Regulations, Ukraine. Geneva: World Health Organization, 2005.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[57] Alice Maria Giani et al. "Long walk to genomics: History and current approaches to genome sequencing and assembly". en. In: *Computational and Structural Biotechnology Journal* 18 (Jan. 2020), pp. 9–19. ISSN: 2001-0370. DOI: 10.1016/j.csbj.

2019.11.002. URL: https://www.sciencedirect.com/science/article/pii/S2001037019303277 (visited on 02/08/2022).

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

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

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

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

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

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

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

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

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

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

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

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

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

[71] Silvia Argimón et al. “Microreact: visualizing and sharing data for genomic epidemiology and phylogeography”. In: *Microbial Genomics* 2.11 (). Publisher: Microbiology Society, e000093. ISSN: 2057-5858, DOI: 10.1099/mgen.0.000093. URL: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000093 (visited on 02/18/2022).

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

[73] Brett E. Pickett et al. “Virus Pathogen Database and Analysis Resource (ViPR): A Comprehensive Bioinformatics Database and Analysis Resource for the Coronavirus Research Community”. en. In: *Viruses* 4.11 (Nov. 2012). Number: 11 Publisher:

Molecular Diversity Preservation International, pp. 3209–3226. ISSN: 1999-4915. DOI: 10.3390/v4113209. URL: https://www.mdpi.com/1999-4915/4/11/3209 (visited on 02/18/2022).

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

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

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

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

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

[79] Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. "Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans". In: *New England Journal of Medicine* 360.25 (June 2009). Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa0903810, pp. 2605–2615. ISSN: 0028-4793. DOI: 10.1056/NEJMoa0903810. URL: https://doi.org/10.1056/NEJMoa0903810 (visited on 02/23/2022).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[112] Heather A. Carleton et al. “Metagenomic Approaches for Public Health Surveillance of Foodborne Infections: Opportunities and Challenges”. In: *Foodborne Pathogens and Disease* 16.7 (July 2019). Publisher: Mary Ann Liebert, Inc., publishers,

pp. 474–479. ISSN: 1535-3141. DOI: 10.1089/fpd.2019.2636. URL: https://www.liebertpub.com/doi/10.1089/fpd.2019.2636 (visited on 02/28/2022).

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

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

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

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

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

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

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

[120] Alexander Sczyrba et al. "Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software". en. In: *Nature Methods* 14.11 (Nov. 2017). Number: 11 Publisher: Nature Publishing Group, pp. 1063–1071. ISSN: 1548-7105.

DOI: 10.1038/nmeth.4458. URL: https://www.nature.com/articles/nmeth.4458 (visited on 03/20/2022).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[135] Torsten Seemann. "Prokka: rapid prokaryotic genome annotation". eng. In: *Bioinformatics (Oxford, England)* 30.14 (July 2014), pp. 2068–2069. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu153.

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

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

[138] e. "IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969)". en. In: *Biochemistry* 9.18 (Sept. 1970), pp. 3471–3479. ISSN: 0006-2960, 1520-4995. DOI: 10.1021/bi00820a001. URL: https://pubs.acs.org/doi/abs/10.1021/bi00820a001 (visited on 03/28/2022).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[155] Nancy Manchanda et al. “GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations”. en. In: *BMC Genomics* 21.1 (Dec. 2020), p. 193. ISSN: 1471-2164. DOI: 10.1186/s12864-020-6568-2. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6568-2 (visited on 08/11/2021).

[156] Jason A. Papin et al. “Improving reproducibility in computational biology research”. en. In: *PLOS Computational Biology* 16.5 (May 2020). Publisher: Public Library of Science, e1007881. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007881. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007881 (visited on 04/01/2022).

[157] Jon F. Claerbout and Martin Karrenbach. “Electronic documents give reproducible research a new meaning”. en. In: *SEG Technical Program Expanded Abstracts 1992*. Society of Exploration Geophysicists, Jan. 1992, pp. 601–604. DOI: 10.1190/1.1822162. URL: http://library.seg.org/doi/abs/10.1190/1.1822162 (visited on 04/01/2022).

[158] Victoria Stodden, Jennifer Seiler, and Zhaokun Ma. “An empirical analysis of journal policy effectiveness for computational reproducibility”. en. In: *Proceedings of the National Academy of Sciences* 115.11 (Mar. 2018), pp. 2584–2589. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1708290115. URL: https://pnas.org/doi/full/10.1073/pnas.1708290115 (visited on 04/01/2022).

[159] Keith A. Baggerly and Kevin R. Coombes. “Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology”. In: *The Annals of Applied Statistics* 3.4 (Dec. 2009). ISSN: 1932-6157. DOI: 10.1214/09-AOAS291. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-3/issue-4/Deriving-chemosensitivity-from-cell-lines--Forensic-bioinformatics-and-reproducible/10.1214/09-AOAS291.full (visited on 04/01/2022).

[160] Yang-Min Kim, Jean-Baptiste Poline, and Guillaume Dumas. “Experimenting with reproducibility: a case study of robustness in bioinformatics”. In: *GigaScience* 7.7 (July 2018), giy077. ISSN: 2047-217X. DOI: 10.1093/gigascience/giy077. URL: https://doi.org/10.1093/gigascience/giy077 (visited on 04/01/2022).

[161] Yunda Huang and Raphael Gottardo. “Comparability and reproducibility of biomedical data”. eng. In: *Briefings in Bioinformatics* 14.4 (July 2013), pp. 391–401. ISSN: 1477-4054. DOI: 10.1093/bib/bbs078.

[162] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. en. In: *Scientific Data* 3.1 (Mar. 2016). Number: 1 Publisher: Nature Publishing Group, p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: https://www.nature.com/articles/sdata201618 (visited on 04/01/2022).

[163] Karthik Ram. “Git can facilitate greater reproducibility and increased transparency in science”. en. In: *Source Code for Biology and Medicine* 8.1 (Dec. 2013), p. 7. ISSN: 1751-0473. DOI: 10.1186/1751-0473-8-7. URL: https://scfbm.biomedcentral.com/articles/10.1186/1751-0473-8-7 (visited on 04/01/2022).

[164] Stephen R. Piccolo and Michael B. Frampton. “Tools and techniques for computational reproducibility”. en. In: *GigaScience* 5.1 (Dec. 2016), p. 30. ISSN: 2047-217X. DOI: 10.1186/s13742-016-0135-4. URL: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-016-0135-4 (visited on 04/01/2022).

[165] Carl Boettiger. “An introduction to Docker for reproducible research”. en. In: *ACM SIGOPS Operating Systems Review* 49.1 (Jan. 2015), pp. 71–79. ISSN: 0163-5980. DOI: 10.1145/2723872.2723882. URL: https://dl.acm.org/doi/10.1145/2723872.2723882 (visited on 04/01/2022).

[166] Geir Kjetil Sandve et al. “Ten Simple Rules for Reproducible Computational Research”. en. In: *PLOS Computational Biology* 9.10 (Oct. 2013). Publisher: Public Library of Science, e1003285. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003285. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285 (visited on 04/01/2022).

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

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

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

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

[171] Matthew Krafczyk et al. “Scientific Tests and Continuous Integration Strategies to Enhance Reproducibility in the Scientific Software Context”. en. In: *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer*

Systems - P-RECS '19. Phoenix, AZ, USA: ACM Press, 2019, pp. 23–28. ISBN: 978-1-4503-6756-1. DOI: 10.1145/3322790.3330595. URL: http://dl.acm.org/citation.cfm?doid=3322790.3330595 (visited on 04/01/2022).

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

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

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

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

[176] Moira Marizzoni et al. "Comparison of Bioinformatics Pipelines and Operating Systems for the Analyses of 16S rRNA Gene Amplicon Sequences in Human Fecal Samples". In: *Frontiers in Microbiology* 11 (2020). ISSN: 1664-302X. URL: https://www.frontiersin.org/article/10.3389/fmicb.2020.01262 (visited on 03/04/2022).

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

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

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

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

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

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

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

[184] Yu-Wei Wu, Blake A. Simmons, and Steven W. Singer. “MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets”. eng. In: *Bioinformatics (Oxford, England)* 32.4 (Feb. 2016), pp. 605–607. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv638.

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

Chapter 2

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

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

This chapter is a reproduction of the following publication:

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

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

As mentioned in Chapter 1, section 1.2.3.2, SMg approaches have been a growing interest to deliver clinically relevant results without *a priori* knowledge of what to expect from a particular clinical sample or patient. The capacity to detect all potential pathogens in a sample has great potential utility in the diagnosis of infectious disease. However, it is unclear how the variety of available methods impacts the end results.

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

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

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

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

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

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

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

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

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

2.1 Abstract

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

2.2 Introduction

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

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

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

DNA [6].

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

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

2.3 Methods

2.3.1 Sample collection

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

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

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

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

2.3.2 Classic culturing and susceptibility testing

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

2.3.3 DNA extraction, library preparation and sequencing

The DNA for metagenomic sequencing was isolated using the Ultra-Deep Microbiome Prep (Molzym Life Science, Bremen, Germany), Micro-Dx™kit (Molzym Life Science) or QIAamp DNA Microbiome Kit (Qiagen, Hilden, Germany) directly from the clinical samples and a negative control consisting of a mock sample of DNA and RNA free water (Table 2.1). These kits include human DNA depletion steps. The QIAamp DNA Micro-

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

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

2.3.4 Bioinformatics analyses

In order to evaluate and compare the accuracy and reliability of the bioinformatics analyses in providing the closest results to culture and WGS of any cultured isolates, three different pipelines (two commercially and one freely available) were used (Figure 2.1). Different tools to perform raw read quality control, filtering and trimming were used and reads were mapped against the human genome (hg19) before performing taxonomic classification. Reads mapping to hg19 where removed from the analysis to increase the efficiency of the bioinformatics tools. Typing (MLST), phylogenetic analysis, plasmid analysis, detection of antimicrobial resistance and virulence genes was performed. To determine the appropriateness of SMg as predictor of the WGS (chromosome and plasmids), SMg results obtained were compared with the results of WGS of any bacterial isolates obtained from culturing the sample.

All the parameters used in each approach are available in Supplementary Table 1 (see 2.8.1).

2.3.4.1 Unix-based approach

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

Plasmid detection was achieved by running the script PlasmidCoverage‡, using the plasmid sequences downloaded from NCBI RefSeq (`ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plasmid/`, as of May 11, 2017). The script uses Bowtie 2 v2.2.9 [9], to map the pre-processed input reads against the plasmid database (Bowtie2 index for all plasmid sequences). For Bowtie 2 we used the '-k' option, allowing each read to map to as many plasmid sequences as present in the NCBI RefSeq plasmid database (since plasmid sequences are modular) [16, 17]. Then, this pipeline used SAMtools v1.3.1 [10] to estimate the coverage for each position, and reported the length of plasmid sequence covered (in percentage) and average depth (mean number of reads mapped against a given position in each plasmid). Plasmids with less than 80% of its length covered were excluded from the final results in line with what has described elsewhere [18]. The pATLAS tool§ was used to visualise which plasmids were present.

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

*`https://github.com/B-UMMI/INNUca/`
†`https://github.com/B-UMMI/ReMatCh/`
‡`https://github.com/tiagofilipe12/PlasmidCoverage`
§`http://www.patlas.site/`

2.3.4.2 Commercial-based approach

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

2.3.4.3 Web-based approaches

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

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

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

The reads were uploaded into CosmosID[**] and Taxonomer[††] [19] directly without any quality trimming. We used the Full Analysis mode in Taxonomer.

2.3.4.4 wgMLST analyses

Typing was done by MLST and wgMLST analyses obtained using Ridom SeqSphere+ v4.0.1. The genomic data (assembled contigs) obtained from SMg was compared to the data obtained through WGS. Since no cg/wg MLST scheme was available for *Escherichia coli*, *Enterococcus faecalis*, *Ochrobactrum intermedium* and *Staphylococcus haemolyticus*, cgMLST and accessory genome schemes were constructed, using Ridom SeqSphere+ cgMLST Target Definer with the following parameters: a minimum length filter that removes all genes smaller than 50 bp; a start codon filter that discards all genes that contain no start codon at the beginning of the gene; a stop codon filter that discards all genes that contain no stop codon or more than one stop codon or that do not have the stop codon at the end of the gene; a homologous gene filter that discards all genes with fragments that occur in multiple copies within a genome (with identity of 90% and >100 bp overlap); and a gene overlap filter that discards the shorter gene from the cgMLST scheme if the two genes affected overlap $>$4 bp. The remaining genes were then used in a pairwise comparison using BLAST version 2.2.12 (parameters used were word size 11, mismatch penalty -1, match reward 1, gap open costs 5, and gap extension costs 2). All genes of the reference genome that were common in all query genomes with a sequence identity of $\geq$ 90% and 100% overlap and, with the default parameter stop codon percentage filter turned on, formed the final cgMLST scheme. The combination of all alleles in each strain formed an allelic profile that was used to generate minimum spanning trees using the parameter "pairwise ignore missing values" during distance calculation [20].

2.3.4.5 Statistical analysis

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

[**] `https://app.cosmosid.com/login`
[††] `https://www.taxonomer.com/`

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

2.4 Results

2.4.1 Classical identification

Nine body fluid samples and one tissue sample from 9 different patients were sequenced, including one sample from peritoneal fluid, five from pus (3 abscesses and 2 empyemas), two from synovial fluid of knees with prosthesis, one from sputum and one from a bone biopsy (Table 2.1). In total 15 different isolates obtained from the 10 samples were considered of possible clinical significance and were selected for species identification and antimicrobial susceptibility testing during routine work up of the samples (Table 2.2,2.3 and 2.4). In samples 2 and 3, only one colony-forming unit (CFU) of *Escherichia coli* and *Staphylococcus epidermidis*, respectively, was detected after 48 hours of incubation. In samples 2 and 5, the anaerobic cultures were mixed to such an extent, that no further characterization of the colonies was performed, and the results were reported as anaerobic mixed culture.

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

There was fungal growth in 2 samples (1 and 5) that included two Candida species (one *Candida glabrata* and one *Candida albicans*). The different bacterial and fungal species identified in each sample are shown in Tables 2.2, 2.3 and 2.4.

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

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

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

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

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

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

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

[d] miniKraken database was used

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Abbreviations: PPV – positive predictive value

The taxonomic classification results for each sample are presented in Tables 2.2, 2.3 and 2.4. In 8 samples, all the microorganisms identified by classical culture were also identified through metagenomics (using at least one method). In sample 2, two of the bacterial species identified by classical culture, i.e., *E. coli* and one *Enterococcus avium* were not identified through shotgun metagenomics and in sample 3 there was no concordance between the results of MALDI-TOF and the taxonomical classification methods at the species level (Tables 2.2, 2.3 and 2.4). We identified *Ochrobactrum intermedium* in the negative control, but in low amounts (1.0% of the reads mapped to the reference genome with the accession number NZ_ACQA01000002 and only 1.4% of the reference genome was covered). The sensitivity and positive predictive value of each classification method is shown in Table 2.5.

2.4.3 Determination of antimicrobial resistance

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

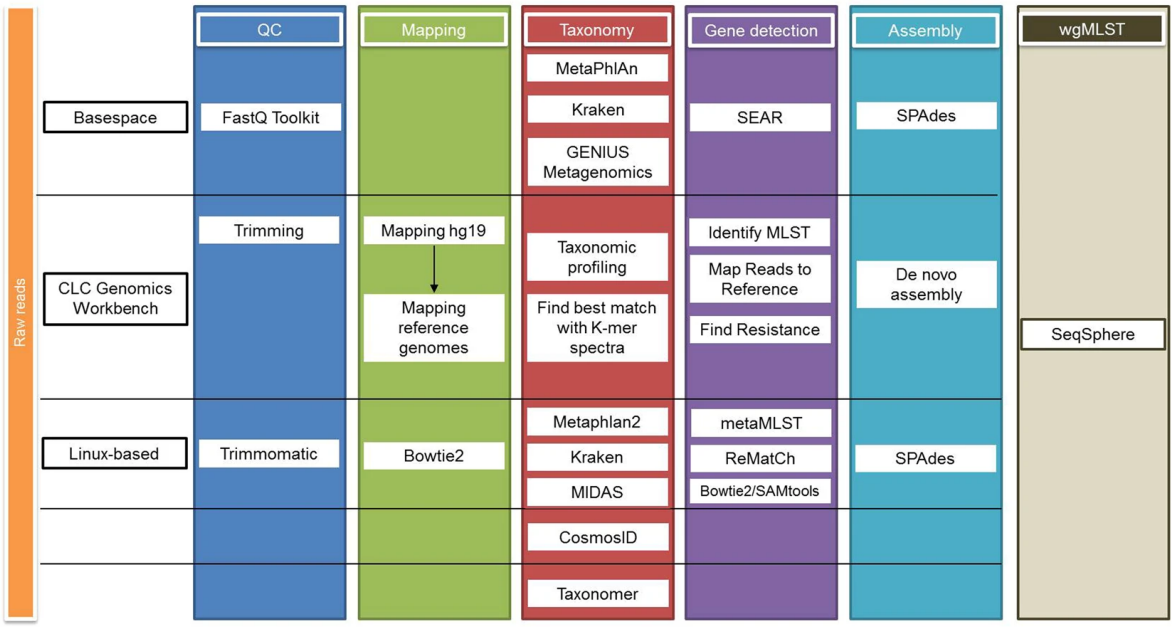


Figure 2.1: Scheme of the bioinformatic analysis of the metagenomics samples.

compared the results with those obtained from WGS and phenotypic resistance testing (Table 2.6).

AMR genes found with CLC Genomics Workbench and ReMatCh in samples 1, 7 and 9 correlated well with phenotypic results. However, in the other 7 samples, not all antimicrobial resistance genes that could explain the phenotypic profile were identified. In addition, in samples 2, 5, 7 and 10, ReMatCh detected different resistance genes compared to those reported by CLC Genomics Workbench (Table 2.6). Some of these differences (genes *norA*, *blaSST*-1, *fusA*) were due to slight differences in the databases used, however, the other resistance genes were present in both databases. Interestingly, in two samples (samples 2 and 5), we were able to identify several antimicrobial resistance genes usually found in anaerobic bacteria. These were not reported by classical microbiology methods, probably because they were not considered relevant pathogens worthy of subsequent susceptibility study (mixed anaerobic culture).

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

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

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

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

[b]Only non-susceptibility is indicated.

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

2.4.4 MLST and wgMLST analysis

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

2.4.5 Characterisation of mobile genetic elements

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

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

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

Abbreviations: ST, sequence type

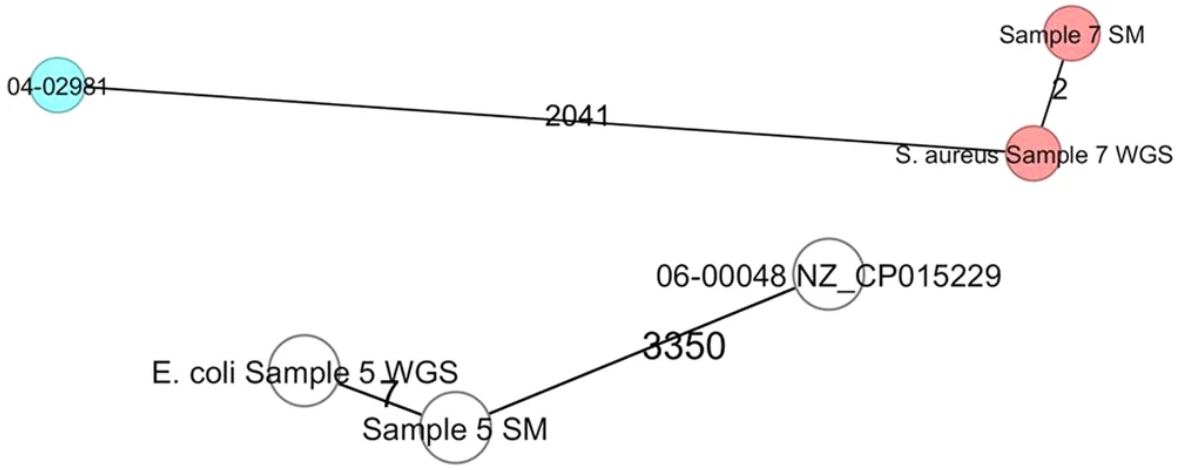


Figure 2.2: Minimum-spanning tree based on wgMLST allelic profiles of 2 S. aureus genomes and 2 E. coli genomes obtained through SM and WGS in comparison to reference strains 04-02981 (GenBank accession number NC_017340) and 06-00048 (NZ_CP015229), respectively. Each circle represents an allelic profile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

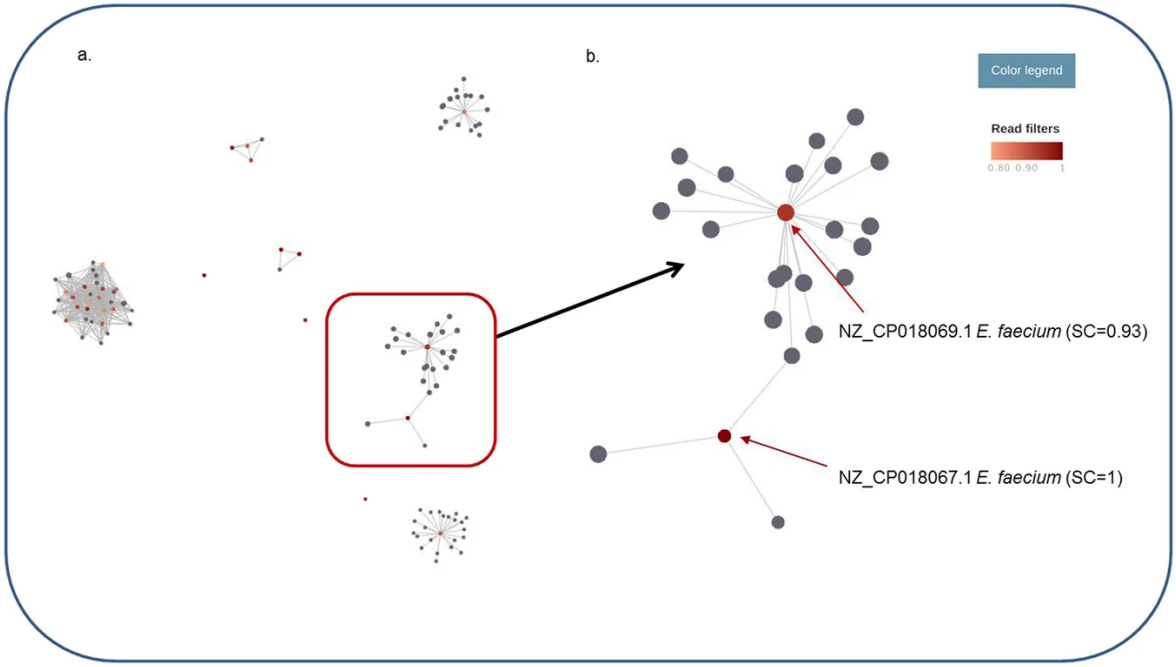


Figure 2.3: (a) Overview of the nodes (representing plasmid sequences) and links between plasmids (connecting similar plasmids) found in Sample 1 (SMg) using the pATLAS tool. (b) A closer look at one of the cloud of plasmids. The colour gradient in each cloud of plasmids represents the plasmid sequence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

2.5 Discussion

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

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

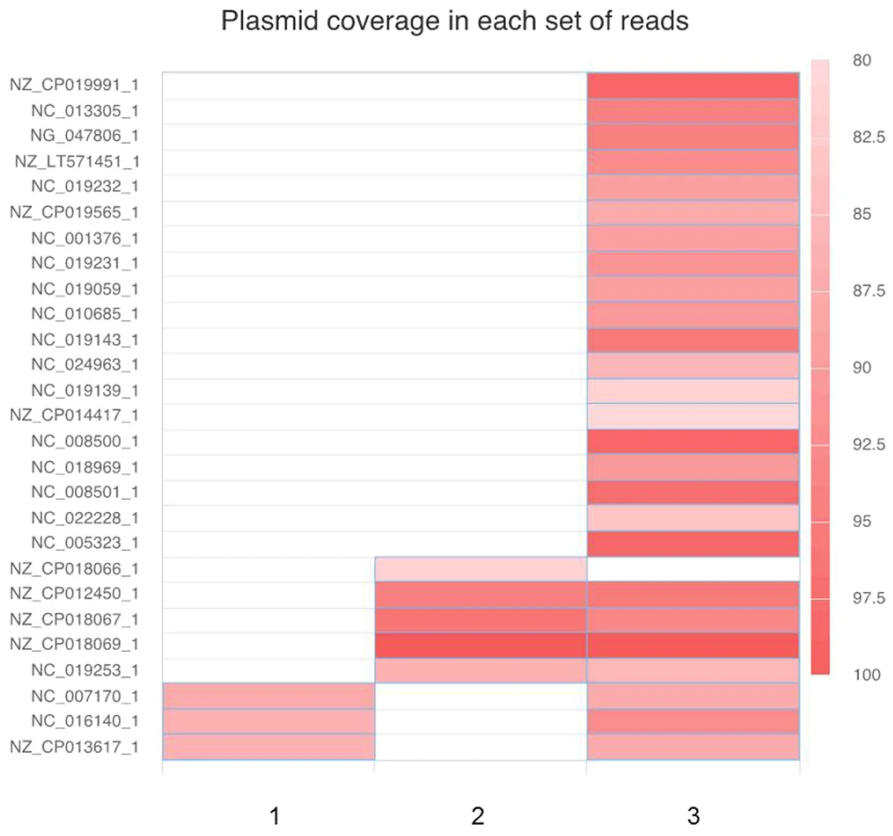


Figure 2.4: A heatmap comparing the identified plasmids using bowtie2 in *S. haemolyticus* WGS (1), *E. faecium* WGS (2) and in the SMg dataset (3) isolated from sample 1.

