Queen Mary University London and Public Health England

Multiple Displacement Amplification and Whole Genome Sequencing for the Diagnosis of Infectious Diseases

Submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy'

Catherine Anscombe
[07/07/2016]

Statement of Originality

I, Catherine Joan Anscombe, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature:

Date:

Funding

This project was funded by Roche Diagnostics as a scientific studentship award and completed at Public Health England Centre for Infection.

Abstract

Next-generation sequencing technologies are revolutionising our ability to characterise and investigate infectious diseases. Utilising the power of high throughput sequencing, this study reports, the development of a sensitive, non-PCR based, unbiased amplification method. Which allows the rapid and accurate sequencing of multiple microbial pathogens directly from clinical samples.

The method employs ϕ29 DNA polymerase, a highly efficient enzyme able to produce strand displacement during the polymerisation process with high fidelity. Problems with DNA secondary structure were overcome and the method optimised to produce sufficient DNA to sequence from a single bacterial cell in two hours. Evidence was also found that the enzyme requires at least six bases of single stranded DNA to initiate replication, and is not capable of amplification from nicks. ϕ29 multiple displacement amplification was shown to be suitable for a range of GC contents and bacterial cell wall types as well as for viral pathogens. The method was shown to be able to provide relative quantification of mixed cells, and a method for quantification of viruses using a known standard was developed.

To complement the novel molecular biology workflow, a data analysis pipeline was developed to allow pathogen identification and characterisation without prior knowledge of input. The use of *de novo* assemblies for annotation was shown to be equivalent to the use of polished reference genomes. Single cell ϕ29 MDA samples had better assembly and annotation than non-amplification controls, a novel finding which, when combined with the very long DNA fragments produced, has interesting implications for a variety of analytical procedures.

A sampling process was developed to allow isolation and amplification of pathogens directly from clinical samples, with good concordance shown between this method and traditional testing. The process was tested on a variety of modelled and real clinical samples showing good application to sterile site infections, particularly bacteraemia models. Within these samples multiple bacterial, viral and parasitic pathogens were identified, showing good application across multiple infection types. Emerging pathogens were identified including *Onchocerca volvulus* within a CSF sample, and *Sneathia sanguinegens* within an STI sample.

Use of ϕ29 MDA allows rapid and accurate amplification of whole pathogen genomes. When this is coupled with the sample processing developed here it is possible to detect the presence of pathogens in sterile sites with a sensitivity of a single genome copy.

Table of Contents

Acknowledgements

"Nothing in the world is worth having or worth doing unless it means effort, pain, difficulty"

Theodore Roosevelt

Firstly, I wish to thank my PhD supervisors, Prof Saheer Gharbia, Dr Raju Misra and Prof Armine Sefton, without whom this thesis would never have existed.

"If I have seen further it is by standing on the shoulders of Giants" Sir Isaac Newton

Secondly, I would like to thank the numerous scientific colleagues who have provided resources, advice and support, this thesis would not have progressed without you.

"It's really clear that the most precious resource we all have is time" Steve Jobs

Additionally, thanks needs to go to Dr David J Allen who gave me the time and support to finally finish this thesis.

"I love people who make me laugh. I honestly think it's the thing I like most, to laugh. It cures a multitude of ills. It's probably the most important thing in a person." Audrey Hepburn

I would also like to thank my friends who have helped me maintain a sense of humour throughout this PhD, finishing this thesis would have been much harder without you.

"I can no other answer make, but, thanks, and thanks." William Shakespeare

A huge thank you to my family for a lifetime's worth of encouragement and understanding, I would be nothing without them.

"There is no passion to be found playing small - in settling for a life that is less than the one you are capable of living." Nelson Mandela

Finally, it gives me great pleasure to dedicate this thesis to my future husband whose enthusiasm for science helped me remember my own. This work would never have been completed without his constant encouragement, understanding and support.

There are not enough words to express my gratitude.

Introduction

1. Introduction

Infectious diseases are still a major cause of morbidity and mortality, even in developed countries. According to the CDC (Centre for Disease Control and Prevention) in the US in 2010 infectious diseases accounted for 23.6 million visits to primary physicians. With global travel becoming increasingly popular and affordable the epidemiology of infections is changing, as demonstrated by the 2009 H1N1 influenza outbreak which rapidly spread across the globe. Additionally, the emergence of drug resistance such as New Delhi Metallo-beta-lactamase-1 (NDM-1) carrying Enterobacteriaceae, meticillin resistant *Staphylococcus aureus* (MRSA) and drug resistant HIV are becoming more common. Diagnosis and surveillance of pathogens is increasingly important as human populations grow and people live in increasingly densely populated areas. Increasing medical interventions, such as the use of chemotherapy and care in intensive care units is leading to a higher burden of resistant pathogens. Opportunistic infections are becoming more of a concern, often being serious and sometimes fatal. In the face of rapidly changing global epidemiology, improved diagnostics and understanding of infectious diseases will become increasingly important for global mortality, morbidity and health economics.

Much of the processes in modern microbiology diagnostics can be traced back to 1890 when Koch established a causative relationship between specific pathogens and disease. The formation of Koch's postulates informed much of the basis of culture based diagnostics. As understanding of infectious disease increased, multiple exceptions to the postulates were observed, such as asymptomatic carriage and differing host susceptibilities to infectious agents. Despite this evidence based culture methods continued to be the main tools used for bacterial infection diagnosis. Automation of culture based methods have allowed increased accuracy and throughput, allowing better diagnosis of bacterial infections. Use of nucleic acid for detection and identification of infectious diseases allowed a shift in the diagnostic process. The development of PCR in 1983 has allowed increasingly rapid identification of infectious agents without the need for culture. This development had significant impact on diagnosis of pathogens which are slow growing, difficult to culture, or with no known culture method. Diagnosis by led to further development of Koch's postulates by Fredrick and Relman in 1996 to augment interpretation of microbial detection using more modern techniques and understanding of infections. However, these still centred around demonstrating reproducible and evidence based results associated with infectious diseases. Combining PCR with fluorescence allowed simultaneous amplification and detection of products in the form of real-time PCR, which increased the speed and ease of PCR. Additionally, the use of different reporting fluoroforms allowed multiplexing of assays made it

possible to simultaneously detect multiple targets in a single test. The invention of Sanger sequencing in 1977 allowed the genetic code of pathogens to be used both for identification and characterisation. In parallel to the advancing of PCR, sequencing technologies became increasingly cost effective thereby increasing the use of sequencing in diagnostic and research microbiology. Another major milestone for diagnosis of infectious disease was the introduction of the MALDI-TOF into microbiology, which allowed rapid identification of many pathogenic bacteria changing the management of infectious disease.

For reliable identification and characterisation of infections, several biological problems need to be addressed. The number of pathogens within clinical samples such as cerebrospinal fluid (CSF), blood and tissue is often low, especially in comparison to human cells. In none sterile site infections pathogens may have low numbers compared to regional flora. Whilst it is sometimes possible to visualise bacteria in samples using staining techniques, this often has low sensitivity and only informs about the morphology of the bacteria. Therefore, most diagnostics rely on increasing either the number of cells, or the number of copies of a specific gene. Traditionally selective media and tissue cultures have been used to increase the number of cells to a level that is detectible. Selective culture will allow pathogens of interest to be preferentially grown, allowing separation from flora. Biochemical test such as carbohydrate metabolism and the presence of enzymes such as oxidase and catalase are used in identification of bacteria, but often involve overnight incubation. Introduction of the MALDI-TOF (Matrix-assisted laser desorption/ionization Time of Flight) mass spectrometers to diagnostic laboratories has allowed rapid identification of bacteria from cultured samples. Identification using MALDI-TOF is achieved by comparing a mass-charge ratio spectrum produced from protein ionisation against a reference database.

Advances in molecular techniques, in particular application of polymerase chain reaction (PCR), has allowed targeted amplification of pathogen signals to quickly detect pathogens with a high sensitivity. Further advances have allowed multiplexing, allowing multiple targets to be detected in a single reaction. Characterisation of pathogens has mainly centred on resistance testing, toxin production and the relatedness of pathogens to each other. Most resistance testing for bacteria is currently phenotypic, such as serial dilutions, disc and E-test. Viral resistance testing is achieved by detecting mutations in drug targets, giving a genotypic prediction of the expected phenotypic reaction to anti-virals. Toxins are either detected directly using ELISA (enzyme-linked immunosorbent assay) or the genes encoding the toxins are detected using polymerase chain reaction (PCR). Serotyping uses antigens presented on the surface of bacterial cells and viral particles to provide further information about the infectious agent. This may be used to predict disease severity, potential treatment options or for epidemiological tracking. Serotyping, phage

typing and an organism's antibiogram have traditionally been use to establish relatedness of infections. Molecular tests allow the use of amplicon detection, single nucleotide polymorphisms (SNP) detection and targeted sequencing, increasing discrimination in a more rapid and reproducible manor. During sample processing for pathogen detection and characterisation there are biases towards known and common aetiologies of infections. There are also frequent cases where, although there is a strong clinical suspicion of infection all tests are negative. Many contributing factors have been reported, such as previous antibiotic treatment, poor sample processing and novel or unknown aetiologies. Furthermore, pathogens are often investigated in isolation with no consideration to interaction with flora or host response, with little explanation about how low virulence pathogens can potentially cause serious disease. Workflows are often multi-factorial and time consuming with several different tests being used to identify and characterise pathogens. Whilst improvements have been made, each technique used is subject to inherent biases, which may contribute to the number of infections which are clinically suspected but the primary pathogen never identified. This is particularly apparent in central nervous system (CNS) infections, one study identified the aetiology in only 28% of cases in 334 patients with encephalitis[1].

Ultimately there are three key challenges to be faced for improving the diagnosis of infectious diseases:

- Low pathogen numbers and signals amongst host and flora,
- Accurate characterisation of pathogens allowing suitable treatment and effective epidemiology
- Diagnosis of infections due to unknown pathogens and pathogens with novel characteristics

1.1 Analysis of Low Level Pathogens

Detection of low level pathogens is essential, particularly where bacteria or viral pathogens cause acute or invasive human illness. For the identification of such pathogens signals must be recognised above both the background host and flora. There are two primary methods to achieve this, increasing the cell number of pathogens (e.g. culture) or detection of increased pathogen signals using molecular methods (e.g. PCR).

Most diagnostics in bacteriology are based on the ability to culture viable organisms, which is achieved through a combination of selective and none selective media. For some samples, such as those collected from sterile sites where there is no regional bacterial flora, non-selective media and enrichment broths are often used to capture all bacteria present in the sample. However, some samples, such as stool samples, have high levels of regional flora. For these cases selective media is used to enrich for predefined pathogenic bacteria over the normal flora. In some cases, only a single bacterial species with specific characteristics is isolated, such as MRSA screens which provides rapid identification of MRSA amongst flora including non- resistant *S. aureus*. Separation of flora and pathogen is more difficult in some samples types, especially when a single species can be classed as both normal flora and a pathogen. For example *Streptococcus pneumoniae* is a major cause of pneumonia, but is also present as a carriage organism in 4% of adults and 53% of children[2]. Semi-quantitative culture methods are used to differentiate infection from carriage, these are open to individual interpretation. Culturing of bacteria is highly labour intensive, but automation, such as blood cultures, within the laboratory has lowered the hands on time required for pathogen detection. Bacteraemia has a very high associated mortality rate, up to 34% [3], but can have very low bacterial load, and no specific clinical symptoms. Automated liquid culture methods such as the Bactec (Becton-Dickinson), allows constant monitoring of growth using CO_2 sensors, which means only positive cultures are processed by hand. However these systems only provide a positive result with no identification, the need to subculture these samples delays the identification by up to 72 hours [4]. Time to positivity of more fastidious organism, particularly those belonging to the HACEK (*Haemophilus*, *Aggregatibacter* (previously *Actinobacillus*), *Cardiobacterium*, *Eikenella corrodens*, *Kingella*) group associated with endocarditis, can take up to 21 days using culture techniques[5].

After isolation, the bacteria of interest need to be identified and characterised. Traditional identification techniques such as biochemistry and serology require high numbers of bacterial cells, with automated biochemical methods such as the Phoenix needing over 1×10^9 bacterial cells[6]. Biochemical methods rely on detection of reactions with a number of substrates, such as fermentation of sugars, production of by products such as indole or H_2S, or presence of enzymes

such as oxidase. Most biochemical panels need to be incubated overnight before results can be interpreted. With the introduction of MALDI-TOF technology into laboratories time to identification has been reduced from 18 hours, to minutes. Only a single colony is needed, which allows rapid identification of bacteria, and allows colonies on mixed plates to be identified without the need of purity plates. The technology has recently been used to identify isolates directly from positive blood cultures[7], saving at least 18 hours. However high numbers of bacteria are needed for this, at least 10^7 CFU/ml [4], limiting its use for detection of pathogens directly from clinical specimens. In addition, pure cultures are required for sensitivity testing. MALDI-TOF also had difficulty resolving the *Streptococcus mitis* group, which is of particular importance as *S. pneumoniae* is a major cause of infections and is associated with high mortality. It is also poor at identifying mixed infections directly from blood cultures.

The traditional culture based work-flow in diagnosing bacterial infections is often time consuming and labour intensive. Minimum time to identification is 18 hours from receipt of sample, and is often longer when pathogens have to be separated from flora. Identification depends on having a database to compare results to, such as a biochemical profile or MALDI-TOF spectra. For novel or unusual isolates, the database may be unreliable or insufficient. Antibiotic sensitivities take at least a further 24 hours after identification, with other characterisations taking up to three days. Overall a fundamental shortfall of the culturing methods is that it relies on both viable organisms, and a suitable culture medium. Viability can be affected by antibiotic use, poor sample transport or sampling technique. It also selects for those organisms which grow quickly and with relatively simple growth requirements that can be met by basic growth media. As a result, organisms which are intracellular, or grow within biofilms are often overlooked as pathogens.

Some viral pathogens are diagnosed by detecting the body's response to the infection through the use of serology. This can be used to detect immunity to a virus such as anti-natal screens and for diagnosis of a current infection. In chronic infections such as Hepatitis B Virus (HBV), serology can be used to distinguish between chronic and acute infections. The detection of antibodies from serum has allowed consolidation of sample processing, allowing high levels of automation. Improvement in data management systems have also allowed automated interpretation and reporting of results. However, serology is an indirect mechanism that relies on a detectable and predictable response to a known infection. This would be difficult to apply to a novel pathogen, and test development relies on understanding immunological response to an infection.

Viruses are obligate intracellular organisms requiring living cells to replicate in. Cell cultures using mammalian cells, was traditionally the main method by which viruses were diagnosed, with specific cell lines allowing pathogens to be detected above flora. Once samples are inoculated into susceptible cells lines, the samples need to checked every other day for cytopathic effects (CPE). Cell cultures need to be kept for up to two weeks for CPE to be observed[8] .Longer incubation times are required for some viruses such as *Cytomegalovirus (*CMV*)*. Some viruses such as Influenza don't ordinarily induce CPE and so secondary indicators are needed, such as interaction with added red blood cells. For final identification, immunofluorescence, neutralization, haemadsorption inhibition, electron microscopy, or molecular tests are normally carried out.

Disadvantages of cell culture include a relative lack of sensitivity for some viruses, the inability to culture viruses of considerable medical importance (i.e. Hepatitis B Virus (HBV) and Hepatitis C (HCV) Virus). Tissue culture is labour intensive with a requirement for specific tissue samples (i.e. foetal lung). For CPE to be observed there is often a need for very high viral numbers, this leads to increased potential for laboratory based infection. For these reasons, cell culture is now not used in most diagnostic laboratories and has been replaced by molecular technologies.

Improvements in molecular techniques have allowed improved diagnosis of infections where specific pathogens can be targeted. Identification of infections through the amplification and detection of specific nucleic acid signals is a major diagnostic tool in virology and has been gaining momentum in bacteriology. The use of PCR allows the amplification and detection of pathogen signals from very low copy numbers, giving highly sensitive results. Real time PCR allows the coupling of amplification and detection, lowering turnaround times, especially when it is performed directly from specimens. The use of specific primers with or without a specific probe also allows for high specificity, selectively amplifying pathogen signals over flora and host. Molecular methods also negate the need for viable organism, and allow safer sample handling. Using specific primers along with whole nucleic acid extraction techniques applicable to all samples allows the consolidation of sample processing. Increases in automation has lowered both hands on and turnaround times. It has had particular advantages for slow growing, difficult to culture or un-culturable pathogens.

Reduced turnaround times for both positive and negative results, has a major impact on patient treatment. This is particularly important for prevention of transmission and appropriate therapy prescription. One example where the introduction of molecular techniques has improved diagnosis and treatment indication is M*ycobacterium tuberculosis (M. tuberculosis), which is the*

causative agent of tuberculosis. *M. tuberculosis* is a slow growing organism, with liquid cultures of smear negative sputum take around can take up to six weeks[9]. Treatment for TB requires a multi-drug regime with a minimum treatment time of six months[10]. Traditional diagnostics of TB have relied on microscopy, liquid and solid culture, detection limits of these methods b eing 10000, 10-50 and 100 CFU/ml respectively[9]. Multiple genes have been targeted for the detection of *M. tuberculosis*, including the *rpoB* gene, IS6110 and 16S rRNA genes as well as multiple housekeeping genes such as those involved in phosphate transport [11]. Both commercial and in house assays have been developed, with detection levels as low as 1 fg and sensitivity varying from 75-100% in studies [11–13]. The Xpert MTB/RIF test (Cepheid) [14], is the most widely used system and allows the detection of TB from samples in two hours. The use of a second probe targeting the *rpoB* gene, where 95% of rifampicin resistance mutations occur can be used to determine whether an isolate is likely to be rifampicin resistant [9]. This test consists of a single modular test, for both extraction and detection, which require little hands on time (20 minutes) or specialised skills. The detection limit is five copies of purified DNA or 131 CFU/ml spiked into sputum[15]. Sensitivity in smear positive respiratory samples has been demonstrated to be up to 100%, whereas sensitivity in smear negative respiratory samples is as low as 75%. In paediatric and extra pulmonary TB the sensitivity is lower, with sensitivities as low as 65% and 25% respectively [9]. Use of this test has increased the number of patients who are appropriately treated for multi -drug resistant tuberculosis (MDR TB). However, this method relies on detecting resistance using a single gene and is not useful for differentiating between MDR, XDR (extensively drug-resistant *tuberculosis*) and TDR TB (totally drug-resistant tuberculosis).

CNS infections often have a rapid onset and are still associated with high morbidity and mortality[16]. Rapid detection of the aetiological agent of the infection allows better case management and will eliminate inappropriate antibiotic and antiviral treatment. Due to the severity of clinical presentation, broad-spectrum antibiotics may be administered prior to sample collection, which lowers the recovery of bacteria reducing the chance of detecting the pathogen. Enteroviruses are the cause of up to 90% of aseptic meningitis, where the cause is known [17]. The use of molecular testing for Enteroviruses has been linked to better patient management, decreased time of stay in hospital and less use of broad spectrum antibiotics [18,19]. Other known viral aetiologies are also commonly tested for in routine or reference laboratories. However even with increased virus target panels, the detection rate can be as low as 14.4% [20], in patients where there was strong symptomatic evidence of a CNS infection. The diagnostic gold standard for bacterial meningitis is still culture; however, the recovery of bacteria is greatly hampered by the previous administration of antibiotics. Real-time PCR has been shown to have a sensitivity of over

95%, when looking for the three most common aetiologies of bacterial CNS infections, (*Haemophilus* influenzae, S. *pnuemoniae* and *Neisseria meningitidis*).[16]

Detection of pathogens above regional flora is often a difficult and subjective process as pathogens may be present in low numbers, be slow growing or have complex growth requirements. The upper respiratory tract is a complex, dynamic eco-system of bacteria, and pneumonia is a major cause of mortality and morbidity, meaning identification of pathogens above this flora is very important**.** Atypical pneumonia is mainly differentiated from typical pneumonias by the presence of extra-pulmonary symptoms. Six specific aetiologies have been described as causing this syndrome (*Chlamydophila psittaci, Francisella tularensis, Coxiella burnetii*, *Chlamydophila pneumonaei*, *Mycoplasma pnuemoniae* and *Legionella pneumonaie*). Typical pneumonia isolates are often grown on selective media plates from sputum, and a semi-quantitative method is used to separate carriage from infection. However, this is not a practical approach with the aetiologies of atypical pneumonia, as the bacteria are difficult to culture due to their intracellular or very fastidious nature. Differentiating the cause of pneumonia is particularly important, as treatment varies depending upon the aetiology. Antigen tests are often used for diagnosis of atypical infections; however often these need to be paired samples to demonstrate a change in titre or conversion to IgG and rates of obtaining second serum sample (convalescent serum) are as low as 10% [21]. Real-time PCR test are sometimes used to support the diagnosis of these pathogens but the sensitivity is often low, around 52% [21] although when used in addition to serology results for *M*. *pneumoniae* RT-PCR increased case detection by over 20% and so a combination test here is recommended.

The use of accurate real-time PCR along with standards allows the number of copies of a virus (viral load) to be established. This is of particular importance for chronic blood borne viruses, such as HIV, where it is used to monitor disease progression. By monitoring the viral load of patients, indications of treatment failure can be identified faster than simply relying on clinical indications.

Diagnosis and disease monitoring in microbiology has been made faster, simpler and safer through the introduction of molecular techniques. Automation of extraction and reaction techniques has consolidated sample processing, increasing sample throughput and lowered the user errors which delay results. Use of specific primers has been very successful at quickly and accurately detecting some pathogens above flora. However, this is limited when the same species of bacteria may be present as commensals but also have the ability to become pathogens, such as S. *pneumoniae* and *Staphylococcus aureus*, which are both major causes of infections in multiple sites. Although the use of molecular diagnostics has reduced the turnaround time to diagnosis,

especially for slow growing organisms such *M. tuberculosis*, it has also reduced the amount of information that we gain about the infection. PCR techniques will amplify where primers bind, whether the organism is viable or not, this is an advantage in cases where previous antibiotic therapy hinders culture, however the method cannot be used as a test of cure, as segments of DNA/RNA may persist long after the pathogen has been killed. Multiplexing of PCRs has been a powerful tool, allowing symptomatic detection panels to be put together, so a single test can detect multiple known aetiologies. However, for molecular diagnostics to be successful, the possible aetiologies must already be known. For example amongst CNS disease whilst some aetiologies are known, it is generally accepted that that vast majority are unknown [20]. As for all targeted molecular diagnostics, results are limited to the selected targets potentially missing cases with unusual pathogens. Specific primers need to be designed and the targets need careful selection to ensure binding to stable areas of the genome, preventing primer drop outs. However, this needs to be balanced with specificity to prevent false positives. Multiplexing of targets may restrict the selection of primers, as the binding temperature and amplicon length need to be similar. Most assays simply detect a certain genomic region, giving positive or negative results, limiting the information about characteristics and transmission. Viruses often have a high mutation rate and some have mechanisms for re-assortment of genes marking target selection and monitoring very important. Many emerging infectious agents are viruses, with potential to spread quickly around the world. With current targeted methods, novel infections would either be missed, or diagnosed as a routine pathogen. Targeted detection would miss gene and plasmid acquisition as well as gene re-assortments such as those seen in Influenza. Ultimately the short fall of current targeted molecular test is the need for knowledge of the targets. For molecular techniques to offer insight into novel pathogens more information needs to gained, in a way that is less targeting than current techniques.

1.2 Accurate Characterisation of Pathogens Allowing Suitable Treatment and Effective Epidemiology

After identification of a pathogen further characterisation might be needed to identify the best patient management, including the best treatment options. Additionally, characterisation may inform patient isolation, epidemiological or public health studies.

Traditionally phenotypic methods have been used to look at antimicrobial resistances. This can include dilution broths, discs or E-tests. Automation with machines such as the Phoenix (BD) or Vitek (bioMérieux) systems have allowed higher throughput and more accurate interpretation of minimum inhibitory concentrations (MICs), as well as automated interpretation based on the isolate being tested. However, for highly resistant organisms more manual methods are needed such as E-tests, which provide a more accurate MIC. As for all phenotypic tests large numbers of pure bacteria are required. Whilst a phenotypic test gives valuable information for treatment guidance it gives no indication of mechanisms, and is not discriminatory enough to provide in depth epidemiological insight.

A valuable use of molecular tests their potential to provide treatment guidelines in a rapid manner, such as resistance indicators, or pre-clinical indications of treatment failures. Among Enterobacteriaceae there is increasing concern about drug resistance, particularly to beta-lactam antibiotics. Extended spectrum beta lactamases (ESBLs) are particularly concerning as they limit the number of possible therapies for treating infection, they are also highly mobile and able to spread between multiple species. ESBL prevalence has been increasing with 1% of all clinically isolated Enterobacteriaceae carried an ESBL[22]. Traditional phenotypic methods often take 48 hours after isolation to confirm the presence of an ESBL, which potentially leads to inappropriate treatment. Use of PCR for the detection of these genes allows quicker detection of ESBLs and improves patient treatment. There are several types of ESBLs, with the most common being TEM, OXO, SHV and CTX[22], due to the high number of types and subtypes, multiplex PCRs are often required to detect their presence. PCR results have been shown to be concordant with phenotypic methods, as well as decreasing time to detection and potentially being a cheaper method [23]. Although most laboratories use in-house tests, commercial tests are available which have been shown to have a sensitivity of nearly 99% and a specificity of 100% [24].There is also potential that PCR could be used for direct detection of bacterial resistance genes directly from clinical samples, reducing turnaround times. However, targeted resistance detection is not able to replace phenotypic tests due to the complex nature of resistance acquisition especially within Gram negative organisms. For example, when looking at detection of resistance due to ESBLS, new or unusual ESBL genes will not be detected, as they won't be included in the targets. Some resistant

bacteria employ multifactorial mechanisms, such as over expression of genes and efflux pumps leading to increases in MICs. PCR genotype and phenotype do not always correlate and false negatives may arise from novel mechanism or mutations in primer sites. Additionally, false positives may occur from gene drop-outs, or disruptions where the primer sites are still intact. Use of targeted assays does not give information about these more complex resistance mechanisms.

Sequencing is used for detection of drug resistant mutations in chronic viral infections such as HIV, Hepatitis C and B. Current guidelines for HIV drug testing in Europe recommend testing patients when they are first infected and when there is suspicion of treatment failure (usually a rise in viral load) [25]. Retro viruses have high mutation rates, with some genes being as high as 10^{-3} [26], this leads to high variations in the population of HIV even within a single person. Drug targets are under particular pressure to mutate, and accumulation of mutations in drug targets can lead to drug failure. It has been estimated that every possible single point mutation occurs between 10^4 and 10^5 times per day in an untreated HIV-1-infected individual [27]. Because of the high variation in the HIV genome, several sets of primers, including redundant bases may have to be used to gain successful amplification. It is often a complex test which requires special training and large amounts of hands-on time. Because of the presence of quasi species Sanger sequences are often interpreted by hand, before the consensus sequence can be loaded into prediction databases. The ability to reliably call quasi species is limited to those that make up at least 20% [28] of the total viral population. The interpretation of the resistant mutations is complicated and often relies on public resources which have to be updated often [29 25] to reflect discovery of new clinically important mutations.

It is often important to be able to differentiate between an outbreak and sporadic cases of infectious diseases. Identification of outbreaks allows infection control and public health measures to be implemented to prevent the further spread of infection in the community. Techniques need to discriminate between isolates of the same species, whilst targeting stable elements present in all isolates. Typing schemes should be rapid, reproducible, comparable and have high resolution. At present there are several different schemes to type bacteria including phenotypic tests, such as serotyping, biotyping and phage typing. However, these often have low sensitivity and are only useful to rule out pathogen links. Molecular typing tests are now commonly used in reference laboratories to identify potential epidemiological links and several different methods are used. Current typing schemes are often species specific, labour intensive and highly specialist. Choosing a typing scheme currently depends on the pathogen, the epidemiological context and the level of discrimination required.

The most commonly used method is Pulsed-field gel electrophoresis (PFGE)[30], and is used for many bacterial species, including *Escherichia coli*[31], *Streptococcus agalactiae*[32] *and Pseudomonas aurginosa*[33]. The method briefly consists of using a rare cutting restriction enzyme to digest the genome of the bacteria. The product is then separated on a gel and the banding patterns used to type the bacteria. The method takes into account whole genomes, and will show any large insertion or deletion events. There are protocols that have tried to establish schemes that can be compared between laboratories, such as Pulsenet[34]. This allows emerging clones to be identified in multiple sites and countries. However, this technique is labour intensive, time consuming and has limited resolution. The technique also requires the bacterial to be cultured to high numbers.

Alternative typing methods take advantage of repetitive elements in bacterial genomes, in the form of variable number tandem repeat (VNTR) typing or Repetitive-element PCR typing. Both of these methods involve amplifying the DNA from the repetitive elements and using a gel to distinguish between the strains. These methods are often cheaper and quicker than PFGE and don't require such high amounts of starting material. However, inter-laboratory reproducibility can be poor due to the low –resolution of agarose gels and overall reproducibility is still questionable[35].Although lower starting material is required, a minimum of 25 ng is recommended for the Repetitive-element PCR[35]which still equates to around $5.5x10^8$ copies of the bacterial genome. The typing schemes are also tailored to each pathogen, with different primers and grading schemes used.

To overcome some of the problems with reproducibility using fragment size analysis, sequencing typing methods are widely used, either based on a single or multiple loci on the genome. The main advantage being robust standardised international databases for many pathogens, with their own the nomenclatures allowing multiple sites to compare strains, which allows clones to be identified across multiple countries[36]. Hence although the method lacks some of the discrimination of VNTR analysis, its standardisation has allowed it to be used across multiple sites and monitoring of clones over a time period. For other bacteria, such as *N. meningitidis,* multiple genes targets are selected, usually 7 housekeeping genes[37]. For each gene a number is assigned to each unique sequence, giving a 7-digit code that is the multi-locus sequence type (MLST) for that isolate. These can then be compared to other isolates.

The use of DNA microarrays allows greater number of loci to be examined by attaching complementary probes to a solid surface[38]. The results are recorded using an automated scanner, removing human based interpretation errors. The technology has been used to look for SNPs, and plasmids which are often ignored by other typing methods. This has allowed for the identification

of multiple virulence gene variants [39]. It has been particularly successful in identifying virulence targets in *E. coli*[40], which is known for its high genetic diversity, allowing novel combinations of virulence genes to be detected. The use of microarrays allows a large number of targets to be used, increasing flexibility and discrimination. However, a prior knowledge of gene targets is required, making them inappropriate for totally novel outbreaks or large scale gene shuffling. They also require specialist equipment and training and will also only detect presence or absence or genes or mutations present on the chip.

Determining how closely related isolates are is important for determining the epidemiology of infections. Effective methods can be used to trace infections globally, and be used to inform infection control practices. However most of the current methods used are complex, time consuming and expensive. The typing schemes are often highly focused to a single organism, are often difficult to compare across sites, and have limited discrimination. Current schemes offer limited flexibility to adapt to outbreaks in novel organism or genotypes. Clonal typing is also important in reconstructing evolutionary changes in pathogens and acquisition of genes such as pathogenicity islands and targeted typing techniques may provide limited data on this.

1.3 Diagnosis of Infections of Unknown Pathogens and Pathogens with Novel Characteristics

Determining the nucleic acid sequence of a segment of a genome can be used to identify, characterise or find novel properties of infectious agents**.** A major application of sequencing has been to use the technology to detect bacteria in samples which were culture negative, or from colonies which were difficult to identify by traditional methods. Prior use of antibiotics, poor sampling or fastidious organisms can lead to cultures being negative despite a strong suspicion of an infection. With increasing medical intervention and immunosuppression, opportunistic pathogens are increasingly being isolated, which may be more difficult to identify.

A major tool for identifying bacteria is the use of the variable region of the ribosomal gene, known as 16S rRNA. This target has been extensively used due to the presence of highly variable regions, and highly conserved region which can serve as primer targets. This has been used for both single isolate identifications within clinical laboratories and for identifying bacterial mixes in bacterial community profiling studies. Diagnosis using 16S rRNA has been used for a variety of clinical samples including tissue and fluid samples from patients with endocarditis, CNS and bone and joint infections [41]. It has often been shown to increase the sensitivity, and decrease the time to identification. This gain is particularly notable in cases of fastidious organisms, especially in the case of endocarditis. One study[42] identified a pathogen in 42.9% of culture negative samples,

with the process being particularly good when previous antibiotic therapy had been administered. The sequence of the 16S rRNA gene is used to identify bacteria that are not commonly encountered, such as fastidious Gram negative rods where traditional methods, such as biochemical assays, give little or no insight into the bacteria identity. In one study 80% of identification of fastidious Gram negative rods was improved using 16s sequencing[43]. Other major reports have shown the use of 16S rRNA sequencing improves identification in cases where cultures are negative, but there is high suspicion of bacterial infection[42]. Previous antibiotic therapy is stated as a major reason for culture negative results, the use of 16S rRNA sequencing has been shown to allow detection of bacteria even under these circumstances [42]. 16S rRNA sequencing has also proved useful in identifying novel pathogens in groups of patients which are immune-compromised such as the detection of *Mycobacterium genavense* [44]. Complex communities can also be studied using the 16S gene; such as studying the oral bacteria community. Culturing these bacteria often proves difficult, with bias towards the faster growing less fastidious bacteria, often the communities exist due to interactions between bacteria, which is difficult to mimic in vitro. A study looking at bacteria associated with periodontal disease [45] identified 596 bacterial species, over half of the species identified were associated with periodontal disease and were previously uncultivated. Several studies have been performed to try and ascertain the core microbiome of humans [46,47,48] showing variation over time, disease state and with the use of antibiotics. However, the information given by use of the 16S gene is limited to identification of the genus or species of bacteria, and often relies on public databases which often have sequencing errors. The technique is limited to those organisms that have a binding site for the selected primers, leading to bias in the identification technique.

Whilst 16s rRNA sequencing has had significant impact on detection of uncultured bacteria in both clinical and metagenomic studies, it lacks resolution to distinguish between some bacterial groups, particularly the Enterobacteriaceae. Alternative gene targets have been investigated, such as the *rpoB* gene, which has both highly conserved and highly variable regions, so with careful selection of primers it can be used to identify bacterial species. A 512 base pair region was used to identify 14 Enterobacteriaceae strains, within this group the rpoB gene differed from 2-22%[49]. It was particularly good at distinguishing between *Escherichia* species which all have identical 16s rRNA sequences. It is also only found in a single copy in genomes, unlike 16s rRNA, which may have several copies complicating the interpretation of sequence.

Targeted amplification using housekeeping genes allows for detection and sequencing of bacterial species and has been used for both diagnosis of infection as well as for metagenomic studies[49–51]. However, the quality of results depends on the selection of primers. No primer set is

without some bias, even a 'universal' primer such as those targeting the 16s rRNA gene, studies have shown different primers give differences in taxonomic compositions from various bacterial communities [52]. Using a bacterial target will ignore any viral or eukaryotes within the sample, the high variability of viruses means there are no universal markers for viral detection. For identification of truly unknown viruses, either random priming or none amplification techniques need to be used. Random priming has been used to look at the virome of the respiratory tract and has identified more than 39 species including a novel Rhinovirus which would have been missed with specific primers used in diagnositics[53]. Using none amplification techniques has also proved a strong method for viral discovery, samples from the oropharyngeal were sequenced and homologues to an algae infecting virus were found [54]. This showed a rare cross-kingdom infecting virus. Further studies showed infection with this virus was associated with impaired cognitive function. Demonstrating the potential power of diagnostics using none specific priming methods.

Recent advances in sequencing technologies have allowed dramatically increased sequence data output, for decreasing costs. In 1977 dideoxy chain termination sequencing was inveted[55], which led to the first completed bacterial genome to be published in 1997[56] . In 2005 one of the first commercialised high through put sequencers, the Roche 454 Gs20 was released. Two years later the Solexa system (now Illumina) was launched. These new technologies gave massive improvements in data output, which permitted a study of more than 3000 *S. pneumoniae* genomes to be conducted in 2014[57]. Analysis of genomes in such high numbers allows comparative analysis between isolates, but also allows bacterial genomes not just to be seen in isolation but also as an interacting community. There are three systems currently used in microbiology laboratories the Roche 454 system, the Ion Torrent system from Life Technologies and Illumina platforms. Each of the sequencing systems has both bench top and large scale versions of their systems. These next generation sequencing (NGS) platforms have three key stages for sequencing, which vary in mechanisms depending upon the platform.

The first stage is the preparation of a sequencing library. This includes the extraction, (possible amplification), fragmentation and tagging of the DNA to be sequenced. The input amount required varies, between platforms and library preparation method, from 1 ng total DNA to 500 ng total DNA. Fragmenting the library can be achieved through numerous methods, including physical sheering followed by end repairing, or enzymatic shearing. The first is often quicker and cheaper, but requires higher DNA input and may result in more loss of material. The later allows less input and fragment size can be controlled by varying the incubation time but is more likely to produce insertion and deletions [58–60]. Adaptors are then added, varying by sequencing technology.

The second stage for sequencing is the amplification of the template molecules. This is achieved through attaching the molecules to a solid surface. In the case of Ion torrent and the Roche 454 system beads are used, these are then enclosed in an aqueous phase micro reaction (emulsion PCR). Initially a single DNA molecule is added to each bead and then amplified, so many copies of each molecule is present. The beads are then concentrated and cleaned before being loaded onto the sequencing chips and sequenced. For the Illumina systems a flow cell is used, the amplification is automated and performed on the sequencing instrument itself[59].

The sequencing chemistries are the main variation between the platforms. The 454 and ion torrent systems flow a single base across the cell at a time, if the base in complementary and added a phosphate and a hydrogen ion are released. On the 454 system the released phosphate group is detected through the reaction with luciferin to create osylucuferin, producing visible light. The intensity and order of the light is recorded and converted into a DNA sequence. The ion torrent system uses a similar method, but the release of hydrogen ions is detected directly through pH change. Both of these methods are prone to errors in homopolymer stretches. The read length of the 454 system is around 1000 bp on the FLX system and 500 bp on the Junior system. Ion torrent read lengths are around 200 bp. The Illumina system uses reversible termination nucleotides, which are fluorescently labelled. Upon incorporation of a base an image is taken and then the signal is quenched and the terminator removed allowing further base additions. Read lengths are around 250 base pairs, using paired end (2x250)[59,60]. The output of these sequencers vary from 100,000 reads (454 junior) to 3 billion reads (Mi Seq on high output mode). Sequencing run times vary from 3 hours (ion Personal Genome Machine (PGM)) to 11 days (Hi seq high output mode). Selection of platforms is dependent on the required outcome, including turnaround time, number of reads required and sample through put.

Although short read technologies are the most used techniques for producing bacterial sequencing, difficulty in placing repetitive elements and fully mapping chromosomes means that it is not always possible to fully identify novel resistance and virulent islands in pathogens. Long read technologies such as the MinION nanopore have been used to fully characterise these gene structures[61]. An alternative long read technology for closing gaps in bacterial genomes is the PacBio sequencer[62]. The PacBio system from Pacific Bioscience allows read lengths of up to 15,000 base pairs to be achieved by direct detection of base incorporation in single copies of the DNA. As labelled bases are incorporated the light is detected and recorded, in contrast to other platforms, no amplification of the library is required as the incorporation on a single molecule can be detected.

Next generation sequencers have already been used within microbiology for many studies, including epidemiological, resistance detection and pathogen discovery. The additional data produced using these technologies have allowed the whole genome of bacteria to be studied, increasing the resolution of typing and allowing novel pathogenic mechanism to be classified. Next generation sequencing technologies have been used to track outbreaks between countries, such as the 2010 outbreak of *Vibrio cholera* in Haiti, which allowed its origin to be determined as Nepal [63], other studies into the lineage of *V. cholera* have tracked its global spread using 154 whole genome isolates from varying temporal and global points [64]. This study supported the theory that the outbreak had multiple waves of spread, rather than a single point outbreak.

Beginning in May 2011 there was a large outbreak of *E. coli,* which caused a high number of haemolytic uremic syndrome (HUS) cases (over 900) [65]. It was highly virulent, antibiotic resistant and was sorbitol fermenting, meaning it was not easily identified in routine laboratories, (where selection for non-sorbitol fermenting 0157 is usually performed). Multiple sites used next generation sequencing to allow identification of the pathogen. One study used a metagenomic approach to identify the pathogen[66]. DNA was extracted directly from stools, fragmented and sequenced using the Illumina HiSeq platform. Sequences present in in a least 20 outbreak stools, but not present in healthy controls, were selected as gene tags to indicate infection. These were then used to identify and isolate signals in the infected patients. A draft genome was produced using the sequence data gathered, which could then be used to improve isolation of pathogens signals from host background. Using this culture independent method, other pathogens could also be identified, such as *Campylobacter jejuni*, which allowed these samples to be removed from the study. The sensitivity was 67% in this study, showing the potential for non-culture based methods to detect novel pathogens. Next generation sequencing technologies were also used to identify the lineage of the outbreak, [65,67,68]. Using isolates and varying sequencing technologies including Illumina and Ion Torrent, the progenitor was identified and gene acquisition mapped. One of the studies sped up the analyses by allowing the data to be open source [67],demonstrating the power of data sharing allowing simultaneous multi-site analysis.

Whole genome sequencing (WGS) has been used to predict the antibiograms for a variety of organisms including, *S. aureus* [69], *M. tuberculosis* [70] and *E. coli* [6]. The advantage of WGS is that additional information such as phylogeny is also available as well as the prediction of methods of acquisition of new resistance genes [69]. Enterobacteriaceae spp**.** pose a serious threat due to their ability to cause infections in multiple sites, and for the ability to spread and acquire new resistance mechanisms. Therefore, the ability to rapidly identify the best treatment for infections would be advantageous. However resistant pathogens often have multiple resistance mechanism

spread across their genome and plasmids. A study comparing phenotypic and genotypic results for susceptibility to seven antibiotics showed good concordance for *E. coli* and *K. pneumoniae*[6]. The advantage of the genotypic method was the mechanisms of resistance were also identified. A reference data base was required for the study, and the results will only be as good as the reference database allows. Online databases are being developed that allow the prediction of antibiotic resistances such as ResFinder [71]. Online databases have the advantage of being accessible to anybody, but can only be updated by specific people allowing quality to be maintained. In a study of 200 isolates using the Resfinder server high concordance (99. 74%) with phenotypic results was found, with 6 out of 7 disagreements being for a single drug in a single bacterial species (spectinomycin in *E. coli*) [72]. Whole genome sequencing can also be used to determine the methods of resistance that are not just associated with the presence or absence of a gene. For example a study of *Acinetobacter baumannii,* identified factors such as a SNP in a signal transduction histidine kinase, which is implicated in the up regulation of an efflux system[73]. This was associated with an increase MIC to tigeclyine. Another study, of the whole genome *S. pneumioniae* [74] showed mutations upstream of a ATP-binding cassette transporters (ABC) transporter gene was associated with resistance to linezolid. They also found mutations in a hypothetical protein was also associated with resistance to linezliod and a further study showed that the mutations were in a region predicted to contain an RNA methyltransferease. An ortholog of this gene was found to also be mutated in a resistant *S. aureus* isolate.

While NGS technologies have improved our ability to investigate infectious diseases, the identification of these pathogens remains restricted by the need for initial culture and use of selective methods. This has biased the use of NGS to infectious agents to those that can be readily cultured and harvested from the media. To truly take advantage of the technologies available, methods for detection and sequencing of pathogens direct from clinical samples need to be developed. Many clinical samples have a low number of the infecting bacterial cells and so amplification of the target organism's DNA will be required before sequencing will be possible. Most molecular techniques used in clinical laboratories currently involve amplification of specific targets, to either identify known targets, or amplifying conserved regions within highly diverse genes to allow identification through sequencing (e.g. 16S rRNA). To be able to diagnose clinical infections the amplification method would need to be able to amplify without prior knowledge of the bacteria being looked for.

One of the major challenges of using next generation sequence technologies is reconstructing a consensus sequence for the produced short reads

Initial stages focus on the raw data produced by the sequencer, which in most cases is in fastq format, however when using the 454 technology the raw data is written into a SFF (standard flowgram format). Preliminary investigation focusses around the quality of data produced by the sequencer, including investigating the read length and read quality. The quality of a base call within a read is measured using a Phred score. The Phred score (often shown as a Q score), indicates the probability that the base was called incorrectly. For example, if a base had a Phred score of Q20 there would be a 1:100 chance that the base was incorrect, giving it a base call accuracy of 99%. The first stages of an analysis pipeline is often to remove poor quality bases, either by removing whole reads of poor quality, or trimming regions of reads with poor quality bases.

The majority of data analysis when using next generation sequencing data for identification is focused around piecing the short reads together, known as assembly. Assembly can be based around a known reference genome, or can only use the read information itself. *De novo* assemblers try to piece together these segments without a reference to guide them, which can be computationally challenging. Several different algorithms have been developed and applied to meet this challenge[107]. One of the earliest assemblers applied to next generation sequence data was SSAKE (Short Sequence Assembly by progressive *K*-mer search and 3' read Extension)[110] the algorithm used by SSAKE is known as the greedy graph algorithm. Firstly, the reads are entered into a hash table with each read given a unique keyed sequence. A prefix tree is then built using the first 11 bases of the 5' end of the reads. Each unassembled read is used in turn and the highest scoring overlap is used to make the next join.

Newbler assembler was released by Roche, which is an overlap layout consensus assembler[109]. Initially seeds of the reads are produced in the form of 16mers, each seed being 12 bases from the previous seed. If two reads have identical seeds the programmed tries to overlap them, with a minimum overlap of 40 base pairs and a minimum stringency of 90%. This then form a consensus known as a unitig, which is a preliminary highly confident contig or mini-assembly. These unitigs are then joined by pairwise over laps to form a larger consensus.

Many recently developed assemblers apply the de Bruijn graph approach to assemble genomes. Each read is decomposed into substrings of a given length (k) known as a kmer, each of these kmers overlaps the previous Kmer by k-1. Each kmer then becomes a node in the graph, with overlapping kmers connection nodes at their edges. Bubbles may form in the graph due to errors, repeats or genome variations. The graph is then compacted by reducing linear chains to single nodes. Paths are then mapped across the graph, with the interpretation of the graph being

dependant on the specific algorithm used by each assembler. Mostly a Eularian approach is adopted, which is the route across the graph which crosses all nodes without re-visiting any. Abyss[111] is an example of an assembler that used this method. Spades[112] is also uses the de Bruijn graph approach as its core algorithm, but has further adaptations for resolving bubbles in the graphs, using a multi-sized de Bruijn graph approach and distance histograms, which improves consensus accuracy. Ray[113] uses a two stage approach that initially uses kmers to construct the graph and then uses the full read information to improve the accuracy of the assembler.

Mira-4[114] initially establishes the relationship between reads using a bitmap algorithm, which states whither a given string contains a substring which is approximately equal to a given pattern. All the potential overlaps are then evaluated based on the length and quality of the overlap and given an alignment weight. This information is then used to form a weighted graph. A pathfinder algorithm is then used to form contigs, it starts at the node edge which has the highest weighting, and moves through the graph using the highest weighted route.

When assembly is completed there are many ways in which to judge its quality, and usually a combination of factors is used. The proportion of the genome covered is a good indication of sequencing success, however coverage will be impacted by the closeness of the reference to the strain actually sequenced. The Lander-Waterman theory assumes that reads map randomly across the genome and coverage is dependant on the read length, number of reads and the size of the genome itself. Another major consideration is the depth of coverage, which is how many reads are assembled the same point in the genome. As depth at a single point increases the accuracy of the consensus also increases, when looking for SNPs compared to a reference, depth of coverage is particularly important to provide evidence of a true SNP. Additionally, when comparing mixed populations increased depth is needed to separate variants from sequencing errors. Other factors when assessing the quality of the assembly includes the number of contigs the genome was assembled into, a contig refers to over lapping sequencing that have been assembled with no gaps. When there is high genome coverage fewer contigs indicates a better genome assembly. Another metric that can be measured is the accuracy of the assembly, especially when using *de novo* techniques, this is achieved by mapping the assembly back to a reference and isolating areas when a misassembly has occurred, again this metric depends upon how related the strain is to the reference.

1.4 Unbiased Sample Preparation for Next Generation Sequencing

To be able to effectively use whole genome sequencing for diagnosis of infections a rapid, robust and unbiased sample preparation for NGS is needed. Current applications of next generation sequencing are usually organism specific, using either cultured bacteria or amplicons. To sequence a genome, at least 1ng is needed, which is the equivalent of over a 200,000 copies of the *E. coli* K12 genome. Mostly this is achieved by culturing the organism and then extracting the DNA, however this is labour intensive and culturing is not always possible. Mixed infections are also a problem for many current protocols, which rely on isolation of pathogens before sequencing. This may mean pathogens present in lower numbers, which are slow growing, or which are not targeted by the culture conditions are not enriched and therefore not sequenced. It is not possible to isolate all infectious agents and so detection and sequencing of nucleic acid is a key diagnostic. There is a need for the development of a rapid, highly sensitive and accurate unbiased method for detection of infectious agents directly from clinical specimens.

The need for rapid and accurate amplification is of particular importance when using the 454 technologies as there is a requirement for a high concentration of DNA to be input into the library preparation, (500ng DNA total). Most current methods for bacterial sequencing involve isolating the bacteria and culturing the bacteria to a high enough number to provide DNA without need for amplification. Where amplification is performed before next generation sequencing it is targeted, amplifying only specific genome regions. By using a strand displacement enzyme to produce genomic DNA for sequencing, the need for culture and targeted amplification can be removed. This will allow production of sequencing that represents the DNA in the sample rather than just the organisms that can be cultured or targeted.

To achieve this goal, the use of novel enzymes will be investigated, as well as novel applications of previously used enzymes. An important application for unbiased whole genome sequencing will be for emerging infectious diseases. Of the many emerging infections detected over the past 20 years most have been viral and one third have been RNA viruses[75]. The most common method for viral detection, for both DNA and RNA, is the use of PCR with specific targets. However, detection of RNA viruses requires additional processing, often relying on random conversion of RNA to DNA using a reverse transcription (RT) enzyme, followed by specific target detection using PCR. RNA conversion is often a limiting step for detection, in terms of both the sensitivity of the assay and the additional time needed for cDNA production. A key

requirement for enzymatic RT is thermostabilty enabling for better transcription through secondary structures, which may cause conversion issues. One of the most commonly used reverse transcription enzymes is SuperScript III (Thermo Fisher Scientific) which produces longer transcripts than other RT enzymes. The sensitivity is also higher, which is quoted as 0.1 pg of RNA[76], this is equivalent to 25654 copies of the HIV virus. Low viral load is around 40 copies/ml, depending on the detection limit of tests used, which equates to 0.000156 pg. Superscript IV (Thermo Fisher Scientific), a modified version of Superscript III, is a more rapid reverse transcriptase, which is less susceptible to sample inhibitors. SuperScript IV (SSIV) is quoted as having better sensitivity than SSIII, yet the company only quotes lower Ct values at 1 pg RNA input compared to SSIII, to better determine their sensitivity and suitability for WGS both of these enzymes will be investigated for detection of targets at very low RNA input.

A study in Yellow Stone National Park[77], identified a thermostable polymerase, which functions as both a reverse transcriptase and DNA polymerase. The enzyme known as PyroPhage 3137 has a processivity score (number of bases added in a single binding reaction) of 47nt[78], which is high in comparison to Taq (9), although much lower than ϕ29 (>70,000). The enzyme fidelity of $8x10^4$ is comparable to Taq ($1.4x10^4$), and in addition to its one step conversion of RNA to DNA and amplification of DNA the enzyme has strand displacement ability, along with the capability to extend from Nicks in the DNA. A study performed by Moser et al[78], indicated that the performance of PyroPhage 3137 was comparable to the two-enzyme system of SuperScript III, with the advantage of a single enzyme protocol. Two versions of the enzyme are commercially available, the wild type, and an exonuclease knock-out version, which is more amenable to reverse transcription, but has a slightly reduced fidelity at $0.9x10^4$. Both versions of the enzyme will be compared for sensitivity and suitability alongside the SuperScript III and IV enzymes.

Several distinct groups of DNA polymerases have been identified (Family A, B, C, D, X and Y), often having a similar structure, whilst having little or no sequence homology, showing convergent evolution and demonstrating their importance in the maintenance and replication of DNA[79]. Different polymerase families have been associated with different functions possessing different enzymatic attributes. Family A polymerases are mostly single unit with highly functional exonucleases involved in DNA repair and Okazaki fragment maturation[80]. Whereas Family B polymerases often consist of multiple subunits and perform the bulk of DNA replication, they also contain exonucleases and have the highest fidelity and processivity[81]. They have also been found to be the main replication enzymes within bacteriophages. Group C and D are less characterised than other DNA polymerase families, but have been shown to be replicative in prokaryotic and eukaryotic cells respectively. Group X polymerases have gap filling properties [82], and family Y are

responsible for replication of damaged and degraded DNA, and lack features such as exonuclease activity[83].

Recent advancements in molecular techniques has allowed modification of polymerase enzymes, to produce more desirable characteristics and increased production, allowing greater commercial availability for a variety of polymerases. Characteristics that have been particularly important for use in diagnostics include fidelity, processivity and thermostabilty that allow cycling with specific primers to be used. Due to its thermostabilty *Taq* polymerase has been the most commonly used DNA polymerase in molecular microbiology, however it has a relatively low processivity, incorporating 9 nucleotides per binding event, and high error rate $5x10^{4}$ [78], which for simple detection of short targets has little impact on result quality. However, for amplification and sequencing of whole genomes other enzymes which have lower fidelity and higher processivity need to be investigated.

Strand displacement enzymes have the potential to decrease our reliance on primers and increase the lengths of DNA produced through amplification. Multiple displacement amplification (MDA) is an alternative method to PCR for the production of DNA in high enough amounts for sequencing. This method has the advantage of being able to produce large lengths of DNA with lower errors than conventional PCR. MDA has been applied to samples with very low starting DNA amounts from single cells (both prokaryotic and eukaryotic) and provided DNA in levels high enough to perform sequencing [84].

MDA utilises a DNA polymerase initially isolated from a bacteriophage in *Bacillus subtilis*, ϕ29 DNA polymerase has two catalytic regions an exonuclease region at the N-terminal region and the polymerase in the C-terminal domain[85]. ϕ29 DNA polymerase is a highly efficient enzyme able to produce strand displacement during polymerisation process without the need for accessory proteins[86]. The DNA is copied repeatedly in a branching mechanism, with the ϕ29 DNA polymerase extending from random hexamer primers. The polymerase also displaces previously made copies of the genome, producing branching DNA[84]. The enzyme is very stable and can continue to work for over 12hours at 30°C. The ϕ29 DNA polymerase possesses a high processivity 3'-5' exonuclease which removes incorrect nucleotide incorporation. The high fidelity and 3'–5' proofreading activity reduces the amplification error rate to 1 in 10^{6}–10^{7} bases compared to conventional *Taq* polymerase with a reported error rate of 1 in 90,000 [87]. A single binding reaction can incorporate over 70kb [86], meaning that this method is not limited by the length of the initial target like conventional PCR. The use of exonuclease resistant hexamer primers increases the amount of amplified product produced by 20x, giving 10,000-fold

amplification of circular ssDNA[88]. It has been demonstrated that this method is suitable for use with single bacterial cells[89]. By using the high processivity combined with the low error rate of ϕ 29 very low copy numbers can be amplified quickly to produce enough DNA to sequence.

Other strand displacement enzymes will also be investigated such as *Bst* which is used for whole genome amplifications[90] as well as loop mediated isothermal amplification (LAMP), which has been used to increase speed of target amplification[91,92]. Other engineered enzymes such as Vent exo- and Klenow fragment, will be investigated, however often the strand displacement properties have been engineered by removing exonuclease portions increasing enzyme error rates. A recently discovered polymerase PyroPhage 3173, that has both DNA polymerase and reverse transcription activity [78] will also be studied. It has been shown to have DNA polymerase fidelity of $8x10^4$, which is comparable to *Taq* polymerase. It is a thermostable reverse transcription enzyme, with its most efficient working temperature being 72°C, compared to Superscript III (Life technologies), which has a maximum temperature of 55°C. Raising the temperature of the reaction with both increase the speed of the reaction and also lower the interference from secondary structure.