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

In a clinical setting, a combination of high sensitivity and high PPV of any new method is key. Popular software designed for bacterial identification, can predict dozens to hundreds of species in in vitro generated bacterial communities of known composition [8]. We observed the same when using Kraken and Taxonomer when comparing to culture-based methods. For both Kraken and Taxonomer, relative abundance cut-off values may be required to limit the number of species identified. However, which cut-off values should be used are a matter of debate, since in some cases, even if applying a cut-off value as low as 1.0% (comparable to what was found in the negative control) would have resulted in decreased sensitivity (e.g. the *Streptococcus anginosus* identified by culture in Sample 5 would have been disregarded). The methods that employ several parameters to infer microbial identification are superior, because they not only rely on the relative abundance of bacterial species, but also on the genome coverage and on the proportion of the genome that was covered. On the other hand, in some cases SMg may be more sensitive than culture in identifying pathogens, reflecting the higher sensitivity or the capacity to detect bacterial species which are non-culturable in the conditions used or that are no longer culturable, such as due to prior antimicrobial therapy. In such cases, other methods like 16S rDNA sequencing or the recently described 16S-23S rDNA sequencing method [21] may be used for discrepancy analyses. However,

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

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

In terms of antimicrobial resistance gene detection, ReMatCh (Unix) and the CLC Genomics Workbench Find Resistance tool gave comparable results. Since ReMatCh (Unix) performs the analysis at the read level, while CLC Genomics Workbench performs it at the contig level, we suggest that both strategies should be employed in parallel when looking for antimicrobial resistance genes. It is also important to emphasise that the contig-level approach employed by CLC Genomics Workbench may give negative results if the sequence coverage is set to a high percentage (e.g. above 80%). This is due to the assembly method, which may split the antimicrobial resistance genes into different contigs, when the number of reads is too low. This phenomenon was observed in Sample 1, for the *aac(6')-aph(2")* gene, which was split into 3 different contigs, each part corresponding to less than 40% of the gene. Only when applying a cut-off value of *geq* 20% for sequence coverage could we identify all three parts of the gene, which in total corresponded to 89% of the entire sequence. Finally, it is important to point out that the ResFinder database (used here), and other databases, focus on acquired genes, not including chromosomal point mutations resulting in antimicrobial resistance. However, a recently developed tool, PointFinder, was added to ResFinder for the detection of chromosomal point mutations associated with antimicrobial resistance [22] and an updated database will be available soon.

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

(chromosome or plasmid). The study of mobile genetic elements, including plasmids, carrying antimicrobial resistance genes present in clinical samples is important to predict possible treatment failures and the spread of resistance within and across bacterial species. When performing bacterial isolation followed by WGS, information on polymicrobial infections may be lost. This is mainly driven by a bottleneck in culture, where some bacterial species are not isolated with standard work up protocols (frequently anaerobes and slow-growing organisms). The presence of antimicrobial resistance genes in plasmids of bacteria other than those isolated through culture poses a risk since they are not identified by conventional methods but could potentially be horizontally transmitted to pathogenic bacteria under the antimicrobial selective pressure of treatment. Antimicrobial administration may also select minority populations where these resistance determinants are found. Furthermore, the understanding of how plasmids are shared by different bacteria in a bacterial community (e.g. within an infection site or in the gut) can improve our understanding of how these elements disseminate across species and from patient to patient11. The SMg approach is clearly more efficient than culture in identifying the "cloud" of plasmids present in a given sample (Figure 2.4) and which can be potentially transferred to more pathogenic species generating problems of resistance, as was the case with the emerge of vancomycin resistance *S. aureus* [23].

Whole-genome sequencing has been used extensively for several purposes [24] and is considered to have the potential of playing an important role in clinical microbiology [25]. It is the ongoing goal of medical molecular microbiology to develop faster typing methods that can be used for outbreak surveillance. For this purpose, we assembled the metagenomics data and compared it with the assemblies given by WGS. Surprisingly, the assemblies provided by SPAdes in BaseSpace were closer to the assemblies provided by WGS. When comparing the genomes obtained through WGS and SMg, we could see that in 4 out of 8 bacterial isolates the number of different alleles was leq 7. This showed the potential of SMg to draw phylogenetic relationships from uncultured bacterial genomes, although more potentially limited than those obtained using WGS data from axenic cultures. As for the detection of resistance genes, a key limiting factor may be the number of bacterial reads, reflected in a lower genome coverage (e.g. samples 4 and 6). In these cases, we would have to either improve the human-DNA depletion step, improve the microbial enrichment or perform sequencing at a higher sequencing depth to have enough microbial reads to be able to get a more appropriate genome coverage. Yet, this last step will severely raise the sequencing costs, which might render the methodology unfeasible for routine application.

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

selves, they offered no direct support. Contacting the developers (via email and posting on their forum) does not guarantee a solution to the issues in a time frame compatible with a routine clinical microbiology laboratory work. The dependence and no direct control over a third party to resolve software bugs and provide a stable platform illustrates a disadvantage of a cloud-based system like BaseSpace. Finally, the Unix-based pipeline complemented the data on antimicrobial resistance genes but did not offer better results in terms of microbial identification and MLST typing. However, many more freely available tools for this last purpose could have been used, potentially improving on the results obtained. Reference-guided assembly approaches, taking advantage of the species information derived in the first steps of our analysis pipelines, will deserve further study in the future since these may provide higher quality assemblies from metagenomics data. The main advantage of an open-source approach is its flexibility since it allows the user to choose the most adequate method for each desired outcome. There were several limitations to this study. First, the number of samples included was low and some of the bacterial isolates were not available for further WGS analysis. However, the extended data analyses performed in each sample limited the number of samples to be included. It is our intention to move forward with the most adequate pipelines for each purpose and apply them to additional patients' samples. Second, the samples differed greatly from each other. However, in our point of view, this was beneficial to the study, since it did not bias the analyses as it could have happened if only one type of sample had been used. Finally, we used three different extraction methods that could have influenced the final results. Yet, as can be seen in Table 2.1, the number of human reads differed between samples, even when using the same extraction kit. This suggests none of the kits is clearly superior to the others and that the ratio between host and microbial DNA or other individual sample characteristics will be the major determinants of the proportion of microbial reads.

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

2.6 Acknowledgements

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

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

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

2.7 Author contributions statement

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

2.8 Additional information

2.8.1 Accession codes

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

2.8.2 Competing financial interests

The authors declare that they have no conflict of interest.

2.9 Supplemental Material

Table 2.8: Supplementary table 1.

FastQ Toolkit v2.2.0	
Minimum read length	32
Sub-sampling	FALSE
Adapter trim stringency	0.9
Select respective adapters	TRUE
Quality trimming	FALSE
Poly-A/T Trimming	FALSE
Read Filtering	FALSE
Modify Reads	FALSE
Fix Format	FALSE
FastQC v1.0.0	
Kmer Size	5
Use Conatminant Filter	TRUE
Kraken Metagenomics v1.0.0	
Host Filter	TRUE RefSeqhg19
Classification Database	MiniKraken 20141208 (latest)
Filter Threshold	0
Metaphlan v1.0.0	
Sensitivity options for read-marker similarity (as descibed by BowTie2)	Very Sensitive
SPAdes Genome Assembler v3.9.0	
Running Mode	Error Correction & Assembly
Dataset type	Multi Cell
Careful Mode	Disable
k-mer lengths	Auto
SEAR: Antibiotic Resistance v1.0.0	
Read length cutoff (bases)	70
Read quality score cutoff	20
Read subtraction against E.coli reference genome (K12)?	No
Clustering stringency (express % as a decimal)	0.98
Annotation stringency (% length of reference ARG sequence mapped to by sequencing reads	40
GENIUS Metagenomics: Know Now v1.1.0	
Can't set any settings in BaseSpace	

Table 2.9: Supplementary table 2.

Trim Reads	
Trimmomatic v0.36 (INNUca v2.6 initial module)	
Quality trim	TRUE
Phred Quality limit	05:20
Trim adapter list	Illumina adapters
Remove 5' terminal nucleotides	TRUE
Number of 5' terminal nucleotides	3
Remove 3' terminal nucleotides	TRUE
Number of 3' terminal nucleotides	3
Discard short reads	TRUE
Minimum number of nucleotides in reads	55
Map Reads to Reference	
Bowtie2 v2.3.2	
References	Homo sapiens (hg19) index
Mode	end-to-end
Mode option	sensitive
Collect unmapped reads	FALSE
Taxonomic classification	
Kraken v0.10.5-beta	
References	miniKraken database (Dec. 8, 2014)
K-mer length	35
MIDAS	
References	midas_db_v1.2 (May 9, 2018)
Word size for blast	28
Alignment coverage	0.75
MetaPhlAn2 v2.0	
References	default database (May 9, 2018)
Minimum total nucleotide length for the markers	2000
Quantile value for robust average	0.1
Statistical approach for converting marker abundances into clade abundances	clade global
Analysis type	profiling a metagenome in terms of relative abundance
Identify MLST	
metaMLST v1.1	
References	metamlstDB_2017
Bowtie2 mode	local
Bowtie2 mode option	very sensitive local
Collect unmapped reads	FALSE
Search for and report all alignment	TRUE
Find Resistance Genes	
ReMatCh v3.2	
References	ResFinder database (29-06-2017)
Minimum coverage to consider a position as present	1
Minimum coverage depth to perform a basecall	1
Minimum gene coverage (%)	80
Minimum gene identity (%)	70
De novo assembly	
SPAdes v3.10.1	
Mode	careful
Error correction	FALSE
Read coverage cut-off value	2
List of K-mers	21,33,55,67,77
Plasmid Detection	
Bowtie2 v2.3.2	
References	NCBI RefSeq (May 11, 2017)
Mode	end-to-end
Mode option	sensitive
Collect unmapped reads	FALSE
Multiple alignment	TRUE

Table 2.10: Supplementary table 3.

Illumina	
Discard sequence names	FALSE
Discard quality scores	FALSE
Selected files	
Paired-end reads	TRUE
Read Orientation	Forward Reverse
minimum distance	1
maximum distance	1000
Remove failed reads	TRUE
Quality score	NCBI/Sanger or Illumina Pipeline 1.8 and later
MiSeq de-multiplexing	FALSE
Illumina trim	FALSE
Trim Reads	
Quality trim	TRUE
Quality limit	0.05
Ambiguous trim	TRUE
Ambiguous limit	2
Trim adapter list	Illumina adapters
Use colorspace	FALSE
Remove 5’ terminal nucleotides	FALSE
Number of 5’ terminal nucleotides	1
Remove 3’ terminal nucleotides	FALSE
Number of 3’ terminal nucleotides	1
Discard short reads	TRUE
Minimum number of nucleotides in reads	30
Discard long reads	FALSE
Maximum number of nucleotides in reads	1000
Map Reads to Reference	
References	Homo sapiens (hg19) sequence
Masking mode	No masking
Masking track	
Match score	1
Mismatch cost	2
Cost of insertions and deletions	Linear gap cost
Insertion cost	3
Deletion cost	3
Insertion open cost	6
Insertion extend cost	1
Deletion open cost	6

Deletion extend cost	1
Length fraction	0.5
Similarity fraction	0.8
Global alignment	FALSE
Color space alignment	TRUE
Color error cost	3
Auto-detect paired distances	TRUE
Non-specific match handling	Map randomly
Find Best Matches using K-mer Spectra	
References	NCBI references (2017-07-08)
K-mer length	16
Only index k-mers with prefix	ATGAC
Check for low quality and contamination	TRUE
Fraction of unmapped reads for quality check	0.1
De Novo Assembly	
Mapping mode	Create simple contig sequences (fast)
Update contigs	TRUE
Mismatch cost	2
Insertion cost	3
Deletion cost	3
Colorspace error cost	3
Length fraction	0.5
Similarity fraction	0.8
Colorspace alignment	TRUE
Alignment mode	local
Match mode	random
Create list of un-mapped reads	FALSE
Automatic bubble size	TRUE
Bubble size	50
Automatic word size	TRUE
Word size	20
Minimum contig length	200
Guidance only reads	
Perform scaffolding	TRUE
Auto-detect paired distances	TRUE
Create report	TRUE
Find Resistance	
DB	Database for Find Resistance (2018-02-02)

Minimum identity %	70
Minimum length %	20
Filter overlaps	TRUE
Local Realignment	
Realign unaligned ends	TRUE
Multi-pass realignment	2
Guidance-variant track	
Maximum guidance-variant length	100
Force realignment to guidance-variants	FALSE
InDels and Structural Variants (2)	
P-Value threshold	1.00E-04
Maximum number of mismatches	3
Ignore broken pairs	TRUE
Filter variants	FALSE
Minimum number of reads	2
Minimum relative consensus coverage	0
Minimum quality score	0
Restrict calling to target regions	
Local Realignment (2)	
Realign unaligned ends	TRUE
Multi-pass realignment	2
Guidance-variant track	Defined by: InDels and Structural Variants (2)
Maximum guidance-variant length	100
Force realignment to guidance-variants	FALSE
Identify MLST Scheme from Genomes	
Schemes	PubMLST (04-03-2017)
Identify MLST	
Scheme	Defined by: Identify MLST Scheme from Genomes
Low coverage reported when below	

2.10 References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[17] Maria Pilar Garcillán-Barcia, Andrés Alvarado, and Fernando de la Cruz. "Identification of bacterial plasmids based on mobility and plasmid population biology". eng. In: *FEMS microbiology reviews* 35.5 (Sept. 2011), pp. 936–956. ISSN: 1574-6976. DOI: 10.1111/j.1574-6976.2011.00291.x.

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

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

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

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

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

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

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

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

Chapter 3

Detection of a novel *mcr-5.4* gene variant in hospital tap water by shotgun metagenomic sequencing

3. DETECTION OF A NOVEL *MCR-5.4* GENE VARIANT IN HOSPITAL TAP WATER BY SHOTGUN METAGENOMIC SEQUENCING

This chapter is a reproduction of the following publication:

G. Fleres, N. Couto, L. Schuele, M. A. Chlebowicz, C. I. Mendes, L. W. M. van der Sluis, J. W. A. Rossen, A. W Friedrich, S. García-Cobos, Detection of a novel mcr-5.4 gene variant in hospital tap water by shotgun metagenomic sequencing, Journal of Antimicrobial Chemotherapy, Volume 74, Issue 12, December 2019, Pages 3626–3628, DOI: `https://doi.org/10.1093/jac/dkz363`

As referenced in Chapter 1, section 1.1.2, sequencing has become a common tool in surveillance and infection prevention, when combined with epidemiological data, have undoubtedly provided immeasurable insights regarding identification of potential sources of pathogenicity and transmission pathways. Shotgun metagenomic (SMg) approaches, just like in a clinical setting, have been a growing interest to deliver relevant results without a priori knowledge of what to expect from a particular environmental sample.

In this publication, second (see Chapter 1, section 1.2.1.2) and third (see Chapter 1, section 1.2.1.3) generation sequencing SMg has been applied to eight concentrated water samples collected the University Medical Center Groningen. In one of the samples, the novel detection of an *mcr-5* gene, named *mcr-5.4*, is reported. To the best of our knowledge, this is the first time that this gene, a mobile colistin resistance (*mcr*) determinant, has been recovered from a hospital water environment, with analysis suggesting the order of *Pseudomonadales* as the most probable host.

My contribution to this publication included the bioinformatics analysis of the *mcr-5.4* carrying sample thorough hybrid assembly using metaSPAdes. The resulting assembled contings were binned with the MaxBin2 tool and the bin having the sequence carrying the gene of interest was taxonomically characterised with Kraken2.

Detection of a novel *mcr-5.4* gene variant in hospital tap water by shotgun metagenomic sequencing

Giuseppe Fleres[1], Natacha Couto[1], Leonard Schuele[1], Monika A Chlebowicz[1], Catarina I Mendes[1], Luc W M van der Sluis[2], John W A Rossen[1], Alex W Friedrich[1], Silvia García-Cobos[1]

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

[2] Center of Dentistry and Oral Hygiene, University Medical Center Groningen, 9712 CP Groningen, The Netherlands

3.1 Letter

Sir,

Colistin is considered a last-resort antibiotic for treating serious infections caused by MDR Gram-negative bacteria. The efficacy of this antibiotic is challenged by the emergence and global spread of mobile colistin resistance (*mcr*) determinants, which threaten human, animal and environmental health. The first mobile colistin resistance gene (*mcr-1*) was reported in 2015 and since then up to eight different variants have been described [1]. In 2017, Borowiak et al.[2] described a new transposon-associated phosphoethanolamine transferase mediating colistin resistance, named *mcr-5*, in d-tartrate-fermenting *Salmonella enterica* subsp. enterica serovar Paratyphi B isolated from poultry. The *mcr-5.3* variant has been recently reported in *Stenotrophomonas* spp. from sewage water [3]. Here we report for the first time (to the best of our knowledge) the detection of an *mcr-5* gene in a hospital water environment using short-read metagenomic sequencing (SRMseq) and subsequent characterization using long-read metagenomic sequencing (LRMseq) to reveal its genetic environment.

In June 2017, eight tap-water samples (900 mL) were collected at the University Medical Center Groningen. Water samples were filtered (0.2 lm) and after DNA extraction (PowerWater DNA Extraction Kit, QIAGEN), SRMseq was performed on a MiSeq instrument (500 cycles) (Illumina). Antibiotic resistance genes were identified in the metagenome assemblies (CLC Genomics Workbench v10.1.1, QIAGEN) using ABRicate-0.7 (https://github. com/tseemann/abricate) and applying the following thresholds: .70% identity and .80% coverage. One sample contained an *mcr*-type gene (5% sequencing depth), with the nucleotide change 313C.T (amino acid change F105L) with respect to the original *mcr-5.1* gene, which was designated *mcr-5.4* by NCBI (accession no. MK965519). This sample was selected for LRMseq; the DNA libraries were prepared

using the Rapid PCR Barcoding Kit (SQK-RPB004) from Oxford Nanopore Technologies (ONT) and loaded into a FLO-MIN106 R9.4 flow cell. The run was performed on a MinION device (ONT) and it proceeded for 24 h. The data were basecalled using Albacore (https://github.com/rrwick/Basecallingcomparison) and further processed with Poretools [4] and Porechop (https://github.com/rrwick/Porechop). Trimmed reads from SRMseq and LRMseq were used for hybrid-assembly analysis by metaSPAdes-3.13.0 [5]. After a BLAST search using the hybrid contig containing the *mcr-5.4* gene, the plasmid pSE13-SA01718 (accession no. KY807921.1) was listed as one of the hits with the highest identity and we used it as a reference for genome comparison with the Artemis Comparison Tool (ACT) v1.0 [6]. The *mcr-5.4*-carrying contig from the hybrid assembly was annotated using PATRIC v3.5.27 [7]. Trimmed reads from SRMseq were used to investigate the bacterial composition by OneCodex [8]. Finally, in order to predict the bacterial host of the *mcr-5.4* gene, a contig-binning analysis of the hybrid-assembled metagenome was performed using MaxBin2 v2.2.4 (https://sourceforge.net/projects/maxbin2/), probability threshold 0.9 and minimum contig length 1000 bp. The resulting bin containing the *mcr-5.4* gene was selected for taxonomy classification using Kraken2 (https://github.com/DerrickWood/kraken2) (minikraken2 DB v1).

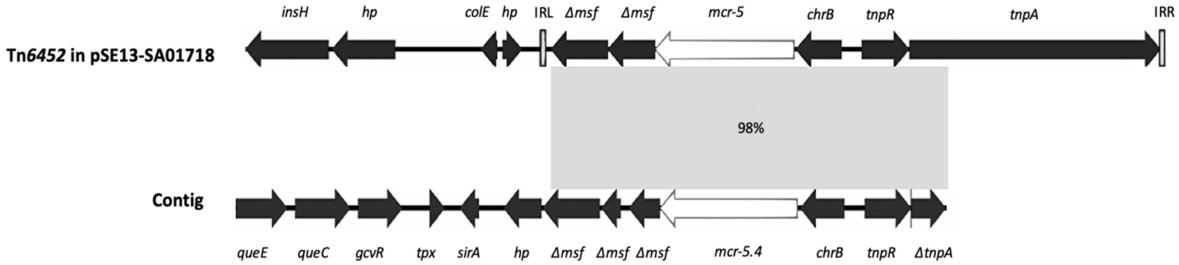


Figure 3.1: Comparative analysis of the genetic environment of *mcr-5* between the reference plasmid pSE13-SA01718 (accession no. KY807921.1) and the annotated hybrid metagenome contig (accession no. MK965519). The contig carrying the *mcr-5.4* gene consists of the following putative gene products: 7-carboxy-7-deazaguanine synthase (queE), 7-cyano-7-deazaguanine synthase (queC), glycine cleavage system transcriptional antiactivator GcvR (gcvR), thiol peroxidase (tpx), sulphurtransferase TusA family protein (sirA), hypothetical protein (hp), truncated MFS-type transporter (Δmsf), lipid A phosphoethanolamine transferase (*mcr-5.4*), ChrB domain protein (chrB), transposon resolvase (tnpR) and truncated transposon transposase (ΔtnpA). Areas with 98% identity between sequences are represented in light grey. Arrows indicate the position and direction of the genes. The transposon Tn6452 sequence in the reference plasmid pSE13-SA01718 is bounded by inverted repeats: IRL and IRR.

SRMseq showed the *mcr-5.4* gene detected in a contig of 2113 bp flanked by two truncated protein-coding sequences (CDSs), encoding the ChrB domain protein (involved in chromate resistance) and the Major Facilitator Superfamily (MFS) transporter. The hybrid-assembly analysis resulted in a contig of 8456 bp consisting of nine CDSs and four truncated CDSs (Figure 3.1). Comparative analysis of the genetic environment of the *mcr-5* gene, between the annotated hybrid metagenome contig and the reference plasmid pSE13-SA01718, showed a region of 4670 bp with 98% identity, corresponding to the backbone of the Tn6452 transposon (Figure 3.1). We observed three truncated CDSs for the MFS-type transporter in our contig instead of two as previously described in the reference sequence pSE13-SA01718. These differences did not appear to be due to sequencing errors when we

checked the sequence MK965519, (i) using pilon (https://github.com/broadinstitute/pilon) to correct for errors in short-read sequencing data and (ii) using CLC Genomic Workbench to update the hybrid contig by mapping both long and short reads against the hybrid contig. We also observed a region of 3786 bp, with no identity either with the reference plasmid pSE13-SA01718 (Figure 3.1) or with any other sequence in the GenBank database.

Species previously described to harbour an *mcr-5* gene are *Escherichia coli*, *Pseudomonas aeruginosa*, *Salmonella enterica*, *Aeromonas hydrophila* and *Cupriavidus gilardii*. The bacterial composition analysis of the water sample using SRMseq showed the presence of *Pseudomonas* spp. (relative abundance: 0.004%), *Cupriavidus* spp. (relative abundance: 0.001%) and *Aeromonas* spp. (relative abundance: 0.0003%). The binning analysis produced a bin positive for the *mcr-5.4* gene consisting of 1336 contigs (genome size: 5 175 285 bp; genome completeness: 68.2%). This bin was taxonomically classified as bacteria (70.73%) and proteobacteria (64.90%), and from this the most abundant class was *Gammaproteobacteria* (37.20%) (order *Pseudomonadales*, 15.57%), followed by *Betaproteobacteria* (14.90%) (order *Burkholderiales*, 10.63%).

Colistin resistance determinants (*mcr*) have been rarely reported in water environments; *mcr-1* has been detected in both hospital sewage and in environmental water streams and *mcr-3* in environmental water [9, 10]. To the best of our knowledge, this is the first-time description of an *mcr-5* gene in an indoor and healthcare water environment. Despite the fact that the comparative analysis showed the hybrid contig covering a large region of Tn6452, neither the left inverted repeat (IRL) nor the right inverted repeat (IRR) have been found. In addition, the lack of the right transposon region does not allow us to search for other possible inverted repeats. Thus, it is not possible to conclude whether the described *mcr-5.4* gene is transferable or not. Taxonomic analysis suggested the order of *Pseudomonadales* as the most probable host of the *mcr-5.4* gene in the water sample. Further studies are needed to determine the frequency of this gene in hospital water and other water environments and to evaluate the potential risks for patients and healthcare workers.

3.2 Acknowledgements

We would like to thank Erwin C. Raangs for technical assistance.

3.3 Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie SklodowskaCurie grant agreement 713660 (MSCA-COFUND-2015-DP 'Pronkjewail'), which includes in-kind contributions by com-

mercial partners. None of the commercial partners had any influence on interpretation of reviewed data and conclusions drawn, or on drafting of the manuscript. This work was partly supported by the INTERREG VA (202085)-funded project EurHealth-1Health, part of a Dutch–German cross-border network supported by the European Commission, the Dutch Ministry of Health, Welfare and Sport (VWS), the Ministry of Economy, Innovation, Digitalization and Energy of the German Federal State of North RhineWestphalia and the German Federal State of Lower Saxony.

3.4 Transparency declarations

None to declare.

3.5 References

[1] Xiaoming Wang et al. "Emergence of a novel mobile colistin resistance gene, *mcr-8* , in NDM-producing *Klebsiella pneumoniae*". en. In: *Emerging Microbes & Infections* 7.1 (Dec. 2018), pp. 1–9. ISSN: 2222-1751. DOI: 10.1038/s41426-018-0124-z. URL: https://www.tandfonline.com/doi/full/10.1038/s41426-018-0124-z (visited on 01/24/2022).

[2] Maria Borowiak et al. "Identification of a novel transposon-associated phosphoethanolamine transferase gene, mcr-5, conferring colistin resistance in d-tartrate fermenting Salmonella enterica subsp. enterica serovar Paratyphi B". en. In: *Journal of Antimicrobial Chemotherapy* 72.12 (Dec. 2017), pp. 3317–3324. ISSN: 0305-7453, 1460-2091. DOI: 10.1093/jac/dkx327. URL: https://academic.oup.com/jac/article/72/12/3317/4161410 (visited on 01/24/2022).

[3] Jun Li et al. "Co-Occurrence of Colistin and Meropenem Resistance Determinants in a *Stenotrophomonas* Strain Isolated from Sewage Water". en. In: *Microbial Drug Resistance* 25.3 (Apr. 2019), pp. 317–325. ISSN: 1076-6294, 1931-8448. DOI: 10.1089/mdr.2018.0418. URL: https://www.liebertpub.com/doi/10.1089/mdr.2018.0418 (visited on 01/24/2022).

[4] N. J. Loman and A. R. Quinlan. "Poretools: a toolkit for analyzing nanopore sequence data". en. In: *Bioinformatics* 30.23 (Dec. 2014), pp. 3399–3401. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btu555. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu555 (visited on 01/24/2022).

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

[6] T. J. Carver et al. "ACT: the Artemis comparison tool". en. In: *Bioinformatics* 21.16 (Aug. 2005), pp. 3422–3423. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bti553. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti553 (visited on 01/24/2022).

[7] Alice R. Wattam et al. "Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center". en. In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D535–D542. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw1017. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1017 (visited on 01/24/2022).

[8] Samuel S Minot, Niklas Krumm, and Nicholas B Greenfield. *One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification*. en. preprint. Bioinformatics, Sept. 2015. DOI: 10.1101/027607. URL: http://biorxiv.org/lookup/doi/10.1101/027607 (visited on 01/24/2022).

[9] Feifei Zhao et al. "IncP Plasmid Carrying Colistin Resistance Gene *mcr-1* in Klebsiella pneumoniae from Hospital Sewage". en. In: *Antimicrobial Agents and Chemotherapy* 61.2 (Feb. 2017). ISSN: 0066-4804, 1098-6596. DOI: 10.1128/AAC.02229-16. URL: https://journals.asm.org/doi/10.1128/AAC.02229-16 (visited on 01/24/2022).

[10] Hongmei Tuo et al. "The Prevalence of Colistin Resistant Strains and Antibiotic Resistance Gene Profiles in Funan River, China". In: *Frontiers in Microbiology* 9 (Dec. 2018), p. 3094. ISSN: 1664-302X. DOI: 10.3389/fmicb.2018.03094. URL: https://www.frontiersin.org/article/10.3389/fmicb.2018.03094/full (visited on 01/24/2022).

Chapter 4

DEN-IM: Dengue virus genotyping from shotgun and targeted metagenomics

This chapter is a reproduction of the following publication:

C. I. Mendes, E. Lizarazo, M. P. Machado, D. N. Silva, A. Tami, M. Ramirez, N. Couto, J. W. A. Rossen, J. A. Carriço, DEN-IM: dengue virus genotyping from amplicon and shotgun metagenomic sequencing. Microbial Genomics, Volume 6, Issue 3, March 2020. DOI: `https://doi.org/10.1099/mgen.0.000328`

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

Dengue virus (DENV) represents a public health threat and economic burden in affected countries. The risk of exposure to DENV is increasing, not only because of travel to endemic regions, but also due to the broader dissemination of the mosquito vector, making the burden of dengue very significant.

The availability of genomic data is key to understanding viral evolution and dynamics, supporting improved control strategies. Currently, the use of second-generation sequencing technologies, which can be applied both directly to patient samples (shotgun metagenomics) and to PCR-amplified viral sequences (amplicon sequencing), is the most informative approach to monitor viral dissemination and genetic diversity by providing, in a single methodological step, identification and characterization of the whole viral genome at the nucleotide level. This makes DENV identification and characterization through genomic analysis by developing a software where the lessons learned in Chapters 2 and 3 are applied.

We have developed DEN-IM, a one-stop, user-friendly, containerised and reproducible workflow for the analysis of Dengue virus short-read sequencing data from both amplicon and shotgun metagenomics approaches. EN-IM was designed to perform a comprehen-

sive analysis in order to generate either assemblies or consensus of full DENV coding sequences and to identify their serotype and genotype. DEN-IM can also detect all four DENV serotypes and the respective genotypes present in a spiked sample, raising the possibility that DEN-IM can play a role in the identification of co-infection cases whose prevalence is increasingly perceived in highly endemic areas.

My contribution to this publication included the design, implementation and optimisation of the DEN-IM the workflow, including the creation of the Docker containers for all dependencies. Two databases, one comprising 3830 DENV sequences for the retrieval of the reads of interest from the input samples, and a second comprising of 161 sequences representing the genetic diversity of all DENV sero and genotypes were constructed by me. Additionally, I've also wrote the manuscript.

DEN-IM: dengue virus genotyping from amplicon and shotgun metagenomic sequencing

Catarina I Mendes[1,2,*], Erley Lizarazo[2,*], Miguel P Machado[1], Diogo N Silva[1], Adiana Tami[2], Mário Ramirez[1], Natacha Couto[2], John W A Rossen[2] and João A Carriço[1]

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

[2]University of Groningen, University Medical Center Groningen, Department of Medical Microbiology and Infection Prevention, Groningen, The Netherlands

[*]Contributed equally

4.1 Abstract

Dengue virus (DENV) represents a public health and economic burden in affected countries. The availability of genomic data is key to understanding viral evolution and dynamics, supporting improved control strategies. Currently, the use of High Throughput Sequencing (HTS) technologies, which can be applied both directly to patient samples (shotgun metagenomics) and PCR amplified viral sequences (targeted metagenomics), is the most informative approach to monitor the viral dissemination and genetic diversity.

Despite many advantages, these technologies require bioinformatics expertise and appropriate infrastructure for the analysis and interpretation of the resulting data. In addition, the many software solutions available can hamper reproducibility and comparison of results. Here we present DEN-IM, a one-stop, user-friendly, containerised and reproducible workflow for the analysis of DENV sequencing data, both from shotgun and targeted metagenomics approaches. It is able to infer the DENV coding sequence (CDS), identify the serotype and genotype, and generate a phylogenetic tree. It can easily be run on any UNIX-like system, from local machines to high-performance computing clusters, performing a comprehensive analysis without the requirement of extensive bioinformatics expertise.

Using DEN-IM, we successfully analysed two DENV datasets. The first comprised 25 shotgun metagenomic sequencing samples of variable serotype and genotype, including an in vitro spiked sample containing the four known serotypes. The second dataset consisted of 106 targeted metagenomic sequences of DENV 3 genotype III where DEN-IM allowed detection of the intra-genotype diversity. The DEN-IM workflow, parameters and execution configuration files, and documentation are freely available at `https://github.com/B-UMMI/DEN-IM`.

4.1.1 Keywords

dengue virus, surveillance, metagenomics, reproducibility, workflow, containerization, scalability

4.2 Author Notes

All supporting data, code and protocols have been provided within the article or through supplementary data files.

Metagenomic sequencing data available under BioProject PRJNA474413. DEN-IM reports for the analysed datasets are available in Figshare under https://doi.org/10.6084/m9.figshare.11316599.v1. Phylogeny inference trees for the dengue virus typing database available in Figshare at https://doi.org/10.6084/m9.figshare.11316599.v1. The supplemental material is available in Figshare at https://doi.org/10.6084/m9.figshare.11316599.v1. DEN-IM's source code and documentation available at https://github.com/B-UMMI/DEN-IM.

4.3 Abbreviations

Coding Sequence (CDS); Dengue virus (DENV); high-performance computing (HPC); high-throughput sequencing (HTS); non-coding region (NCR); quality control (QC); Reverse Transcription Polymerase Chain Reaction (RT-PCR)

4.4 Data Summary

1. The supplemental material and tables are available at Figshare under https://doi.org/10.6084/m9.figshare.9963812
2. The 106 DENV-3 targeted metagenomics paired-end short-read datasets are available under BioProject PRJNA394021. The 25 shotgun metagenomics dataset is available under BioProject PRJNA474413. The accession number for all the samples in the shotgun metagenomics dataset are available in the Supplementary material
3. The accession numbers for the 41 samples, belonging to zika virus, chikungunya virus and yellow fever virus shotgun and targeted metagenomic datasets are available in the Supplementary material.

4. DEN-IM reports for the analysed datasets are available at Figshare (https://doi.org/10.6084/m9.figshare.9318851).

5. Phylogeny inference trees for the dengue virus typing database available at Figshare (https://doi.org/10.6084/m9.figshare.9331826).

6. Code for the DEN-IM workflow is available at `https://github.com/B-UMMI/DEN-IM` and documentation, including step-by-step tutorials, is available at `https://github.com/B-UMMI/DEN-IM/wiki`.

4.5 Impact Statement

The risk of exposure to DENV is increasing not only by travelling to endemic regions, but also due to the broader dissemination of the mosquito, making the burden of dengue very significant.

The decreasing costs and wider availability of HTS makes it an ideal technology to monitor DENV's transmission. Metagenomics approaches decrease the time to obtain nearly complete DENV sequences without the need for time-consuming viral culture through the direct processing and sequencing of patient samples. A ready to use bioinformatics workflow, enabling the reproducible analysis of DENV, is therefore particularly relevant for the development of a straightforward HTS workflow.

DEN-IM was designed to perform a comprehensive analysis in order to generate either assemblies or consensus of full DENV CDSs and to identify their serotype and genotype. DEN-IM can also detect all four DENV genotypes present in a spiked sample, raising the possibility that DEN-IM can play a role in the identification of co-infection cases whose prevalence is increasingly appreciated in highly endemic areas. Although being ready-to-use, the DEN-IM workflow can be easily customised to the user's needs.

DEN-IM enables reproducible and collaborative research, being accessible to a wide group of researchers regardless of their computational expertise and resources available.

4.6 Introduction

The Dengue virus (DENV), a single-stranded positive-sense RNA virus belonging to the Flavivirus genus, is one of the most prevalent arboviruses and is mainly concentrated in tropical and subtropical regions. Infection with DENV results in symptoms ranging from mild fever to haemorrhagic fever and shock syndrome [1]. Transmission to humans occurs through the bite of Aedes mosquitoes, namely Aedes aegypti and Aedes albopictus [2]. In

2010, it was predicted that the burden of dengue disease reached 390 million cases/year worldwide [3]. The high morbidity and mortality of dengue makes it the arbovirus with the highest clinical significance [4]. DENV is a significant public health challenge in countries where the infection is endemic due to the high health and economic burden. Despite the emergence of novel therapies and ecological strategies to control the mosquito vector, there are still important knowledge gaps in the virus biology and its epidemiology [2].

The viral genome of $\sim$11,000 nucleotides, consists of a CDS of approximately 10.2 Kb that is translated into a single polyprotein encoding three structural proteins (capsid - C, premembrane - prM, envelope - E) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5). Additionally, the genome contains two Non-Coding Regions (NCRs) at their 5' and 3' ends [5].

DENV can be classified into four serotypes (1, 2, 3 and 4), differing from each other from 25% to 40% at the amino acid level. They are further classified into genotypes that vary by up to 3% at the amino acid level [2]. The DENV-1 serotype comprises five genotypes (I-V), DENV-2 groups six (I-VI, also named American, Cosmopolitan, Asian-American, Asian II, Asian I and Sylvatic), DENV-3 four (I-III and V), and DENV-4 also four (I-IV).

Although real-time reverse transcription polymerase chain reaction (RT-PCR) will probably remain the front line in Dengue etiological diagnosis, the implementation of a surveillance system relying on HTS technologies allows the simultaneous identification and characterization by serotyping and genotyping of DENV cases at the nucleotide level in a single methodological step. Due to the high sensitivity of these technologies, previous studies showed that viral sequences can be directly obtained from patient sera using a shotgun metagenomics approach [6]. Alternatively, HTS can be used in a targeted metagenomics approach in which a PCR step is used to pre-amplify viral sequences before sequencing. In recent years, HTS has been successfully used as a tool for identification of DENV directly from clinical samples [6, 7]. This also allows the rapid identification of the serotype and genotype important for disease management as the genotype may be associated with disease outcome [8].

Several initiatives aim to facilitate the identification of the DENV serotype and genotype from HTS data. The Genome Detective project (https://www.genomedetective.com/) offers an online Dengue Typing Tool (https://www.genomedetective.com/app/typingtool/dengue/) [9] relying on BLAST and phylogenetic methods in order to identify the closest serotype and genotype, but it requires as input assembled genomes in FASTA format. The same project also offers the Genome Detective Typing Tool (https://www.genomedetective.com/app/typingtool/virus/) [10] identifying viruses present in a sample. Additionally, there are several tools available for viral read identification and assembly, such as VIP [11], virusTAP [12] and drVM [13], but none performs genotyping of the identified reads.

We developed DEN-IM as a ready-to-use, one-stop, reproducible bioinformatic analysis workflow for the processing and phylogenetic analysis of DENV using paired-end raw HTS data. DEN-IM is implemented in Nextflow [14], a workflow manager software that uses Docker (https://www.docker.com) containers with pre-installed software for all the workflow tools. The DEN-IM workflow, as well as parameters and documentation, are available at `https://github.com/B-UMMI/DEN-IM`.

4.7 The DEN-IM Workflow

DEN-IM is a user-friendly automated workflow enabling the analysis of shotgun or targeted metagenomics data for the identification, serotyping, genotyping, and phylogenetic analysis of DENV, as represented in Figure 4.1, accepting as input raw paired-end sequencing data (FASTQ files) and informing the user with an interactive and comprehensive HTML report (Supplementary Figure 4.5), as well as providing output files of the whole pipeline.

It is implemented in Nextflow, a workflow management system that allows the effortless deployment and execution of complex distributed computational workflows in any UNIX-based system, from local machines to HPC with a container engine installation, such as Docker (https://www.docker.com/), Shifter [15] or Singularity [16]. DEN-IM integrates Docker containerised images, compatible with other container engines, for all the tools necessary for its execution, ensuring reproducibility and the tracking of both software code and version, regardless of the operating system used.

Users can customise the workflow execution either by using command line options or by modifying the simple plain-text configuration files. To make the execution of the workflow as simple as possible, a set of default parameters and directives is provided. An exhaustive description of each parameter is available as Supplementary material (see 4.13.2).

The local installation of the DEN-IM workflow, including the docker containers with all the tools needed and the curated DENV database, requires 15 Gigabytes (Gb) of free disk space. The minimum requirements to execute the workflow are at least 5 Gb of memory and 4 CPUs. The disk space required for execution depends greatly on the size of the input data, but for the datasets used in this article, DEN-IM generates approximately 5 Gb of data per Gb input data. DEN-IM workflow can be divided into the following components:

4.7.0.1 Quality Control and Trimming

The Quality Control (QC) and Trimming block starts with a process to verify the integrity of the input data. If the sequencing files are corrupted, the execution of the analysis of that sample is terminated. The sequences are then processed by FastQC

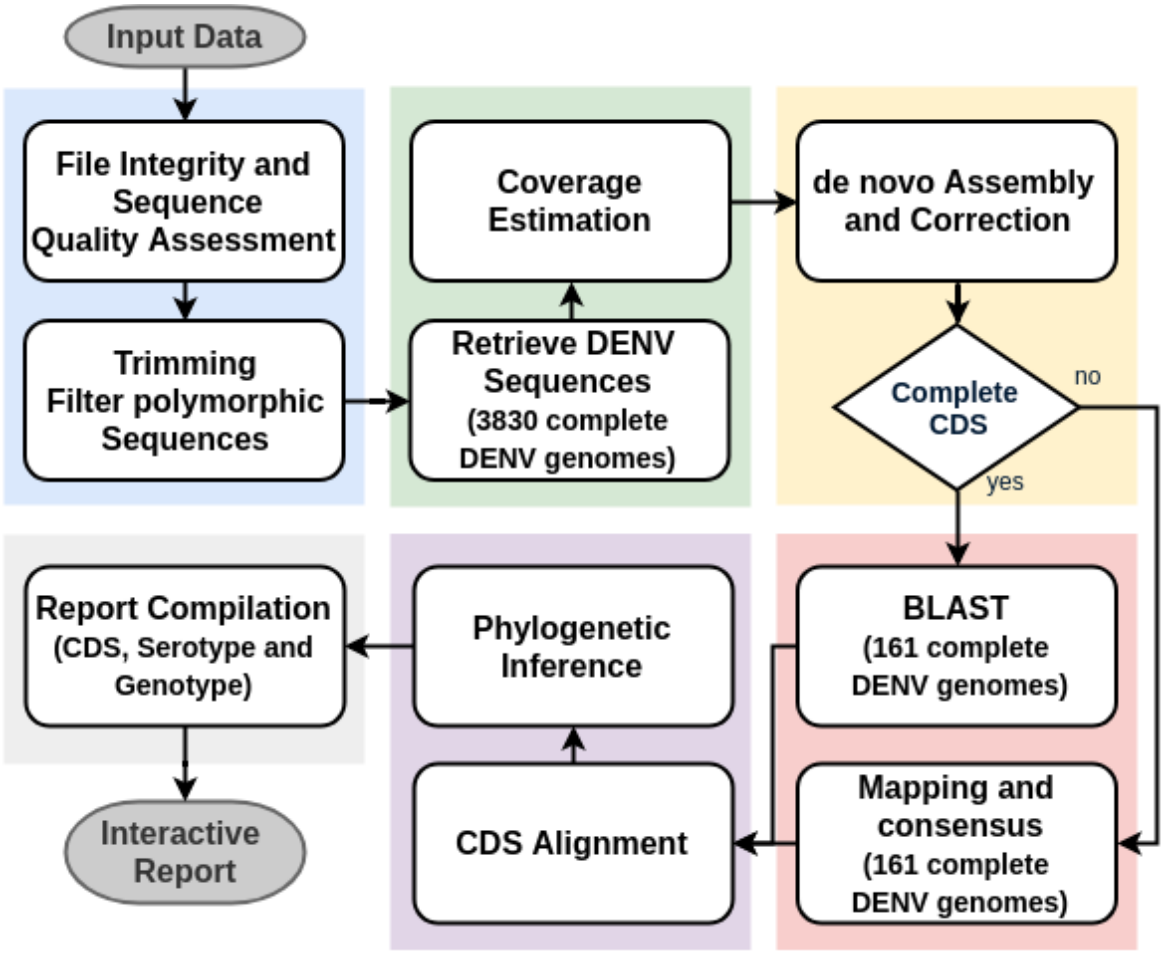


Figure 4.1: The DEN-IM workflow separated into five different components. The raw sequencing reads are provided as input to the first block (in blue), responsible for quality control and elimination of low-quality reads and sequences. After successful preprocessing of the reads, these enter the second block (green) for retrieval of the DENV reads using the mapping database of 3858 complete DENV genomes as a reference. This block also provides an initial estimate of the sequencing depth. After the de novo assembly and assembly correction block (yellow), the CDSs are retrieved and then classified with the reduced-complexity DENV typing database containing 161 sequences representing the known diversity of DENV serotypes and genotypes (red). If a complete CDS fails to be assembled, the reads are mapped against the DENV typing database and a consensus sequence is obtained for classification and phylogenetic inference. All CDSs are aligned and compared in a phylogenetic analysis (purple). Lastly, a report is compiled (grey) with the results of all the blocks of the workflow.