The sensitive of the enzymes will be investigated, using amplification of a single bacterial cell as the target. Enzyme bias will be assessed by using a variety of pathogens, including viruses, bacteria with varying GC contents, and mixed samples.

1.5 Project Hypothesis

Unbiased whole genome amplification and characterisation of pathogens directly from clinical specimens enables more in-depth analysis of pathogen or mixed pathogens infecting sterile samples than traditional methods.

1.6 Project Aims

Inherent biases in sample processing may result in undiagnosed infections, especially where pathogens are novel or unculturable. Current diagnostic techniques are based around isolation techniques, which aim to simplify mixed infections, therefore it is likely low abundant co-infecting microorganisms may be missed. Direct pathogen sampling and partial genome sequencing from infected sites will transform identification, better directing therapeutics and patient care. This could potentially impact on the identification and characterisation of low abundance pathogens, with numerous possible aetiologies. We propose to undertake unbiased amplification of pathogens present in low numbers and or mixed with normal flora from clinical material. A novel approach will be developed and evaluated, resulting in the generation of comprehensive genetic information about the characteristics and genetic composition of pathogens analysed. This information could be used to inform clinicians of the predicted antibiogram, toxin gene presence and virulence determinants, to better guide the decision tree related to treatment and control. Using whole genome analysis enables real time infection control decisions to be made, to help contain outbreaks in a rapid manner.

Objectives

- Develop a method for amplification of very low and mixed starting material that is rapid, limits the introduction of amplification errors and amplifies all regions of the genome (chromosomal and extra-chromosomal). Different enzymes and methods will be tested to find the most suitable mechanism generating high quality, informative data.

- Define the sensitivity of the assay, including single cell work, and low number cell mixes.

- Integrate sample preparation pathways to allow a workflow for unbiased diagnosis of both viral and bacterial pathogens to be created

- To examine approaches to allow pathogen signals to be amplified preferentially to host signals. This will allow unbiased amplification methods to be applied directly to environmental and clinical samples.

- An optimised bioinformatic pipeline will be developed, based on comparison of multiple quality trimmers and genome assemblers. The criteria used to identify the optimal pipeline will be genome assembly quality. The pipeline will aim to have an output that allows identification and characterisation of pathogens, including predictions for novel characteristics.

- A primer free amplification technique will be developed which will prevent the loss of information at genome ends and will incorporate a system to allow simultaneous amplification of RNA and DNA in a representative manner. Methods for tagging of samples will be explored to allow the initial diversity of a sample to be accurately determined.

- Multiple sample types will be tested including low starting number and mixed, GC content and pathogen genome size will be varied, in both control and clinical samples.

The methods developed will aim for a culture independent, unbiased method to produce whole genome sequencing data from mixed and low cell numbers. The method will aim to achieve a consolidated laboratory workflow for sample processing and allow full characterisation of a mixed sample from a single molecular assay. The methods can be applied for the detection of known pathogens, characterisation of novel features and potentially pathogen discovery. Additionally, the method could be applied to assess co-infections and novel pathogen interactions as well changes in flora in health and disease. It can also be applied to large scale metagenomic studies to give whole genome information.

Materials and Methods

2. Materials and Methods

2.1. Culture Conditions and Strain Acquisition for Pathogens used in this Thesis

2.1.1. Bacteria used as Controls

Control bacteria for initial experiments were selected from onsite PHE collections; reasons for selection were availability of completed reference sequences, variable GC content and different cell wall types. The selected bacteria were *Escherichia coli (E. coli) K12 substr. MG1655, Actinomyces naeslundii (A. naeslundii) (NCTC 10301)) and (Pepto) Clostridium difficile* 630 (*C. difficile* 630). For later experiments involving mixed cells, genome size was the main selection criteria, the two bacteria selected were *Haemophilus influenza (H. influenzae) (*NCTC 8143) and *Enterococcus faecalis (E. faecalis) (*NCTC 12201). Bacterial isolates were recovered from long term storage and cultured from beads stored at -80^{o}C and grown at 37^{o}C overnight on Columbia blood agar (CBA), chocolate Columbia blood agar (ChocBA) or Fastidious Anaerobe Agar (FBA), under aerobic, 5% CO_2 or anaerobic condition. Culture details and basic genome information is listed below **Table 2-1.** Agar plates are made and quality controlled onsite at PHE Colindale by the media service unit.

Strain	Agar	Incubation atmosphere	GC	Genome size
Escherichia coli K12 substr. MG1655	CBA	O_2	50%	4.6MB
Actinomyces naeslundii (NCTC 10301)	FBA	An O_2	69%	3.0MB
Peptoclostridium difficile 630	FBA	An O_2	29%	4.2MB
***Haemophilus influenza* (NCTC 8143)**	ChocBA	5% C O_2	38%	1.8MB
***Enterococcus faecalis* (NCTC 12201)**	CBA	O_2	37.5%	3.2MB

Table 2-1 Details of the culture conditions and basic genome information of bacterial strains used as controls in this thesis

2.1.2. Clinical Isolate Collection and Growth

Clinical isolates were collected from the long term storage from the microbiology department at the Royal Free Hospital Hampstead. These were sub-cultured from beads stored at -80^{o}C, checked for purity and the identity confirmed by MALDI-TOF before fresh beads were made and transported. Clinical information and isolate sensitivity were gathered based on final reports issued by the laboratory and all patient identifiable information was removed from reports before use. Isolates were selected using a variety of criteria; wherever possible the source was a sterile site, with a focus on blood culture isolates. Firstly the most common blood culture isolates in England according the study produced by the Health Protection Agency[93] were collected. Furthermore, isolates were selected for additional interest including antimicrobial resistance, difficulty in traditional identification and isolates of high clinical impact. The list of is olates with their culture conditions is shown in **Table 2-2**

	Isolate	Hospital ID	Source	Other info	Media	Incubation
1	*Escherichia coli*	15M154850	Blood Culture	ESBL	CBA	O_2
2	*Staphylococcus epidermidis*	15M033693	Fluid Aspirate		CBA	O_2
3	*Staphylococcus aureus*	15M154940	Blood Culture	MSSA	CBA	O_2
4	*Staphylococcus aureus*	15M150748	Blood Culture	MRSA	CBA	O_2
5	*Enterococcus faecium*	15M032581	Bone (Hip)	VanA	CBA	O_2
6	*Klebsiella pneumoniae*	14M181010	Blood Culture	ESBL	CBA	O_2
7	*Streptococcus pneumoniae*	14M180245	Blood Culture		ChocBA	5% C O_2
8	*Pseudomonas aurginosa*	15M154350	Blood Culture		CBA	O_2
9	*Pseudomonas aurginosa*	13M155577	Blood Culture	Resistant to Carbapenems	CBA	O_2
10	*Proteus vulgaris*	15M017752	Tissue (Foot)		CLED	O_2
11	*Enterobacter cloacae*	15M154841	Blood Culture	Inducible AmpC	CBA	O_2
12	*Streptococcus agalactiae*	15M036872	High Vaginal Swab		CBA	5% C O_2
13	*Bacteroides vulgatus*	14M180862	Blood Culture		FBA	An O_2
14	*Streptococcus oralis*	15M150586	Blood Culture		CBA	5% C O_2
15	*Streptococcus mitis*	15M150586	Blood Culture		CBA	5% C O_2
16	*Streptococcus anginosus*	15M152569	Blood Culture		CBA	5% C O_2
17	*Haemophilus influenzae*	15M154514	Blood Culture		ChocBA	5% C O_2
18	*Streptococcus pyogenes*	15M153838	Blood Culture		CBA	O_2
19	*Salmonella sp.*	14M180208	Blood Culture	WGS available	CBA	O_2
20	*Serratia marcescens*	14M154557	Blood Culture	Inducible AmpC	CBA	O_2
21	*Fusobacterium necrophorum*	14M163444	Blood Culture		FBA	An O_2
22	*Listeria monocytogenes*	15M152571	Blood Culture		CBA	O_2
23	*Actinomyces naeslundii*	14M346820	Bone (Femur)		FBA	An O_2
24	*Proprionibacterium acens*	14M37379	Tissue (Shoulder)		FBA	An O_2
25	*Paenibacillus anaericanus*	14M366429	Tissue (Shoulder)	ID by 16S sequencing	FBA	An O_2
26	*Clostridium butyricum*	14M180616	Fluid (hip)		FBA	An O_2
27	*Shigella sonnei*	15M005139	Faeces		CBA	O_2

Table 2-2 Details of clinical isolates collected from the Microbiology Department at the Royal Free Hospital Hampstead, including growth conditions and original infection site.

2.1.3. Tissue Culture of Viral Strains

2.1.3.1. Culture of Adenovirus

Adenovirus was selected as an initial test virus because of its DNA genome, the relative ease of culture and it's none enveloped capsid which allowed testing of the extraction method.

Frozen Human Embryonic Kidney 293 cells were retrieved from storage in liquid nitrogen, and gently warmed to defrost. 9 ml of Minimum Essential Media (Life technologies) supplemented with 10% foetal bovine serum (Sigma-Aldrich) and 1% penicillin/streptomycin (Gibco) was added to 1 ml of frozen cell stock. Cells were gently mixed and then centrifuged at 1000 rpm for 5 minutes. The supernatant was removed and 10ml of warmed media was added to a T25 corning flask (Sigma-Aldrich) and washed cells were added. The cells were checked for healthy morphology before the flask was incubated at 37^0C at 5% CO_2 using vented caps. After 24 hours the cells were 80% confluent and were passaged into T75 flasks. Briefly the Hank's balanced salt solution (HBSS) (Thermo Fisher Scientific) and trypsin 0.25% (Invitrogen) were added and the cells and incubated at room temperature for 2 minutes before removal of the liquid. The flasks were then incubated for 5 minutes at 37^oC, before 5ml of warmed media was added. 10ml of warmed media was added to the new T75 flask before addition of the washed cells. Cells were checked daily and once at 80% confluence, were split into fresh flasks for infection. The flasks were incubated for 72 hours as previously. After cells had reached 90% confluence they were infected with 5 ml supernatant containing Adenovirus 40, Adenovirus 41 or 5ml media for negative control. Samples were checked daily for signs of Cytopathic effect (CPE) and compared to the negative control. Once 100% CPE was observed the flask was frozen at -80^oC and thawed three times and the flask content was transferred to a falcon tube (Thermo Fisher Scientific). The sample was then centrifuged at 5000 rpm, 10 minutes, to pellet the cell debris; aliquots of 1 ml of the virus supernatant were then made and stored at -80^oC.

2.1.3.2. HIV Culture

HIV was used for initial work to investigate application of methods to RNA genomes , strain selection was based availability of a well characterised category two tissue culture model. Additionally, as its an enveloped virus, the sample processing pipeline could be tested on a more fragile virus.

8E5/LAV[94] cells, which are a sub clone of A3.01, a CD4+ CEM derived human T-cell line (HIV-1 subtype B, NIBSC repository reference ARP110), were defrosted by gently warming. The cell line, containing a single defective copy of HIV-1 proviral DNA, was grown in 90% RPMI 1640 (life technologies) and 10% foetal bovine serum (Life Technologies). Initially the cells were grown in

T25 flasks at 5% CO_2 until 80% confluence was achieved. After which the cells were transferred in to T75 flasks under the same conditions. Once 90% confluence was achieved supernatant from the flask was poured into a 50ml falcon tube, and cells pelleted at 6000xg for 20 minutes. The supernatant was then aliquoted into 1ml aliquots. The virus particle number was quantified using The TC20 automated cell counter (Bio Rad).

2.1.4. Viral Pathogens Externally Acquired

Control materials of a variety of viral strains were ordered from the National Institute of Biological Control (NIBSC) collection of Nucleic acid reference materials. The viruses selected were representative of different genome types, nucleic acid organisation and particle sizes. The list and details supplied by NIBSC is shown in **Table 2-3**. These strains were stored at -80^{o}C until use.

NIBSC product number	Virus	Additional information
08/310	Varicella Zoster virus (type B)	This preparation contains both infectious Varicella Zoster virus which has NOT been inactivated and Foetal calf serum. The control consists of a whole virus preparation of Varicella Zoster virus Type B diluted in a buffer comprising 10 mM Tris-HCl pH7.4 and 2% foetal calf serum. **Ct 30**
11/182	HBV DNA working reagent	The preparation contains a dilution of freeze-dried human plasma known to be positive for HBV DNA. The diluent is made from human plasma tested negative for anti-HCV, anti-HIV, and HBsAg. The reagent contains a genotype A isolate diluted in HBV negative plasma. **1140 IFU/ml**
11/208:	Parvovirus B19 DNA working reagent	The preparation contains a dilution of parvovirus B19 in normal human plasma tested and found negative for anti-HBsAg, anti-HCV and anti-HIV 1. the control consists of a dilution of Parvovirus B 19 genotype 1 diluted in normal human plasma **unitage of 800 IFU/ml (range 569-1039)**
07/294:	Norovirus (genogroup GII)	This preparation contains both infectious Norovirus genogroup (G) II which has NOT been inactivated and foetal calf serum. The control consists of a diluted faecal sample containing Human Norovirus genogroup II diluted in a buffer comprising 10mM Tris-Cl pH7.4 and 2% foetal calf serum. **Ct 30**
13/102:	HAV RNA working reagent	The preparation contains a dilution of Hepatitis A virus in normal human plasma tested and found negative for anti HBsAg, anti-HCV and anti-HIV1. The control consists of a dilution of Hepatitis A virus genotype 1B diluted in normal human plasma **Unitage 588 IFU/ml (range 507–680 IFU/ml).**
13/168	Measles (MVi/Moscow.RUS/0.88)	This preparation contains both Measles virus which has been inactivated and foetal Calf serum The control consists of a whole virus preparation of Measles Virus MVi/Moscow.RUS/0.88virus diluted in a buffer comprising 10 mM Tris-HCl pH7.4 and 2% foetal calf serum **Ct 30**
07/298	Influenza A H3N2 (A/Wyoming/3/2003)	This preparation contains both infectious Influenza virus which has NOT been inactivated and foetal calf serum. The control consists of a whole virus preparation of Influenza virus (A/Wyoming/3/2003, H3N2) diluted in a buffer comprising 10 mM Tris-HCl pH7.4 and 2% foetal calf serum. **Ct 30**
02/264:	HCV RNA working reagent	The preparation contains material of human origin, and either the final product or the source materials, from which it is derived, have been tested and found negative for HBsAg, anti-HIV and HAV RNA The reagent contains a genotype 3 isolate diluted in HCV RNA negative plasma. Each vial of the reagent contains 0.5 ml of liquid **assigned a value 1438 IFU/ml**
08/314	human cytomegalovirus (HCMV)	This preparation contains infectious human cytomegalovirus which has NOT been inactivated and foetal calf serum. The reagent consists of a whole virus preparation of HCMV strain AD169 diluted in a buffer comprising 10mM Tris-HCl pH 7.4 and 2% foetal calf serum. **Ct 30**

Table 2-3 Details of viral control media supplied by NIBSC, including viral strain, product number, details given by NISBC regarding the preparation of the control media and any quantification data supplied.

2.2. Evaluation of Potential Whole Genomes Using None PCR Amplification Techniques

Initial experiments focused on finding a suitable alternative to PCR that allowed rapid and accurate amplification of whole genomes without targeting specific genes. ϕ29 was the preliminary focus, due to its ability to accurately amplify DNA into long transcripts. Further work was carried out to attempt to remove the need for primers, as there are inherent biases even in random primers. Experiments were carried out to assess the compatibility of reverse transcription enzymes with DNA amplifying enzymes to allow methods to be applied to RNA pathogens. An additional enzyme was investigated, which is able to both convert RNA and amplify DNA, potentially allowing a one-step process for all pathogen types.

Initial bacterial work was performed using *Escherichia coli K12 substr. MG1655* which is well characterised strain, with a GC content of 50%. There is also a fully constructed reference genome available for this strain. To give a quick evaluation of genome coverage after amplification, a PCR was designed with ten gene targets spread evenly across the genome. Primers for this were selected using the NCBI primer designing tool. When looking at RNA conversion, RNA extracted from tissue culture grown HIV was used (describe in **2.1.3.2**). To investigate the length of transcripts created by the reverse transcription enzymes, a PCR was designed with varying lengths of amplified targets**.**

2.2.1. Use of ϕ29 MDA for Whole Genome Amplification Using Random Primers

2.2.1.1. Preparation of None Amplification Controls

The material from two overnight agar culture plates was suspended in 1 ml of sterile PBS, spun at 6000xG for 5 minutes, the supernatant removed and the pellet resuspended in 400 µl sterile PBS. The DNA was then extracted using the Qiagen QIAamp DNA Mini Kit, using the protocol for extraction from blood, with reactions scaled up from 200 µl to the 400 µl input. Briefly 40 µl of proteinase K and 400 µl buffer AL was added to the suspended bacteria and the mixture vortexed. This was incubated for 10 minutes at 56^{o}C before the addition of 400 µl 100% ethanol. 1240 µl of the product was then added to the spin column and centrifuged at 6000 x g for 1 min and the filtrate discarded; this was repeated for the remaining product. The column was then washed using 500 µl buffer AW1 and centrifuged at 6000 x g for 1 min, with the filtrate being discarded. The column was washed again by adding 500 µl buffer AW2, before an additional centrifugation at 20,000 x g for 3 minutes. The DNA was then eluted by addition of 50 µl of DNase

free water, incubation at room temperature for 2 minutes and eluted into a clean tube by centrifuging the column at 6000 x g for 1 min.

2.2.1.2. Extraction for ϕ29 MDA amplification

Bacterial cells or viral particles were suspended in PBS, and 4 µl was used in the amplification reaction. Extraction was performed using alkaline method; briefly cell suspensions were added to 200 mM potassium hydroxide (Qiagen) and 50mM dithiothreitol (Qiagen) and incubated at 65^{o}C for 10 minutes. The reaction was then neutralised using neutralisation buffer (Qiagen). The sample was then briefly vortex and placed on ice.

2.2.1.3. Amplification reaction for ϕ29 MDA

The ϕ29 MDA was performed using the Repli-g Single Cell Kit (Qiagen). A master mix was prepared in a total volume of 40 µl, with 29 µl reaction buffer, containing endonuclease resistant hexamer primers and 2 µl (40 U) of ϕ29polymerase (Qiagen, REPLI-g Single Cell Kit). The denatured DNA was then added to the master mix and the sample was placed on the thermocycler which ran on the following parameters, 30^{0}C for 16 hours, 65^{0}C for 3 minutes and then a 4^{0}C hold.

2.2.1.4. Quantification of Product

The ϕ29 MDA product was quantified using the Qubit broad range assay (Thermo Fisher Scientific). Briefly a working solution was made by preparing a 1:200 dilution of the Qubit reagent. Two standards were then made by adding 10 µl of standard to 190 µl working solution, and samples were prepared by adding 2 µl of sample to 198 µl working solution. The samples were vortexed and incubated at room temperature for 2 minutes. The tubes were then read using the Qubit 2.0 Fluorometer. The same tube was read three times for each sample (in accordance to the manufacturer's instructions) and an average taken as the DNA concentration.

2.2.1.5. Visualisation of Amplification Product

5 µg of product was visualised using a 0.8% agarose gel stained with SYBR safe, this was ran for up to four hours at 50v to visualise the large product. The product was run alongside a 1kb plus DNA ladder (Life technologies).

2.2.1.6. Assessment of K-12 Genome Coverage

The PCR was performed using GeneAmp PCR System 9700 (Applied Bioscience). The reaction was performed using AmpliTaq Gold (Thermo Fisher Scientific) in a total volume of 25 µl. The reaction consisted of 2.5 µl PCR 10x gold buffer, 2 µl 25mM $MgCl_{2,}$ 0.5 µl DNTP mix (10nM), 0.125

µl DNA polymerase (1.25 Units) and 1.25 µl of 20 µM forward and reverses primer. To this master mix 5 µl of a 1:1000 dilution of ϕ29 MDA product was added. The cycling condition were as follows, 95°C 10 minutes, and 30 cycles 94°C 30 seconds, 59°C 30 seconds, 72°C 30 seconds, hold at 72°C 10 minutes, and a final hold at 4°C. The products were visualised using a 1.5% agarose gel stained with SYBR safe or the Bioanalyser (Agilent) 1000 DNA kit. Details of the primers can be found in **Table 2-4**

Gene	Position	F primer	R primer
yafD	231122	TACCACCTGGGGAACCGTTA	TCACTGACGTTCAGACCACG
nagC	699597	CGGCAGATTAGTGCGAAACG	CGGACAGCGGCAAAATTCAT
fabG	1149893	AGTTATTGGCACTGCGACCA	CGCCGTTCACATGCAAAGTT
abgB	1399834	GGGGGCGCGGATAAGATAAA	TTACCCGCAACGTAGGCAAT
hisG	2088216	ACTTTACCCTGCGTCGTCTG	GAATTGAACTGGCACCCAGC
eutB	2555340	TTGAGGTTCTGGTCAGCGTC	GACGACGTGCAAAGTATCGC
speB	3080899	AAAGCAGGATCCAGGCAGTC	ACTGGGTGATTACTGGCGTG
kefB	3476824	CTTCATGGTGTCTTCCGGCT	CCTCGGGGTGCTTTATACCC
mnmG	3921767	GCAGATCTTCACCACTGGCT	CCGAACGGTATCTCCACCAG
treR	4464322	ACGATCTCCGGATGGAGGAA	ACCCGTCTGGATTCGTTGTC

Table 2-4 Details of PCR targets used in initial assessment of amplification of the *E. coli* K12 genome, including gene name, gene position and primer sequences

2.2.2. Assessment RNA Conversion Enzymes

2.2.2.1. Concentration of HIV Using PEG-High Volume

100 ml of tissue culture supernatant from 8E5/LAV[94] cells was divided into four 25 ml aliquots, and centrifuged at 6000xg to remove host cells. The supernatant was then added to 25 ml of Polyethylene glycol (PEG) 20,000 (Sigma-Aldrich) in 0.9% NaCl to a total of 20% (w/v) PEG, vortexed and incubated overnight at 4°C. The sample was centrifuged at 17860xg for 20 minutes. The pellet was resuspended in 5 ml sterile PBS, vortexed and combined into a single tube and incubated overnight at 4°C. The sample was once again centrifuged at 17860xg for 20 minutes. The sample was resuspended in 1ml sterile PBS. The 1ml was then concentrated as before and resuspended in 20 µl.

2.2.2.2. RNA Extraction

The concentrated virus was then extracted using either alkali extraction or column based extraction. For the alkali extraction the protocol described in **2.2.1.2** was followed but with the volume scaled up to a total reaction volume of 50 µl. Alternatively the concentrated virus was extracted using PureLink® Viral RNA/DNA Kit (Invitrogen), as described in the manual, with the

substitution of the carrier RNA with Linear Acrylamide (LPA) (Invitrogen). Briefly 25 µl Proteinase K, 200 µl lysis buffer and 5 µl LPA was added to the sample, gently vortexed and then incubated at 56°C for 15 minutes. After the addition of 250 µl 100% ethanol the sample was incubated for a further 5 minutes at room temperature. The sample was then added to the viral spin column and centrifuged for 1 minute at 6800xg, and the flow through was discarded. The sample was then washed twice with wash buffer before being eluted in 50 µl RNase free water. The RNA from both extractions was then quantified using Qubit RNA HS Assay Kit or Qubit RNA BR Assay Kit.

After quantification dilutions were prepared to contain 1000, 10, 0.1, 0.01, 0.001 and 0.0001 pg/µl concentrations using RNAse free water containing 2 U/µl RNaseOUT™ Recombinant Ribonuclease Inhibitor (Thermo Fisher Scientific). These dilutions were then used for investigating the sensitivity of reverse transcription enzymes.

2.2.2.3. Addition of RNase and DNase

After viral concentration with PEG, and suspension in 50 µl of sterile PBS 10ng of RNase A (Thermo Fisher Scientific) along with 5 units of DNase I, (Thermo Fisher Scientific) was added. The sample was then briefly vortexed and incubated at 37°C for 20 minutes. After incubation 100 U RiboLock (Thermo Fisher Scientific) was added to the sample and 5µl 0.5M EDTA.

2.2.2.4. RNA Conversion using Superscript III

1 µl of RNA was added to 50 ng of random primers (Thermo Fisher Scientific) and 1 µl 10mM dNTPS mix (Thermo Fisher Scientific) and made up to 13 µl with sterile water. This was then incubated at 65°C for 5 minutes and put on ice for 2 minutes. To this 4 µl 5x first-strand buffer, 1 µl 0.1M DTT, 1 µl RNaseOUT and 1 µl SuperScript III (Thermo Fisher Scientific) was added. The sample was then incubated at 55°C for 60 minutes before inactivation at 70°C 15 minutes

2.2.2.5. RNA Conversion using Superscript IV

1 µl of RNA was added to the following reaction to 50 ng of random primers and 1 µl 10mM dNTP mix and made up to 13 µl with sterile water. This was then incubated at 65°C for 5 minutes before the addition of 4 µl 5 x SSIV buffer, 1 µl 0.1M DTT, 1 µl RNaseOUT and 1 µl SuperScript IV reverse transcriptase (Life technologies). The reaction was then for 10 minutes at 55°C and 10 minutes at 80°C.

2.2.2.6. PCR for Amplification of HIV Transcripts

Using envelope targeting primers[95] one forward primer and five reverse primers in the envelope gene were selected, listed in **Table 2-5.** PCRs were performed for each primer combination in the following mixture 5µl 10X PCR Buffer, 1.5 µl 50 mM MgCl, 1 µl 10mM dNTP mix, 1 µl 20 µM forward and reverse primers, 0.4 µl Taq (5U/ µl), 5 µl first strand cDNA, water to 50 µl total volume. The following cycling conditions were used 94°C 2 minutes, 30 cycles 94°C 30 seconds, 53°C 30 seconds, 72°C 30 seconds, hold at 72°C 10 minutes, and a final hold at 4°C

	Primer	Location	Product length	Sequence(5'–3')
Forward	E70	335		GGGATCAAAGCCTAAAGCCATGTGTAA
Reverse	E03	+218	2283	TAAGTCATTGGTCTTAAAGGTACCTG
	E45	2113	1778	CCTGCCTAACTCTATTCAC
	E65	1568	1233	AGTGCTTCCTGCTGCTCC
	E125	1091	756	CAATTTCTGGGTCCCCTCCTGAGG
	E145	757	422	CAGCAGTTGAGTTGATACTACTGG

Table 2-5 Details of primer name, amplicon size and genome position of primers used to amplify HIV

2.2.2.7. PyroPhage

PyroPhage 3173 (Lucigen) is designed to have a one-step reaction for reverse transcription and PCR amplification. Two forms of the enzyme are available, a wild-type and an exonuclease knock out (exo-); both enzymes were tested using the specific primer combinations described in **Table 2-5**.

Initially the enzyme was used to amplify the cDNA produced by SuperScript IV in **2.2.2.5** using the following reaction using either the wild type or exo- versions of the enzyme. 200 ng cDNA was added to 25 µl PyroPhage 3173 2X PCR Buffer, 1 µl 10mM dNTP mix, 1 µl 10 µM forward and reverse primer, 0.5 µl PyroPhage 3173 and water to a total of 50 µl. The cycling conditions were the same as used in the PCR in **2.2.2.6**

The utility of the enzyme to perform both RNA conversion and DNA amplification in a single step was then investigated. The reaction was set up as follows 1 µg RNA, 25 µl PyroPhage 3173 2X PCR Buffer, 1 µl 10mM dNTP mix, 1 µl 10 µM forward and reverse primer, 0.5 µl PyroPhage 3173 and water to a total of 50 µl.

The following cycling conditions were used 94°C 2 minutes, 30 cycles 94°C 30 seconds, 53°C 30 seconds, 72°C 30 seconds, hold at 72°C 10 minutes, and a final hold at 4°C.

2.2.2.8. Reverse Transcription and MDA

The cDNA library produced using SuperScript III and IV in **2.2.2.4** and **2.2.2.5** were amplified using ϕ29 MDA. Briefly 10 µl of cDNA was added to 40 µl mastermix and amplified for 8 hours. After amplification, the ϕ29 MDA product was quantified and then diluted 1:1000 before addition to the PCR reactions using primers described in **2.2.2.6**.

2.2.3. Whole Genome Amplification Using Nick Extension Amplification

Experiments were undertaken to assess the possibility of producing whole genome amplification by initiating the DNA replication at nicks in DNA. Nt. BstNBI (New England Biolabs) is an endonuclease that cleaves one strand of DNA at a specific site downstream of a 5 base recognition site. The enzyme has previously been used for isothermal amplification from nicking by Moser et [78] using the amplification enzyme PyroPhage 3174 [77].

2.2.3.1. Nicking Reaction with Nt. BstNBI

Initially bioinformatic predictions of the number of nicking events in the three bacteria were performed on the reference genome of, *Escherichia coli K12 substr. MG1655, Actinomyces naeslundii (NCTC 10301)* and *C. difficile.* This was performed by searching the reference sequence for the recognition sites on both strands of DNA. Using the following command

```
Grep -o 'recognition site' <reference_file.fasta> | wc -w
```

Command 2-1 bioinformatic search for nicking enzyme recognition site

Firstly 100 ng of control DNA (as described in **2.2.1.1**) extracted from *E. coli K12, C. difficile* or *A.naeslundii* were incubated with Nt.BstNBI in the following reaction scaled down from a generic restriction enzyme protocol. Input DNA 100 ng, 1 µl 10x buffer (NEB), 1U Nt. BstNBI enzyme (NEB) and water to a total volume of 10 µl.

The reaction was incubated at 55°C for 1 hour. The enzyme was then de activated at 80°C for 20 minutes. The DNA was cleaned up using isopropanol precipitation. Briefly 5 µl 7.5M NH_4OAc (Sigma-Aldrich) and 30 µl 100% isopropanol (Sigma-Aldrich) was added to the samples and incubated at room temperature for 10 minutes. This was then centrifuged at 4°C for 10 minutes at 21000xg. The pellet was then washed with 70% ethanol and centrifuged as before. The ethanol was removed and the sample allowed to air-dry for 15 minutes before being resuspended in 15 µl

of molecular grade water. The sample was then quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and stored at -20°C.

2.2.3.2. MDA Reaction with Nicked DNA

DNA was nicked and cleaned up as described in **2.2.3.1** then four ϕ29 MDA reactions were set up for each isolate,

1. Non-nicked with primer control
2. Non-nicked without primer
3. Nicked DNA with primer
4. Nicked DNA no primers.

The ϕ29 enzyme for all nicking work was supplied by NEB, with specific 10x ϕ 29 DNA polymerase reaction buffer. Samples which included primers were denatured prior to addition to the reaction mix below. 0.5 ng of DNA was added to the reactions. 5 µl of ϕ29 10x buffer, BSA to 0.2 µg/ µl, 200 µM dNTPs, 10U ϕ 29, 250ng Endonuclease resistant random hexamers (where needed) and water to a total volume of 50 µl

The reactions were then incubated at 30°C for 6 hours and DNA was quantified at 1, 2, 3, 4, 5 and 6 hours, and inactivated at 65°C for 10 minutes.

2.2.3.3. Co-nicking and MDA Amplification

The nicking and amplification reactions were combined without a clean-up stage. 0.5 ng of DNA was added to the above nicking reaction (**2.2.3.1**) and incubated for 30 minutes at 55°C, then ramped down 30°C before addition of the MDA reaction mix (**2.2.3.2**). The same four combinations of reactions were used as previously, with the samples including primers denatured before addition to ϕ29 MDA mix.

2.2.3.4. Use of T4 Gene 32 Protein for Amplification Stabilisation

T4 Gene 32 Protein (NEB) is a single stranded binding protein (SSBP), which stabilises single stranded DNA, preventing re-annealing. This was previously reported[96] to increase DNA yields during amplification at a level of 2.5 µg/100 µl. The T4 Gene 32 Protein was supplied at a concentration of 10 mg/ml, in 50 µl reactions 1.5 µg was added (1.5 µl 1:10 dilution).

2.2.4. Combining Displacement Enzymes

Other polymerases with strong stand displacement activity were combined with ϕ29 to investigate the impact on extension from nicks. For assessment of the potential use of enzyme combinations with ϕ29, *E. coli* K12 cultured control DNA was extracted as described in **2.2.1.1** and nicked as described in **2.2.3.1.**

2.2.4.1. Klenow Fragment (3′→5′ exo-)

Klenow fragment has been previously used to produced DNA from nicks [97] and so was investigated in isolation and in combination with ϕ29. The following reaction mix combinations were trialled

1. Klenow Fragment with nicked DNA and random primers
2. Klenow Fragment with nicked DNA
3. Klenow Fragment with nicked DNA and SSBP
4. Klenow Fragment and ϕ29 with nicked DNA and random primers
5. Klenow Fragment and ϕ29 with nicked DNA
6. Klenow Fragment and ϕ29 with nicked DNA and SSBP

Klenow Fragment (NEB) was supplied with 10 x NEB buffer 2 (NEB). 2.5 µl was added to the reaction along with 2.5 µl 10 x ϕ29 DNA Polymerase Reaction Buffer (NEB) to create a combination reaction buffer. To this buffer mix 0.2 µg/ µl of Bovine Serum Albumin (BSA), 200 µM dNTPs and 10 U of Klenow fragment were added along with the required combination of the following components:

- 250ng Endonuclease resistant random hexamers
- 1.5 µg T4 Gene 32 Protein (NEB)
- 10U of ϕ29

10 ng nicked DNA was added and reactions made up to a total volume of 50 µl with water. If samples had primers added DNA was denatured by heating to 95°C for 2 minutes and held on ice for one minute before addition to the reaction mix. All reactions were then incubated at 30°C for six hours.

2.2.4.2. Deep Vent$_R$™ DNA Polymerase

Deep Vent$_R$ has known strand displacement activity, but is normally used for thermo cycling production of DNA, where there are high levels of secondary structure. Initially a thermocycling reaction was set up using the *E. coli* K12 primers described in **2.2.1.6**

A 50 µl reaction was set up, using Deep Vent$_R$ DNA Polymerase (NEB). The mix consisted of 5 µl ThermoPol Reaction Buffer (10X), 1 µl dNTP mix, 1.25 µl 20µM forward and reverse primer, 10 ng denatured DNA, 0.5 µl Deep Vent DNA Polymerase and water to 50 µl. The cycling condition were as follows, 95°C 5 minutes, and 30 cycles 94°C 30 seconds, 59°C 30 seconds, 72°C 30 seconds, hold at 72°C 10 minutes, and a final hold at 4°C.

To assess the potential for isothermal amplification the same reaction mix was used, with 250 ng endonuclease resistant random hexamers used in place of the specific primers. 10 ng denatured DNA was added to the reaction which was held at 30°C for 1 minute, and either held at 30°C for six hours or the temperature was ramped up to 40°C, 50°C, 60°C, 70°C or 80°C for six hours.

To test for ability to extend from nicks, 10 ng of nicked DNA was added to the reaction mix without primers, with or without the addition of 1.5 µg T4 Gene 32 Protein. This was then incubated at 70°C or 80°C for six hours.

2.2.4.3. Bst DNA Polymerase, Large Fragment

50 µl reactions using *Bst* DNA Polymerase, Large Fragment and 10X ThermoPol Reaction Buffer (NEB).). 5 µl ThermoPol Reaction Buffer (10X), 3 µl $MgSO_4$ (100 mM), 2 µl dNTP mix, 10 ng nicked DNA, 10 U *Bst* DNA polymerase and water to 50 µl. Where necessary 1.5 µg of T4 Gene 32 Protein (NEB) was added. The sample was then held for six hours at 30°C, 40°C, 50°C, 60°C, or 70°C.

When combined with ϕ29 2.5 µl ThermoPol Reaction Buffer (10X) and 2.5 µl 10 x ϕ29 DNA Polymerase Reaction Buffer was used in the above reaction mix. For the primer control 250 ng endonuclease resistant random hexamers were added. Before addition of the ϕ29 to the mixture (including Bst) was held at 50°C, 60°C, or 70°C for one hour before cooling to 30°C and the ϕ29 (10U) added.

2.2.5. Amplification of DNA using DNA tagging

Addition of nucleic acid tags to the 3' end of nucleic acids has the potential to stabilise the nucleic acid and produce more full length transcripts by allowing replication to initiate upstream

of the start of the target DNA. By limiting the start of replication to a single sight amplification bias and over amplification of regions will be limited. Addition of tags also gives the potential to add barcodes, allowing initial abundance in samples to be calculated.

Tag Design

Tags were designed using the basic hairpin structure used by Ding et al [98], with additional bases added to increase the melting temperature, and to include the recognition site for the restriction enzyme BamHI (NEB). Over hangs were added to act as priming sites, of either five, ten or fifteen bases, the secondary structure was predicted using http://eu.idtdna.com/calc/analyzer. Sequences of the tag designs are detailed bellow

Basic hairpin

5' GGATCCGCGCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGCGCGGATCC 3'

5 base 5' overhang

5' TCGTAGGATCCGCGCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGCGCGGATCC 3'

10 base 5' overhang

5'AAGTATCGTAGGATCCGCGCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGCGCGGATCC 3'

15 base 5' overhang

5'AGTTCAAGTATCGTAGGATCCGCGCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGCGCGGATCC 3'

2.2.5.1. Tagging with RNA Segments

RNA bases were added along the length of the stem and loop to allow better degradation of the stem loop using RNAses. The positions of the RNA bases are highlighted below.

5' GGATCCGCGCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGCGCGGATCC 3'

2.2.5.2. Formation of Stem-loop

The stem loop designs were ordered from Eurogentec, in dried formation after being purified by Polyacrylamide Gel Electrophoresis. The DNA tags were reconstituted in sterile water to a concentration of 200 µM. The DNA was phosphorylated using T4 Polynucleotide Kinase (3' phosphatase minus) (NEB) using the following reaction mix, 10 ng tag DNA, 1 µl 10x buffer, 1 µl 10

mM ATP, and water to 10 µl. This was incubated at 37°C for 30 minutes and heat deactivated at 65°C for 20 minutes. The DNA tag was then incubated at 95°C for one minute before being placed on ice for 2 minutes and then placed at room temperature for 20 minutes to aid hairpin formation. The stem loops with RNA segments were placed on ice for two minutes after enzyme inactivation. The stem loops were then ran on a 2% agarose gel to assess size of products produced and check for double stranded DNA formation.

2.2.5.3. Attachment of Stem-loop using t4 RNA Ligase

T4 RNA Ligase 1 (NEB) catalyses the ligation of a 5′ phosphoryl-terminated nucleic acid donor to a 3′ hydroxyl-terminated nucleic acceptor, through the formation of a 3′ → 5′ phosphodiester bond. To prevent self-ligation, the DNA was dephosphorylated using Alkaline Phosphatase, Calf Intestinal (CIP) (NEB). Variable DNA and RNA input amounts were dephosphorylated by addition of 1U CIP for every 1µg of nucleic acid and incubated at 37°C for 30 minutes.

Stem-loops were attached in 20 µl reactions by addition of 10 U TD RNA ligase 1 along with 1mM ATP, and incubated at 37°C overnight.

2.2.5.4. DNA Amplification with Tagged DNA

10 ng of tagged DNA was incubated in a 50 µl reaction with 10U ϕ29, 200 µM dNTPs and 5 µl 10x ϕ29 buffer. The sample was then incubated for four or six hours at 30°C.

2.2.5.5. RNA Conversion with Tagged RNA

10 ng of RNA with tag attached was added 1µl 10mM dNTP mix, 4 µl 5 x SSIV buffer, 1 µl 0.1M DTT, 1 µl RNaseOUT and 1 µl SuperScript IV reverse transcriptase (Life technologies) and water to a total volume of 20 µl. The reaction was then incubated for 15 minutes at 55°C and 10minutes at 80°C.

2.3. Development of ϕ29 MDA for Rapid and Sensitive Whole Genome Amplification of Bacterial Isolates

The second study of the thesis evaluated the suitability of the enzyme ϕ29 for detection and characterisation of low abundant microbial genomes production, for the application of whole genome sequencing. The first question addressed was whether ϕ29 could produce DNA that was suitable for library preparation for next generation sequencing, and how this compared to the DNA produced by current culture based methods. Additionally, the sensitivity of the method was evaluated and the processivity investigated. The quality of the sequencing data produced was assessed and compared to that produced by culture based methods. To establish these properties *Escherichia coli K12 substr. MG1655* was used as the test organism. The impact of GC content and cell wall type was also investigated, along with the ability to detect and analyse extra chromosomal elements using *Actinomyces naeslundii (NCTC 10301)* and *Peptoclostridium difficile* 630. To further test the enzyme a mixed bacterial sample was amplified to look for amplification bias, using *Haemophilus influenza (*NCTC 8143) and *Enterococcus faecalis (*NCTC 12201). Additionally, DNA viral genomes from Adenovirus 40 and 41 were amplified to investigate the application to smaller genomes and the extraction of none enveloped viruses. Finally, negative samples were sequenced to monitor contamination, randomly produced DNA and sequencing artefacts.

2.3.1. Development of ϕ29 MDA for whole genome Sequencing

2.3.1.1. Preparation of None Amplification Controls

Overnight cultures of bacteria were extracted using Qiagen QIAamp DNA mini kit as described in **2.2.1.1**. The DNA was quantified using Qubit BR kit and 500 ng of DNA taken forward to sequencing.

2.3.2. Fragmenting DNA for sequencing

The recommended method for fragmenting DNA for sequencing on the 454 Junior was physical shearing for 1 minute using nebulisation. However due to the long lengths and presence of secondary structure in DNA produced by ϕ29 MDA alternative methods were investigated. DNA extracted from cultured bacteria was fragmented using the recommended parameters 30psi for 60 seconds. Three biological replicates of the culture based control were sequenced

2.3.2.1. Using Nebulisation

500ng of DNA was made up to a final volume of 100 µl with TE buffer. The mixture was added to an assembled nebulisation cup (Roche) to which 500 µl of nebulisation buffer was added (Roche). The sample was then nebulised at 30 psi (pounds per square inch) for 60, 120 or 180 seconds.

2.3.2.2. De-branching of Product

3 µg of ϕ29 MDA DNA was de-branched using S1 nuclease in a 90 µl reaction as follows, 3 µl 10x buffer, 3 µl 0.5M NaCl, 10 µl S1 nuclease (1U/ µl) with water to make the volume to 90µl. The digestion reaction was left at room temperature for 30 minutes and the enzyme deactivated by incubating at 70°C with 6 µl 0.5M EDTA. This was then nebulised at 30psi for 60, 120 or 180 seconds.

2.3.2.3. Using Fragmentase

NEBNext® dsDNA Fragmentase (NEB) generates dsDNA breaks in a time-dependent manner to yield 50–1,000 bp DNA fragments. 1 µg of ϕ29 MDA was added to the following reaction mix, 2µl 10x Fragmentase buffer, 2 µl 10x BSA, and made up to a total volume of 18 µl using the supplied sterile water. The reaction was then incubated on ice for 5 minutes, before the addition of 2 µl dsDNA Fragmentase. After which the reaction was vortexed and incubated at 37°C 10, 15 or 20 minutes. After incubation 5 µl 0.5M EDTA was added to stop the reaction.

2.3.2.4. Combination of S1 Nuclease and Fragmentase

5 µg of DNA was added to the following reaction mix, 10 µl Fragmentase 10x buffer, 10 µl 10x BSA, 24 µl water. This was then incubated on ice for 5 minutes before the addition of the following, 10 µl dsDNA Fragmentase, 10 µl S1 10x buffer, 10 µl 0.5M NaCl, 30 µl S1 nuclease (1U/ml). Followed by 30-minute incubation at 37°C, after which 2.5 µl 5M EDTA was added to stop the reaction.

2.3.2.5. Sample Purification

The sample was purified using the Qiagen MinElute PCR purification kit, as described in the Roche Rapid Library Preparation method. Briefly, 2.5ml of PB buffer was added to the sample and mixed. 750 µl of this mixture was added to the spin column and centrifuged for 15 seconds and the through flow discarded. This was repeated until the entire sample had been added. The sample was then washed with 750 µl PE buffer and eluted in 20 µl of TE buffer. After which the DNA was quantified using Qubit broad range kit as previously described in **2.2.1.4,** 500ng of DNA in a total volume of 16µl in TE buffer was added then taken forward for sequencing.

2.3.3. Sample sequencing

For sequencing the following Roche 454 Junior protocols were followed:

2.3.3.1. Rapid Library Preparation Method Manual-Match 2012

Briefly, 500ng of fragmented DNA was end repaired and adapters added, the DNA was then size selected using AMPure beads. Library quality was assessed using the Agilent Bioanalyser where successful library preparation consisted of the average read length being between 600 and 900bp. The library was quantified using the QuantiFluor ST Fluorometer (Promega) fluorescence was measured against the standards provided in the library preparation kit. Samples were then diluted to working stocks of $1x10^7$ molecules/µl, in TE Buffer.

2.3.3.2. EmPCR Amplification Method Manual -Lib-L-March 212

Firstly, an emulsion PCR was set up, aiming at four or two copies per bead. The emulsion was then dived into 100 µl aliquots in a PCR plate. The emulsion PCR used the following parameters in a thermocycler with heated lid turned on. A hold at 94°C for 4 minutes, 50 cycles of 30 seconds at 94°C, 4.5 minutes at 58°C, 30 seconds at 68°C, with a final hold at 10°C. The beads were then harvested and the emulsion broken using isopropanol and ethanol. The beads were then washed using a provided buffer and concentrated into 1 ml. Enrichment primers were then annealed to the libraries and the beads enriched to keep only beads with attached DNA. The recovered beads were then visually quantified before a sequencing primer was attached to the libraries.

2.3.3.3. Sequencing Method Manual – January 2013

Firstly, all frozen sequencing reagents were defrosted in the dark by submersion in water. Then the instruments were primed using the provided pre-wash buffer. Four bead layers were prepared as instructed and added to the Pico Titer Plate (PTP) in the order shown in **Table 2-6.** The PTP was then loaded onto the sequencer and sequenced usng 200 cycles and full processing for shotgun libraries. Run time was 9 hours and 20 minutes.

Bead layer	Bead type
Layer 1	Enzyme beads pre-layer
Layer 2	DNA and packing beads
Layer 3	Enzyme beads post-layer
Layer 4	PPiase beads

Table 2-6 details of bead layers in the PTP for DNA sequencing on the Junior 454

2.3.4. Determining Sensitivity and Processivity of Technique in the Context of Bacterial Genomes

2.3.4.1. Lowering Reaction Volumes

Reaction volumes were scaled down to total volumes of 25 µl and 12.5 µl with two sets of biological replicates being sequenced for each reaction volume.

2.3.4.2. Preparation of Single Cells for Sequencing

Single colonies from overnight cultures were suspended in 1ml phosphate-buffered saline (PBS) to produce a 10^{-3} dilution, serial 1:10 dilutions were then prepared until 10^{-10} was achieved. A blood agar plate was divided into four and 10 µl of the 10^{-7} – 10^{-10}; dilutions were plated four times on each plate. These were then incubated overnight and a colony count performed. This was then used to calculate the average number of cells in the PBS and the required volume of diluted cells was added to the extraction for ϕ29 MDA amplification. Three biological replicates of estimated single cells were sequenced.

2.3.4.3. Reducing Incubation Time

Incubation times of eight, four, two and one hours were investigated using single cell extracts in 50 µl reactions. Bacteria were extracted as described in **2.2.1.2** , and amplified in the reaction as described in **2.2.1.3** before incubation for the required time at 30°C and inactivation at 65°C for 10 minutes. Two biological replicates were performed per incubation time.

2.3.5. Basic Analysis of Sequencing Data Produced

After 454 junior sequencing run completion run statistics were viewed using the Roche 'runviewer' programme, where information on raw reads, reads filter passed and read length were obtained. According to Roche sequencing guidelines, a successful run follows the following characteristics,

- Raw reads above 200,000 reads
- Passed filter wells >100,000 (>50%)
- Average read length 400-500 base pairs

After sequencing if these minimum parameters were achieved the sequencing was considered successful and further analysis performed.

2.3.5.1. *Reference Assembly*

Initially the SFF files (Standard flowgram format) were mapped against the appropriate reference as listed in **Table 2-7.** Reference sequences were downloaded in fasta format from the National Centre for Biotechnology Information (NCBI) or Integrated Microbial Genome (IMG) database. The sequencing files were assembled using the standard parameters in Newbler GS Reference Mapper (Roche Diagnostics). This was used for initial run assessment to determine genome coverage and number of mapped reads. The software allows both SFF and FASTQ files to be input and assembled, using **Command 2-2**. From the output file 'NewblerMetrics.txt', reference assembly information was extracted, including the number of reads which mapped to the reference, the number and size of contigs and the percentage of the reference which was covered. Once the reference assembly was completed, the resulting bam file was sorted and indexed using Sam tools [99]. The assembly was then viewed against the reference in Artemis 15.0.0[100].

```
~/runMapping -o <output_folder_name> <referencefile.fasta> <inputfile.fasta/SFF>
```

Command 2-2 reference assembly of SFF using Newbler

Reference name	Identification
***Enterococcus faecalis* V583 chromosome, complete genome**	NCBI Reference Sequence: NC_004668.1
***Haemophilus influenzae* 10810, complete genome**	NCBI Reference Sequence: NC_016809.1
***Escherichia coli* str. K-12 substr. MG1655 strain K-12 cont1.1, whole genome shotgun sequence**	NCBI Reference Sequence: NZ_AYEK01000001.1
***Actinomyces naeslundii* MG1**	IMG reference 2502171150
***Peptoclostridium difficile* 630**	GenBank: AM180355.1
Human adenovirus 41 isolate Tak, complete genome	GenBank: DQ315364.2
Human adenovirus F, complete genome (human adeno virus serotype 40)	NCBI Reference Sequence: NC_001454.1
***Escherichia coli* K-12 plasmid F DNA, complete sequence**	GenBank: AP001918.1
***Peptoclostridium difficile* 630 plasmid pCD630**	NCBI Reference Sequence: NC_008226.1

Table 2-7 details of reference files used in initial assessment of the of ϕ29 MDA to amplify whole genomes including extra chromosomal elements.

2.3.5.2. *De Novo* Assembly

Initial *de novo* assemblies were performed using Newbler GS *De novo* Assembler (Roche Diagnostics) using default parameters, inputting either SFF files or FASTQ files as detailed below in **Command 2-3.** The *de novo* assembly output was then assessed using QUAST: Quality Assessment Tool for Genome Assemblies [101], **Command 2-4**, along with the relevant reference sequence from **Table 2-7**. From the output file quast_results.tsv, information on contig size, reference coverage and misassemblies were extracted.

```
~/runAssembler -o <output_folder_name> <inputfile.fastq/SFF>
```

Command 2-3 *de novo* assembly of SFF file using Newbler

```
~/Quast.py -R <reference.fa> -o <output_folder> <inputfile.fna>
```

Command 2-4 assessment of *de novo* assembly quality using QUAST

2.3.5.3. Statistical Analysis of Run Data

Statistical tests were performed using the inbuilt functions in Microsoft Excel (2010). Two-tailed T-tests were used to determine if two sets of data were significantly different from each other. Firstly, an F test was performed to assess whether the distribution of data was equal. After this a two tailed t test was performed in excel using the output of the F test, either equal or unequal distribution. A cut off of 0.05 was used to interpret the information.

2.3.5.4. Extra Chromosomal Elements

For known and expected extra chromosomal elements, reference mapping was used as described in **2.3.5.1** using the appropriate reference file from **Table 2-7**.

2.3.5.5. Extraction of Unmapped Reads

A custom python script was written and named 'unmapped.py' to identify the sequencing reads that were unmapped during reference assembly (**2.3.5.1**), the read identifications were then written to a text file **Command 2-5**. The list of read headings was then used to extract those reads from the original SFF using sfffile tool, (GS junior suite) as detailed in **Command 2-6**

```
for line in open("454ReadStatus.txt"):
    columns = line.split()
    if len(columns) > 1:
      if columns[1].startswith('Unmapped'):
        print columns[0]
```

Command 2-5 Custom python script for identifying reads which were unmapped after a reference assembly using Newbler

```
~/sfffile -i  <inputfile.txt>  -o <new_file.sff>  <original_file.sff>
```

Command 2-6 use of sfffile to create a new SFF file with unmapped reads removed, using the text output file from Command 2-5

2.3.5.6. Conversion of SFF file

SFF files were converted to fastq format using sff2fastq (The Genome Institute at Washington University, St. Louis, MO), **Command 2-7,** which allows the extraction of read information from SFF files and output of quality scores and sequences into a FASTQ format.

```
~/sff2fastq -n <inputfile.sff> > <output.fastq>
```

Command 2-7 conversion of SFF file to a fastq using sff2fastq

2.3.5.7. Identification of Read Taxonomy

To identify the taxonomy of the reads a Blastn (Basic Local Alignment Search Tool, nucleotide) search was used against the non-redundant nucleotide NCBI database. Firstly fastq files were converted to fasta format using seqtk version 1.0-r82, and then a local Blastn search was performed **Command 2-8.**

```
~/seqtk fq2fa <inputfile.fastq> > <output.fasta>

~/Blastn -query <inputfile.fasta> -db <nr_database> -out <output.nt.out>
```

Command 2-8 conversion of fastq to fasta using seqtk, use of Blastn to identify reads

2.3.5.8. Lowest Common Ancestor Analysis

MEGAN5 - MEtaGenome ANalyzer[102], was used to visualise the results of the Blastn output, using the following Lowest common ancestor (LCA) parameters.

- The Min_Score of 150 (minimum threshold for the bit score of hits (indication of alignment quality)).
- The Max_Expected of 0.01 (maximum threshold for the expected value of hits, E-value)
- The Min_Support_Percentage of 0.01 (minimum percentage of reads a hit needs)

The corresponding fasta was also loaded into MEGAN, allowing reads to be extracted and written to a new fasta file. After the reads had been extracted in to fasta file, the read IDs were extracted into a text file, and this list was used to create a fastq file with extracted read and quality data **Command 2-9**.

```
~/perl -ne 'if(/^>(\S+)/){print "$1\n"}'  <megan.fasta> > <megan_list.txt>

~/seqtk subseq  <inputfile.fastq>  <megan_list.txt> > <megan.fastq>
```

Command 2-9 extraction of read headings from the fasta file output by MEGAN using Perl script and the creation of a fastq of the corresponding reads from the original fastq

2.3.6. Impact of GC content on sequencing production

48 hour cultures of *Peptoclostridium difficile* 630 and *Actinomyces naeslundii (NCTC 10301)* were extracted as described in **2.2.1.1** and 500 ng of DNA was sequenced using the 454 Junior and manufacturers recommended method **(2.3.3),** with two replicates for each bacteria.

Single cells for *C. difficile* and *A. naeslundii* were prepared as described in **2.3.4.2** with the exception of cultures being incubated for 48 hours. For ϕ29 MDA reactions single cells were extracted using the alkali method described in **2.2.1.2** and amplified in 50 µl for two hours. These were then prepared for sequencing by use of S1 nuclease and 2 minutes nebulisation. Before

sequencing on the 454 Junior using manufacturers recommended method **(2.3.3)**. Three biological replicates of each bacterium were performed

2.3.7. Application to low level mixed bacteria

Serial dilutions of *Haemophilus influenza (*NCTC 8143) and *Enterococcus faecalis (*NCTC 12201) were prepared as described in **2.3.4.2**. Mixes of the two bacteria were then set up varying the cell input. A single cell of *E. faecalis* was mixed with a single, ten, 100 or 1000 *H. influenzae* cells. These mixes were then extracted, amplified using ϕ29 MDA for two hours and sequenced.

2.3.8. Amplification of DNA virus

Tissue culture grown Adenovirus 40 and 41 were defrosted and centrifuged at 4000x g for 10 minutes to pellet any eukaryotic cells. A one in ten dilution of the viral supernatant was prepared and 3 µl of this was extracted and amplified using ϕ29 MDA for two hours. Three biological replicates were performed for each serotype.

A further three biological replicates of Adeno41 were performed with an additional DNase stage to remove tissue culture DNA. After centrifugation to pellet host cells, 2 µl of supernatant was added to 2 µl 10X DNase I Reaction Buffer (NEB), 1 µl DNase I (NEB) and 15 µl water. This was then incubated for 10 minutes at 37^{o}C before the addition of 1 µl 0.5M EDTA and heat inactivated at 75^{o}C.