(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, version 0.11.7) to determine the quality of the individual base pairs of the raw data files. The low-quality bases and adapter sequences are trimmed by Trimmomatic [17] (version 0.36). In addition, paired-end reads with a read length shorter than 55 nucleotides after trimming are removed from further analyses. Lastly, the low complexity sequences, containing over 50% of poly-A, poly-N or poly-T nucleotides, are filtered out of the raw data using PrinSeq [17] (version 0.10.4).

4.7.0.2 Retrieval of DENV sequences

In the second step, DENV sequences are selected from the sample using Bowtie2 [18] (version 2.2.9) and Samtools [18] (version 1.4.1). As a reference we provide the DENV mapping database, a curated DENV database composed of 3830 complete DENV genomes. An in-depth description of this database is available as Supplementary material (see 4.13.1). A permissive approach is followed by allowing for mates to be kept in the sample even when only one read maps to the database in order to keep as many DENV derived reads as possible. The output of this block is a set of processed reads of putative DENV origin.

4.7.0.3 Assembly

DEN-IM applies a two-assembler approach to generate assemblies of the DENV CDS. To obtain a high confidence assembly, the processed reads are first de novo assembled with SPAdes [19] (version 3.12.0). If the full CDS fails to be assembled into a single contig, the data is re-assembled with the MEGAHIT assembler [20] (version 1.1.3), a more permissive assembler developed to retrieve longer sequences from metagenomics data. The resulting assemblies are corrected with Pilon [21] (version 1.22) after mapping the processed reads to the assemblies with Bowtie2.

If more than one complete CDS is present in a sample, each of the sequences will follow the rest of the DEN-IM workflow independently. If no full CDS is assembled neither with SPAdes nor with MEGAHIT, the processed reads are passed on to the next module for consensus generation by mapping, effectively constituting DEN-IM's two-pronged approach using both assemblers and mapping.

4.7.0.4 Typing

For each DENV complete CDS, the serotype and genotype is determined with the Seq_Typing tool (`https://github.com/B-UMMI/seq_typing`, version 2.0) [22] using BLAST [23] and the custom Typing database of DENV containing 161 complete sequences (see 4.13.1). The tool determines which reference sequence is more closely related to the query based on the identity and length of the sequence covered, returning the serotype and genotype of the reference sequence

If a complete CDS fails to be obtained through the assembly process, the processed reads are mapped against the same DENV typing database, with Bowtie2, using the Seq_Typing tool, with similar criteria for coverage and identity to those used with the BLAST approach. If a type is determined, the consensus sequence obtained follows through to the next step in the workflow. Otherwise, the sample is classified as Non-Typable and its process terminated.

4.7.0.5 Phylogeny

All DENV complete CDSs and consensus sequences analysed in a workflow execution are aligned with MAFFT [24] (version 7.402). By default, or if the number of samples analysed is less than 4, four representative sequences for each DENV serotype (1 to 4) from NCBI are also included in the alignment. The NCBI references included are NC_001477.1 (DENV-1), NC_001474.2 (DENV-2), NC_001475.2 (DENV-3) and NC_002640.1 (DENV-4). The closest reference sequence to each analysed sample in the DENV typing database to each analysed sample can also be retrieved and included in the alignment. With the resulting alignment, a Maximum Likelihood tree is constructed with RaXML [25] (version 8.2.11).

4.7.0.6 Output and Report

The output files of all tools in DEN-IM's workflow are stored in the 'results' folder in the directory of DEN-IM's execution, as well as the execution log file DEN-IM and for each component.

The HTML report (Supplementary Figure 4.5), stored in the 'pipeline_results' directory contains all results divided into four sections: report overview, tables, charts and phylogenetic tree. The report overview and all tables allow for selection, filtering and highlighting of particular samples in the analysis. All tables have information on if a sample failed or passed the quality control metrics highlighted by green, yellow or red signs for pass, warning and fail messages, respectively.

The *in silico* typing table contains the results of the serotype and genotype of each CDS analysed, as well as identity, coverage and GenBank ID of the closest reference in the DENV typing database. The quality control table shows information regarding the number of raw base pairs and number of reads in the raw input files and the percentage of trimmed reads. The mapping table includes the results for the mapping of the trimmed reads to the DENV mapping database, including the overall alignment rate, and an estimation of the sequence depth including only the DENV reads. For the assembly statistics table, the number of CDSs in each sample, the number of contigs and the number of assembled base pairs generated by either SPAdes or MEGAHIT assemblers is included. The number of contigs and assembled base pairs after correction with Pilon is also presented in the table. The assembled contig size distribution scatter plot is available in the chart section, showing the contig size distribution for the Pilon corrected assembled CDSs.

Lastly, a phylogenetic tree is included, rooted at midpoint for visualisation purposes, and with each tip coloured according to the genotyping results. If the option to retrieve the closest typing reference is selected, these sequences are also included in the tree with respective typing metadata. The tree can be displayed in several conformations provided by Phylocanvas JavaScript library (`http://phylocanvas.net`, version 2.8.1) and it is possible

to zoom in or collapse selected branches. The support bootstrap values of the branches can be displayed, and the tree can be exported as a Newick tree file or as a PNG image.

4.8 Software comparison

DEN-IM offers a core assembly functionality, leveraging a de novo and consensus assembly approach, to obtain a full CDS sequence to perform geno- and serotyping, followed by phylogenetic positioning of the samples analysed. This results in a phylogenetic tree showing the genotyping results, presented in an HTML file.

There are several alternative tools, both command line and online based, capable of identifying DENV reads and performing assembly (Table 4.1). VIP and drVM are both standalone pipelines, like DEN-IM, and several components overlap with DEN-IM's but the retrieval of viral sequences is not targeted for DENV, and no serotyping and genotyping is performed. VIP performs a phylogenetic analysis against the reference database. VirusTAP is a web server for the identification of viral reads using the ViPR and IRD databases, or alternatively with the RefSeq Virus database. GenomeDetective is also a web service that provides two tools, one for the assembly of viral sequences from raw data (Virus tool) and another for serotyping and genotyping of DENV fasta sequences (Dengue Typing tool). Both tools need to be run consecutively, with the Virus Tool providing a link to redirect to the Dengue Typing tool when a DENV sequence is identified.

Table 4.1: DEN-IM's workflow comparison with different tools for the identification and genotyping of DENV from sequencing data.

Tool	Quality Control	DENV Sequence Retrieval	Assembly	Typing	Phylogeny	Report
DEN-IM	✓	✓	✓	✓	✓	✓(one report with all samples analysed)
VIP	✓	✓[1]	✓	X	✓	✓
VirusTAP	✓	✓[1]	✓	X	X	✓(web-based, one per sample, downloadable)
drVM	✓	✓[1]	✓	X	X	X
GenomeDetective Virus Tool	✓	X	✓	X	X	✓(web-based, one per sample)
GenomeDetective Dengue Typing Tool	X	X	X	✓[2]	X	✓(web-based, one per sample)

[1] Targeted for viral sequences, but not specific for DENV

[2] Sequence file can be received from GenomeDetective Virus Tool, as well as independently uploaded

Of all the tools listed in Table 4.1, only Genome Detective offers a tool to determine the DENV sero- and genotype from a fasta sequence, but the need to run their virus identification tool prior to obtain a sequence from the raw sequencing data increases the time to obtain a typing result, especially when a large number of sequences needs to be analysed. Moreover,

these tools are not open source, so we are unable to compare the methodology used with our own. Additionally, there might be privacy issues in submitting data to external services, like VirusTAP and GenomeDetective, especially when handling metagenomics data that contain human sequences subjected to strict privacy laws in most countries. Therefore, a stand-alone tool is preferable for these analyses since these can be run in secure local environments. DEN-IM's main advantage when compared to web-based platforms is the ability to analyse batches of samples in a scalable manner, obtaining a report summarizing all the samples analysed and a phylogeny analysis of all DENV CDSs recovered.

4.9 Results

To evaluate the DEN-IM workflow performance, we analysed three datasets, one containing shotgun metagenomics sequencing data of patient samples (see Table 4.2), a second with amplicon sequencing data, a set with 106 paired-end samples obtained from Parameswaran et al [26] and another set with 78 single-end samples available under BioProject PRJNA321963, and a third dataset of publicly available sequences, both from amplicon and shotgun metagenomics, containing 45 chikungunya virus (CHIKV) samples, 66 zika virus (ZKV), and 21 yellow fever virus (YFV) samples (see Table 4.3). All analyses were executed with the default resources and parameters (available at `https://github.com/B-UMMI/DEN-IM`). In the shotgun metagenomics and the single-end amplicon sequencing datasets the closest typing reference in the final tree and the NCBI DENV references for each serotype were included in the phylogenetic analysis. The resulting reports for each dataset are available on Figshare at `https://doi.org/10.6084/m9.figshare.9318851`.

4.9.0.1 Shotgun metagenomics dataset

We analysed a dataset containing 22 shotgun metagenomics paired-end short-read Illumina sequencing samples from positive dengue cases, one positive control (purified from a DENV culture), one negative control (blank), and an in vitro spiked sample containing the 4 DENV serotypes (see 4.13.3). On average, each sample took 7 minutes to analyse. A total of 75 CPU hours were used to analyse the 25 samples, with a total of 17 Gb in size. This analysis resulted in 69 Gb of data. The negative control and the 92-1001 sample had no reads after trimming and filtering of low complexity reads, therefore they were removed from further analysis (see 4.4). When mapping to the DENV mapping database, the percentage of DENV reads in the 21 clinical samples, positive control and spiked sample passing QC ranged from 0.01% (sample UCUG0186) to 85.38% (sample Positive Control - PC). After coverage depth estimation, the analysis of the samples 91-0115 and UCUG0186 was terminated due to a low proportion of DENV reads (0.05% and 0.01% respectively). Therefore, they failed to meet the threshold criterion of having an estimated depth of coverage of $\geq$10x (estimated cover-

ages of 3.17x and 5.65x, respectively). Sequence data of sample 91-0106 contained only 960 DENV reads (0.03%) but these were successfully assembled into a CDS with an estimated depth of coverage of 14.71x.

In the assembly module, the remaining 19 samples, the spiked sample and the PC were assembled with DEN-IM's two assembler approach. Twenty-four full CDS were assembled (see 4.6), even in samples originally having DENV read content as low as 0.03% of the total reads. Sixteen samples, including the spiked sample and the positive control, were assembled in the first step with the SPAdes assembler, and five in the second with the MEGAHIT assembler. In the spiked sample, all four CDSs were successfully assembled and recovered.

Serotype and genotype were successfully determined for the 24 DENV CDSs by BLAST (see 4.6). The most common were serotype 2 genotype III (Asian American) and serotype 4 genotype II, with 8 samples each (33%), followed by serotype 3 genotype III (n=5, 21%), serotype 1 genotype V (n=2, 8%) and serotype 2 genotype V (Asian I) (n=1, 4%). All CDSs recovered and the respective closest reference genome in the typing database were aligned and a maximum likelihood phylogenetic tree was obtained to visualise the relationship between the samples (Figure 4.2). There was a perfect concordance between the results of serotyping and genotyping and the major groups in the tree. Four distinct CDSs were assembled for the spiked sample that resulted in different coverages of each serotype CDS (2032x times coverage for DENV-2, 229x coverage for DENV-1, 76x coverage for DENV-3 and 30x times coverage for DENV-4), in accordance with the ranking order of the real-time RT-PCR results (see 4.13.3).

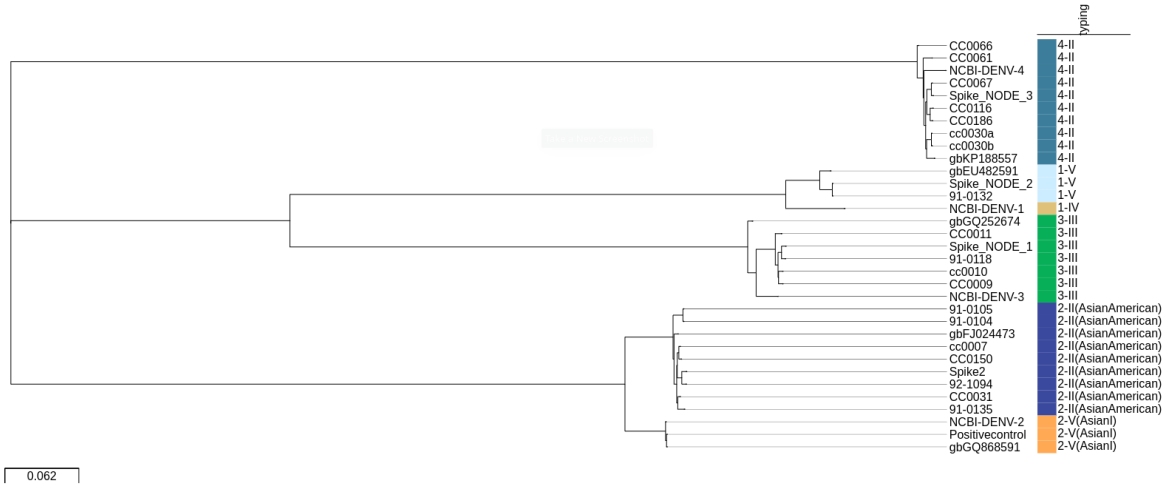


Figure 4.2: Phylogenetic reconstruction of the shotgun metagenomic dataset. Maximum Likelihood tree in the DEN-IM report for the 24 complete CDSs (n=21 samples) obtained with the metagenomics dataset, the respective closest references in the typing database (identified by their GenBank ID), and the NCBI DENV references for each serotype (NCBI-DENV-1: NC_001477.1, NCBI-DENV-2: NC_001474.2, NCBI-DENV-3: NC_001475.2, NCBI-DENV-4: NC_002640.1). The tree is midpoint rooted for visualisation purposes and the scale represents average substitutions per site. The colours depict the DENV genotyping results.

4.9.0.2 The Amplicon Sequencing Dataset

To validate DEN-IM's performance in a amplicon sequencing approach, a dataset of 106 paired-end HTS samples of PCR products using primers targeting DENV-3 (27) were analysed (see 4.13.4). On average, each sample took 5 minutes to analyse. The 106 samples, with 51 Gb in size, took 3622 CPU hours to be analysed, resulting in 424 Gb of data.

No samples failed the quality control block (see Table 4.5). The proportion of DENV reads ranged from 24.72% (SRR5821236) to 99.81% (SRR5821254) of the total processed reads. The samples with less than 70% DENV DNA were taxonomically profiled with Kraken2 (28) and the minikraken2_v2 database (`ftp://ftp.ccb.jhu.edu/pub/data/kraken2_dbs/minikraken2_v2_8GB_201904_UPDATE.tgz`) and the source of contamination was determined to have come largely from Human DNA (see Table 4.6).

Of the 106 samples, 43 (41%) managed to assemble a complete CDS sequence (see Table 4.5) whereas a mapping approach was used for the remaining 63 samples (60%) and a consensus CDS was generated. For the assembled CDSs, all but one were assembled with MEGAHIT after not producing a full CDS with SPAdes. Moreover, pronounced variation on the size of the assembled contigs is evident in the contig size distribution plot (see 4.7).

All 106 CDSs recovered belonged to serotype 3 genotype III. Despite the same classification, the maximum likelihood tree indicates that there is detectable genetic diversity within the dataset (486 SNPs in 10237 nucleotides) (Figure 4.3).

A second amplicon dataset, containing 78 DENV-1 single-end samples recovered from different Aedes aegypti isofemale hosts were analysed (see 4.13.4). On average, each sample took 3 minutes to analyse. The 78 samples, with 19 Gb in size, took 278 CPU hours to be analysed, resulting in 203 Gb of data.

No samples failed the quality control block and the proportion of DENV reads ranged from 59% (SRR3539343) to 96% (SRR3539408) of the total processed reads (see Table 4.7). Of the 78 samples, 53 (68%) assembled a complete CDS sequence and in the remaining 25 (32%) the complete CDS was obtained through mapping. All CDSs recovered, the respective closest reference genome in the typing database and NCBI's references for each DENV serotype were aligned and a maximum likelihood phylogenetic tree was obtained (Figure 4.4). All 78 samples belong to serotype 1 genotype I and, similarly to the previous dataset of 106 samples, there was detectable genetic diversity within the dataset (651 SNPs in 10808 nucleotides excluding reference sequences).

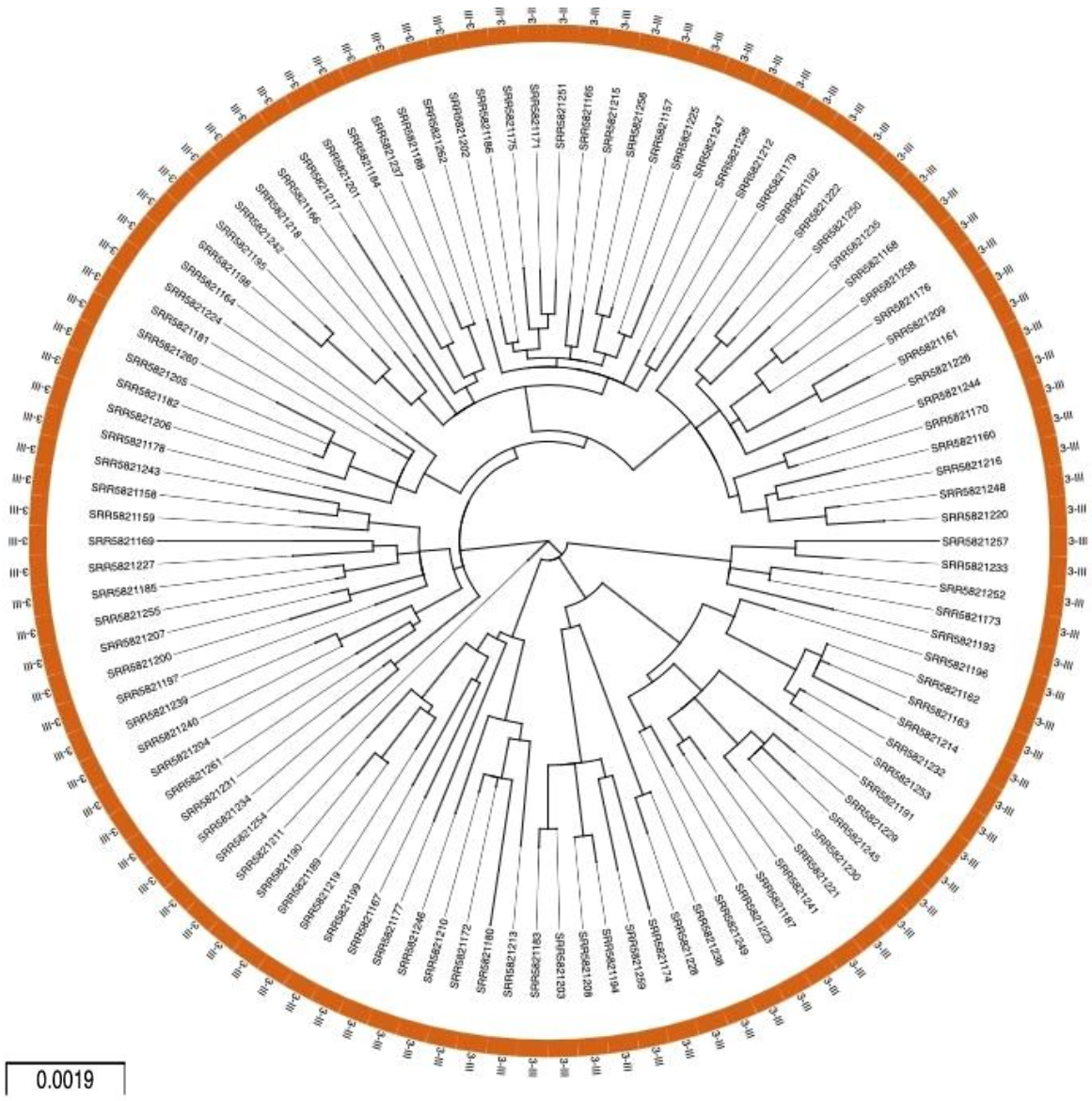


Figure 4.3: Phylogenetic reconstruction of the paired-end targeted metagenomic dataset. Maximum likelihood circular tree in the DEN-IM report for the 106 complete CDSs obtained with the targeted metagenomics dataset (n=106). All samples belong to serotype 3 genotype III. The scale represents average substitutions per site.

4.9.0.3 The Non-DENV Arbovirus Dataset

In order to evaluate DEN-IM's specificity to DENV sequences, a third dataset of publicly available sequences of arbovirus other than DENV, both from amplicon and shotgun metagenomics, was analysed containing 45 CHIKV samples, 66 ZKV, and 21 YFV samples (see Table 4.3). All 132 samples failed DEN-IM's workflow, 16 due to not enough sequencing data remaining after quality trimming, and the remaining 116 due to very low estimated coverage of the DENV genome (less than 0.01x), as expected.