2.3.9. Development of contamination library

A negative library was sequenced to allow removal of reads from kit contamination. ϕ29 MDA reaction was set up as above using the PBS used to suspend the cells. Negative libraries were sequenced periodically to monitor potential contamination.

2.4. Development of a data analysis pipeline for processing data produced using ϕ29 MDA

2.4.1. Quality control

The methods adopted in the project aimed to be a balance of the laboratory and bioinformatic techniques to produce the highest quality output possible.

2.4.1.1. Removal of Human Contamination

Large amounts of human signal in the sequencing run increases the sequencing cost by wasting data with none pathogen reads. Human reads may interfere with the assembly of pathogens and may even lead to erroneous gene identification. However harsh laboratory methods for removal of host contamination may cause degradation of the pathogen signals. Removing the reads at an early stage of data processing will lower the size of the file, increasing the speed of downstream application.

Using Newbler GS Reference Mapper (Roche Diagnostics), SFF files were reference mapped to the human genome reference, a concatenated file of all chromosomes listed in **Table 2-8.** The reads not mapped to the human genome were then extracted and written to a new SFF as described in **2.3.5.5**.

Type	Name	RefSeq		Type	Name	RefSeq
Chr	1	NC_000001.11		Chr	13	NC_000013.11
Chr	2	NC_000002.12		Chr	14	NC_000014.9
Chr	3	NC_000003.12		Chr	15	NC_000015.10
Chr	4	NC_000004.12		Chr	16	NC_000016.10
Chr	5	NC_000005.10		Chr	17	NC_000017.11
Chr	6	NC_000006.12		Chr	18	NC_000018.10
Chr	7	NC_000007.14		Chr	19	NC_000019.10
Chr	8	NC_000008.11		Chr	20	NC_000020.11
Chr	9	NC_000009.12		Chr	21	NC_000021.9
Chr	10	NC_000010.11		Chr	22	NC_000022.11
Chr	11	NC_000011.10		Chr	X	NC_000023.11
Chr	12	NC_000012.12		Chr	Y	NC_000024.10

Table 2-8 list of accessions for human genome chromosome reference files from NCBI, which were concatenated to create the reference file to remove human signals

2.4.1.2. Removal of Kit Contamination

The problem of environmental and kit contamination was recently highlighted in the literature [103], when looking at high sensitivity assays, especially in the context of cross laboratory studies. Multiple negative libraries produced over time allowed temporal monitoring of the environment, kit lot numbers and changing laboratory practices. A negative library database was developed and used to remove known nonsense reads from the sequencing library.

Reads in the negative libraries were first checked for identifiable reads using Blastn as described in **2.3.5.7** and the results visualised using MEGAN**.** Any reads that didn't match to known bacteria were concatenated into a single fasta file, to create the negative library.

Sequenced reads were then mapped against this database, and any sequence matches were removed using an adapted version of the python script for identification of unmapped reads (**Command 2-5**) shown in **Command 2-10** and a new SFF file was created with reads that didn't have hits in the negative library.

```
for line in open("454ReadStatus.txt"):
    columns = line.split()
    if len(columns) > 1:
      if columns[1].startswith('Full'):
        print columns[0]
```

Command 2-10 Python script for identifying reads which fully mapped to the reference after reference assembly using Newbler

2.4.2. Error Trimming

As suggested by Fabbro et al [104]the best trimming algorithm and software is highly dependent on the data set, downstream analysis and user defined trade-offs. The associated Phred, or Q score with each read dictates the confidence in the base calling at that point. Poor quality bases add potentially unreliable sequences into the data, possibly causing false interpretations during downstream analysis. Erroneous bases can cause particular problems for kmer based assemblers by producing false kmers, increasing the complexity of the assembly and leading to false assembly. By removing erroneous bases, the overall amount of data is decreased but the

reliability is increased, speeding up analysis and allowing more confidence in the results. Different methods for read error trimming were investigated to find the most suitable for the data produced.

2.4.2.1. Prinseq

PRINSEQ (PReprocessing and INformation of SEQuences) [105] is a window based trimming tool with a variety of trimming options, four of these were investigated. The first was the removal of known index sequences at the start of the reads, for this the -trim_left function was used with the value 4. This command was used in all further algorithms. The input and output sequence file was a fastq file (out_put 3); a log file for each parameter was also created.

The programme has two trimming parameters which were tested with different quality cut offs. The first parameter tested was filtering reads based on an average Q score using the -min_qual_mean <integer> function. The quality scores tested were Q35, Q30, Q25 and Q20. The second parameter tested was an end trimming command, removing poor quality bases from the 5' and 3' end of the read. Using the commands, -trim_qual_left <integer> and -trim_qual_right <integer> functions with values Q35, Q30, Q25 and Q20, using the same value to trim both ends. The final parameter to be investigated was the removal of reads that had a length shorter than 99, using -min_len <integer> function.

Finally, combinations of filtering and trimming functions used above were investigated. The -min_qual_mean <integer> was fixed at the Q20 parameter and the end trimming (-trim_qual_left <integer> and -trim_qual_right <integer> functions) were varied using the values Q30, Q28, Q25 and Q20, with a fixed minimum read length of 99 base pairs **Command 2-11**

```
perl prinseq-lite.pl -fastq <input.fastq> -out_format 3 min_length 99 -trim_left 4
-trim_qual_left <20-35> -trim_qual_right <20-35>   -min_qual_mean <20-35>  -log
 <log_name> -out_good <output_file>
```

Command 2-11 Prinseq command for error trimming of a fastq file

2.4.2.2. Cutadapt

Cutadapt[106] is a running sum error trimmer. Firstly, the known tag sequence was removed using –u command, which removes a fixed number of bases from the beginning of reads. Low quality ends were removed to Q35, Q30, Q25 or Q20, using the –q command. Minimum reads

length was set at 99 using the –m function. The input and output format for this trimmer was fastq **Command 2-12**

```
cutadapt -u 5 -q <20-35> -m 99 -o <output_file.fastq> <input_file.fastq>
```

Command 2-12 Cutadapt command for error trimming of a fastq file

The produced FASTQ files were then assembled using both Newbler GS Reference Mapper and Newbler GS *De novo* Assembler as described above. *De novo* assemblies were assessed using QUAST (Quality Assessment Tool for Genome Assemblies)[101].

2.4.3. Abundance Filtering

2.4.3.1. Depth Analysis

Samtools depth in combination with awk was used to assess the depth of the reference assembly **Command 2-13.**

```
# No coverage
samtools depth <inputfile.bam> | awk '$3==0 {print}'
#less than 5
samtools depth <inputfile.bam> | awk '$3<5 {print}'
#total number of times tehre was 0 coverage
samtools depth <inputfile.bam> | awk 'BEGIN {total=0} {if ($3==0) total =total+1} END {print total}'
#average depth of the assembly
samtools depth <inputfile.bam> | awk '{sum+=$3{ END {print sum/NR}'
#the peak depth of assembly
samtools depth <inputfile.bam> | awk 'BEGIN {max=0} {if ($3>max) max=$3} END {print max} '
```

Command 2-13 depth analysis commands using Samtools and awk for identification of the number of points in the reference assembly with no coverage, coverage depth less than five, average depth of coverage and peak depth of coverage

2.4.3.2. Khmer

Data was digitally normalised to reduce the dynamic range of the coverage depth, using Khmer[107]. The aim of digital normalisation was to decrease the data size by discarding redundant reads, whist maintaining the information within the file. This software also has the additional benefit of removing those reads that appear at very low numbers, which are most likely due to random sequencing errors.

The programme uses a reference free algorithm to estimate the genome coverage of the data by looking at the abundance distribution of kmers on the assumption that kmers tend to have similar abundances within a read. (The more times a piece of DNA is copied the higher the kme r abundance will be in that region). In the absence of errors, the average kmer number can be used to estimate the depth of coverage. Additionally, any kmers that overlap errors will have low abundance, which is the basis for the additional error trimming.

Reads were normalized using The khmer software package [107] known as "digital normalization". Firstly, the fastq files were normalised by median, to a depth of 60, using the *normalize-by-median.py* function of the programme, K values were variable (20, 50, 80, 110 and 150), C represents the depth cut off, x informs the amount of memory to use , saving the hash table speeds up downstream work. After digital normalisation to 60, abundance trimming was performed using abundance-*dist.py* with the depth cut off at 2 which will trim at k-mers below this abundance. The V command was used to indicate variable coverage, meaning the programme will only trim low-abundance k-mers from sequences that have high coverage. The digital normalisation was then repeated with the depth cut off of 50. Details of the commands used are below in **Command 2-14.**

```
#digital normalisation

source ~/khmerEnv/bin/activate

#normalized data by median cutoff 60
python ~/khmerEnv/bin/normalize-by-median.py -k <int> -x 2e8 -C 60 <input_file>.fastq --
-savehash <input_file>-dn.kh

#abundance filter
python ~/khmerEnv/bin/filter-abund.py -C 2 -V <input_file>_pure-dn.kh <input_file>.fastq
.keep


#normalize data by median cutoff 50

python ~/khmerEnv/bin/normalize-by-median.py -k <int> -x 2e8 - C 50 <input_file>.fastq
.keep.abundfilt

#output
cp <input_file>.fastq.keep.abundfilt.keep  <input_file>_ab.fastq
```

Command 2-14 Khmer commands for a three pass digital normalisation with low abundance error trimming aiming at a final coverage depth output of 50

2.4.4. De novo Assembly

Several different algorithms were trialled to find the best *de novo* algorithm for this set of data to find the most suitable one. Four data sets were selected to compare the *de novo* assemblers, single cell amplifications from *E. coli* k12, *C. difficile* 630 and *A. naeslundii,* along with adenovirus post DNAse treatment, were assembled using each software. Firstly the SFF files were converted to fastqs before abundance and error trimmed as described **2.4.2** and **2.4.3.2.**

2.4.5. Assembly with SSAKE

The version of SSAKE[110] used was V3-8-3, SSAKE requires the input to be in fasta format so the resulting trimmed fastq files were converted as described previously**.** The files were then assembled using the shell script in **Command 2-15**. With –f indicating the input file, -w being the lowest depth of overlap allowed and –b indicating the name of the output files.

```
#!/bin/sh

../../programmes/ssake_v3-8-3/SSAKE -f ../actino_k_trimmed.fasta -w 2 -b actino
../../programmes/ssake_v3-8-3/SSAKE -f ../adeno_k_trimmed.fasta -w 2 -b adeno
../../programmes/ssake_v3-8-3/SSAKE -f ../c_diff_k_trimmed.fasta -w 2 -b c_diff
../../programmes/ssake_v3-8-3/SSAKE -f ../e_coli_k_trimmed.fasta -w 2 -b e_coli
```

Command 2-15 Shell script for assembling reads using the *de novo* assembler SSAKE of four selected genomes

2.4.6. Assembly using Abyss

The version of Abyss[111] used to assemble the data was 1.9.0.Firstly the kmer length used in the assembly used needed to be optimised, the maximum possible kmer length was 96, and the length suggested as a start point in the manual was 20 and so the script in **Command 2-16** (A) was used to assess the best possible kmer size for this data set, using the *E. coli* k12 dataset. The files were then assembled using the best K value found using the shell script in **Command 2-16** (B). With k representing the kmer length and -o being the output name. abyss-fac produces basic information on the assembly quiality. A shell script was then used to run through the assembly files usign a kmer value of 96.

A

```
#!/bin/sh
export k
  for k in {20,40,60,80,96}; do
    mkdir k$k
    cd k$k
    ABYSS  -k $k -o ecoli-contigs.fa ../e_coli_k_trimmed.fastq
    cd ..
  done
  abyss-fac k*/ecoli-contigs.fa
```

B

```
#!/bin/sh

ABYSS  -k 96 -o actino_abyss.fa ..actino_k_trimmed.fastq
ABYSS  -k 96 -o adeno_abyss.fa ../adeno_k_trimmed.fastq
ABYSS  -k 96 -o c_diff_abyss.fa ../c_diff_k_trimmed.fastq
ABYSS  -k 96 -o e_coli_abyss.fa ../e_coli_k_trimmed.fastq
```

Command 2-16 ABYss *de novo* assembly, A) k-mer optimisation script B) shell script for assembly of four selected genomes

2.4.7. Assembly with Spades

The version of SPades[112] used was 3.5.0, the shell script in **Command 2-17** was used to asseble inout reads. With –o specifying the output directory to be created, -s indicates the reads are unpaired.

```
#!/bin/sh

../SPAdes-3.5.0-Linux/bin/spades.py -o actino -s ../actino_k_trimmed.fastq
../SPAdes-3.5.0-Linux/bin/spades.py -o adeno  -s ../adeno_k_trimmed.fastq
../SPAdes-3.5.0-Linux/bin/spades.py -o c_diff -s ../c_diff_k_trimmed.fastq
../SPAdes-3.5.0-Linux/bin/spades.py -o e_coli -s ../e_coli_k_trimmed.fastq
```

Command 2-17 shell script for *de novo* assembly of four selected genomes with SPades

2.4.8. Asembly using Ray

The version of Ray[113] used was 2.3.1, firstly the optimal kmer size was investigated using the script shown in **Command 2-18 (A).** The maximum kmer size is 149, a range of kmer values from 69 to 149 were trialled against the *E. coli* k12 reads to assess the best Kmer value for this assembler. The reads were then assembled using the optimsed kmer size using the shell script in in **Command 2-18 (B).** Where -n is the number of machines available for Ray to use, -o names the output directory to be made and –s is the sequecne input file. A kmer size of 149 was selected

for the files to be assembled with.

A

```
#!/bin/sh
export k
  for k in (69,89,109,129,149); do
    mkdir k$k
    cd k$k
    mpiexec -n 7 Ray -k$k -o e_coli_Ray -s ../e_coli_k_trimmed.fastq
    cd ..
  done
```

B

```
#!/bin/sh

mpiexec -n 7 Ray -k149 -o actino_Ray -s ../actino_k_trimmed.fastq
mpiexec -n 7 Ray -k149 -o adeno_Ray -s ../adeno_k_trimmed.fastq
mpiexec -n 7 Ray -k149 -o c_diff_Ray -s ../c_diff_k_trimmed.fastq
mpiexec -n 7 Ray -k149 -o e_coli_Ray -s ../e_coli_k_trimmed.fastq
```

Command 2-18 *de novo* assembly with Ray (A) kmer optimisation script (B) assembly shell script

2.4.9. Assembly with Mira

The version of Mira used was 4.0.2, before assembly with Mira a mainfest file was prepared for each set of reads, shown in **Figure 2-1.** Within this file the following parameters were specified, project is the name of the files that will be produced. Job describes what the input data is, and what processes are required. In this case the type of data is genome data, the asssembly is *de novo* and an accurate assembly is required. The manifest also specifies that the reads were produced on the 454 platform. And data gives the location of the file that neds to be assembled. The assembly is then performed by identify the manifest file that contains the information.

```
project = ecoli_mira
job = genome,denovo,accurate
parameters = 454_SETTINGS

readgroup = ecoli_MDA
data = ../e_coli_k_trimmed.fastq
technology = 454
```

Figure 2-1 manifest file for Mira

2.4.10. Genome Annotation

Prokka[115] was used to annotate the *de novo* assembly of both the none amplification control and the ϕ29 MDA single cell of *E. coli*, *A. naeslundii* and *C. difficile*. It was also used to annotate the Adenovirus ϕ29 MDA sample which had had a DNAse treatment prior to sequencing. Alongside the sequencing files, the reference for each genome was annotated.

Sequence files were prepared by abundance filtering as described in **section 2.4.3** using a kmer size of 140, normalized to a depth of 50. The resulting sequences were then error trimmed using Prinseq as described in **section 2.4.2.1** using end trimming cut off of Q20 and average quality of Q20 and the minimum read length of 99. The files were then assembled using Spades as described in **section 2.4.7**. The version of Prokka used was 1.9. The full commands used are detailed in **Command 2-19.** The name of the created output directory is specified using --outdir, and the prefix of the files within this folder is defined using --prefix. The command to add gene annotations is –addgenes. By default, Prokka assumes the input file is a bacterial assembly, in the case of the viral annotations – kingdom command is used to specify viral input. If known the input genus and species can be specified, which is only used in headers in the report. The final command is to specify the location and name of the input fasta file, in this case the output from the Spades assembly.

```
#bacterial annotation
$prokka --outdir <dir_name> --prefix <file_prefix>  --addgenes --genus <bacteria_genus>
--species <species_name>   --gram <+/-> <input.fasta>
#viral annotation
$prokka --outdir <dir_name>  --prefix <file_prefix>  --kingdom viruses  --addgenes
<input.fasta>
```

Command 2-19 Prokka commands for genome annotations of bacteria and viruses

2.4.10.1. Gene Extraction

The total number of genes was extracted from the GenBank file created by Prokka, by searching for the phrase “product=”. The results were input into a text file, and the number of genes predicted was then calculated by counting the number of lines. The number of hypothetical proteins was then calculated by searching for “hypothetical proteins” in the resulting text file, which was output into another text file and again the number of lines were counted. The extracted lists of genes were then entered into jvenn[116] which filters the lists for repetition and draws a Venn diagram to allow the identification of common items on each list. The information contained in each section of the Venn diagram was then downloaded in csv format and analysed for patterns.

```
grep -i "product=" <input.gbk> > <output.txt>
wc -l <output.txt>
grep -i "*.hypothetical protein.*" <output.txt> > <output_hypo.txt>
wc -l <output_hypo.txt>
```

Command 2-20 gene extraction and counting from Prokka output

2.4.11. Virulence Factor Detection

Rapidly detecting virulence factors such as toxins using genome data will allow prediction of disease states caused by the pathogen

2.4.11.1. Virulence Factor Data Base (VFDB)

The virulence factor database[117–119] was downloaded and a local blast library was produced. The results of the *de novo* assembly were then investigated using Blastn and this database.

2.4.11.2. Virulence Factors from Prokka Annotation

The product list extracted from Prokka as described in **2.4.10.1** was searched for the following phrases phage, plasmid, toxin, pathogen, virulence, adhesion, transposon and conjugative. The results were output into a text file and investigated.

2.4.12. Resistance Prediction

The use of genomic data to predict resistance in bacteria is an on-going challenge; matching genotype to phenotype is a major development, with many species specific studies being performed with the aim to address this challenge. In this study generic methods were investigated in order to provide a single pipeline for all species identified using this method.

- Resistance
- Drug
- Efflux
- Beta-lac
- Antibiotic
- Cillian
- Mycin
- Cef
- Penem
- Oxacin
- Bactam
- Cyclin
- Etheprim
- Kacin
- Lothin
- Lisitn
- Furatoin
- Icin
- Fusaric

2.4.12.1. Using ardbAnno.pl

A test file was created detailing the fasta files that were products of the *de novo* assembly, then the *ardbAnno.pl* [120] script was ran (Perl *ardbAnno.pl*). This produced genomeList.tab which was viewed using excel to assess resistance factors.

2.4.12.2. Using ResFinder

The online version of ResFinder 2.1 [71] was used, the fasta file produced by the *de novo* assembly was uploaded to the site.

2.4.13. Final pipeline

The final pipeline uses all the best parameters identified in this chapter to allow automation in the genome assessment. The final pipeline is split into two segments, each automated by simply changing the input file. The break in the automation allows visualisation of the sample, and allows mixed cells to be separated and analysed separately in the second half of the pipeline.

2.4.14. Part 1

The first part of the data analysis pipeline, has the raw reads as the input file, removes host signals as well as any reads that map to the negative library. The pipeline then abundance and error trims the data before it is *de novo* assembled. The taxonomy of the reads is then identified using blast and the resulting file is opened in Megan viewer. This is then visually investigated before the second part of the pipeline was implemented.

```
#!/bin/sh

#map against human

~/runMapping -o human /mnt/data123/data1/cat/h_genome/human.fa <input_file>.sff

#move unmapped.txt

cp human/454ReadStatus.txt human.txt

#remove unmapped reads

python ~/unmapped_human.py human.txt > unmapped_human.txt

~/gsSeqTools/bin/sfffile -i  unmapped_human.txt -o unmapped_human.sff <input_file>.sff


#map against crap library

~/runMapping -o contaminant ~/neg_lib.fasta  unmapped_human.sff

cp contaminant/454ReadStatus.txt contaminant.txt

python ~/unmapped_contamination.py contaminant.txt > unmapped_contaminant.txt

~/gsSeqTools/bin/sfffile -i  unmapped_contaminant.txt -o unmapped_contaminant.sff
 unmapped_human.sff

#convert file to fastq

~/sff2fastq -n unmapped_contaminant.sff >  <input_file>_pure.fastq

#digital normalisation

source ~/khmerEnv/bin/activate

#normalized data by median
python ~/khmerEnv/bin/normalize-by-median.py -k 140 -x 2e8 -C 60 <input_file>_pure
.fastq --savehash <input_file>_pure-dn.kh

#abundance filter
python ~/khmerEnv/bin/filter-abund.py -C 2 -V <input_file>_pure-dn.kh <input_file>
>_pure.fastq.keep


#normalize

python ~/khmerEnv/bin/normalize-by-median.py -k 140 -x 2e8 -C 50 <input_file>_pure
.fastq.keep.abundfilt

#output
cp <input_file>_pure.fastq.keep.abundfilt.keep  <input_file>_pure_ab.fastq

perl ~/prinseq-lite.pl -fastq <input_file>_pure_ab.fastq -out_format 3 min_length 99 -
-trim_left 4 -trim_qual_left 20 -trim_qual_right 20  -min_qual_mean 20 -log pure -
-out_good <input_file>_pure_ab_trimmed

#convert to fasta

seqtk fq2fa <input_file>_pure_ab_trimmed.fastq  > <input_file>_pure_ab_trimmed.fasta

#blast

blastn -query <input_file>_pure_ab_trimmed.fasta -db ~/blast/nt/nt -out <input_file>
>_pure_ab_trimmed.blastn.nt.out -num_threads 6

#open in megan
~/MEGAN  -g false -x "import blastfile=<input_file>_pure_ab_trimmed.blastn.nt.out
 readfile=<input_file>_pure_ab_trimmed.fasta  meganfile=<input_file>_pure_ab_trimmed
.blastn.rna minscore=150   "
```

Command 2-21 Pipeline part 1- for automated preparation of reads by removal of host and environmental contaminations, followed by error and abundance trimming. Reads are then identified using Blastn before being visualised using LCA analysis on MEGAN.

2.4.15. Part 2

The second half of the pipeline runs using the extracted reads from megan which are in fasta format. It extracts the reads from the trimmed fastq file to give a fastq of the extracted reads. The reads are then assembled, and the resulting file annoatated and information extracted.

```
#!/bin/sh

#extract read IDs
perl -ne 'if(/^>(\S+)/){print "$1\n"}'  <megan_input_file>.fasta > <megan_input_file>_list.txt

#extract from fastq

seqtk subseq  <input_file>_pure_ab_trimmed.fastq  <megan_input_file>_list.txt > <megan_input_file>.fastq
#assembly
python ~/SPAdes-3.5.0-Linux/bin/spades.py -o <megan_input_file> -s <megan_input_file>.fastq
#rename assembly
cp <megan_input_file>/contigs.fasta <megan_input_file>_assembled.fasta
#annotate
prokka --outdir <megan_input_file>_prokka --prefix <megan_input_file>_prokka --addgenes <megan_input_file>_assembled.fasta

#extract gene IDs

grep -i "product=" <megan_input_file>_prokka/<megan_input_file>_prokka.gbk > <megan_input_file>_prokka_all_products.txt

grep -i "drug\|beta-lac\|betalac\|antibiotic\|mycin\|icin\|cillin\|cef\|penem\|oxacin\|bactam\|cyclin\|etheprim\|kacin\|lothin\|listin\|nitrofuratoin"                <megan_input_file>_prokka_all_products.txt > <megan_input_file>_prokka_resistance.txt


grep -i "phage\|plasmid\|toxin\|virulence\|pathogen\|adhesin\|transposon\|plasmid "  <megan_input_file>_prokka_all_products.txt > <megan_input_file>_prokka_virulence.txt

```

Command 2-22 Second part of the automated pipeline, which extracts reads identified by LCA analysis on MEGAN into a fastq file. This fastq is then de novo assembled and annotated.

2.5. Further Application of ϕ29 MDA to viral pathogens

2.5.1. Sequencing of DNA and RNA extracted from viral tissue culture.

Methods for amplification of pathogens need to be applicable to both RNA and DNA, ideally with a single sample processing pathway capable of detecting both nucleic acid types. This will need to combine both reverse transcription and DNA amplification. Along with a compatible concentration and extraction technique, with added stages for removing any host genetic material. The aim is to have a very sensitive method for pathogen detection, this section will concentrate on the ability to characterise a variety of viral pathogens with a single work-flow. Initial work focused on the compatibility of the process for both RNA and DNA virus, with both an enveloped and none enveloped virus being used (HIV and Adenovirus 41). After this a more complex viral mixture was set up, with viruses representing multiple genome types, viral parcel size and particle type.

2.5.1.1. Concentration of Tissue Culture Material for HIV and Adenovirus

The supernatant from tissue culture of 8E5/LAV and Adenovirus 41 were defrosted at 37°C before being placed on ice. The aliquots were then centrifuged at 6000xg to pellet host cells, before 500 µl of the supernatants were concentrated using PEG as described by Kohno et al [121]. PEG 20,000 was dissolved in 0.9% NaCl to give a PEG total of 20% (w/v). A 1:1 volume ratio of sample and PEG was used, and the samples were then incubated at 4°C for 16 hours or one hour or on ice for 1 hour. The samples were then centrifuged at 17860xg for 20 minutes, the supernatant removed and the pellet resuspended in 50µl PBS.

2.5.1.2. Addition of RNase and DNase

After concentration with PEG, and suspension in 50µl of sterile PBS 10 units of RNase A (Thermo Fisher Scientific) along with 5 units of DNase I, (Thermo Fisher Scientific) was added. The sample was then briefly vortexed and incubated at 37°C for 20 minutes. After incubation 100U RiboLock (Thermo Fisher Scientific) was added to the sample and 5µl 0.5M EDTA.

2.5.1.3. Extraction of RNA and DNA using Column Extraction Method

The RNA and DNA were then extracted using the alkali method described in 2.4.1 or by using PureLink® Viral RNA/DNA Mini Kit (Thermo Fisher Scientific), as described in the manual . Carrier RNA was substituted with Linear Acrylamide (LPA) (Invitrogen). Briefly 25µl Proteinase K, 200 µl lysis buffer and 5 µl LPA was added to the sample, gently vortexed and then incubated at 56°C for

15 minutes. After the addition of 250 µl 100% ethanol the sample was incubated for a further 5 minutes at room temperature. The sample was then added to the viral spin column and centrifuged for 1 minute at 6800xg, and the flow through was discarded. The sample was then washed twice with wash buffer before being eluted in 25 µl RNase free water.

Any nucleic acids not immediately used were stored at -80°C. The RNA was quantified using the Qubit High sensitivity RNA assay. The DNA was quantified using Qubit High sensitivity DNA assay.

2.5.1.4. Dilution of RNA and DNA

After quantification the RNA and DNA was diluted to a level that would represent 25 viral per µl for each virus. Calculations of viral particles based on nucleic acid quantification was performed using End Memo online tool for copy number calculator. Calculations were performed based on HIV having two copies of the ssRNA genome per Virion and Adenovirus having a single copy of a dsDNA genome. Dilutions were made using RNAse free water containing 2 U/µl RNaseOUT™ Recombinant Ribonuclease Inhibitor. The DNA extracted from the adenovirus tissue culture was diluted to 0.001 pg/µl. This was added to the equivalent of 25 virions of HIV (0.0002 pg/ µl).

2.5.1.5. Reverse Transcription, ϕ29 Amplification and Sequencing

The mixture of the extracted RNA and DNA was added to the reverse transcription reaction with SuperScript IV in a half volume reaction. 1 µl of each viral extract dilution was added to 0.5 µl 50 µM random hexamers (Life Technologies), 0.5 µl 10mM dNTP mix (Life Technologies), and 3.5 µl nuclease-free water. This mix was vortexed and centrifuged before incubation at 65°C for 10 minutes and then on ice for 2 minutes. Whilst the reaction was on ice the following components were added, 2 µl SSIV busser (Life Technologies), 0.5 µl 100mM DTT, 0.5 µl RNASE OUT, and 0.5 µl SSIV. This was then gently vortexed, centrifuged and incubated at 23°C for 10 minutes, 55°C for 10 minutes. After completion of the reverse transcription reaction the product was then heated to 95°C for two minutes and then held on ice for two minutes. The resulting DNA was then amplified using ϕ29 MDA for two hours. The product was then fragmented using S1 nuclease treatment and nebulisation for two minutes before library preparation as described by Roche protocols and sequenced on the 454 Junior.

2.5.2. Adenovirus and HIV-Viral Particles

Dilutions of the supernatant of HIV and Adenovirus tissue cultures were made so as to contain an estimated 25 viral particles. These were pooled and then the total volume made up to 10 ml using PBS. This was then concentrated with PEG 20,000 incubated overnight at 4°C as described in **2.5.1.1** , host nucleic acid removed **2.5.1.2** and the sample was extracted using column extraction **2.5.1.3**. The resulting eluted nucleic acid was then added to a reaction with SuperScript IV using random primers **2.5.1.5** before being heated to 95°C for two minutes followed by 2 minutes on ice. This was then amplified in a two hour ϕ29 MDA reaction. The resulting DNA was quantified, fragmented and sequenced on the 454 junior as previously described.

2.5.3. Amplification of Mixed Viruses

Viruses detailed in **Table 2-3** were obtained from NIBSC and stored at -80°C until use. These were selected to represent a wide range of viruses, including different viral morphology, size and particle type. A review of concentration methods was undertaken, with most focusing on a concentration method optimised for a single virus group. Details of the results of this review are shown in **Table 2-9**. After consideration of all the concentration protocols, it was decided to use the highest molecular weight PEG, which should allow concentration of even the smallest viruses, and has previously been used for even delicate enveloped viruses. As this method is focusing on finding a rapid method of pathogen detection, it was decided to incubate the samples on ice for an hour. Centrifuge speed was selected as the highest available on common bench top centrifuges 17860xg.

Family	Virion Size	Enveloped?	Shape	PEG	Incubation	Spin	Ref
Herpesviridae	150-200nm	yes	spherical to pleomorphic	20% PEG 8000	Ice 30 min	11,000xg 20 min 4°C	[122,123]
Adenoviridae	90nm	no	Icosahedral	7% PEG 8000	2 hours 4°C	10, 000 x g 30 minutes	[124]
Parvoviridae	18-26nm	no	Icosahedral	7.5% PEG 8000	16 hours 4°C	15,000xg 45 min 4°C	[125]
Hepadnaviridae	42nm	yes	spherical	10% 8000 PEG	4°C I hour	925 xg 20 min	[126,127]
Retroviridae	80-100nm	Yes	spherical to pleomorphic	20% PEG 20,000	4°C 16 hours	17, 860x g 20 min	[121]
Paramyxoviridae	150nm	yes		10% PEG 8000	4°C 2 hours	10,000x g 1 hour	[128]
Orthomyxoviridae	80-120	yes	Usually rounded but can be filamentous	50% PEG 6000	4°C 2 hours	10,000 x g 30 minutes	[129]
Picornaviridae	30nm	no	spherical	8% PEG 8000	4°C overnight	10,000x g 1 hour	[130,131]
Caliciviridae	27-40nm	no	icosahedral	8% PEG 8000	2 hours 4°C	10,000xg 20 mins	[132]
Flaviviridae	50nm	yes	spherical	6% PEG 6000		7500 x g 30 min	[133]

Table 2-9 Details of previously used concentration methods for the viruses to be studied, including details of the virus size and particle type.

2.5.3.1. Creating Mixed Viral media

To test the concentration and amplification methods a low input viral mix was created using the viruses obtained from NIBSC. Where the standards were assigned a value in terms of infectious units per millilitre (IFU/ml) the volume was calculated which would contain approximately 25 viral particles, which were 22 µl HBV, 31 µl Parvovirus 19, 43 µl HAV and 17 µl HCV. For samples where a Ct value was given 20 µl of sample was added. These were pooled, along with the addition of the equivalent of 25 virions of HIV and Adenovirus from tissue culture. Before pooling of viral samples all tubes were centrifuged at 6000xg to pellet host cells.

NIBSC product number	Virus	Quantification information	µl required for estimated 25 IFU
08/310	Varicella Zoster virus	Ct 30	20
11/182	HBV	1140 IFU/ml	22
11/208:	Parvovirus B19	unitage of 800 IFU/ml (range 569-1039)	31
07/294:	Norovirus	Ct 30	20
13/102:	HAV	Unitage 588 IFU/ml (range 507–680 IFU/ml).	43
13/168	Measles	Ct 30	20
07/298	Influenza	Ct 30	20
02/264:	HCV	assigned a value 1438 IFU/ml	17

Table 2-10 details of quantification measurements for viruses used in the mixed viral media along with volume input of each virus into the initial viral mix

2.5.4. Preparation of DNA for sequencing from mixed viral media

Pooled virus media was made up to a total volume of 10ml in HBSS. 10 ml of 20% (w/v) PEG 20,000 in 0.9% NaCl was added to the viral mix. This was then gently shaken at room temperature for 5 minutes before incubation on ice for 1 hour. The sample was then centrifuged at 17860xg for 20 minutes, the supernatant removed and the pellet resuspended in 50µl HBSS by vigorous vortexing.

The sample was then treated with DNAse and RNAse as described in **2.5.1.2** before being extracted using column extraction as described in **2.5.1.3**. After extraction the nucleic acid was reverse transcribed and amplified using a 2 hour ϕ29 MDA reaction as described in **2.5.1.5**. Once amplified the sample was quantified and sequence using the 454 Junior.

2.5.4.1. Improvements in Mixed Viral Media

After reviewing of initial sequencing results, the input amounts of each viral pathogen were adjusted to create a more even mixture. The VZV reference material was diluted 1:100 in HBSS and 20 µl added. The amount of the HBV reference material added was lowered to 6 µl and the Parvovirus was diluted 1:10 and 30 µl added. The Norovirus was diluted 1:100 and 12.5 µl added. The amount of HAV virus added was increased to 65 µl, the amount of HCV added was lowered to 3.5 µl. the amount of measles reference material added was increased to 200 µl. a 1:100 dilution of influenza was prepared and 17 µl of this was added to the viral mix. The amount of HIV and Adenovirus remained the same. The mixed media was made up to a total volume of 10ml and prepared for sequencing as described in **2.5.4**.

2.6. Development of a Bacteraemia Model to Simulate Highly Sensitive WGS of Pathogens from Sterile Clinical Specimens

2.6.1. Survival of Bacterial in Horse Blood and 2% Saponin De Novo Assembly

Freshly grown plates from all of the clinical isolates in **Table 2-2** were suspended in 1 ml PBS and diluted in a 1:10 series and 10µl of each dilution plated out in triplicate. After 24 or 48 hours incubation the number of cells was counted and an average calculated. 500 cells were then added to 1ml PBS, 1ml defibrinated horse blood (TCS Bioscience) or 1ml 2% saponin (Sigma) and vortexed, before being incubated for 2 hours at room temperature. The samples were then vortexed again and two aliquots of 100µl from each sample were plated on appropriate media and incubated in appropriate conditions for 24 or 48 hours. Colonies were then counted and survival rates calculated.

2.6.2. Bacterial Isolation Using Density Gradients

Percoll (Sigma) density gradients were set up using 1.5M NaCl at the following densities, 1.13 g/ml (*E. coli*), 1.09 (erythrocytes) and 1.1 for a small layer in between to catch un-separated cells. The volume calculation shown in **Equation 1** was used and the layers prepared as shown in **Figure 2-2** to form a discontinuous gradient. Briefly the highest density gradient was added first using a wide bore pipette, then subsequent layers were added by slowly pipetting the layer with the tip against the edge of the tube. To this a 2ml mixture of blood and PBS (50:50) spiked with 50 cells of wither *S. aureus* or *E. coli* was added gently to the top of the tube. The sample was then centrifuged at 800xg for 25 minutes. The different segments were then plated out as follows, the top layer was split into 3x1ml and the layer with the red blood cells was plated in the following order 2x500 µl, 200 µl and 100 µl and the final 500µl clear zone was also plated out. A control of the 2ml horse blood and PBS spiked with *S. aureus* or *E. coli* was also plated out to assess the bacterial input.

Volume Calculation

$$V_0 = V \times \frac{\rho - 0.1\rho_{10} - 0.9}{\rho_0 - 1}$$

V_o = Volume of undiluted Percoll/Percoll PLUS required in ml
V = Volume of final working solution in ml
ρ = Desired density of final working solution
ρ_o = Density of Percoll/Percoll PLUS undiluted
ρ_{10} = Density of 1.5 M NaCl (1.058 g/ml) or 2.5 M sucrose (1.316 g/ml)

Equation 1 Percoll volume equation to calculate formation of required density gradient

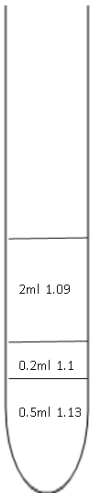


Figure 2-2 prepared Percoll layers including volume and density of each layer

2.6.3. Erythrocyte Depletion Using HetaSep®

2.6.3.1. Centrifugation vs. Gravity.

There are two options for use of HetaSep® (Stemcell technologies), either separation through centrifugation or by gravity, both were investigated. Samples were set up using 2ml 50:50 horse blood and PBS spiked with either *E. coli* or *S. aureus*, or with 1ml horse blood spiked with bacteria. To this 400µl HetaSep® was added to the PBS and horse blood mix, or 200µl HetaSep® was added to the horse blood sample. Samples were then either incubated at 37°C for 20 minutes or centrifuged at 90xg for1 minute and then left at room temperature for 10 minutes. The top and bottom fractions were then plated on blood agar and incubated overnight. A control was also plated which was spiked blood or blood PBS mixed incubated at 37°C for 20 minutes and then the whole sample was plated out.

2.6.3.2. Single vs. Double Incubation

Samples were set up using 2ml PBS and horse blood (50:50) spiked with either *E. coli* or *S. aureus* with the addition of 500µl HetaSep®. Set A was the incubated for 20 minutes at 37°C as previous. Three segments were then plated the top segment, the segment at the interphase of the red blood cells and supernatant, and the pellet containing the red blood cells. Set B was incubated at 37°C for 10 minutes and the top layer removed, an additional 500µl HetaSep® was

added and the sample incubated for a further 10 minutes. The same segments were plated as above with the addition of the pellet from the first HetaSep® incubation.

2.6.3.3. Different HetaSep® Ratios and Shortened Incubation Time

Different ratios of blood to HetaSep® were investigated, along with the impact of PBS dilution with a shorter incubation time. The following ratios of blood to HetaSep® were set up, 1:0.2 (recommended), 1:.4, 1:0.6, 1:0.8, 1:1. The same ratios were also set up with the blood and PBS mix. The samples were then incubated at 37°C for 10 minutes. The top and bottom segments for each were plated, along with a control that had no HetaSep® added.

2.6.4. Bacterial isolation and sequencing from Horse blood

Blood models were set up with the addition of 10 bacterial cells (*S. aureus and E. coli*) to 1ml horse blood. The following workflow was then applied, with samples being cultured at each stage to assess bacterial survival. 200 µl HetaSep® was added and the sample vortexed and incubated at 37°C for 10 minutes. 500 µl supernatant was removed and 200µl 5% was added to a final 2% solution, and incubated at room temperature for 5 minutes. 700µl sterile water was added for a water shock and incubated at room temperature for 3 minutes before salt restoration with the addition of 21µl 5M NaCl. The sample was then centrifuged for 5 minutes at 3000xg and the supernatant removed and discarded. The pellet was then washed in lowering volumes of PBS, initially 200µl then 100µl followed by 20µl with each stage being span at 6000xg fir three minutes, and the final pellet being resuspended in 4µl PBS **Figure 2-3-A.**

2.6.4.1. Bacterial Isolation from Horse Blood-Improvements

The work flow from above was repeated with the following points of improvements, the volume of supernatant removed from the HetaSep® incubation was increased from 500µl to 550µl. After this the volumes of Saponin and water and salt restoration were increased in ratio with the increased volume. The water shock was also reduced to 30 seconds. The initial centrifugation after salt restoration was increased from 3000xg to 4000xg. The final centrifugation during the wash stages was increased from 3 minutes to 5 minutes **Figure 2-3-B.**

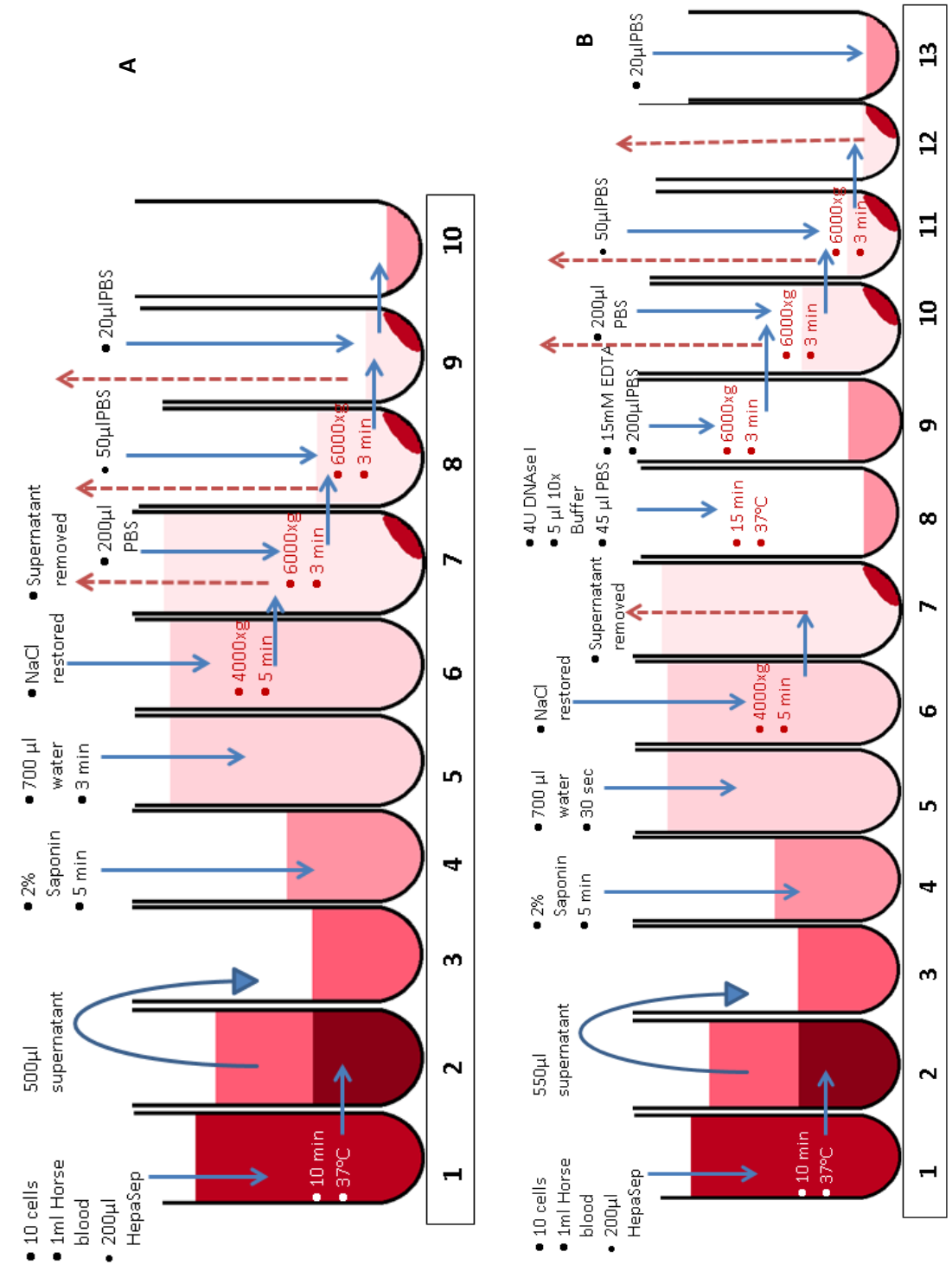


Figure 2-3 Work flow for isolation of bacterial cells from whole blood including removal of RBCs using HetaSep, selective eukaryotic lysis using saponin and water shock followed by salt restoration and washes with PBS (A) original workflow (B) workflow after improvements to include shorter water shock and DNAse treatment.

2.6.4.2. Sequencing from Horse Blood

The *E. coli* and *S. aureus* isolated using the method in **2.6.4.1** were amplified and then sequenced.

2.6.4.3. Addition of DNAse Stage

A DNase treatment was added after the first spin at 4000xg for 5 minutes. 5 µl of 10xbuffer and 2µl turbo DNase 1 (Ambion) was added to the pellet and the sample was vortexed and incubated at 37^{o}C for 15 minutes. EDTA was then added to a final concentration of 15nM and the sample was then either heated to 75^{o}C before entering the washing phase, or was taken straight in to the washing phase. To assess the inactivation of the DNase1 after washing 1µg of DNA was added to the pellet and incubated at 37^{o}C for 30 minutes and then the DNA was quantified.

2.6.5. Process for Isolation, Amplification and Sequencing of Pathogens in Clinical Specimens

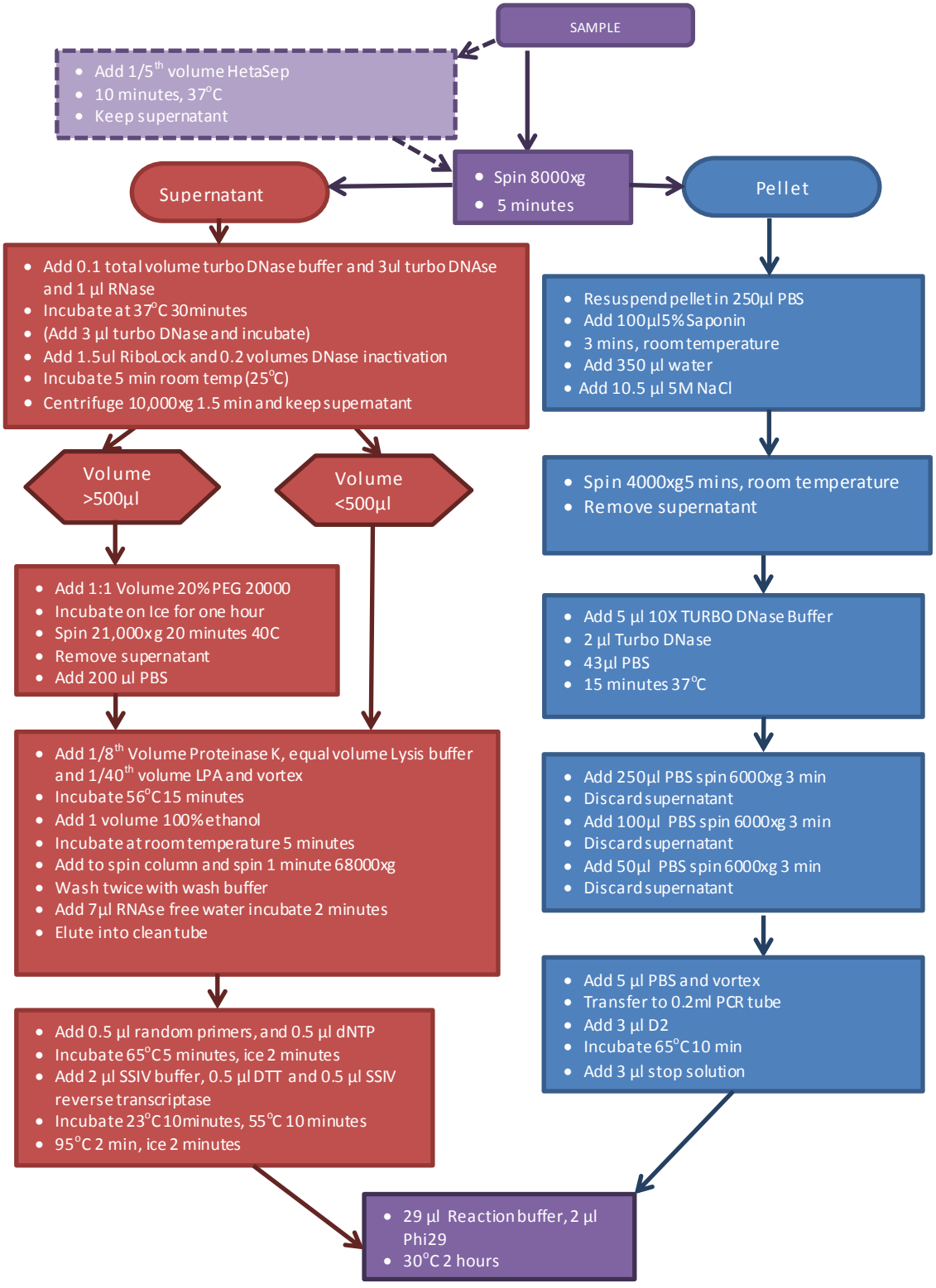


Figure 2-4 final sample processing pipeline for whole genome sequencing from sterile site infections with two processing pathways for bacteria and viruses. Processing includes, isolation concentration, extraction and amplification of pathogen signals

2.7. Sequencing using Illumina Technology

The MDA products were treated with S1 nuclease as described in **section 2.3.2.2,** nebulised for 15 seconds before being purified using the MinElute PCR Purification Kit (Qiagen). The sample was then diluted to 0.2ng/µl and 5µl was taken forward to the library preparation.

2.7.1. Library Preparation

Nextera XT DNA library preparation Kit (Illumina) was used to prepare the libraries for sequencing. The first stage involves simultaneous fragmentation and tagging of the DNA (Tagmentation), using an engineered transposon. Briefly, buffer TD, the DNA and the enzymes were mixed in a 20µl reaction and incubated at 55°C for 5 minutes, before being neutralised. Indexes were then added to the samples before PCR amplification, using the following parameters: 72°C hold 3 minutes, and 95°C hold 30 seconds, 12 cycles of 95°C 10 seconds, 55°C 30 seconds, 72°C 30 seconds. There is a further hold at 72°C followed by cooling and holding at 10°C. The PCR products are then cleaned up using AMPure XP beads, aiming for a pool size of 300-500bp in length. 1 µl of undiluted library was then QC tested on the Bioanalyser (Agilent) using a High Sensitivity DNA chip. After library normalisation the samples were pooled.

2.7.2. Sequencing

When sequencing on the MiSeq, V2 300 cycle kits were used, when sequencing on the HiSeq, V1 200 cycle kits were used.

2.7.3. Data analysis Adaptations for Illumina data

The pipeline was adapted for use with the shorter paired-end reads produced by illumina platforms. The first alteration was the use of 'seqtk subseq' command (line 35 **Command 2-23**) to remove reads which mapped to the host or contmaination, in place of the custo m script and sfffile command previously used for 454 junior produced data. The value of K used in the digital normalisation stage was reduced to 32 (line 48 **Command 2-23**) to reflect the shorter reads produced on the ilumina platform. Orphan reads were removed (line 58 **Command 2-23**) to mainitain pair-end reads. During the error trimming stage the mimimun read length was reduced to 25 (line 65 **Command 2-23**). When converting the fastq produced to a fasta file for use in Blastn analysis the comman 'Seqtk seq –A' was used, reflecting an updated version of the Seqtk software (release 1.2).

```
#!/bin/sh
mkdir <input_file>
cd  <input_file>
#map against human

~/runMapping -o human_R1 /srv/data1/cat/h_genome/human.fa ../<input_file>_R1.fastq &
~/runMapping -o human_R2 /srv/data1/cat/h_genome/human.fa ../<input_file>_R2.fastq
wait
#move unmapped.txt

cp human_R1/454ReadStatus.txt human_R1.txt
cp human_R1/454ReadStatus.txt human_R2.txt

#remove unmapped reads
python ~/unmapped_human_R1.py human_R1.txt > unmapped_human_R1.txt
python ~/unmapped_human_R2.py human_R2.txt > unmapped_human_R2.txt


seqtk subseq  ../<input_file>_R1.fastq  unmapped_human_R1.txt > unmapped_human_R1
.fastq
seqtk subseq  ../<input_file>_R2.fastq  unmapped_human_R2.txt > unmapped_human_R2
.fastq

#map against contaminant library

~/runMapping -o contaminant_R1 ~/neg_lib.fasta  unmapped_human_R1.fastq & ~/runMapping
 -o contaminant_R2 ~/neg_lib.fasta  unmapped_human_R2.fastq
wait

cp contaminant_R1/454ReadStatus.txt contaminant_R1.txt
cp contaminant_R1/454ReadStatus.txt contaminant_R2.txt


python ~/unmapped_contamination_R1.py contaminant_R1.txt > unmapped_contaminant_R1.txt

python ~/unmapped_contamination_R2.py contaminant_R2.txt > unmapped_contaminant_R2.txt


seqtk subseq  ../<input_file>_R1.fastq  unmapped_contaminant_R1.txt > <input_file>
>_R1_pure.fastq
seqtk subseq  ../<input_file>_R2.fastq  unmapped_contaminant_R2.txt > <input_file>
>_R2_pure.fastq

```

```
#digital normalisation

source ~/khmerEnv/bin/activate
#merge paris
python ~/khmerEnv/bin/interleave-reads.py <input_file>_R1_pure.fastq <input_file>_R2_pure.fastq > <input_file>_pure.fastq


#normalized data by median
python ~/khmerEnv/bin/normalize-by-median.py -k 32 -x 2e8 -C 60 -p <input_file>_pure.fastq  -s <input_file>_pure-dn.kh
#abundance filter
python ~//khmerEnv/bin/filter-abund.py -C 2 -V <input_file>_pure-dn.kh  <input_file>_pure.fastq.keep


#normalize

python ~/khmerEnv/bin/normalize-by-median.py -k 32 -x 2e8 -C 50 <input_file>_pure.fastq.keep.abundfilt

# remove orphan reads
python ~/khmerEnv/bin/extract-paired-reads.py <input_file>_pure.fastq.keep.abundfilt.keep
# separate pairs
python ~/khmerEnv/bin/split-paired-reads.py <input_file>_pure.fastq.keep.abundfilt.keep.pe
mv <input_file>_pure.fastq.keep.abundfilt.keep.pe.1 <input_file>_pure_ab_L.fastq
mv <input_file>_pure.fastq.keep.abundfilt.keep.pe.2 <input_file>_pure_ab_R.fastq
deactivate

prinseq-lite.pl -fastq <input_file>_pure_ab_L.fastq   -out_format 3 min_length 25
-trim_left 4 -trim_qual_left 20 -trim_qual_right 20  -min_qual_mean 20 -log trim_L
-out_good <input_file>_pure_ab_trimmed_L & prinseq-lite.pl -fastq <input_file>_pure_ab_R.fastq   -out_format 3 min_length 25 -trim_left 4 -trim_qual_left 20
-trim_qual_right 20  -min_qual_mean 20 -log trim_R -out_good <input_file>_pure_ab_trimmed_R

wait

seqtk seq -A <input_file>_pure_ab_trimmed_L.fastq  > <input_file>_pure_ab_trimmed_L.fasta
seqtk seq -A <input_file>_pure_ab_trimmed_R.fastq  > <input_file>_pure_ab_trimmed_R.fasta

blastn -query <input_file>_pure_ab_trimmed_L.fasta -db ~/blast_maintained/nt/nt -out
 <input_file>_pure_ab_trimmed_L.fasta.nt.out -num_threads 6
blastn -query <input_file>_pure_ab_trimmed_R.fasta -db ~/blast_maintained/nt/nt -out
<input_file>_pure_ab_trimmed_R.fasta.nt.out -num_threads 6

cd ..
```

Command 2-23 pipeline part one adapted for Illumina sequencing, including adaptations for shorter reads in the abundance and error trimming and removal of orphaned reads

2.8. Application of Methods Developed to Real and Modelled Clinical Samples

Several sample types, including real clinical samples and infection models were prepared for sequencing in parallel. This allowed the sample processing developed in **2.6.4** to be tested for utility on several sample types simultaneously. All samples were sequenced in a single sequencing run on the HiSeq. A summary of all samples can be found in **Table 2-12**

2.8.1. Negative Samples

An extraction negative control was set up by starting with 1ml of PBS in place of sample and the entire sample processing including reverse transcription and ϕ29 MDA was performed, and the products sequenced. The second negative control was water which was used in a reverse transcription and ϕ29 MDA reaction.