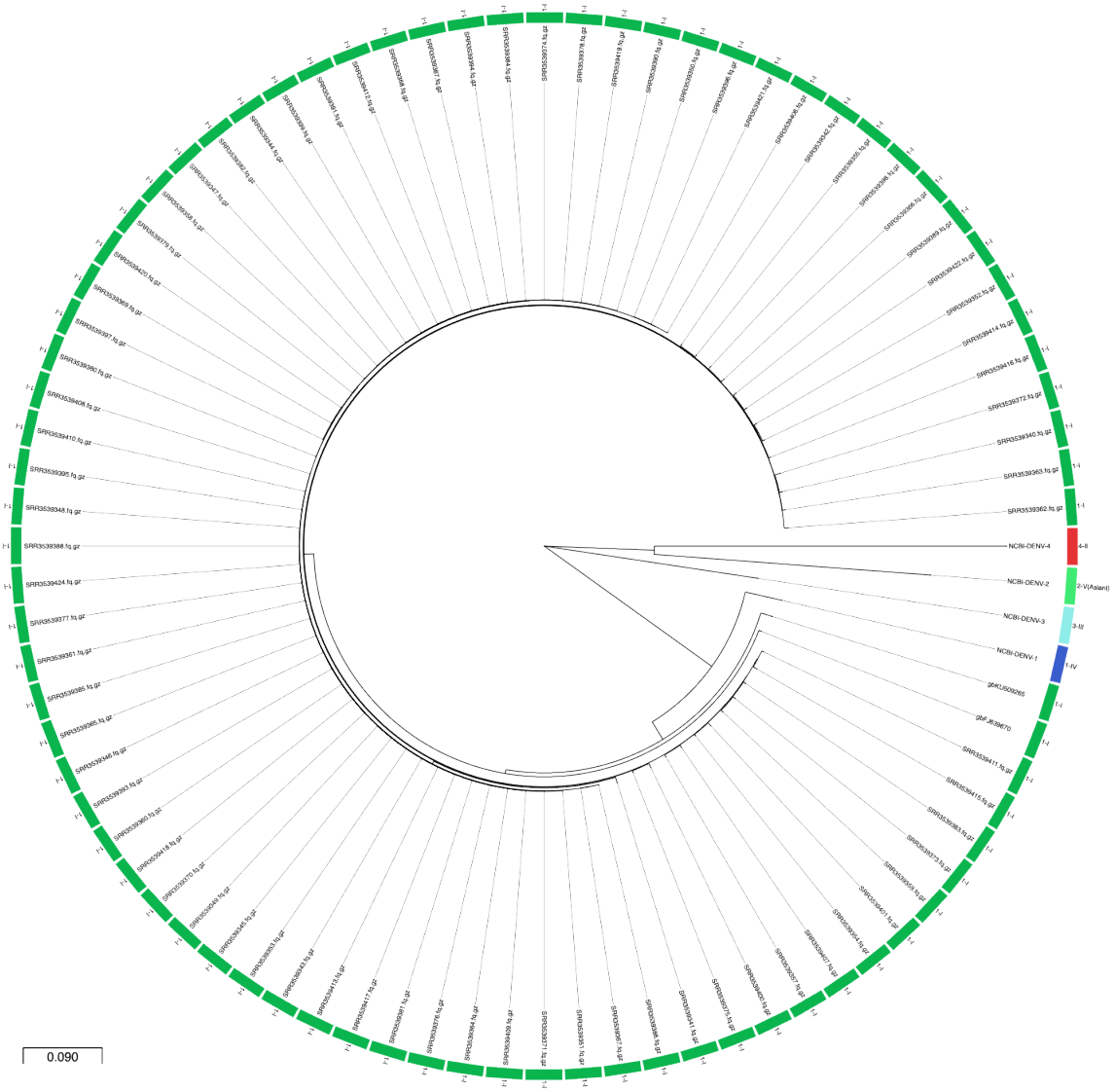


Figure 4.4: Phylogenetic reconstruction of the single-end targeted metagenomic dataset. Maximum likelihood circular tree in the DEN-IM report for the 78 complete CDSs obtained with the targeted metagenomics dataset (n=78) and the NCBI DENV references for each serotype (NCBI-DENV-1: NC_001477.1, NCBI-DENV-2: NC_001474.2, NCBI-DENV-3: NC_001475.2, NCBI-DENV-4: NC_002640.1). All samples belong to serotype 1 genotype I. The scale represents average substitutions per site.

4.10 Conclusion

We have successfully analysed two DENV datasets, one comprising 25 shotgun metagenomics sequencing samples and a second of 106 paired-end and 78 single-end targeted metagenomics samples.

In the first dataset, we recovered 24 CDSs from 19 clinical samples, including a spiked sample and a positive control that were correctly serotyped and genotyped. Besides the negative control, 3 samples did not return typing information due to failing quality checks.

The proportion of DENV reads in the metagenomics samples was highly variable. This

may reflect the viral load in patients in which DENV was detected by real-time RT-PCR. In the spiked sample, containing 4 distinct DENV serotypes, all four were correctly detected despite not being present in equal concentrations, highlighting the potential of the DEN-IM workflow to accurately detect and recover multiple DENV genomes from samples with DENV co-infection, even if the serotypes are present in low abundance. Indeed, recent studies from areas of high endemicity suggest that co-infection with multiple DENV serotypes may frequently occur [27, 28] and the co-circulation of different DENV strains of the same serotype, but distinct genotypes, in these areas [27] raises the possibility of simultaneous infection with more than one genotype.

When analysing the 106 paired-end targeted metagenomics dataset, only 43 CDS samples were de novo assembled. For the remaining 63 samples, consensus sequences were obtained through mapping. In all samples DENV 3-III was correctly identified. Similar results were obtained for the 78 single-end samples where 53 CDS were de novo assembled, and 25 consensus sequences were obtained through mapping. All samples were identified as DENV-1 I. These two datasets demonstrate the success of DEN-IM's two-pronged approach of combining assembler and mapping. DEN-IM's specificity was shown when it found no false positive results when analysing a dataset containing arboviruses other than DENV.

DEN-IM is built with modularity and containerisation as keystones, leveraging the parallelization of processes and guaranteeing reproducible analyses across platforms. The modular design allows for new modules to be easily added and tools that become outdated to be easily updated, ensuring DEN-IM's sustainability. The software versions are also described in the Nextflow script and configuration files, and in the dockerfiles for each container, allowing the traceability of each step of data processing.

Being developed in Nextflow, DEN-IM runs on any UNIX-like system and provides out-of-the-box support for several job schedulers (e.g., PBS, SGE, SLURM) and integration with containerised software like Docker or Singularity. While it has been developed to be ready to use by non-experts, not requiring any software installation or parameter tuning, it can still be easily customised through the configuration files.

The interactive HTML reports (see 4.5) provide an intuitive platform for data exploration, allowing the user to highlight specific samples, filter and re-order the data tables, and export the plots as needed.

Together with the workflow and software containers, a database containing 3858 complete DENV genomes for DENV sequence retrieval and a subset database with 161 curated DENV genomes for serotyping and genotyping are provided. While constructing these databases, the obstacles reported by Cuypers et al [29] were apparent, namely the lack of formal definition of a DENV genotype and the lack of a standardised classification procedure that could assign sequences to a previously defined genotypic/sub-genotypic clade [29]. Discrepancies between the phylogenetic relationship and the genotype assignment were fre-

quent and, throughout this study, the classification of some strains within the ViPR database [30] was updated. As suggested previously [29], further evaluation of the DENV classification will benefit future research and investigation into the population dynamics of this virus. Our typing approach was designed to use the currently accepted DENV classification. However, DEN-IM can be easily modified if a new DENV classification system is to be established in the future.

DEN-IM provides a user-friendly workflow that makes it possible to analyse short-read raw sequencing data from shotgun or targeted metagenomics for the presence, typing and phylogenetic analysis of DENV. The use of containerised workflows, together with shareable reports, will allow an easier comparison of results globally, promoting collaborations that can benefit the populations where DENV is endemic. The DEN-IM source code is freely available in the DEN-IM GitHub repository (`https://github.com/B-UMMI/DEN-IM`), which includes a wiki with full documentation and easy to follow instructions.

4.11 Author Statements

4.11.1 Authors and contributions

C.I.M., E.L., N.C., M.R., J.A.C. and J.W.A.R. designed the workflow. C.I.M implemented and optimised the workflow, created the Docker containers, and wrote the manuscript. M.P.M. implemented the DENV genotyping module in the workflow and D.N.S. contributed to the development of DEN-IM's HTML report. E.L., A. T., and N.C. provided the shotgun metagenomics data used to test and validate the workflow and wrote the manuscript. A.T., N.C., M.R., J.A.C. and J.W.A.R. critically revised the article. All authors read, commented on, and approved the final manuscript.

4.11.2 Conflict of interest

The authors declare that they have no competing interests.

4.11.3 Funding information

C.I.M. was supported by the Fundação para a Ciência e Tecnologia (grant SFRH/BD/129483/2017). Erley Lizarazo received the Abel Tasman Talent Program grant from the UMCG, University of Groningen, Groningen, The Netherlands. This work was partly supported by the ONEIDA project (LISBOA-01-0145-FEDER-016417) co-funded by FEEI–Fundos Europeus Estruturais e de Investimento from Programa Operacional Regional

Lisboa 2020 and by national funds from FCT–Fundação para a Ciência e a Tecnologia and by UID/BIM/50005/2019, project funded by Fundação para a Ciência e a Tecnologia (FCT)/ Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) through Fundos do Orçamento de Estado.

4.11.4 Ethical approval

This study followed international standards for the ethical conduct of research involving human subjects. Data and sample collection was carried out within the DENVEN and IDAMS (International Research Consortium on Dengue Risk Assessment, Management and Surveillance) projects. The study was approved by the Ethics Review Committee of the Biomedical Research Institute, Carabobo University (Aval Bioetico #CBIIB(UC)-014 and CBIIB-(UC)-2013-1), Maracay, Venezuela; the Ethics, Bioethics and Biodiversity Committee (CEBioBio) of the National Foundation for Science, Technology and Innovation (FONACIT) of the Ministry of Science, Technology and Innovation, Caracas, Venezuela; the regional Health authorities of Aragua state (CORPOSALUD Aragua) and Carabobo State (INSALUD); and by the Ethics Committee of the Medical Faculty of Heidelberg University and the Oxford University Tropical Research Ethics Committee.

4.11.5 Consent for publication

All individuals, or a parent or legal guardian if under 16 years of age, whose sample and data were collected have given consent to participate in the study.

4.11.6 Acknowledgements

The authors would like to thank Tiago F. Jesus and Bruno Ribeiro-Gonçalves for their invaluable help with the Nextflow implementation. We would also like to thank Erwin C. Raangs from the UMCG for his assistance in the sequencing of the shotgun metagenomics dataset. Additionally, the authors thank Lize Cuypers, Krystof Theys, Pieter Libin and Gilberto Santiago for their discussions on DENV nomenclature and classification. This work was done in collaboration with the ESCMID Study Group on Molecular and Genomic Diagnostics (ESGMD), Basel, Switzerland.

4.12 Data Bibliography

- Catarina Inês Mendes. DEN-IM supplemental material and tables are deposited at Figshare with DOI 10.6084/m9.figshare.9963812 (https://doi.org/10.6084/m9.figshare.9963812.v3).
- Catarina Inês Mendes. DEN-IM reports for the analysed datasets tables are deposited at Figshare with DOI 0.6084/m9.figshare.9318851 (https://doi.org/10.6084/m9.figshare.9318851).
- Catarina Inês Mendes. Phylogeny inference trees for the dengue virus typing database are deposited at Figshare with DOI 10.6084/m9.figshare.9331826 (https://doi.org/10.6084/m9.figshare.9331826).
- Catarina Inês Mendes. Code for the DEN-IM workflow (https://github.com/B-UMMI/DEN-IM).

4.13 Supplementary Material

4.13.1 Dengue virus reference databases

We have compiled a database of 3858 complete DENV genomes obtained from the NIAID Virus Pathogen Database and Analysis Resource (ViPR) in October 2019 [30] (`http://www.viprbrc.org/`).The sequences were distributed unevenly throughout the four DENV serotypes, with DENV-1 being the most represented with 1636 sequences (42.72%), followed by DENV-2 with 1067 sequences (27.86%), DENV-3 with 807 sequences (21.07%), and DENV-4 with 320 sequences (8.36%). The selection criteria for the search were as follows: a) complete genome sequence only, b) human or mosquito host, c) collection year (1950-2018). Data available from all countries was included and duplicated sequences were removed and only the sequences with sub-type data were kept. A representative of DENV serotype 1 genotype III was introduced (EF457905, recovered from monkey) as no representatives were available with the search criteria used. This genotype is sylvatic and considered extinct [31, 32]. Additionally, any sample with IUPAC codes in the sequence provided were excluded.

In order to recover the maximum number of DENV reads from the input HTS data in the first mapping step (Figure 2.1), we maintained the database with the 3858 complete DENV genomes to retain as much diversity as possible. This database is referred as DENV mapping database and is available on GitHub at `https://github.com/B-UMMI/DEN-IM/blob/master/ref/DENV_MAPPING_V3.fasta`.

For typing purposes, overly similar sequences in the collection were removed from the database by clustering the sequences in each serotype at 98% nucleotide similarity with CD-HIT [33], leaving 161 representative sequences of all described DENV serotypes and genotypes, with 46 DENV-1 sequences (Table 4.8), 63 DENV-2 (Table 4.9), 25 DENV-3 (Tables 4.10) and 27 DENV-4 (Table 4.11). This database is referred as DENV typing database and is available on GitHub at `https://github.com/B-UMMI/DEN-IM/blob/master/ref/DENV_TYPING_V3.fasta`. This step is necessary to speed up the classification step for genotyping.

Phylogenetic analysis of typing collection was performed by aligning the full reference genomes with MAFFT [24], in auto mode and with automatic sequence orientation adjustment. A phylogenetic tree was inferred with RAxML (version 8.12.11) [25] using the GTR-Γ substitution model and 500 times bootstrap. Additionally, the same analysis was performed with the envelope protein (E) only, as this region has been used traditionally for sero- and genotyping [34–40], and continues to be the standard in many laboratories for genotyping. The resulting trees are available as supplemental material (Figures 4.8 to 4.11) and on Figshare (`https://10.6084/m9.figshare.9331826`).

The sequence JF459993 from the DENV-1 collection, as of April 2019, was annotated in ViPR as belonging to genotype IV, but in our analysis, it clustered within genotype I clade (Figure 4.8). The classification of DENV-1 I was also obtained from GenomeDetective Dengue Subtyping Tool (`https://www.genomedetective.com/app/typingtool/dengue/`), so we proceeded to alter the annotation of this particular sample (Table 4.7). In order to harmonise dengue nomenclature, the system uses Roman-numeric labels to identify the genotype, with the exception of Serotype 2 (Table 4.5), which used both Roman-numeric and geographic origin due to the widespread adoption of the latter.

4.13.2 Workflow parameters

The short-read data is passed as input through the “–fastq” parameter, that by default is set to match all files in the “fastq” folder that match the pattern “*_R1,2*”. Both paired and single-end sequencing data can be passed through with the “–fastq” parameter, as defined by the pattern used.

In the process to verify the integrity of the short-read raw sequencing data, the integrity of the input files is assessed by attempting to decompress and read the files. An estimation of the depth of coverage is also performed. By default, the input size ("—genomeSize") is set to 0.012 Mb and the minimum coverage depth ("—minCoverage") is set to 10. If any input file is found to be corrupt, its progression in the workflow is aborted.

In the FastQC and Trimmomatic module, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is run with the parame-

ters "–extract –nogroup –format fastq". FastQC will inform Trimmomatic [41] on how many bases to trim from the 3'and 5' ends of the raw reads. By default, Trimmomatic uses the default set of Illumina adapters provided with the workflow but this behaviour can be overwritten with the "—adapters" parameter. The additional Trimmomatic parameters "-–trimSlidingWindow", "—trimLeading", "—trimTrailing" and "—trimMinLength"can all be set to different values.

The removal of low complexity sequences is done with PrinSeq [17] using a custom parameter ("–pattern"), which by default is set to the value "A 50%; T 50%; N 50%", removing sequences whose content is at least half composed of a polymeric sequence (A, T or N).

To retrieve the reads that map to the DENV reference database, Bowtie2 [18] is run with default parameters with the DENV mapping database as a reference. For paired-end data, the reads and their mates that map to the reference are retrieved with "samtools view -buh -F 12" and "samtools fastq" commands. In single-end reads, all mapped reads are retrieved with "samtools view -buh -F 4" and "samtools fastq". The DENV mapping database can be altered with the "—reference" parameter, or alternatively, a Bowtie2 index can be provided with the "—index" parameter. This allows for the workflow to work with other databases obtained through public and owned DENV genomes. The coverage estimation step is performed on the retrieved DENV reads with the same parameters are the first estimation ("—genomeSize=0.012" and "—minCoverage=10").

In the assembly process, the retrieved DENV reads are firstly assembled with SPAdes Genome Assembler [19] with the options "–careful –only-assembler –cov-cutoff". The coverage cut-off if dictated by the "—spadesMinCoverage" and "—spadesMinKmerCoverage" parameters, set to 2 by default. If the assembly with SPAdes fails to produce a contig equal or greater than the value defined in the "–minimumContigSize" parameter (default of 10000), the data is re-assembled with the MEGAHIT assembler [20] with default parameters. By default, the k-mers to be used in the assembly in both tools ("–spadesKmers" and "–megahitKmers") are automatically determined depending on the read size. If the maximum read length is equal or greater than 175 nucleotides, the assembly is done with the k-mers "55, 77, 99, 113, 127", otherwise the k-mers "21, 33, 55, 67, 77" are used.

To correct the assemblies produced, the Pilon tool [21] is run after mapping the QC'ed reads back to the assembly with Bowtie2 and "samtools sort". This process also verifies the coverage and the number of contigs produced in the assembly. The behaviour can be altered with the parameters "–minAssemblyCoverage", "–AMaxContigs" and "–genomeSize", set to "auto", 1000 and 0.01 Mb by default. The first parameter, when set to 'auto', the minimum assembly coverage for each contig required is set to the 1/3 of the assembly mean coverage or to a minimum of 10x. The ratio of contig number per genome MB is calculated based on the genome size estimation for the samples. The contigs larger than the value defined in the "–size" parameter (default of 10000 nucleotides) are considered to be complete CDSs and follow the rest to the workflow independently. If no complete CDS is recovered, the

QC'ed read data is passed to the mapping to module that does the DENV typing database and consensus generation.

The serotyping and genotyping are performed with the Seq_Typing tool [22] with the command "seq_typing.py assembly" or "seq_typing.py reads", using as reference the provided curated DENV typing database. It is possible to retrieve the genomes of the closest references and include them in the downstream analysis by changing the "–get_reference" option to "true". By default, this is not included in the analysis.

The CDSs, and the reference sequences if requested, are aligned with the MAFFT tool [24] with the options "–adjustdirection –auto". By default, four representative sequences for each DENV serotype (1 to 4) from NCBI is also included in the alignment. This option can be turned off by changing the value of "–includeNCBI" to "false". If the number of sequences in the alignment is less than 4 these are automatically added.

A maximum likelihood phylogenetic tree is obtained with the RaXML tool [25] with the options "-p 12345 -f -a". Additionally, and by default, the substitution model ("–substitutionModel") is set to "GTRGAMMA", the bootstrap is set to 500 ("–bootstrap") and the seed to "12345" ("–seedNumber").

4.13.3 Shotgun Metagenomics Sequencing Data

Samples of plasma (n=9) and serum samples (n=13) from confirmed dengue symptomatic patients were collected in Venezuela between 2010-2015 (Table S2) (see Availability of supporting materials). DENV positivity was confirmed by either RT-qPCR [42] or nested RT-PCR [36].

As a positive control sample, the supernatant of a viral culture containing DENV-2 strain 16681 was used. The negative control sample consisted of DNA- and RNA-free water (Sigma-Aldrich, St. Louis, MO, USA).

A spiked sample was produced consisting of a mixture of four 5 µl of cDNA isolated from clinical samples including all DENV serotypes (DENV-1 to -4). The viral cDNA for these samples was not in equal concentration and the viral copy number in the clinical samples was assessed by RT-PCR [36]. The results were as follow: DENV-2 with 1070000 copies/µl, DENV-1 with 117830 copies/µl, DENV-3 with 44300 copies/µl and DENV-4 with 6600 copies/µl.

The cDNA libraries were generated using either the NEBNext® RNA First and Second strand modules and the Nextera XT DNA library preparation kit (NXT), or the TruSeq RNA V2 library preparation kit (TS). The libraries were sequenced in MiSeq and NextSeq instruments using 300-cycles v2 paired-end cartridges.

The DEN-IM workflow was executed with the raw sequencing data using the default parameters and resources in an HPC cluster with 300 Cores/600 Threads of Processing Power and 3 TB RAM divided through 15 computational nodes, 9 with 254 GB Ram and 6 with 126GB RAM.

4.13.4 Amplicon Sequencing Data

The accession numbers for the 106 DENV-3 paired-end amplicon sequencing paired-end short-read datasets are available under BioProject PRJNA394021. The accession numbers for the 78 DENV-1 amplicon sequencing single-end short-read datasets are available under BioProject PRJNA321963. The Run Accession IDs for both sets were obtained with NCBI's RunSelector and the raw data was downloaded with the GetSeqENA tool (https://github.com/B-UMMI/getSeqENA).

The DEN-IM workflow was executed with the raw sequencing data with default parameters and resources in the same HPC cluster as the shotgun metagenomics dataset.

4.13.5 Non-DENV Arbovirus Data

The accession numbers for the 132 samples, belonging to zika virus (ZKV), chikungunya virus (CHIKV) and yellow fever virus (YFV) amplicon and metagenomic datasets are available as supplemental material (Table S4). As with the amplicon sequencing dataset, the list of Run Accession IDs was obtained with NCBI's RunSelector and the raw data was downloaded with the GetSeqENA tool (https://github.com/B-UMMI/getSeqENA).

The DEN-IM workflow was executed with default parameters and resources in the same HPC cluster as the amplicon and shotgun metagenomics datasets.

4.13.6 Supplemental Tables

Table 4.2: Collection date, serotype confirmation and run accession identifier for the metagenomic sequencing dataset.

Sample	Collection Date	Source	Serotype (qPCR)	Serotype	Genotype	Run Accession
91-0104	21/9/2015	plasma	2	2	III(AsianAmerican)	SRR8842525
91-0105	22/9/2015	plasma	2	2	III(AsianAmerican)	SRR7252349
91-0115	30/9/2015	plasma	3	-	-	SRR7252368
91-0118	5/10/2015	plasma	3	3	III	SRR7252362
91-0132	19/10/2015	plasma	1	1	V	SRR8883926
91-0135	27/10/2015	plasma	2	2	III(AsianAmerican)	SRR9004764
92-1001	2/10/2015	plasma	1	-	-	SRR7252337
92-1094	16/10/2015	plasma	2	2	III(AsianAmerican)	SRR8842524
CC0007	31/8/2010	serum	2	2	III(AsianAmerican)	SRR7252354
CC0009	31/8/2010	serum	3	3	III	SRR8842527
CC0010	27/8/2010	serum	3	3	III	SRR7252358
CC0011	27/8/2010	serum	3	3	III	SRR8842526
CC0030a	1/9/2010	serum	4	4	II	SRR7252356
CC0030b	1/9/2010	serum	4	4	II	SRR7252355
CC0031	2/9/2010	serum	2	2	III(AsianAmerican)	SRR8842521
CC0061	20/1/2011	serum	4	4	II	SRR8842520
CC0066	11/10/2011	serum	4	4	II	SRR8842523
CC0067	18/10/2011	serum	4	4	II	SRR8842522
CC0116	29/3/2012	serum	4	4	II	SRR8842519
CC0150	9/5/2012	serum	2	2	III(AsianAmerican)	SRR8842518
CC0186	17/7/2012	serum	4	4	II	SRR9004763
UCUG0186	30/8/2010	serum	4	4	II	SRR8842528
Negative Control	-	-	-	-	-	SRR8842530
Positive Control	-	-	2	2	V(AsianI)	SRR8886136
Spiked sample	-	-	1,2,3,4	1,2,3,4	V,III(Asian American),III,II	SRR8842529

Table 4.3: Run accession ID, BioProject SRA Study ID, source and organism present for each sample of the negative control dataset (ZKV – zika virus, CHIKV – chikungunya virus, YFV – yellow fever virus).