2.8.2. Further Blood Models

Four further bacteraemia models were created to model more challenging diagnosis. The first was a mixture of *S. Pyogenes* (clinical isolate 15M153838) and Influenza virus (NIBSC control media Influenza A H3N2 07/298) to model cases of secondary bacteraemia following viral infection. The second model was a *Shigella,* which many methods (including MALDI-TOF) cannot differentiate from *E. coli*. A model using non Typhi invasive *Salmonella* was used, as whole genome sequencing was available from the *Salmonella* Reference Unit at PHE Colindale, and this provided an opportunity to compare straight from sample methods to culture based sequence library preparation. The final model was a *P. aeruginosa*, which was highly resistant, including to carbapenems.

Dilutions of bacteria were prepared as described in **2.3.4.2** and single cells of *S.Pyogenes*, *Salmonella sp* and *P. aeruginosa* were inoculated into 1 ml horse blood. Ten cells of *S. sonnei* were inoculated into 1 ml horse blood, due to poor survival rates shown in **2.6.1.** For the Influenza in the mixed model, 20 µl of a 1:100 dilutions of the control material was used, to aim at 25 viral copies, as described in **2.5.4.1.**

2.8.3. Tissue Model

Post mortem none human primate tissue (rhesus macaque) was obtained from PHE Porton Down. Roughly 1cm^3 tissue samples were collected and were stored in RNAlater (Sigma) until use. The first tissue model was to simulate endocarditis, by inoculating heart tissue with *H. influenzae*

*(*clinical isolate 15M154514) dilutions of an overnight culture were prepared as described in **2.3.4.2** and five cells were inoculated into the tissue. The second model was to simulate viral hepatitis, using Hepatitis A virus control media, the aim was to inoculate 25 viral particles, and so 65 µl of the control media was added (as in **2.5.4.1**) to the tissue. The third tissue was a model of a mixed infection using Bacteroides *vulgatus (*14M180862), *Streptococcus oralis (*15M150586), *Streptococcus mitis (*15M150586) and *Streptococcus anginosus (*15M152569). Five cells of each bacterium were mixed into a total volume of 50 µl and inoculated into the brain tissue. The final model was a viral kidney infection, for this CMV was used, 20 µl of 1:100 control media was inoculated into the tissue, this value was aiming at ~25 viral particles based on previous work using VZV virus in **2.5.4.1**

Samples were inoculated using, extra-long 10 µl pipette tips (VWR®) and incubated at 37°C for one hour. The tissue samples were first ground using Dounce tissue grinder set (Sigma-Aldrich), with the addition of 500 µl sterile PBS. 50 U of collagenase (Sigma) was added and the sample was incubated on a shaker at 37°C. 250 µl of 2% 2-mercaptoethanol (Sigma) was then added and the sample incubated a further hour at room temperature on a shaker. The sample was then centrifuge at 6000 xg for 5 minutes and the supernatant discarded. The sample was then washed three times using 3 ml PBS and centrifuged at 6000 xg. The final pellet was resuspended in 5ml of PBS. This was then vacuum filtered using a 100µm filter (Millipore). The filter was discarded and the filtered sample was used as the input for the previously developed sample process.

2.8.4. Urines

Urine samples were sent from Addenbrookes hospital for deep sequencing analysis as part of an on-going project looking at genotyping BK virus directly from samples. Samples from symptomatic patients were included in the project and four of these urines were randomly chosen and 50 µl used for processing in this thesis.

2.8.5. CSF

CSF samples with unknown aetiology from symptomatic patients were collected as part of a PHE historic study. 20 µl of four of these CSF were processed, however due to the uncertain age and handling of these samples some adaptations of the sample processing was made. DNAse and RNAse stages weren't carried out, to be certain of the recovery of pathogen DNA which may have been degraded.

2.8.6. STI swabs

Four blinded samples were used from sexually transmitted bacteria reference unit (STBRU) at PHE Colindale full sample details are in **Table 2-11**. These were samples which had been referred to the reference laboratory for diagnostic confirmation. The samples consisted of swabs which had been placed in buffer and vortexed to deposit material into liquid. After completion of STBRU processing 40 µl of sample was processed as shown in **Figure 2-4**

	Swab type	Buffer type	Reference lab results
1	RECTAL	BD VIPER	C. trachomatis + (Ct@23.02) LGV + (Ct@23.41/23.63)
2	VAGINAL	BD VIPER	C. trachomatis + (CT@23.56) LGV -
3	VAGINAL	COBAS	N. gonorrhoeae + (Ct @29.14/29.14)
4	VAGINAL	COBAS	N. gonorrhoeae + (Ct @32.64/30.36)

Table 2-11 Details of STBRU samples processed in this thesis, including swab site, buffer and STBRU results LGV= Lymphogranuloma venereum

2.8.7. *Treponema pallidum* Samples

In collaboration with St Marys Hospital six samples with presumed *T. pallidum* infections were tested, two sets of matched samples from patients positive for *Treponema pallidum* (whole blood and ulcer samples) and two further blood samples from an additional two patients. Samples were retried from the Imperial College London Communicable Disease Research Tissue Bank. Ulcer samples were shipped in RNA later solution (Sigma).

Sample number	Sample type	Input	Input amount
1	Control	Negative extract	1 ml PBS
2	Control	Negative amplification	10 µl water
3	Bacteraemia model	• *Streptococcus Pyogenes* (15M153838) • Influenza A H3N2 (07/298)	Single cell 20 µl 1:100 dilution of control media (~25 VP)
4	Bacteraemia model	*Shigella sonnei* (15M005139)	Ten cells
5	Bacteraemia model	*Salmonella Sp.* (14M180208)	Single cell
6	Bacteraemia model	*Pseudomonas aurginosa* (13M155577)	Single cell
7	Endocarditis model	*Haemophilus influenzae* (15M154514)	Five cells
8	Viral Hepatitis	HAV RNA (13/102)	65 µl control media (~25 VP)
9	Mixed brain abscess	• *Bacteroides vulgatus* (14M180862) • *Streptococcus oralis* (15M150586) • *Streptococcus mitis* (15M150586) • *Streptococcus anginosus* (15M152569)	Five cells of each
10	Viral kidney infection	human cytomegalovirus (HCMV) (08/314)	20 µl 1:100 dilution of control media (~25 VP)
11	Clinical Urine		50 µl clinical sample
12	Clinical Urine		50 µl clinical sample
13	Clinical Urine		50 µl clinical sample
14	Clinical Urine		50 µl clinical sample
15	Clinical CSF		20 µl clinical sample
16	Clinical CSF		20 µl clinical sample
17	Clinical CSF		20 µl clinical sample
18	Clinical CSF		20 µl clinical sample
19	STI swab		40 µl buffer
20	STI swab		40 µl buffer
21	STI swab		40 µl buffer
22	STI swab		40 µl buffer
23	*T. Pallidum* blood		1ml blood
24	*T. Pallidum* blood		1ml blood
25	*T. Pallidum* blood		1ml blood
26	*T. Pallidum* blood		1ml blood
27	*T. Pallidum* ulcer		1ml of buffer containing swab
28	*T. Pallidum* ulcer		1ml of buffer containing swab

Table 2-12 Summary of real and modelled samples processed using developed sample extraction and amplified by ϕ29 MDA

RESULTS

3. Results: Development of Non-PCR Amplification Techniques for Rapid and Sensitive Whole Genome Amplification of Pathogens

Current diagnostic approaches are based on either the ability to culture pathogens, or the detection of a specific amplified gene target. These approaches instil a diagnostic bias toward known causes of infection, potentially missing a large variety of unculturable and/or novel pathogens. Use of an unbiased amplification method to amplify the whole genome of a pathogen without prior knowledge of the pathogens, both bacterial and viral (DNA and RNA), could be a powerful tool for infectious disease diagnosis. The potential of ϕ29 DNA for pathogen amplification was investigated to assess its suitability as a highly sensitive, accurate and rapid method for whole genome sequencing production.

Current reverse transcription and DNA amplification techniques rely heavily on the use of primers to initiate replication. Primers introduce bias into whole genome amplification, which can be a particular problem in extreme GC contents. Reaction conditions need to be optimised to allow primer binding, including reaction temperature and salt concentration of the reaction mix. Additionally, amplification will only occur downstream of where the primer binds, which can be an issue with short or degraded nucleic acids. This is often a particular problem in RNA viruses, with the 3' ends of viral genomes often being difficult to sequence. This is because the primer effect is doubled due to two rounds of reactions with primers. Two alternatives to the use of primers will be investigated. The first will be introduction of nicks into double stranded DNA, and extension from these nicks. The second will involve addition of a tag to the nucleic acid, and extension from this tag.

Several previous studies have successfully copied DNA from nicking points[97,134,78,135], however the product lengths were often limited, with Sequenase 2.0 producing fragments of 250 bp[134], and 3173 Pol and Klenow producing products of a maximum of 400 bp[78,135]. It is hoped that by using ϕ29 which has very strong strand displacement activity and high productivity, larger fragments suitable for WGS will be produced. The nicking enzyme Bst.NBI (New England Bioscience) was used by Moser et al[78] to extend from nicks using PyroPhage 3137 on pUC19 DNA. Additionally, Bst.NBI has a rare binding site and is not affected by DNA methylation. This enzyme will be the starting point for investigating DNA extension from nicks using ϕ29.

Secondly, a looping extension tag will be designed and ligated on to the end of DNA to investigate the potential DNA production from these tags. The tag will contain a recognition site for a nicking enzyme, allowing its removal after extension. An alternative tag will be designed that also has small RNA segments which will allow degradation of the tag. Additionally, the tag will be ligated onto RNA molecules to investigate its potential use for RNA conversion. The advantage of this method is the additional control of where amplification begins, which is external to the region of interest itself.

After establishing the most suitable way to prepare untargeted whole genome amplification this method will be compared to currently used methods for library preparation (culture based). Additionally, the sensitivity and processivity of the enzyme in the context of amplification of bacterial DNA for WGS will investigated. Further analysis will be undertaken to evaluate the impact of genomic GC content on DNA production as well as considering the application of the technique to mixed bacterial samples.

An important application for unbiased whole genome sequencing will be for emerging infectious diseases. Of the many emerging infections detected over the past 20 years most have been viral and one third have been RNA viruses[75]. Detection of viruses in clinical samples can be hindered by low viral loads and difficulty in concentrating viruses, which vary greatly in the ir morphology and particle sizes. Viral particles show a huge variety of morphologies, size and surface type. In order to efficiently isolate, extract and identify viruses, concentration methods need to be employed. Polyethylene glycol (PEG) is a nontoxic water-soluble synthetic polymer which has multiple uses in biological studies. Addition of high weight PEG allows the concentration of viruses using centrifugation without the need for ultracentrifuges, PEG is an inert chemical which doesn't interfere with viral viability. Multiple virus specific PEG concentration methods have been published, **Table 2-9**, this study will aim to identify a suitable PEG concentration method for multiple viruses. To achieve this, a mixture of viruses will be produced, representing multiple viral morphologies and varying genome types.

3.1 Use of ϕ29 MDA for Whole Genome Amplification using Random Primers

3.1.1 Quantification and Visualisation

A single colony of *E. coli* K12 MG1655 from an overnight culture was suspended in 4 µl sterile PBS, extracted and amplified as described in method sections **2.2.1.2** and **2.2.1.3.** Three ϕ29 MDA replicates were performed, which produced DNA concentrations of 1.12x103 ng/µl, 998 ng/ µl and $1.24x10^3$ ng/ µl. The ϕ29 MDA products were visualised on an agarose gel, after three hours a smear of DNA could be seen**, Figure 3-1.** The smallest product was 200bp, and some DNA products greater than 12 kb in length failed to migrate through the gel.

A PCR for ten loci spread across the *E. coli* K12 MG1655 genome was used as a rapid indication of the presence of the whole genome in the ϕ29 MDA product. The PCR described in **section 2.2.1.6** was performed using 1:1000 dilutions of the ϕ29 MDA products (as recommended by the ϕ29 manufacturer). All ten PCR products from the *E. coli* k12 MG1655 genome were detected in all three ϕ29 MDA, with all bands at the expected size. Three negative sample were also amplified using ϕ29 MDA alongside the *E. coli* samples, producing 765 ng/ µl, 693 ng/ µl and 936 ng/ µl of DNA. Despite this, all ten PCR amplifications to detect *E. coli* were negative.

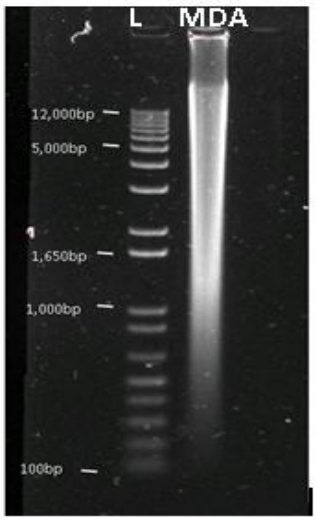


Figure 3-1 Gel Image Showing the Size of the ϕ29 MDA (MDA) Product Produced, Along Side the Ladder (L)

3.2 Assessment of RNA Conversion Enzymes

3.2.1 HIV Concentration and RNA Extraction

HIV was grown in tissue culture as described in **2.1.3.2**, when the supernatant was investigated the number of virions per ml was determined to be $2x10^8$. 100 ml of the supernatant from the HIV tissue culture was concentrated via serial centrifugation using PEG 20,000 as described in **2.2.2.1**. The final supernatant was removed, suspended and alkali extracted as described in **2.2.1.2,** with the exception that the volume was scaled up to a total reaction volume of 50 µl. The RNA and DNA were then quantified and a final concentration of 1.62 ng/µl of RNA and 1.87 ng/µl of DNA was produced. The high DNA concentration, strongly suggests the large presence of host contamination. To resolve this, a DNase and RNase treatment stage was added, described in **2.2.2.3.** Following this treatment, the final RNA concentration was 1.14 ng/µl and the DNA concentration was below detection (<10 pg/µL). Based upon this, RNase and DNase treatment was used for all other extractions. To reduce assay time, shortened PEG incubation times were investigated. After incubation with PEG for one hour at 4°C the amount of RNA extracted was 0.76 ng/µl, and after one hour on ice the RNA extracted was 1.07 ng/µl.

3.2.2 HIV Amplicon Amplification

To investigate the sensitivity of three reverse transcription enzymes the HIV extract was diluted to 1000, 10, 0.1, 0.01, 0.001 and 0.00001 pg/µl. Three replicates of these dilutions were then added to the reaction with SuperScript III (SSIII) and SuperScript IV (SSIV) with random primers (**2.2.2.4** and **2.2.2.5**), followed by amplification with specific HIV primers described in **2.2.2.6.** The dilutions were also added to reactions with PyroPhage 3137 wild-type with the same specific primers (Error! Reference source not found.).

Varying lengths of PCR product were investigated to assess the ability of the enzymes to produce long transcripts. A single forward primer was used, alongside five different reverse primers to produce products of different lengths, (422, 756, 1233, 1778 and 2283 base pairs).

All five products were observed when using SSII and SSIV, at input amounts of 1000 pg, 10 pg and 0.1 pg. SSIII was unable to amplify from less than 0.1 pg, whereas when using SSIV, two products were observed at 0.01 pg (1233 bp and 756 bp), no other amplification products were observed less than 0.1 pg. PyroPhage 3137 enzyme no products were observed at any of the RNA input amounts (**Table 3-1).**

Reverse primer	E03 (2283bp)			E05 (1778bp)			E65 (1233bp)			E125 (756bp)			E145 (422bp)		
Input (pg)	III	IV	PP	III	IV	PP	III	IV	PP	III	IV	PP	III	IV	PP
1000	✓	✓	×	✓	✓	×	✓	✓	×	✓	✓	×	✓	✓	×
10	✓	✓	×	✓	✓	×	✓	✓	×	✓	✓	×	✓	✓	×
0.1	✓	✓	×	✓	✓	×	✓	✓	×	✓	✓	×	✓	✓	×
0.01	×	×	×	×	×	×	×	✓	×	×	✓	×	×	×	×
0.001	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
0.0001	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

Table 3-1 results of HIV specific PCR following reverse transcription with SuperScript III, SuperScript IV and PyroPhage after PEG concentration, SuperScript III=III, SuperScript IV=IV and PyroPhage 3137= PP, ✓ =PCR positive, ✖ = PCR negative

3.2.3 HIV Random Genome Amplification

To check the compatibility of the reverse transcription enzymes with ϕ29, 10 µl of the cDNA produce by the SuperScript III and IV was added to the ϕ29 MDA reaction. The product was then diluted 1:1000 before being amplified by the specific PCR described before. When using SSIII, the specific PCR only detected products where the input RNA was 1000 pg. When using SSIV all products were detected in the 1000 pg input and in the 10 pg input the 756 bp and 422 bp products were detected. No other products were detected. The results are summarised in **Table 3-2**

Reverse primer	E03 (2283bp)		E05 (1778bp)		E65 (1233bp)		E125 (756bp)		E145 (422bp)	
Input (pg)	III	IV	III	IV	III	IV	III	IV	III	IV
1000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	×	×	×	×	×	×	×	✓	×	✓
0.1	×	×	×	×	×	×	×	×	×	×
0.01	×	×	×	×	×	×	×	×	×	×
0.001	×	×	×	×	×	×	×	×	×	×
0.0001	×	×	×	×	×	×	×	×	×	×

Table 3-2 Specific HIV PCR results after PEG concentration, reverse transcription with SSIII (III) or SSIV (IV) and amplification using ϕ29 MDA. ✓ =PCR positive, ✖ = PCR negative

3.2.4 HIV Random Genome Amplification-After Column Extraction

After concentration using PEG, samples were DNase and RNase treated and extracted using a column based method as described in **2.2.2.2** with an elution volume of 50 µl. The average output of the extraction was 1.8 ng/µl RNA. This RNA was diluted to produce samples at 1000, 10, 0.1, 0.01, 0.001 and 0.00001 pg/µl. These were converted to cDNA using SuperScript III or IV before the reaction was put into the ϕ29 MDA reaction, diluted 1:1000 and then amplified using the HIV specific targets.

After the RT reaction with SSIII, the PCR was able to detect products at all sizes in the 1000, 10, 0.1, and 0.01 pg input samples. There were also products at 2283, 1778 and 1233 bp in the 0.001 pg input sample. No products were detected in the 0.0001 pg input.

When RT was per formed using SSIV, the PCR was able to detect products at all sizes in the 1000, 10, 0.1, 0.01 and 0.001 pg input samples. There were also products at 2283, 1778 and 1233 bp in the 0.0001 pg input sample, **Table 3-3**

Reverse primer	E03 (2283bp)		E05 (1778bp)		E65 (1233bp)		E125 (756bp)		E145 (422bp)	
Input (pg)	III	IV	III	IV	III	IV	III	IV	III	IV
1000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
0.1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
0.01	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
0.001	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓
0.0001	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗

Table 3-3 Specific HIV PCR results after PEG concentration, column extraction, RT using SSIII or SSIV and ϕ29 MDA amplification ✓ =PCR positive, ✗ = PCR negative

3.2.5 PyroPhage 3137 after Column Extraction

After extraction using the column method the resulting RNA dilutions were used in reactions with the specific HIV primers using both PyroPhage 3137 wild type and the PyroPhage 3137 exonuclease knock out. No products above 422 base pairs were detected using either enzyme. The PCR amplification of the wild type enzyme RT was only positive in the 1000 pg input. The exonuclease knock-out produced 422 bp results with input levels of 1000, 10, 0.1 and 0.01 pgs. These results are illustrated in **Table 3-4**

Reverse primer	E03 (2283bp)		E05 (1778bp)		E65 (1233bp)		E125 (756bp)		E145 (422bp)	
Input (pg)	WT	Exo⁻	WT	Exo⁻	WT	Exo⁻	WT	Exo⁻	WT	Exo⁻
1000	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
10	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
0.1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
0.01	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
0.001	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
0.0001	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

Table 3-4 Results of specific HIV amplification using PyroPhage WT or exo- after PEG concentration and column extraction ✓ =PCR positive, ✗ = PCR negative

3.3 DNA Amplification for Nicks

3.3.1 Nicking Prediction Results

Four genomes (three bacterial with varying GC content and a viral genome) were investigated for the frequency of the nicking site, both in the forward strand and in the reverse strand. *E. coli* had 1698 nicking sites in the forward and 1661 sites in the reverse strand resulting in an average gap between nicks of 2732 and 2793 bps. *C. difficile* had 938 and 963 recognition sites in the forward and reverse strand leading to nicking every 45735 and 44548 bases. The *A. naeslundii* had 3232 and 3129 predicted nicking sites, with a gap between nicks of 941 and 972 bps. The adenovirus genome had 17 and 21 predicted nicking sites, with the average size of products being 1666 and 2058 bases.

3.3.2 Amplification of Nicked DNA using ϕ29

Amplification of three bacterial genomes was investigated, *E. coli, C. difficile* and *A. naeslundii*. 100 ng of extracted bacterial DNA (**2.2.1.1**) was nicked using Bst.NBI in 10 µl (**2.2.3.1**) reactions, which were incubated for one hour before the enzyme was inactivated. ϕ29 MDA reactions were then set up in total volumes of 50 µl, with 10 ng input. These were incubated for up to six hours, with the DNA being quantified after one, two, four and six hours. For each bacteria four reactions were used, the first with none nicked control DNA with primers (positive control), the second with none nicked DNA and no primers (negative control). The third reaction had nicked DNA with the addition of primers, and the final reaction had nicked DNA without primers. Where samples had primers the DNA was denatured using heat before addition to the reaction mix. A negative sample was also included, consisting of sterile water in place of input DNA incubated with or without primers.

The positive control for *E. coli* produced 19.3, 100, 1000 and >1200 ng/µl after one, two, four and six hours. *C. difficile* and *A. naeslundii* gave similar patterns, with the *C. difficile* sample producing 15.6, 104, 922 and >1200 ng/µl, and the *A. naeslundii*. producing 3.72, 106, 588 and >1200 ng/µl. DNA production the negative control reaction varied between 0.236 and 0.376 ng/µl across all bacteria and negative sample.

The *E. coli* sample which had been nicked but also had primers included produced 5.56, 25.8, 36.8 and 52.4 ng/µl after one, two, four and six hours. Similar results were obtained for the *C. difficile* and *A. naeslundii* (3.87, 31.2, 42.3 and 45.0 ng/µl for *C. difficile* and 1.81, 18.0, 21.9 and 40.9 ng/µl for *A. naeslundii*).

The amount of DNA produced by the nicked DNA without primers ranged from 0.234 and 0.643 ng/µl. The full DNA quantification results can be found in **Table 3-5,** the quantification of *E. coli* DNA across all reaction conditions are shown in **Figure 3-2-A**

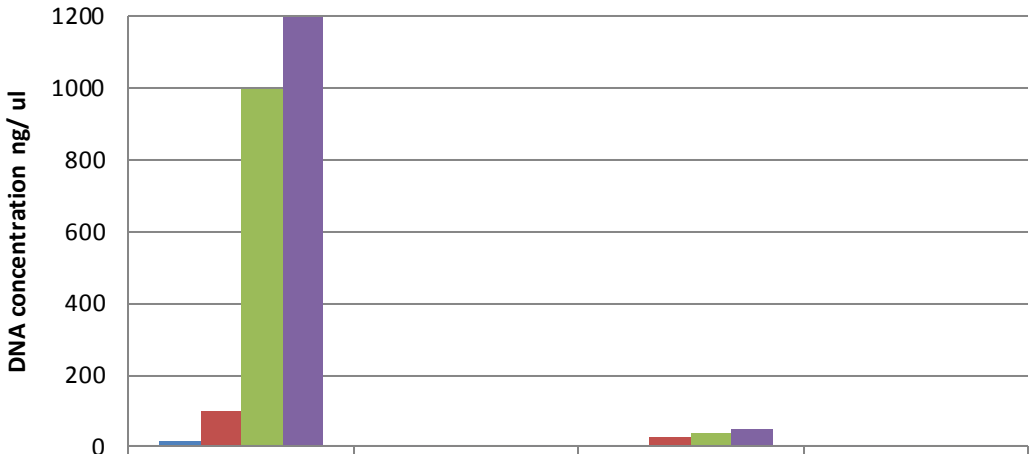
3.3.2.1 Amplification of Nicked DNA after Clean up

E. coli DNA was nicked and the resulting DNA cleaned up using isopropanol precipitation described in **2.2.3.1**. 10 ng was incubated in ϕ29 MDA reactions using the same conditions as previously described in **3.3.2.** The positive control produced 21.8, 97.9, 998 and > 1200 ng/µl after one, two, four and six hours. The negative control gave 0.236, 0.212, 0.258 and 0.314 ng/µl. The nicked DNA with the addition of a primer produced 18.8, 60.7, 154 and 490 ng/µl after one two, four and six hours. The nicked DNA without primers quantified as 0.356, 0.490, 0.398 and 0.411 ng/µl. The results are illustrated in **Figure 3-2-B.**

Time	Condition	*E. coli* (ng/µl)	*C. difficile* (ng/µl)	*A. naeslundii* (ng/µl)	Neg (ng/µl)
I hour	Not nicked and primers	19.3	15.6	3.72	ND
	Not nicked no primer	0.240	0.292	0.234	ND
	Nicked and primer	5.67	3.87	1.81	
	Nicked no primer	0.564	0.346	0.413	
2 hours	Not nicked and primers	100	104	106	36.8
	Not nicked no primer	0.272	0.236	0.336	ND
	Nicked and primer	25.8	31.2	18.0	
	Nicked no primer	0.465	0.643	0.329	
4 hours	Not nicked and primers	1000	922	588	276
	Not nicked no primer	0.368	0.326	0.376	ND
	Nicked and primer	36.8	42.3	21.9	
	Nicked no primer	0.432	0.398	0.392	
6 hours	Not nicked and primers	>1200	>1200	>1200	798
	Not nicked no primer	0.362	0.360	0.334	0.270
	Nicked and primer	52.4	45.0	40.9	
	Nicked no primer	0.389	0.431	0.234	

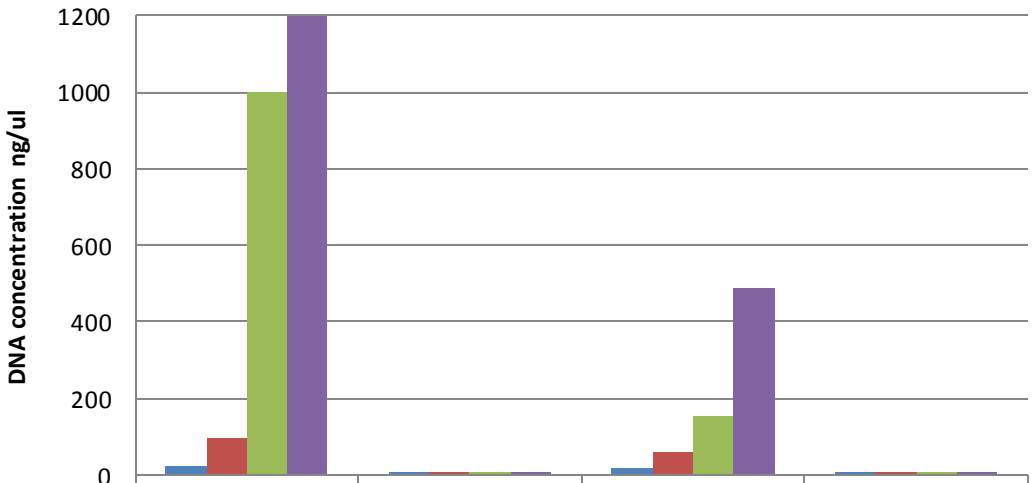
Table 3-5 DNA quantification post ϕ29 MDA amplification of *E. coli, C. difficile* and *A. naeslundii* DNA using either whole or nicked DNA, incubated with or without primers. ND= not detected (<10 pg/µL assay detection limit)

A



	Not nicked and primers	Not nicked no primer	Nicked and primer	Nicked no primer
I hour	19.3	0.24	5.67	0.564
2 hours	100.2	0.272	25.8	0.465
4 hours	1000	0.368	36.8	0.432
6 hours	1200	0.362	52.4	0.389

B



	Not nicked and primers	Not nicked no primer	Nicked and primer	Nicked no primer
I hour	21.8	0.236	18.8	0.356
2 hours	97.2	0.212	60.7	0.49
4 hours	998	0.258	154	0.398
6 hours	1200	0.314	490	0.411

Figure 3-2 DNA quantification post ϕ29 MDA amplification using either whole or nicked *E. coli* DNA, incubated with or without primers (A) before and (B) after isopropanol clean-up

3.3.2.2 Co-nicking and Amplification

T4 Gene 32 Protein binds to single stranded DNA thereby stabilising single stranded DNA regions, (known as a single stranded binding protein, SSBP), nicked DNA was incubated with and without this SSBP before addition to the ϕ29 MDA reaction **(described in 2.2.3.3**).