Run ID	Bioproject	SRA Study	Source	Organism
SRR8031152	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR8062732	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR8031153	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR8063606	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR8063603	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR8063605	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR8031155	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR8031154	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR8063604	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR8062733	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR7985391	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR7985394	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR7985620	PRJNA494391	SRP163225	Shotgun Metagenomic	CHIKV
SRR7985390	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR7985392	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR7985621	PRJNA494391	SRP163225	Shotgun Metagenomic	CHIKV
SRR5179639	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179637	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV

SRR5179646	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR7985389	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR7985622	PRJNA494391	SRP163225	Shotgun Metagenomic	CHIKV
SRR7985619	PRJNA494391	SRP163225	Shotgun Metagenomic	CHIKV
SRR5179667	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179653	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR7985393	PRJNA494391	SRP163225	Shotgun Metagenomic	ZKV
SRR5179638	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179636	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179666	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179650	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179649	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179643	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179635	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179645	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179642	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179644	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179647	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179641	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179640	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179652	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179648	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR5179651	PRJNA361543	SRP096859	Amplicon Metagenomics	YFV
SRR9020503	PRJNA541092	SRP195668	Amplicon Metagenomics	CHIKV
SRR9020505	PRJNA541093	SRP195669	Amplicon Metagenomics	CHIKV
SRR9020506	PRJNA541094	SRP195670	Amplicon Metagenomics	CHIKV
SRR9020509	PRJNA541095	SRP195671	Amplicon Metagenomics	CHIKV
SRR9020511	PRJNA541096	SRP195672	Amplicon Metagenomics	CHIKV
SRR9020513	PRJNA541097	SRP195673	Amplicon Metagenomics	CHIKV
SRR9020514	PRJNA541098	SRP195674	Amplicon Metagenomics	CHIKV
SRR9020516	PRJNA541099	SRP195675	Amplicon Metagenomics	CHIKV
SRR9020518	PRJNA541100	SRP195676	Amplicon Metagenomics	CHIKV
SRR9020520	PRJNA541101	SRP195677	Amplicon Metagenomics	CHIKV
SRR9020521	PRJNA541102	SRP195678	Amplicon Metagenomics	CHIKV
SRR9020523	PRJNA541103	SRP195679	Amplicon Metagenomics	CHIKV
SRR9020525	PRJNA541104	SRP195680	Amplicon Metagenomics	CHIKV
SRR9020527	PRJNA541105	SRP195681	Amplicon Metagenomics	CHIKV
SRR9020529	PRJNA541106	SRP195682	Amplicon Metagenomics	CHIKV
SRR9020530	PRJNA541107	SRP195683	Amplicon Metagenomics	CHIKV
SRR9020532	PRJNA541108	SRP195684	Amplicon Metagenomics	CHIKV
SRR9020534	PRJNA541109	SRP195685	Amplicon Metagenomics	CHIKV
SRR9020537	PRJNA541110	SRP195686	Amplicon Metagenomics	CHIKV

SRR9020539	PRJNA541111	SRP195687	Amplicon Metagenomics	CHIKV
SRR9020541	PRJNA541112	SRP195688	Amplicon Metagenomics	CHIKV
SRR9020542	PRJNA541113	SRP195689	Amplicon Metagenomics	CHIKV
SRR9020504	PRJNA541114	SRP195690	Amplicon Metagenomics	CHIKV
SRR9020507	PRJNA541115	SRP195691	Amplicon Metagenomics	CHIKV
SRR9020508	PRJNA541116	SRP195692	Amplicon Metagenomics	CHIKV
SRR9020510	PRJNA541117	SRP195693	Amplicon Metagenomics	CHIKV
SRR9020512	PRJNA541118	SRP195694	Amplicon Metagenomics	CHIKV
SRR9020515	PRJNA541119	SRP195695	Amplicon Metagenomics	CHIKV
SRR9020517	PRJNA541120	SRP195696	Amplicon Metagenomics	CHIKV
SRR9020519	PRJNA541121	SRP195697	Amplicon Metagenomics	CHIKV
SRR9020522	PRJNA541122	SRP195698	Amplicon Metagenomics	CHIKV
SRR9020524	PRJNA541123	SRP195699	Amplicon Metagenomics	CHIKV
SRR9020526	PRJNA541124	SRP195700	Amplicon Metagenomics	CHIKV
SRR9020528	PRJNA541125	SRP195701	Amplicon Metagenomics	CHIKV
SRR9020531	PRJNA541126	SRP195702	Amplicon Metagenomics	CHIKV
SRR9020533	PRJNA541127	SRP195703	Amplicon Metagenomics	CHIKV
SRR9020535	PRJNA541128	SRP195704	Amplicon Metagenomics	CHIKV
SRR9020536	PRJNA541129	SRP195705	Amplicon Metagenomics	CHIKV
SRR9020538	PRJNA541130	SRP195706	Amplicon Metagenomics	CHIKV
SRR9020540	PRJNA541131	SRP195707	Amplicon Metagenomics	CHIKV
SRR7369225	PRJNA47661	SRP150883	Shotgun Metagenomic	ZKV
SRR7369226	PRJNA47661	SRP150883	Shotgun Metagenomic	ZKV
SRR6505781	PRJNA431343	SRP131290	Shotgun Metagenomic	ZKV
SRR8260975	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8260976	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8260977	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8260978	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8260979	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8260980	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261322	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261325	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261326	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261329	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261330	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261331	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261332	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261333	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261335	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261336	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261338	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261341	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV

4. DEN-IM: DENGUE VIRUS GENOTYPING FROM SHOTGUN AND TARGETED METAGENOMICS

SRR8261342	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261343	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261345	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261346	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261347	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261348	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261352	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261353	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261354	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261355	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261356	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261359	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261360	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261361	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261362	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261364	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261365	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261366	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261367	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261369	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261402	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261404	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261407	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261411	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261412	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261413	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261415	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261416	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV
SRR8261417	PRJNA314889	SRP162155	Shotgun Metagenomic	ZKV

Table 4.4: Number of raw base pairs, overall alignment rate against the DENV mapping database, estimated coverage depths and serotype and genotype for 25 shotgun metagenomics sequencing samples.

Sample	Raw Megabases	% DENV Reads	Estimated coverage depth (times)	Serotype	Genotype
91-0104	2193.71	12.46	5944.67	2	III (AsianAmerican)
91-0105	191.37	4.01	495.97	2	III (AsianAmerican)
91-0115	179.24	0.05	3.74	-	-
91-0118	195.27	1.69	86.53	3	III
91-0132	378.21	20.02	4698.12	1	V
91-0135	91.71	21.45	1287.52	2	III (AsianAmerican)
92-1001 a)	163.44	-	-	-	-
92-1094	1197.92	8.48	4032.21	2	III (AsianAmerican)
CC0007	252.97	3.79	383.77	2	III (AsianAmerican)
CC0009	2055.13	9.48	8226.27	3	III
CC0010	368.64	5.68	1197.58	3	III
CC0011	924.69	8.38	3016.17	3	III
CC0030a	261.12	52.52	2914.87	4	II
CC0030b	399.04	10.51	677.96	4	II
CC0031	1572.1	68.91	52318.33	2	III (AsianAmerican)
CC0061	1262.83	8.97	5120.4	4	II
CC0066	1087.45	2.8	569.7	4	II
CC0067	1022.06	5.55	2548.84	4	II
CC0116	773.31	6.72	2313.99	4	II
CC0150	1403.69	17.41	12065.81	2	III (AsianAmerican)
CC0186	671.78	0.03	14.71	4	II
UCUG0186 b)	1116.67	0.01	5.65	-	-
Negative Control a)	163.67	-	-	-	-
Positive Control	443.93	85.38	19362.07	2	V (Asian I)
				3	III
Spike	1518.93	41.7	22289.98	1	V
				2	III (AsianAmerican)
				4	II

a) Failed quality control - No sequence data after quality trimming.

b) Failed quality control - Low sequence depth (<10x).

Table 4.5: Number of raw base pairs, overall alignment rate, in percentage, for the mapping against the DENV database, number of ORFs recovered, and respective serotype and genotype for 106 paired-end amplicon sequencing samples.

Sample	Raw Megabases	% DENV DNA	CDS Assembly	Serotype	Genotype
SRR5821157	439.35	82.56	consensus	3	III
SRR5821158	77.34	85.19	consensus	3	III
SRR5821159	68.00	91.11	consensus	3	III
SRR5821160	119.54	97.77	consensus	3	III
SRR5821161	53.40	92.76	consensus	3	III
SRR5821162	49.59	99.39	consensus	3	III
SRR5821163	66.43	97.78	consensus	3	III
SRR5821164	69.96	99.18	consensus	3	III
SRR5821165	75.48	98.38	consensus	3	III
SRR5821166	38.99	62.03	de novo	3	III
SRR5821167	73.15	49.19	de novo	3	III
SRR5821168	49.59	99.63	consensus	3	III
SRR5821169	119.39	99.74	de novo	3	III

SRR5821170	61.45	99.09	consensus	3	III
SRR5821171	61.63	98.92	consensus	3	III
SRR5821172	69.86	98.96	de novo	3	III
SRR5821173	80.37	97.59	de novo	3	III
SRR5821174	37.58	76.69	de novo	3	III
SRR5821175	112.70	75.55	de novo	3	III
SRR5821176	139.34	99.03	de novo	3	III
SRR5821177	41.19	44.56	de novo	3	III
SRR5821178	59.03	81.06	de novo	3	III
SRR5821179	95.59	84.7	de novo	3	III
SRR5821180	48.75	98.15	consensus	3	III
SRR5821181	64.45	99.3	consensus	3	III
SRR5821182	64.40	98.88	consensus	3	III
SRR5821183	115.14	95.61	consensus	3	III
SRR5821184	170.72	94.11	de novo	3	III
SRR5821185	181.75	98.19	de novo	3	III
SRR5821186	246.98	96.4	de novo	3	III
SRR5821187	55.62	99.74	consensus	3	III
SRR5821188	70.95	99.39	consensus	3	III
SRR5821189	82.61	99.27	de novo	3	III
SRR5821190	138.58	98.81	consensus	3	III
SRR5821191	59.92	99.72	de novo	3	III
SRR5821192	40.53	36.88	consensus	3	III
SRR5821193	92.08	98.9	de novo	3	III
SRR5821194	58.69	98.53	consensus	3	III
SRR5821195	127.80	99.64	consensus	3	III
SRR5821196	59.30	86.62	de novo	3	III
SRR5821197	87.78	99.47	de novo	3	III
SRR5821198	185.55	99.72	de novo	3	III
SRR5821199	83.55	99.62	consensus	3	III
SRR5821200	85.52	99.5	consensus	3	III
SRR5821201	129.77	94.6	consensus	3	III
SRR5821202	56.60	99.81	consensus	3	III
SRR5821203	80.28	99.22	consensus	3	III
SRR5821204	68.46	95.52	de novo	3	III
SRR5821205	44.45	98.53	consensus	3	III
SRR5821206	43.67	97.88	consensus	3	III
SRR5821207	78.93	99.22	de novo	3	III
SRR5821208	87.45	97.72	consensus	3	III
SRR5821209	73.40	94.16	de novo	3	III
SRR5821210	55.86	91.35	de novo	3	III
SRR5821211	75.53	85.6	consensus	3	III

SRR5821212	98.89	99.09	de novo	3	III
SRR5821213	84.85	95.03	de novo	3	III
SRR5821214	15.33	96.28	de novo	3	III
SRR5821215	13.08	96.74	consensus	3	III
SRR5821216	45.07	98.85	de novo	3	III
SRR5821217	161.65	88.94	consensus	3	III
SRR5821218	51.09	95.29	consensus	3	III
SRR5821219	84.68	99.1	de novo	3	III
SRR5821220	88.26	82.64	de novo	3	III
SRR5821221	64.76	86.62	de novo	3	III
SRR5821222	93.47	97.48	consensus	3	III
SRR5821223	86.50	98.99	de novo	3	III
SRR5821224	73.31	26.43	consensus	3	III
SRR5821225	68.85	98.43	consensus	3	III
SRR5821226	67.75	96.67	consensus	3	III
SRR5821227	32.56	99.54	de novo	3	III
SRR5821228	38.73	86.68	consensus	3	III
SRR5821229	77.18	99.69	consensus	3	III
SRR5821230	175.73	99.58	de novo	3	III
SRR5821231	100.82	99.58	de novo	3	III
SRR5821232	86.89	99.47	consensus	3	III
SRR5821233	270.15	99.56	consensus	3	III
SRR5821234	76.07	99.75	consensus	3	III
SRR5821235	32.78	79.78	consensus	3	III
SRR5821236	80.19	24.72	de novo	3	III
SRR5821237	50.59	97.38	consensus	3	III
SRR5821238	63.56	97.63	de novo	3	III
SRR5821239	29.66	41.15	consensus	3	III
SRR5821240	62.61	94.64	de novo	3	III
SRR5821241	17.52	98.03	consensus	3	III
SRR5821242	58.86	99.25	consensus	3	III
SRR5821243	50.08	93.56	consensus	3	III
SRR5821244	32.67	99.09	consensus	3	III
SRR5821245	64.96	99.77	consensus	3	III
SRR5821246	104.11	90.14	consensus	3	III
SRR5821247	98.64	99.73	consensus	3	III
SRR5821248	129.28	90.73	consensus	3	III
SRR5821249	45.76	93.13	de novo	3	III
SRR5821250	72.54	98.88	de novo	3	III
SRR5821251	115.85	97.7	consensus	3	III
SRR5821252	60.76	94	consensus	3	III
SRR5821253	64.45	99.66	consensus	3	III

SRR5821254	0.27	98.12	consensus	3	III
SRR5821255	62.53	99.55	de novo	3	III
SRR5821256	54.57	99.58	consensus	3	III
SRR5821257	34.90	99.53	de novo	3	III
SRR5821258	68.64	99.6	consensus	3	III
SRR5821259	73.04	98.8	consensus	3	III
SRR5821260	54.60	99.14	consensus	3	III
SRR5821261	55.54	95.5	de novo	3	III
SRR5821262	106.05	91.78	consensus	3	III

Table 4.6: Taxonomic profiling results for the amplicon sequencing samples with less than 70% DENV DNA.

Sample	**Bowtie2**	**Kraken2 (minikraken2_v2 DB)**		
	DENV (%)	**Unclassified**	**Homo sapie**	**DENV (%)**
SRR5821236	24.72	5.47	71.61	19.63
SRR5821224	26.43	7.01	71.06	19.58
SRR5821192	36.88	8.12	61.78	28.73
SRR5821239	41.15	8.29	56.43	33.84
SRR5821167	49.19	14.79	50.16	34.38
SRR5821166	62.03	13.72	37.77	47.97

Table 4.7: Number of raw base pairs, overall alignment rate, in percentage, for the mapping against the DENV database, number of ORFs recovered, and respective serotype and genotype for 78 single-end amplicon sequencing samples.

Sample	Raw Megabases	% DENV DNA	CDS Assembly	Serotype	Genotype
SRR3539340	330365175	83.7	consensus	I	1
SRR3539341	317977866	66.56	consensus	I	1
SRR3539342	406075245	74.2	consensus	I	1
SRR3539343	302220886	59.24	de novo	I	1
SRR3539344	424801129	83.21	de novo	I	1
SRR3539345	345821429	92.58	de novo	I	1
SRR3539346	411918039	90.92	de novo	I	1
SRR3539347	411031278	90.92	de novo	I	1
SRR3539348	469139944	92.45	de novo	I	1
SRR3539349	537372466	90.77	de novo	I	1
SRR3539350	401844325	90.32	de novo	I	1
SRR3539351	401993816	89.76	de novo	I	1
SRR3539352	357846693	88.48	consensus	I	1
SRR3539353	412322289	82.94	de novo	I	1
SRR3539354	398022772	86.07	consensus	I	1
SRR3539355	388552807	92.43	consensus	I	1
SRR3539357	351745878	88.51	consensus	I	1
SRR3539358	398098393	70.74	de novo	I	1
SRR3539359	479640173	91.59	consensus	I	1
SRR3539360	374570187	75.78	de novo	I	1
SRR3539361	370202077	72.56	de novo	I	1
SRR3539362	402201658	83.22	consensus	I	1
SRR3539363	467055595	75.85	consensus	I	1
SRR3539364	312321789	65.93	de novo	I	1
SRR3539365	253871159	88.37	de novo	I	1
SRR3539366	246292055	82.12	consensus	I	1
SRR3539367	228721211	86.98	de novo	I	1
SRR3539368	253255975	89.84	de novo	I	1
SRR3539369	254904463	91.39	de novo	I	1
SRR3539370	256094646	89.77	de novo	I	1
SRR3539371	266981417	93.77	de novo	I	1
SRR3539372	195098066	82.5	consensus	I	1
SRR3539373	237636237	84.54	consensus	I	1
SRR3539374	202624880	91.98	de novo	I	1
SRR3539375	399641302	87.95	consensus	I	1
SRR3539376	209424800	92.58	de novo	I	1
SRR3539377	278160288	89.75	de novo	I	1
SRR3539378	328706147	87.33	de novo	I	1

SRR3539379	370640534	88.93	de novo	I	1
SRR3539380	313475971	66.56	de novo	I	1
SRR3539381	327213068	89.39	de novo	I	1
SRR3539382	295317021	78.49	de novo	I	1
SRR3539383	335941236	81.98	consensus	I	1
SRR3539384	383785104	90.79	de novo	I	1
SRR3539385	330006204	88.86	de novo	I	1
SRR3539386	454412182	87.3	de novo	I	1
SRR3539387	321847824	92.31	de novo	I	1
SRR3539388	354652844	92.31	de novo	I	1
SRR3539389	345354321	88.38	consensus	I	1
SRR3539390	412365081	84.83	de novo	I	1
SRR3539391	374367976	84.5	de novo	I	1
SRR3539393	428999734	82.5	de novo	I	1
SRR3539394	323218873	91.91	de novo	I	1
SRR3539395	375283202	91.7	de novo	I	1
SRR3539396	434756338	84.96	de novo	I	1
SRR3539397	361928373	93.07	de novo	I	1
SRR3539398	462599218	80.7	consensus	I	1
SRR3539399	379115053	86.74	de novo	I	1
SRR3539400	404747525	93.34	consensus	I	1
SRR3539401	327849624	94.64	consensus	I	1
SRR3539406	209992112	78.25	de novo	I	1
SRR3539407	370290249	91.86	consensus	I	1
SRR3539408	191269315	95.58	de novo	I	1
SRR3539409	398058055	91.69	de novo	I	1
SRR3539410	393229460	94.48	de novo	I	1
SRR3539411	387469496	93.21	consensus	I	1
SRR3539412	53752250	82.32	de novo	I	1
SRR3539413	347547808	85.47	de novo	I	1
SRR3539414	355980530	84.16	consensus	I	1
SRR3539415	364109410	92.25	consensus	I	1
SRR3539416	341121914	86.11	consensus	I	1
SRR3539417	339098553	84.5	de novo	I	1
SRR3539418	332640627	85.66	de novo	I	1
SRR3539419	360466242	85.65	de novo	I	1
SRR3539420	415554748	84.23	de novo	I	1
SRR3539421	322411348	93.42	de novo	I	1
SRR3539422	387614239	82.17	consensus	I	1
SRR3539424	446656613	84.25	de novo	I	1

Table 4.8: Representative sequences of serotype 1 diversity in the Dengue Virus Typing Database.

Sample	ViPR Classification	Origin	Collection Year
EU482591	DENV-1 V	USA	2006
KU509254	DENV-1 V	Venezuela	2011
MF004384	DENV-1 V	France	2014
GU131956	DENV-1 V	Mexico	2006
AF311956	DENV-1 V	Brazil	1997
FJ205874	DENV-1 V	USA	1995
FJ478457	DENV-1 V	USA	1996
EU482567	DENV-1 V	USA	1998
DQ285559	DENV-1 V	Reunion	2004
JN903578	DENV-1 V	India	2007
KP188548	DENV-1 V	Brazil	2013
JQ922544	DENV-1 V	India	1963
KX380796	DENV-1 V	Singapore	2012
JQ922548	DENV-1 V	India	2005
KP406801	DENV-1 V	South Korea	2004
DQ285562	DENV-1 V	Comoros	1993
JQ922546	DENV-1 V	India	1971
EF457905	DENV-1 III	Malaysia	1972
AF180818	DENV-1 II	Unknown	Unknown
JQ922547	DENV-1 II	Thailand	1960
KY496855	DENV-1 IV	Taiwan	2016
LC128301	DENV-1 IV	Philippines	2016
KX951689	DENV-1 IV	Taiwan	2004
KC762653	DENV-1 IV	Indonesia	2008
KU509261	DENV-1 IV	Indonesia	2010
AB189121	DENV-1 IV	Indonesia	1998
KC762620	DENV-1 IV	Indonesia	2007
EU863650	DENV-1 IV	Chile	2002
AB195673	DENV-1 IV	Japan	2003
AB204803	DENV-1 IV	Japan	2004
JF459993	DENV-1 I	Myanmar	2002
KT827371	DENV-1 I	China	2014
KX620454	DENV-1 I	China	2014
FJ639670	DENV-1 I	Cambodia	2001
KU509250	DENV-1 I	Thailand	2012
KJ755855	DENV-1 I	India	2013
GU131678	DENV-1 I	Viet Nam	2008
KU509265	DENV-1 I	Unknown	2012
KF955446	DENV-1 I	Viet Nam	2008
JF937615	DENV-1 I	Viet Nam	2008

FJ639678	DENV-1 I	Cambodia	2003
EU660395	DENV-1 I	Viet Nam	2007
AB608789	DENV-1 I	Taiwan	1994
GQ868636	DENV-1 I	Cambodia	2008
KY586539	DENV-1 I	Thailand	1995
KU509258	DENV-1 I	Eritrea	2010

Table 4.9: Representative sequences of serotype 2 diversity in the Dengue Virus Typing Database.

Sample	ViPR Classification	Origin	Collection Year
HQ705624	DENV-2 III (AsianAmerican)	Nicaragua	2009
KY977454	DENV-2 III (AsianAmerican)	Panama	2011
KY474330	DENV-2 III (AsianAmerican)	Ecuador	2014
FJ024473	DENV-2 III (AsianAmerican)	Colombia	2005
JX669476	DENV-2 III (AsianAmerican)	Brazil	2010
JN819419	DENV-2 III (AsianAmerican)	Brazil	2000
KF955364	DENV-2 III (AsianAmerican)	Puerto Rico	2006
JX669480	DENV-2 III (AsianAmerican)	Brazil	1995
FJ639699	DENV-2 III (AsianAmerican)	Cambodia	2002
EU482449	DENV-2 III (AsianAmerican)	Viet Nam	2006
EU482778	DENV-2 III (AsianAmerican)	Viet Nam	2003
KY586692	DENV-2 V (AsianI)	Thailand	2001
KY586679	DENV-2 V (AsianI)	Thailand	2001
KY586571	DENV-2 V (AsianI)	Thailand	2006
KY586572	DENV-2 V (AsianI)	Thailand	2006
EU726767	DENV-2 V (AsianI)	Thailand	1994
GQ868591	DENV-2 V (AsianI)	Thailand	1964
KF704356	DENV-2 IV (AsianII)	Cuba	1981
JQ922552	DENV-2 I (American)	India	1960
KJ918750	DENV-2 I (American)	India	2007
JQ922553	DENV-2 I (American)	India	1980
GQ868592	DENV-2 I (American)	Colombia	1986
JX966379	DENV-2 I (American)	Mexico	1994
GQ398257	DENV-2 I (American)	Indonesia	1977
KY923048	DENV-2 VI (Sylvatic)	Malaysia	2015
JF260983	DENV-2 VI (Sylvatic)	Spain	2009
KY937189	DENV-2 II (Cosmopolitan)	China	2015
KY937188	DENV-2 II (Cosmopolitan)	China	2015
KY937187	DENV-2 II (Cosmopolitan)	China	2015
JQ955624	DENV-2 II (Cosmopolitan)	India	2011
KU509271	DENV-2 II (Cosmopolitan)	India	2006
KF041232	DENV-2 II (Cosmopolitan)	Pakistan	2011

JQ922551	DENV-2 II (Cosmopolitan)	India	2005
JX475906	DENV-2 II (Cosmopolitan)	India	2009
MG779194	DENV-2 II (Cosmopolitan)	Kenya	2017
FJ882602	DENV-2 II (Cosmopolitan)	Sri Lanka	1996
EU056810	DENV-2 II (Cosmopolitan)	Burkina Faso	1983
KY627763	DENV-2 II (Cosmopolitan)	Burkina Faso	2016
KM279515	DENV-2 II (Cosmopolitan)	Singapore	2011
KX452015	DENV-2 II (Cosmopolitan)	Malaysia	2014
KC762662	DENV-2 II (Cosmopolitan)	Indonesia	2007
KU509270	DENV-2 II (Cosmopolitan)	Unknown	2012
KP012546	DENV-2 II (Cosmopolitan)	China	2014
KX452034	DENV-2 II (Cosmopolitan)	Malaysia	2014
KX452048	DENV-2 II (Cosmopolitan)	Malaysia	2014
KX452044	DENV-2 II (Cosmopolitan)	Malaysia	2014
HM488257	DENV-2 II (Cosmopolitan)	Guam	2001
KU509277	DENV-2 II (Cosmopolitan)	Philippines	2010
KU509269	DENV-2 II (Cosmopolitan)	Philippines	2009
KU509274	DENV-2 II (Cosmopolitan)	Philippines	2010
GQ398263	DENV-2 II (Cosmopolitan)	Indonesia	1975

Table 4.10: Representative sequences of serotype 3 diversity in the Dengue Virus Typing Database.

Sample	ViPR Classification	Origin	Collection Year
KF954946	DENV-3-III	China	2013
JQ922557	DENV-3 III	India	2005
KU509286	DENV-3 III	India	2011
EU687233	DENV-3 III	USA	2002
GQ252674	DENV-3 III	Sri Lanka	1997
FJ882573	DENV-3 III	Sri Lanka	1993
GQ199887	DENV-3 III	Sri Lanka	1983
JQ922555	DENV-3 III	India	1966
HM631854	DENV-3 II	Cambodia	2008
KY586703	DENV-3 II	Thailand	2006
KU509280	DENV-3 II	Thailand	2011
FJ744730	DENV-3 II	Thailand	2001
KY586814	DENV-3 II	Thailand	2006
DQ863638	DENV-3 II	Thailand	1973
KC762684	DENV-3 I	Indonesia	2007
KY863456	DENV-3 I	Indonesia	2016
KC762691	DENV-3 I	Indonesia	2008
KC762692	DENV-3 I	Indonesia	2010
KY794787	DENV-3 I	Papua New Guinea	2007

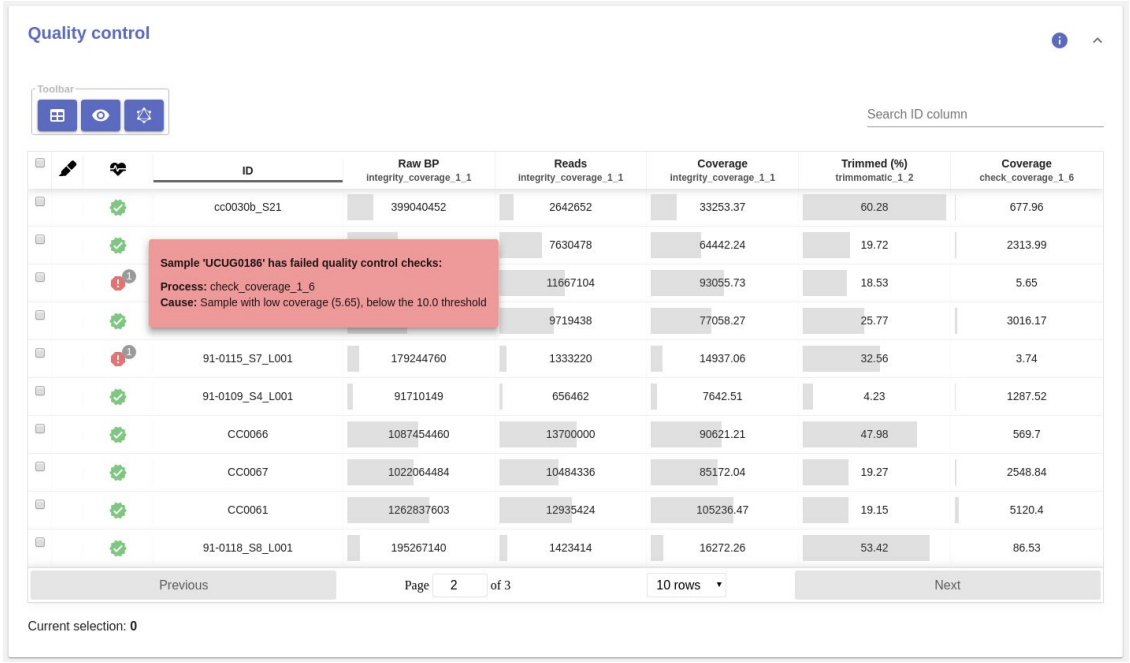
MF004386	DENV-3 I	Malaysia	2012
AB189128	DENV-3 I	Indonesia	1998
KU509279	DENV-3 I	Philippines	2008
FJ898455	DENV-3 I	Cook Islands	1991
KU725666	DENV-3 V	Unkown	Unknown

Table 4.11: Representative sequences of serotype 4 diversity in the Dengue Virus Typing Database.

Sample	ViPR Classification	Origin	Collection Year
MG601754	DENV-4 I	China	2013
KY586839	DENV-4 I	Thailand	1995
KT026308	DENV-4 I	Thailand	2011
JN638572	DENV-4 I	Cambodia	2008
KY586942	DENV-4 I	Thailand	2006
KP792537	DENV-4 I	Singapore	2011
MG272273	DENV-4 I	India	2016
MG272272	DENV-4 I	India	2016
KU509287	DENV-4 I	India	2009
JQ922559	DENV-4 I	India	1979
GQ868594	DENV-4 I	Philippines	1956
JQ922558	DENV-4 I	India	1962
KU523872	DENV-4 II	Indonesia	2015
KP723482	DENV-4 II	China	2010
JX024757	DENV-4 II	Singapore	2010
KC762695	DENV-4 II	Indonesia	2007
JQ915088	DENV-4 II	New Caledonia	2009
GQ398256	DENV-4 II	Singapore	2005
KP188557	DENV-4 II	Brazil	2012
KY474335	DENV-4 II	Ecuador	2014
KT276273	DENV-4 II	Haiti	2014
KF907503	DENV-4 II	Senegal	1953
KY586945	DENV-4 III	Thailand	1998
JF262779	DENV-4 IV	Malaysia	1975

4.13.7 Supplemental Figures

a)



ID	Raw BP integrity_coverage_1_1	Reads integrity_coverage_1_1	Coverage integrity_coverage_1_1	Trimmed (%) trimmomatic_1_2	Coverage check_coverage_1_6
cc0030b_S21	399040452	2642652	33253.37	60.28	677.96
		7630478	64442.24	19.72	2313.99
		11667104	93055.73	18.53	5.65
		9719438	77058.27	25.77	3016.17
91-0115_S7_L001	179244760	1333220	14937.06	32.56	3.74
91-0109_S4_L001	91710149	656462	7642.51	4.23	1287.52
CC0066	1087454460	13700000	90621.21	47.98	569.7
CC0067	1022064484	10484336	85172.04	19.27	2548.84
CC0061	1262837603	12935424	105236.47	19.15	5120.4
91-0118_S8_L001	195267140	1423414	16272.26	53.42	86.53

b)

ID	seqtyping dengue_typing_assembly_1_11	Identity dengue_typing_assembly_1_1	Coverage dengue_typing_assembly_1_11	Reference dengue_typing_assembly_1_11
Spike_NODE_3_length_10199_cov_229.022822_pilon	1-V	98.03	100	gb:EU482591
91-0132_S6_L001_NODE_1_length_10217_cov_2041.464103_pilon	1-V	98.03	100	gb:EU482591
CC0031_k77_16_flag_0_multi_50991.9804_len_10065_pilon	2-III(AsianAmerican)	99.21	98.95	gb:FJ024473
cc0007_S5_L001_NODE_1_length_10200_cov_119.535810_pilon	2-III(AsianAmerican)	99.22	100	gb:FJ024473
91-0105_S2_L001_NODE_1_length_10207_cov_218.928825_pilon	2-III(AsianAmerican)	98.72	100	gb:FJ024473
Spike_NODE_4_length_10192_cov_76.477014_pilon	2-III(AsianAmerican)	98.66	100	gb:FJ024473
CC0150_NODE_1_length_10242_cov_3878.632858_pilon	2-III(AsianAmerican)	99.13	100	gb:FJ024473
91-0109_S4_L001_NODE_1_length_10219_cov_652.125222_pilon	2-III(AsianAmerican)	98.86	100	gb:FJ024473
91-0104_NODE_1_length_10181_cov_326.327573_pilon	2-III(AsianAmerican)	98.72	100	gb:FJ024473
92-1094_NODE_1_length_10194_cov_816.395572_pilon	2-III(AsianAmerican)	98.67	100	gb:FJ024473
Poditivecontrol_S21_L001_k77_1_flag_1_multi_18626.0847_len_10237_pilon	2-V(Asian!)	100	100	gb:GQ868591
CC0011_NODE_1_length_10201_cov_607.828724_pilon	3-III	98.7	100	gb:EU687233
Spike_NODE_1_length_10266_cov_2032.312101_pilon	3-III	98.36	100	gb:EU687233
CC0009_NODE_1_length_10208_cov_2013.867437_pilon	3-III	98.61	99.97	gb:EU687233
91-0118_S8_L001_NODE_1_length_10178_cov_13.815371_pilon	3-III	98.44	99.99	gb:EU687233
cc0010_S8_L001_NODE_1_length_10206_cov_450.729095_pilon	3-III	98.66	100	gb:EU687233
CC0061_k77_1_flag_1_multi_4641.2458_len_10267_pilon	4-II	98.51	100	gb:KP188557
CC0067_NODE_1_length_10197_cov_734.756522_pilon	4-II	98.78	100	gb:KP188557
cc0030a_S12_k77_1_flag_1_multi_2605.9226_len_10163_pilon	4-II	98.92	99.82	gb:KP188557
cc0030b_S21_NODE_1_length_10173_cov_54.900771_pilon	4-II	98.92	100	gb:KP188557
CC0116_k77_2_flag_1_multi_2097.0000_len_10197_pilon	4-II	98.67	100	gb:KP188557
Spike_NODE_2_length_10203_cov_29.787675_pilon	4-II	98.75	99.95	gb:KP188557
CC0066_NODE_1_length_10174_cov_40.432750_pilon	4-II	98.5	100	gb:KP188557
91-0106_S12_L001_k77_17_flag_1_multi_13.3022_len_10127_pilon	4-II	98.72	99.67	gb:KP188557

Figure 4.5: DEN-IM report tables. a) DEN-IM's quality control report containing information of the number of base-pairs and the number of reads for the analysed samples, the estimated coverage depth before and after mapping, and the percentage of reads in the input data that were trimmed. b) DEN-IM's typing report for 24 CDSs recovered from the metagenomic dataset. The ID contains the CDS contig name, the typing result for serotype-genotype, the values for identity and coverage, and the GenBank ID of the closest reference in the Typing Database containing 161 complete DENV genomes.

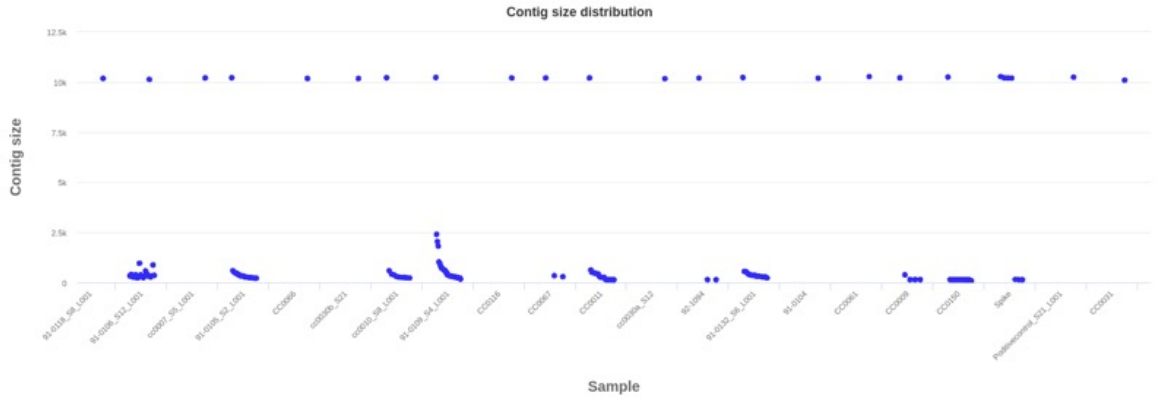


Figure 4.6: Contig size distribution for the shotgun metagenomics sequencing dataset. Each dot depicts an assembled DENV contig. Above the 10Kb are full CDS of DENV.

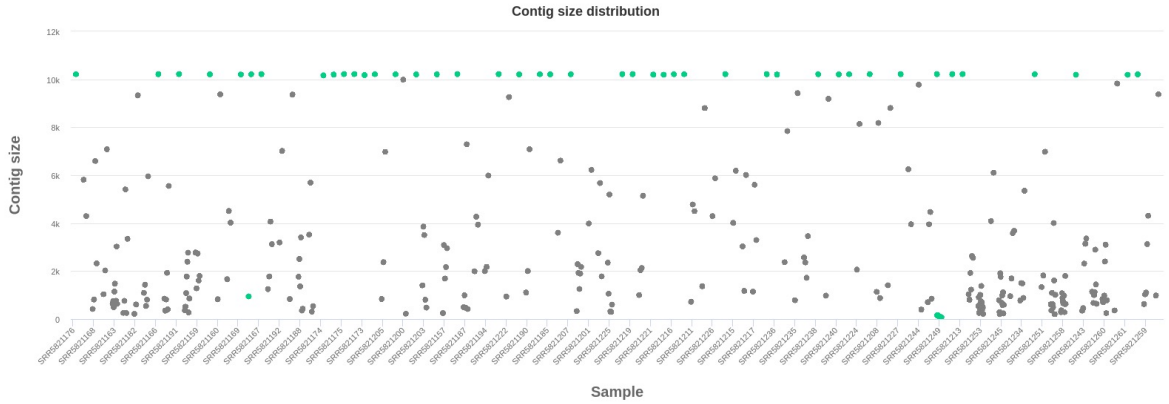


Figure 4.7: Contig size distribution of the amplicon sequencing dataset with 106 paired-end samples. Each dot depicts an assembled DENV contig. Above the 10Kb are full CDS of DENV. Contigs belonging from samples that assembled a complete DENV CDS are highlighted in green, whereas the remaining are coloured in grey.

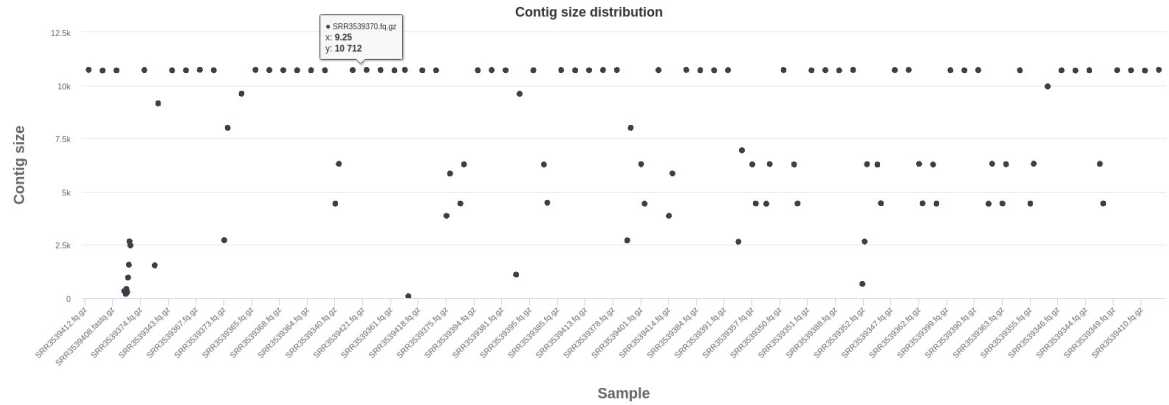


Figure 4.8: Contig size distribution of the amplicon sequencing dataset with 78 single-end samples. Each dot depicts an assembled DENV contig. Above the 10Kb are full CDS of DENV.

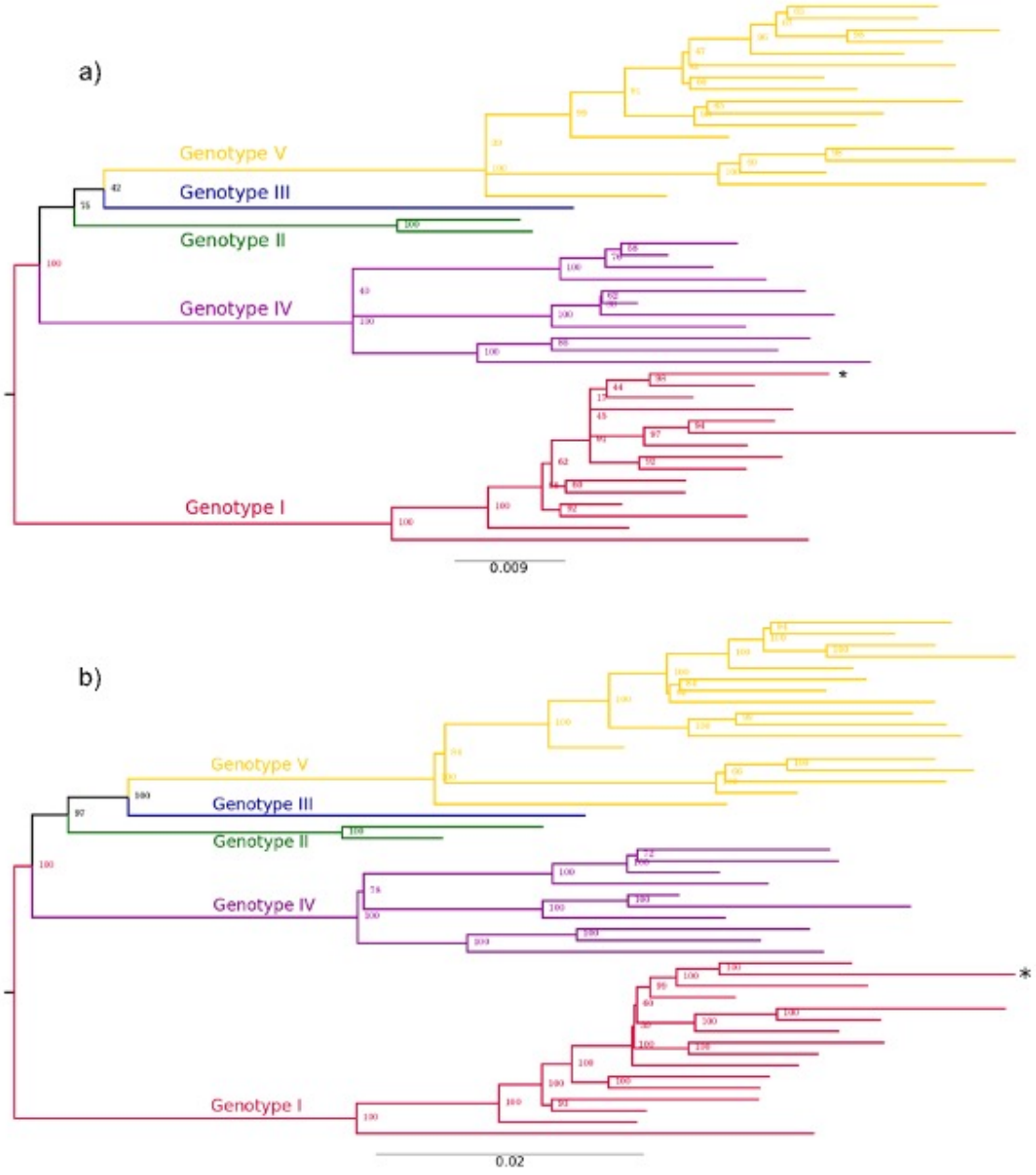


Figure 4.9: Maximum Likelihood inference of the multiple sequence alignment of the 46 DENV-1 complete genomes in the typing dataset, with a) envelope region and b) whole genome sequence. 1635 complete DENV-1 genomes were clustered at 98% nucleotide identity and the representative genomes were aligned with MAFFT. A maximum likelihood tree was inferred with RAxML. The tree is coloured according to genotype (red: genotype I; green: genotype II; blue: genotype III; purple: genotype IV). The sample JF459993, marked with a star, is currently annotated in ViPR as belonging to genotype IV but, given to the good phylogenetic support, it was re-classified as belonging to the genotype I.

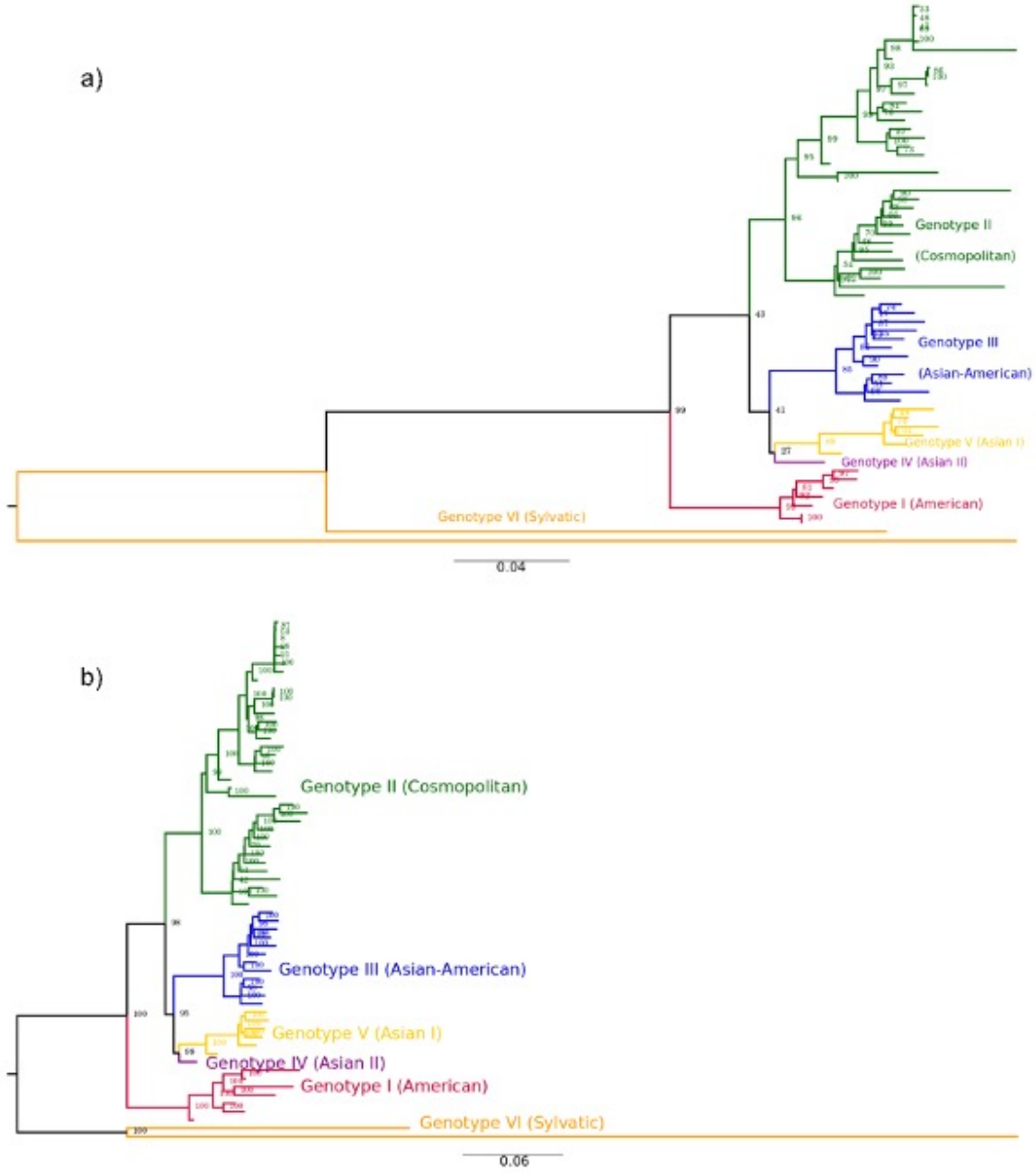


Figure 4.10: Maximum Likelihood inference of the multiple sequence alignment of the 63 DENV-2 complete genomes in the typing dataset, with a) envelope region and b) whole genome sequence. 1067 complete DENV-1 genomes were clustered at 98% nucleotide identity and the representative genomes were aligned with MAFFT. A maximum likelihood tree was inferred with RAxML. The tree is coloured according to genotype (red: genotype I; green: genotype II; blue: genotype III; purple: genotype IV).

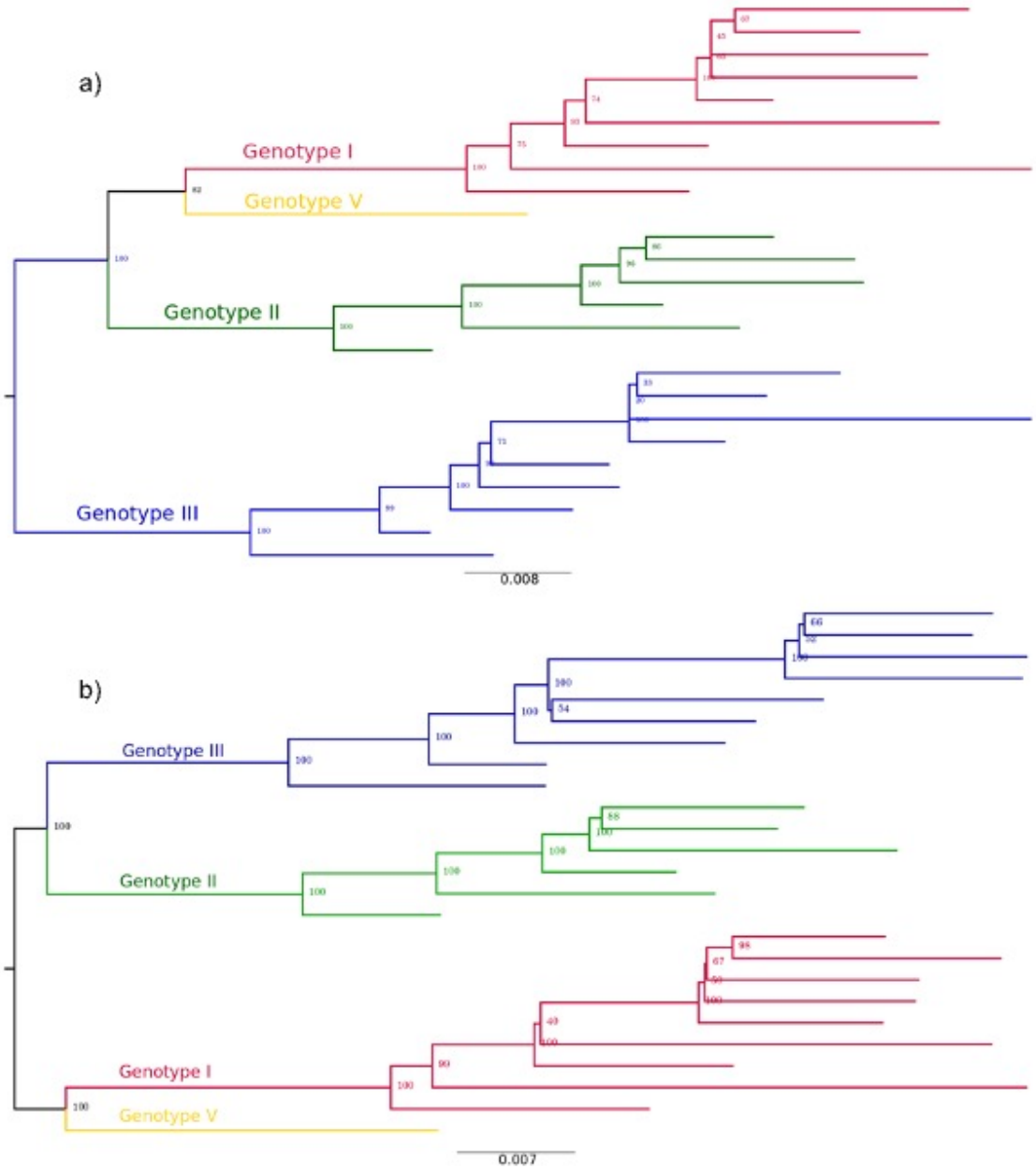


Figure 4.11: Maximum Likelihood inference of the multiple sequence alignment of the 25 DENV-3 complete genomes in the typing dataset, with a) envelope region and b) whole genome sequence. 807 complete DENV-3 genomes were clustered at 98% nucleotide identity and the representative genomes were aligned with MAFFT. A maximum likelihood tree was inferred with RAxML. The tree is coloured according to genotype (red: genotype I; green: genotype II; blue: genotype III; purple: genotype IV).

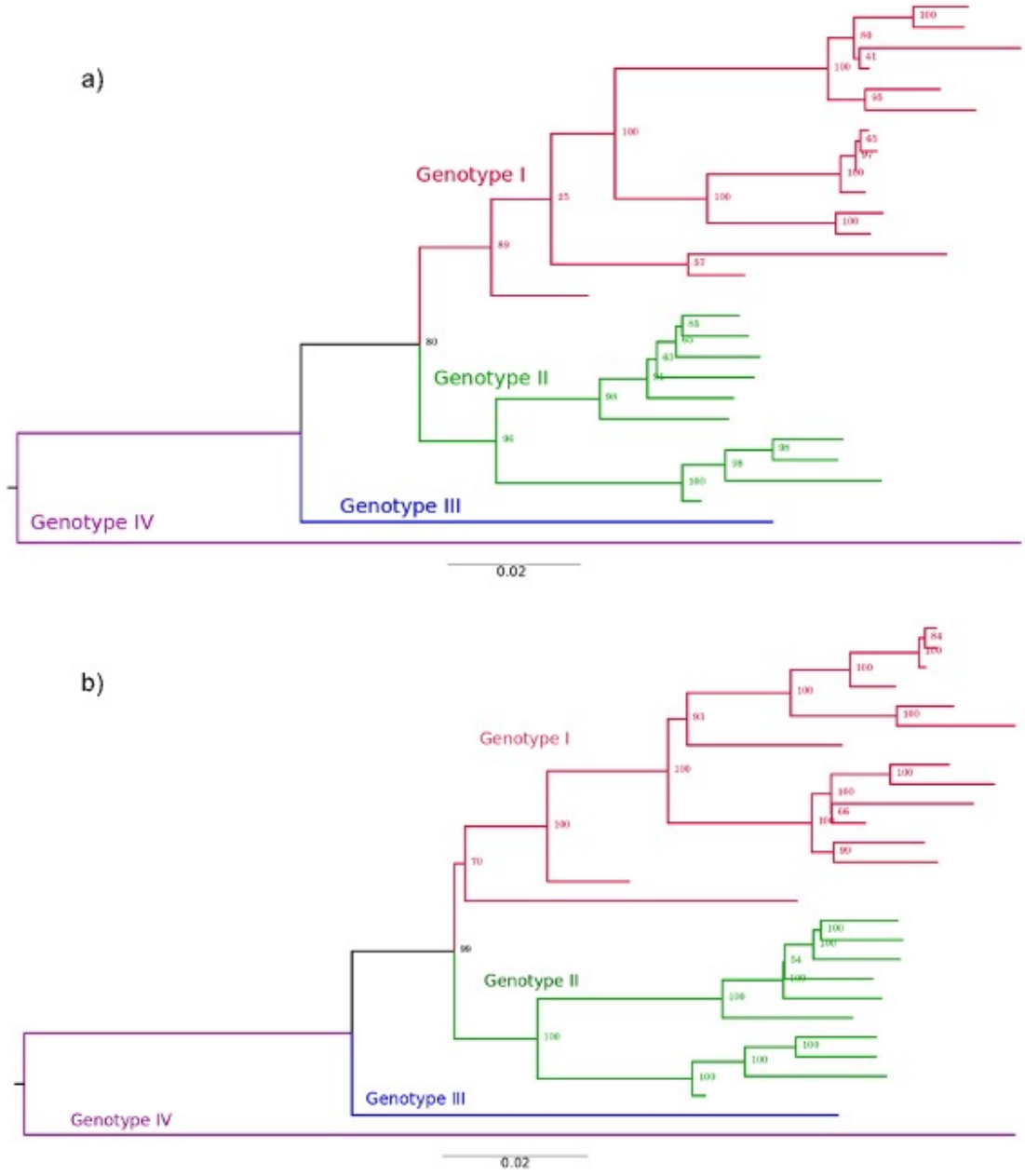


Figure 4.12: Maximum Likelihood inference of the multiple sequence alignment of the 27 DENV-4 complete genomes in the typing dataset, with a) envelope region and b) whole genome sequence. 320 complete DENV-4 genomes were clustered at 98% nucleotide identity and the representative genomes were aligned with MAFFT. A maximum likelihood tree was inferred with RAxML. The tree is coloured according to genotype (red: genotype I; green: genotype II; blue: genotype III; purple: genotype IV).

4.14 References

[1] World Health Organization. *Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control*. en. Google-Books-ID: dlc0YSIyGYwC. World Health Organization, 2009. ISBN: 978-92-4-154787-1.

[2] Michael S. Diamond and Theodore C. Pierson. "Molecular Insight into Dengue Virus Pathogenesis and Its Implications for Disease Control". English. In: *Cell* 162.3 (July 2015). Publisher: Elsevier, pp. 488–492. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.07.005. URL: https://www.cell.com/cell/abstract/S0092-8674(15)00842-9 (visited on 01/20/2021).

[3] Samir Bhatt et al. "The global distribution and burden of dengue". eng. In: *Nature* 496.7446 (Apr. 2013), pp. 504–507. ISSN: 1476-4687. DOI: 10.1038/nature12060.

[4] José Lourenço et al. "Challenges in dengue research: A computational perspective". en. In: *Evolutionary Applications* 11.4 (2018). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/eva.12554, pp. 516–533. ISSN: 1752-4571. DOI: 10.1111/eva.12554. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/eva.12554 (visited on 01/24/2022).

[5] Katrin C. Leitmeyer et al. "Dengue Virus Structural Differences That Correlate with Pathogenesis". en. In: *Journal of Virology* 73.6 (June 1999), pp. 4738–4747. ISSN: 0022-538X, 1098-5514. DOI: 10.1128/JVI.73.6.4738-4747.1999. URL: https://journals.asm.org/doi/10.1128/JVI.73.6.4738-4747.1999 (visited on 03/29/2022).

[6] Nathan L. Yozwiak et al. "Virus Identification in Unknown Tropical Febrile Illness Cases Using Deep Sequencing". en. In: *PLoS Neglected Tropical Diseases* 6.2 (Feb. 2012). Ed. by Rebeca Rico-Hesse, e1485. ISSN: 1935-2735. DOI: 10.1371/journal.pntd.0001485. URL: https://dx.plos.org/10.1371/journal.pntd.0001485 (visited on 06/19/2021).

[7] Chun Kiat Lee et al. "Clinical use of targeted high-throughput whole-genome sequencing for a dengue virus variant". en. In: *Clinical Chemistry and Laboratory Medicine (CCLM)* 55.9 (Sept. 2017). Publisher: De Gruyter, e209–e212. ISSN: 1437-4331. DOI: 10.1515/cclm-2016-0660. URL: https://www.degruyter.com/document/doi/10.1515/cclm-2016-0660/html (visited on 01/24/2022).

[8] Zareen Fatima et al. "Serotype and genotype analysis of dengue virus by sequencing followed by phylogenetic analysis using samples from three mini outbreaks-2007-2009 in Pakistan". In: *BMC Microbiology* 11.1 (2011), p. 200. ISSN: 1471-2180. DOI: 10.1186/1471-2180-11-200. URL: https://doi.org/10.1186/1471-2180-11-200 (visited on 01/24/2022).

[9] Vagner Fonseca et al. "A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes". en. In: *PLOS Neglected Tropical Diseases* 13.5 (2019). Publisher: Public Library of Science, e0007231. ISSN: 1935-2735. DOI: 10.1371/journal.pntd.0007231. URL: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0007231 (visited on 01/20/2021).

[10] Michael Vilsker et al. "Genome Detective: an automated system for virus identification from high-throughput sequencing data". In: *Bioinformatics* 35.5 (Mar. 2019), pp. 871–873. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty695. URL: https://doi.org/10.1093/bioinformatics/bty695 (visited on 01/24/2022).

[11] Yang Li et al. "VIP: an integrated pipeline for metagenomics of virus identification and discovery". en. In: *Scientific Reports* 6.1 (Mar. 2016). Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics;Microbiology Subject_term_id: computational-biology-and-bioinformatics;microbiology, p. 23774. ISSN: 2045-2322. DOI: 10.1038/srep23774. URL: https://www.nature.com/articles/srep23774 (visited on 01/24/2022).

[12] Akifumi Yamashita, Tsuyoshi Sekizuka, and Makoto Kuroda. "VirusTAP: Viral Genome-Targeted Assembly Pipeline". In: *Frontiers in Microbiology* 7 (2016). ISSN: 1664-302X. URL: https://www.frontiersin.org/article/10.3389/fmicb.2016.00032 (visited on 01/24/2022).

[13] Hsin-Hung Lin and Yu-Chieh Liao. "drVM: a new tool for efficient genome assembly of known eukaryotic viruses from metagenomes". en. In: *GigaScience* 6.2 (Feb. 2017). ISSN: 2047-217X. DOI: 10.1093/gigascience/gix003. URL: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/gix003/2929394 (visited on 03/29/2022).

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

[15] Lisa Gerhardt et al. "Shifter: Containers for HPC". In: *Journal of Physics: Conference Series* 898 (Oct. 2017), p. 082021. ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/898/8/082021. URL: https://iopscience.iop.org/article/10.1088/1742-6596/898/8/082021 (visited on 03/24/2021).

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

[17] Robert Schmieder and Robert Edwards. "Quality control and preprocessing of metagenomic datasets". In: *Bioinformatics* 27.6 (Mar. 2011), pp. 863–864. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr026. URL: https://doi.org/10.1093/bioinformatics/btr026 (visited on 01/24/2022).

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

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

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

[21] Bruce J. Walker et al. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement". en. In: *PLOS ONE* 9.11 (Nov. 2014). Publisher: Public Library of Science, e112963. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0112963. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112963 (visited on 01/24/2022).

[22] Miguel Paulo Machado et al. "Epidemiological Surveillance and Typing Methods to Track Antibiotic Resistant Strains Using High Throughput Sequencing". en. In: *Antibiotics: Methods and Protocols*. Ed. by Peter Sass. Methods in Molecular Biology. New York, NY: Springer, 2017, pp. 331–356. ISBN: 978-1-4939-6634-9. DOI: 10.1007/978-1-4939-6634-9_20. URL: https://doi.org/10.1007/978-1-4939-6634-9_20 (visited on 01/24/2022).

[23] S. Altschul. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic Acids Research* 25.17 (Sept. 1997), pp. 3389–3402. ISSN: 13624962. DOI: 10.1093/nar/25.17.3389. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/25.17.3389 (visited on 03/29/2022).

[24] Tsukasa Nakamura et al. "Parallelization of MAFFT for large-scale multiple sequence alignments". In: *Bioinformatics* 34.14 (July 2018), pp. 2490–2492. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty121. URL: https://doi.org/10.1093/bioinformatics/bty121 (visited on 01/20/2021).

[25] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9 (2014), pp. 1312–1313. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu033. URL: https://doi.org/10.1093/bioinformatics/btu033 (visited on 01/24/2022).

[26] Poornima Parameswaran et al. "Intrahost Selection Pressures Drive Rapid Dengue Virus Microevolution in Acute Human Infections". en. In: *Cell Host & Microbe* 22.3 (Sept. 2017), 400–410.e5. ISSN: 19313128. DOI: 10.1016/j.chom.2017.08.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1931312817303359 (visited on 06/19/2021).

[27] Paula Eillanny Silva Marinho et al. "Meningitis Associated with Simultaneous Infection by Multiple Dengue Virus Serotypes in Children, Brazil - Volume 23, Number 1—January 2017 - Emerging Infectious Diseases journal - CDC". en-us. In: (). DOI: 10.3201/eid2301.160817. URL: https://wwwnc.cdc.gov/eid/article/23/1/16-0817_article (visited on 01/24/2022).

[28] Manchala Nageswar Reddy et al. "Occurrence of concurrent infections with multiple serotypes of dengue viruses during 2013–2015 in northern Kerala, India". en. In: *PeerJ* 5 (Mar. 2017), e2970. ISSN: 2167-8359. DOI: 10.7717/peerj.2970. URL: https://peerj.com/articles/2970 (visited on 03/30/2022).

[29] Lize Cuypers et al. "Time to Harmonize Dengue Nomenclature and Classification". en. In: *Viruses* 10.10 (Oct. 2018), p. 569. ISSN: 1999-4915. DOI: 10.3390/v10100569. URL: http://www.mdpi.com/1999-4915/10/10/569 (visited on 06/19/2021).

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

[31] Christian Julián Villabona-Arenas and Paolo Marinho de Andrade Zanotto. "Worldwide spread of Dengue virus type 1". eng. In: *PloS One* 8.5 (2013), e62649. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0062649.

[32] Nikos Vasilakis and Scott C. Weaver. "The history and evolution of human dengue emergence". eng. In: *Advances in Virus Research* 72 (2008), pp. 1–76. ISSN: 0065-3527. DOI: 10.1016/S0065-3527(08)00401-6.

[33] Weizhong Li and Adam Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". eng. In: *Bioinformatics (Oxford, England)* 22.13 (July 2006), pp. 1658–1659. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl158.

[34] R. Rico-Hesse. “Molecular evolution and distribution of dengue viruses type 1 and 2 in nature”. eng. In: *Virology* 174.2 (Feb. 1990), pp. 479–493. ISSN: 0042-6822. DOI: 10.1016/0042-6822(90)90102-w.

[35] Rebeca Rico-Hesse. “Microevolution and virulence of dengue viruses”. eng. In: *Advances in Virus Research* 59 (2003), pp. 315–341. ISSN: 0065-3527. DOI: 10.1016/s0065-3527(03)59009-1.

[36] R. S. Lanciotti et al. “Rapid detection and typing of dengue viruses from clinical samples by using reverse transcriptase-polymerase chain reaction”. eng. In: *Journal of Clinical Microbiology* 30.3 (Mar. 1992), pp. 545–551. ISSN: 0095-1137. DOI: 10.1128/jcm.30.3.545-551.1992.

[37] R. S. Lanciotti, D. J. Gubler, and D. W. Trent. “Molecular evolution and phylogeny of dengue-4 viruses”. eng. In: *The Journal of General Virology* 78 (Pt 9) (Sept. 1997), pp. 2279–2284. ISSN: 0022-1317. DOI: 10.1099/0022-1317-78-9-2279.

[38] Chonticha Klungthong et al. “The molecular epidemiology of dengue virus serotype 4 in Bangkok, Thailand”. eng. In: *Virology* 329.1 (Nov. 2004), pp. 168–179. ISSN: 0042-6822. DOI: 10.1016/j.virol.2004.08.003.

[39] Chunlin Zhang et al. “Clade replacements in dengue virus serotypes 1 and 3 are associated with changing serotype prevalence”. eng. In: *Journal of Virology* 79.24 (Dec. 2005), pp. 15123–15130. ISSN: 0022-538X. DOI: 10.1128/JVI.79.24.15123-15130.2005.

[40] Chunlin Zhang et al. “Structure and age of genetic diversity of dengue virus type 2 in Thailand”. eng. In: *The Journal of General Virology* 87.Pt 4 (Apr. 2006), pp. 873–883. ISSN: 0022-1317. DOI: 10.1099/vir.0.81486-0.

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

[42] Gilberto A. Santiago et al. “Analytical and Clinical Performance of the CDC Real Time RT-PCR Assay for Detection and Typing of Dengue Virus”. In: *PLoS Neglected Tropical Diseases* 7.7 (July 2013), e2311. ISSN: 1935-2727. DOI: 10.1371/journal.pntd.0002311. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3708876/ (visited on 04/11/2022).

Chapter 5

Conclusion

Appendix A

Appendix