10 ng of *E. coli* DNA was incubated with Bst.NBI with or without T4 Gene 32 Protein in 10 µl reactions and incubated at 55°C for 30 minutes. The temperature was then lowered to 30°C before the addition of the ϕ29 MDA reaction mixture. For the primer control samples after 30 minutes of incubation at 55°C the sample temperature was raised to 95°C before being placed on ice and the ϕ29 reaction mixture added along with primers. The results are summarised in **Table 3-6**. The control sample produced 3.44, 19.8, 28.8 and 43.8 ng/µl after one, two, four and six hours without the addition of T4 Gene 32 Protein. The sample incubated with T4 Gene 32 Protein produced 1.89, 5.98, 10.8 and 19.7 ng/µl. The test sample with no primer had DNA concentrations of 0.311, 0.43, 0.299 and 0.310 ng/µl after the same time points.

Time	Condition	Bst.NBI (ng/µl)	Bst.NBI and T4 Gene 32 Protein (ng/µl)
I hour	Nicked and primer	3.44	1.89
	Nicked no primer	0.311	0.289
2 hours	Nicked and primer	19.8	5.98
	Nicked no primer	0.341	0.213
4 hours	Nicked and primer	28.8	10.8
	Nicked no primer	0.299	0.313
6 hours	Nicked and primer	43.8	19.7
	Nicked no primer	0.310	0.287

Table 3-6 DNA concentrations after co-nicking and amplification using ϕ29 of *E. coli* DNA with and without SSBP

3.3.3 Enzyme combinations with ϕ29 for Nicked DNA Amplification

Other polymerases with strong stand displacement activity were combined with ϕ29 to investigate the impact on extension from nicks.

3.3.3.1 Klenow Fragment (3′→5′ exo-)

Klenow exo- has previously been shown to be able to replicate from nicks, 10 ng of nicked DNA was added to a reaction with Klenow fragment with and without the addition of ϕ29 and T4 Gene 32 Protein. Reactions were incubated for six hours at 30°C. Controls for these incubation conditions were also performed using primers for Klenow and for Klenow with ϕ29.

The positive control of Klenow fragment using primers produced 460 ng/µl. Klenow fragment incubated with nicked DNA produced 56 ng/µl, and when the SSBP was added 73 ng/µl of DNA was produced. The positive control using Klenow and ϕ29 (with primers) produced 950 ng/µl. The Klenow and ϕ29 incubated with nicked DNA produced 65 ng/µl DNA and with the addition of SSBP 78 ng/µl of DNA was produced. When the fragments produced using Klenow with nicked DNA were investigated using the Bioanalyser (Agilent Technologies, USA), most of the DNA produced was between 150 and 300 base pairs with a peak at 195 bases, with Klenow alone 72% of the DNA was between 150 and 300 base pairs, with 16% being 300 and 100 base pairs. When SSPB was added, 69% of the reads fell between 150 and 300 bases and 24% fell between 300 and 1000 bases

Condition	DNA Produced
Klenow fragment with primers	**460 ng/µl**
Klenow fragment with nicked DNA	56 ng/µl
Klenow fragment with nicked DNA and SSBP	73 ng/µl
Klenow with ϕ29 and primers	**950 ng/µl**
Klenow fragment and ϕ29 with nicked DNA	65 ng/µl
Klenow fragment and ϕ29 with nicked DNA and SSBP	78 ng/µl

Table 3-7 DNA concentration after nicked *E. coli* DNA was incubated with Klenow fragment or Klenow fragment and ϕ29 with or without SSBP extension from nicks

3.3.3.2 Deep Vent$_R$™ DNA polymerase

Deep Vent$_R$ has strong strand displacement activity, but its usual application is in cycling PCRs for samples which have high GC content or stem loop structures. An initial investigation was undertaken to ascertain if the enzyme was capable of isothermal amplification. Initially a thermocycling reaction was used as a control for the reaction buffer, using K12 specific primers and cycling conditions from **2.2.1.6.** After the positive control was cycled 30 times in 50 µl reactions, an average 12.3 ng/µl of DNA was produced, when the products were visualised on an agarose gel all products were present at expected size (around 650 base pairs)

The ability of the enzyme to perform isothermal amplification was then investigation by incubating DNA at the primer annealing temperature for 1 minute followed by 6 hours at 30^{o}C, 40^{o}C, 50^{o}C, 60^{o}C, 70^{o}C and 80^{o}C. The results of the isothermal reactions at 30^{o}C, 40^{o}C, 50^{o}C, 60^{o}C, 70^{o}C and 80^{o}C after six hours were, 0.189, 0.209, 0.178, 0.452, 2.45 and 2.98 ng/µl.

The ability of the enzyme to extend from nicks was investigated by incubating 10 ng of nicked DNA with Vent$_R$ at 70^{o}C or 80^{o}C for 6 hours with and without the SSBP. The amount of DNA produced without SSBP was 0.211 and 0.450 ng/µl at 70^{o}C and 80^{o}C. When incubated with SSBP 0.231 and 0.399 ng/µl was produced at 70^{o}C and 80^{o}C.

The use of Vent$_R$ alongside ϕ29 was investigated by incubating the Vent$_R$ with and without SSBP for 1 hour at 70^{o}C or 80^{o}C, before the reaction was ramped down to 30^{o}C and the ϕ29 was added. The reactions were then incubated at 30^{o}C for six hours. The amount of DNA produced without SSBP was 0.178 and 0.521 ng/µl at 70^{o}C and 80^{o}C. With the addition of SSBP the amount of DNA produced was 0.299 and 0.510 ng/µl for the 70^{o}C and 80^{o}C reactions. The results of all incubation parameters are shown in **Table 3-8.**

INCUBATION	DNA PRODUCED	
Thermal cycling control	12.3 ng/µl	
Isothermal incubations-with primers		
30°C	0.189 ng/µl	
40°C	0.209 ng/µl	
50°C	0.178 ng/µl	
60°C	0.452 ng/µl	
70°C	2.45 ng/µl	
80°C	2.98 ng/µl	
Isothermal incubations-from nicks		
	$Vent_R$	**$Vent_R$ with SSBP**
70°C	0.211 ng/µl	0.231 ng/µl
80°C	0.450 ng/µl	0. 399 ng/µl
70°C then ϕ29 at 30°C	0.178 ng/µl	0.299 ng/µl
80°C then ϕ29 at 30°C	0.521 ng/µl	0.510 ng/µl

Table 3-8 DNA production using $Vent_R$ in the thermocycling control, isothermal incubation with primers and for reactions using nicked DNA with and without the addition of SSBP

3.3.3.3 Bst DNA polymerase, Large Fragment

Bst has strong strand displacement and is a recommended enzyme for isothermal amplification. The enzyme was incubated with 10 ng of nicked *E. coli* DNA with and without SSBP at 30°C, 40°C, 50°C, 60°C and 70°C for six hours **Table 3-9**. After incubation the sample held at 30°C measured 0.178 ng/µl without and 0.190 ng/µl with the addition of SSBP. The sample held at 40°C produced 2.56 ng/µl without SSBP and 2.07 ng/µl with SSBP. When held at 50°C the *Bst* produced 5.89 ng/µl without SSBP and 6.12 ng/µl with SSBP. At 60°C 20.1 ng/µl was produced without SSBP and produced 22.6 ng/µl with SSBP. At 70°C the *Bst* produced 32.0 ng/µl without SSBP and 29.4 ng/µl with SSBP. When the sample held at 70°C was visualised using the Bioanalyser, the product was around 400 base pairs in length.

Combination with ϕ29 was then investigated. A buffer control with ϕ29, *Bst* and primers was incubated at 30°C for six hours, this produced 750 ng/µl. The *Bst* was incubated with 10 ng of nicked *E. coli* DNA for 1 hour, at 50°C, 60°C or 70°C, which produced 2.61, 2.66 and 3.16 ng/µl respectively. Then the reaction was cooled to 30°C and the ϕ29 added, and the sample incubated for six hours. These reactions produced 2.56 ng/µl, 2.78 ng/µland 3.11 ng/µl respectively. The results of the reaction using *Bst* are summarised in **Table 3-9**

Temperature	*Bst* (six hours) ng/µl	*Bst* and SSBP (six hours) ng/µl	*Bst* (one hour) ng/µl	With ϕ29 (six hours) ng/µl
30°C	0.178	0.190		
40°C	2.56	2.07		
50°C	5.89	6.12	2.61	2.56
60°C	20.1	22.6	2.66	2.78
70°C	32.0	29.4	3.16	3.11

Table 3-9 DNA production from nicks in *E. coli* DNA using isothermal incubation with *Bst* with and without SSBP

3.4 DNA Production using DNA Tagging

3.4.1 Tag Design

The stem loop design was adapted from a stable looping primer from Ding et al (2012) [98], where it was used as part of a single molecule sequencing technique. The basic sequence is as follows.

5'-phosphorylated-GCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGC-3'

This was adapted by the addition of a restriction enzyme recognition site for BamHI (NEB) and with addition bases to increase the melt temperature thereby improving the stability of the stem loop.

5' GGATCCGCGCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGCGCGGATCC 3'

This sequence forms the basic structure of the hair pin tag, **Figure 3-3 A**. Additional bases were then added to the 5' end on a trial and error basis for an additional, 5, 10 or 15 bases **Figure 3-3 C**. The formation of the primer was predicted using on online tool called OligoAnalyzer 3.1 from integrated DNA technologies. A second design was also investigated with RNA segments in the hairpin, to aid in the loop degradation, **Figure 3-3 B**.

The final sequences are as follows

5 bases

5' TCGTAGGATCCGCGCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGCGCGGATCC 3'

10 bases

5'AAGTATCGTAGGATCCGCGCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGCGCGGATCC 3'

15 bases

5'AGTTCAAGTATCGTAGGATCCGCGCCTGATCGTCCACTTTTTTTTTAGTGGACGATCAGGCGCGGATCC 3'

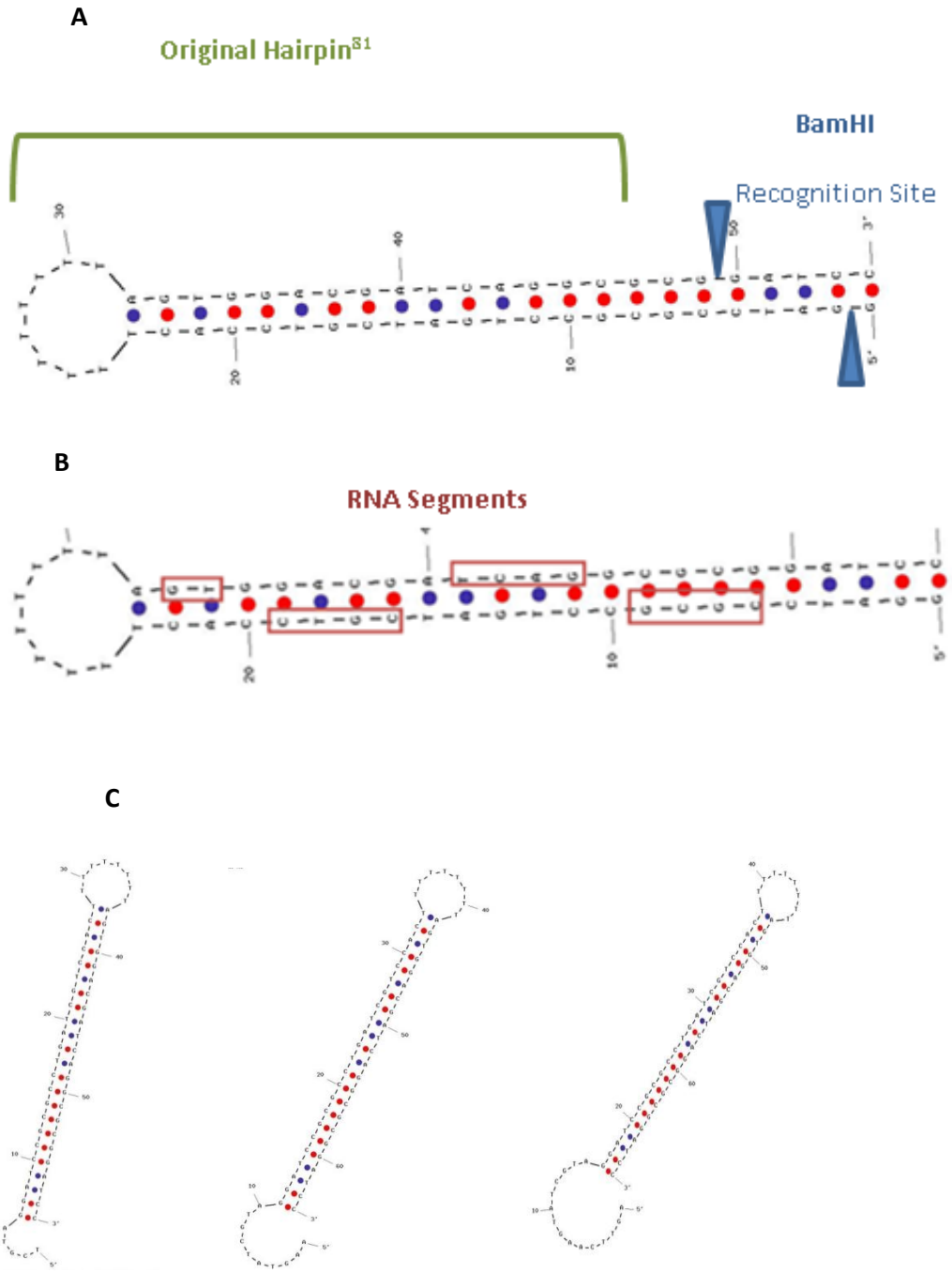


Figure 3-3 visualisation of predicted secondary structure of tag design, (A) stem loop formation and position of restriction enzyme recognition site (B) position of RNA segments (C) additional of 5, 10 and 15 bp overhangs

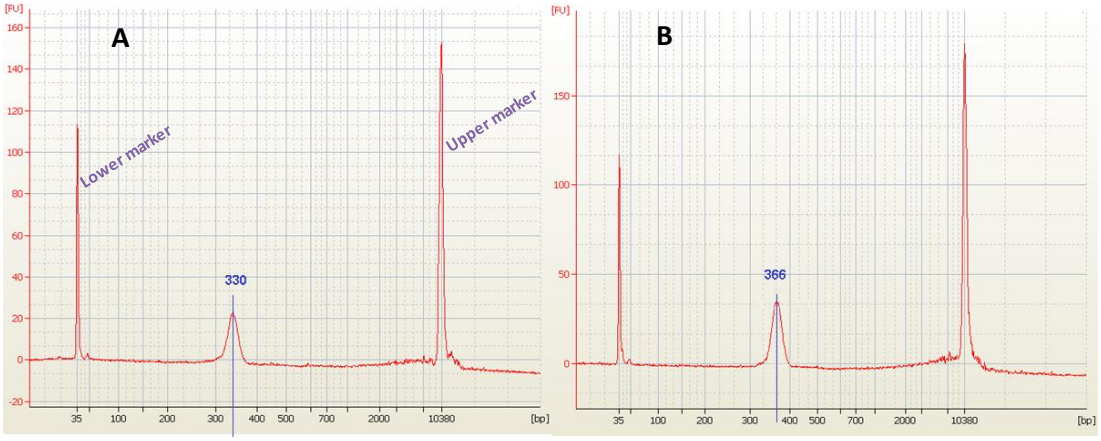
3.4.2 DNA Amplification using Tagging

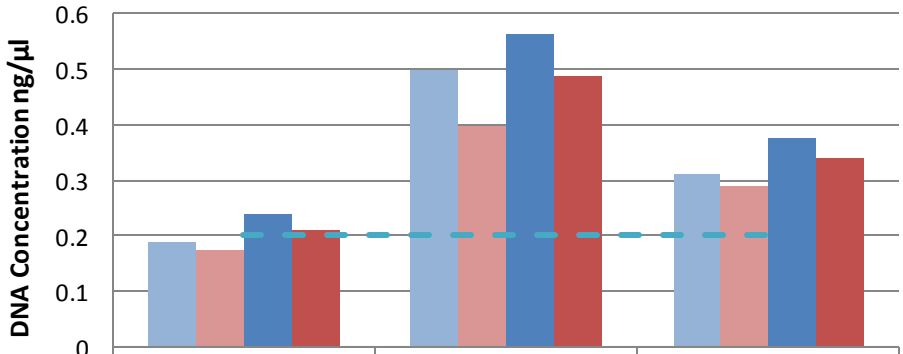
A 330 base pair PCR amplicon was used to assess the ability of ϕ29 to extend from the DNA tag. After attachment of the tag (as described in **2.2.5.3**) the product was visualised using the Bioanalyser. Before attachment the peak was 330 bases, after attachment the peak had shifted to the right, with the peak around 360 bps **Figure 3-4**.

10 ng of tagged DNA was incubated with ϕ29 for four hours in 50 µl reactions **Figure 3-4**. When quantified using the Qubit the five base tag sample had a DNA concentration of 0.189 ng/µl, the ten base tag sample had 0.497 ng/µl and the 15 base tag sample had 0.311 ng/µl. When this experiment was repeated with six hours' incubation the five base tag sample had a DNA concentration of 0.239 ng/µl, the ten base tag sample had 0.565 ng/µl and the 15 base tag sample had 0.375 ng/µl.

3.4.2.1 DNA Amplification using DNA-RNA Hybrid Tag

After the tag containing RNA was attached to the PCR products, 10 ng of tagged DNA was incubated with ϕ29 for four hours in 50 µl reactions. After the incubation the five base tag sample contained 0.176 ng/µl, the ten base tag sample had 0.398 ng/µl and the 15 base sample tag had 0.289 ng/µl. When this experiment was repeated with six hours incubation five base tag sample had a DNA concentration of 0.211 ng/µl, the ten base tag sample had 0.489 ng/µl and the 15 base tag sample had 0.340 ng/µl**. Figure 3-4**





	5 bases	10 bases	15 bases
four hour (nq/ul)	0.189	0.497	0.311
Four hour- RNA (ng/ul)	0.176	0.398	0.289
Six hour (ng/ul)	0.239	0.565	0.375
Six hour-RNA (ng/ul)	0.211	0.489	0.34
input (ng/ul)	0.2	0.2	0.2

Figure 3-4 Bioanalyser trace showing the amplicon size before (A) and after (B) DNA tag addition shown alongside the upper and lower markers. (C) DNA produced after addition of DNA or DNA-RNA hybrid tag to 330 bp amplicon, with 5, 10 or 15 base overhand, incubated with ϕ29 for four or six hours.

3.4.2.2 *Amplification of Longer DNA Fragments*

To investigate if low DNA yields were due to short DNA fragments longer DNA fragments were tagged and amplified. DNA Extracts of *E. coli* K12 overnight cultures were nebulised for 30 seconds, and the DNA purified as described in **2.3.2.5,** when visualised on the Bioanalyser the product sizes peaked at 2000 bases **Figure 3-5 (A)**. After attachment of DNA tags the fragments were incubated for four hours in a 50 µl ϕ29 MDA reaction. The five base tag reaction had a DNA concentration of 0.210 ng/µl, the 10 base tag reaction produced 0.988 ng/µl, and the 15 base tag sample produced 0.417 ng/µl. When the product of the ten base tag amplification was visualised, the peak product size had increased to 5000 bases, with some DNA being longer than the marker at 10800 bps **Figure 3-5 (B)**.

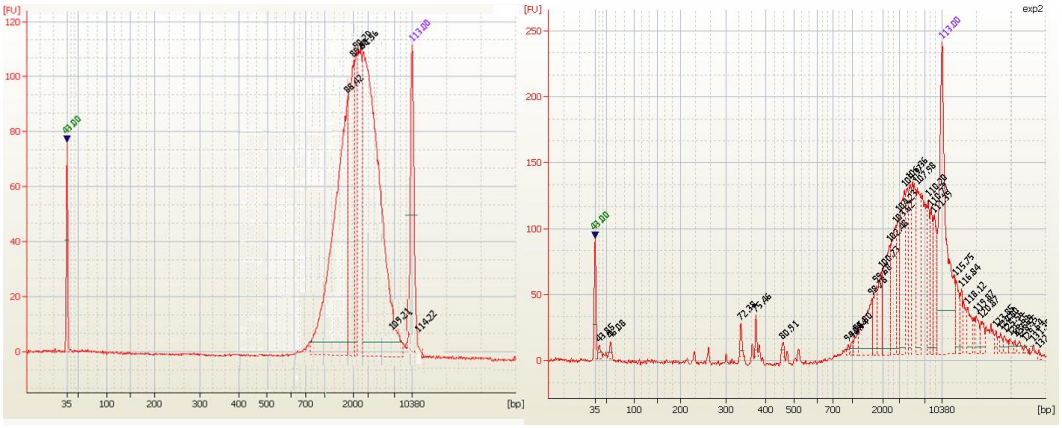


Figure 3-5 Bioanalyser trace of fragmented DNA size (A) before and (B) after attachment and amplification from tag

3.4.3 *RNA Conversion and Amplification using Tagging*

HIV tissue culture supernatant was extracted and treated with RNase and DNAse before the attachment of the tags. The RNA with tags attached was then added to a reverse transcription reaction using SuperScript IV. After this two HIV specific PCR were performed using primers E70 and E145 (product size 422 bases), and E80 and E145 (product size 123 bases). A control RT reaction and PCR was performed with the same extracts using random primers in the RT reaction. The positive control produced products at the expected size and the PCR for all tag sizes produced no product.

3.5 Development of ϕ29 MDA for Whole Genome Sequencing

One *E. coli* ϕ29 MDA product from **0** was randomly selected and taken forward for sequencing. The ϕ29 MDA product was used for library preparation described in **2.3.3.** 500 ng of the product was nebulised for 60 seconds as recommended by the Roche Junior guidelines. After library preparation the quality of the product was assessed, the resulting fragment size was between 600 bp and 2500 bp. During the library preparation the library was quantified by detection of florescent markers present on the indexes attached to the reads. The florescence of the prepared library was 30.73, which equated to 4.32 $x10^8$ molecules/ µl, which was within the acceptable parameters. This was diluted down to $1x10^7$ before continuing with emulsion PCR (EmPCR) setup in **2.3.3.2**

When the EmPCR was prepared the aim was to achieve six library copies per bead. After the EmPCR the beads were harvested, those with attached library were selectively isolated. When these beads were visualised very few beads were present, much fewer than the required 500,000 beads needed for sequencing, and so sequencing was not undertaken.

In an attempt to improve the EmPCR yield the same library was again prepared for EmPCR, but aiming for two copies per bead. Again this failed to produce sufficient beads.

To investigate the possibility of loss of library, the ten loci PCR described in **2.2.1.6** was performed on the product post nebulisation and the $1x10^7$ library. All products were detected at the correct size in both of these samples confirming the presence of expected DNA.

3.5.1 Library Preparation Optimisation

After the library failed to produce sufficient beads for sequencing, the use of S1 nuclease was investigated. S1 removes branches in DNA, lowering secondary DNA structure. This was used in combination with either prolonged physical fragmentation or enzymatic fragmentation.

The ϕ29 MDA product was digested using S1 nuclease as described in **2.3.2.2**. The products produced by the S1 nuclease, were either nebulised (60, 120 or 240 seconds) or fragmented using the Fragmentase enzyme (10, 15 and 20 minute incubations) and a combination of S1 and Fragmentase enzymes was also trialled **Table 3-10**.

After nebulisation for 60 seconds the size range of products was average between 600 bp and 2500 bp, with the peak being around 1800 bp. The 120 second nebulisation product size ranged from 300 bp to 2000 bp, with the peak at 600 bp. The 240 second nebulisation had a size range of 0 bp to 650 bp, with a peak of 350 bp.

When the Fragmentase enzyme was used after S1 treatment, the majority of the products after the 10 and 15 minute incubations were larger than detectable on the Bioanalyser, over 10380 bp. The 20-minute incubation showed a size ranging from 2000 bp to above the upper marker, with a peak at 10,000 bps

When S1 nuclease and Fragmentase were combined into a single reaction the product was too large to be detected on the Bioanalyser and when ran on an agarose gel showed no size difference from the unfragmented ϕ29 MDA product. The results for the different fragmentation conditions are shown in **Table 3-10**

Condition	Smallest fragment size (bp)	Peak fragment (bp)	Largest fragment size (bp)
	Nebulisation		
60 seconds	600	1,800	2,500
120 seconds	300	600	2,000
240 seconds	0	350	650
	Fragmentase		
10 minutes	>10,380	>10,380	>10,380
15 minutes	>10,380	>10,380	>10,380
20 minutes	2,000	10,000	>10,380
	S1 and Fragmentase single reaction		
20 minutes	>10,380	>10,380	>10,380

Table 3-10 DNA fragment sizes produced by different fragmentation methods applied to ϕ29 MDA products of *E. coli* genome.

3.5.2 Sequencing after Fragmentation Optimisation

The ϕ29 MDA product was treated with S1 and then nebulised for two minutes before undergoing library preparation and sequencing. When investigated on the Bioanalyser the size range of the library was 400 bp to 1000 bp with a peak at 600 bp.

When preparing the EmPCR both two and four copies per bead were trialled. When four copies per bead was used considerably more than two million beads were produced, which is too high for sequencing. When two copies per bead were used just fewer than two million beads

were produced, allowing sequencing to be undertaken. Sequencing on the Roche 454 Junior was then performed as described in **2.3.3.3**

The sequencing run was successful, allowing basic analysis of the run to be undertaken as described in **2.3.5.** The number of reads generated in the sequencing run was 238, 827. The number reads that passed all filters was 180,144, totalling 79,350,696 bases, the mean length of reads was 439.16 bp. When assembled using Newbler against the reference (2.3.5.1), 98.44% of the genome was covered in 242 contigs, using 91% of the filter passed reads.

Based on the results obtained, treatment with S1 nuclease and physical fragmentation using 120 second nebulisation were used for further runs.

3.6 Comparison of ϕ29 MDA to Current Methods of Bacterial Sequencing

3.6.1 E. coli K12 Non-Amplification Control

Three replicates of the non-amplification control for *E. coli* K12 were performed as described in **2.3.1** briefly DNA was extracted from overnight cultures and 500 ng was used for library preparation. The DNA was fragmented using 1-minute nebulisation.

After completion of the sequencing runs the average number of reads produced was 194,646 before filter and 158,132 after filtering **Figure 3-6 (A), Table 3-11.** The mean read lengths produced were 468.52, 464.7 and 467.31 bps. After reference assembly reference coverage for each replicate was 98.24%, 98.14% and 94.26%. The total number of contigs in the reference assemblies was 310, 391 and 1281. *De novo* assemblies covered 94.30%, 81.19% and 79.22% of the genomes in 1636, 1502 and 1954 contigs **(Figure 3-6 (B))** the number of *de novo* assembly errors for each replicate was 25, 27 and 29. The largest contig in each *de novo* assembly was 38913, 44816 and 13251, with N50s of 7452, 3601 and 2464.

	Control 1	Control 2	Control 3
No: Raw reads	211,297	201,309	171,333
No: Reads passing filter	177,442	168,106	128,849
% reads passing filter	83.98	83.51	75.20
Mean read length(bps)	468.52	464.7	467.31
Reference assembly			
% ref coverage	98.24	98.14	94.26
% reads mapped	93.76	94.87	93.31
No: Contigs	310	391	1281
***De novo* assembly**			
% ref coverage	94.30	81.19	79.22%
No: contigs	1,636	1,502	1,954
largest contig	38,913	44,816	13,251
N50	7,452	3,601	2,464
#misassemblies	25	27	29

Table 3-11 Sequencing results of the *E. coli* non-amplification controls including reference and *de novo* assembly results

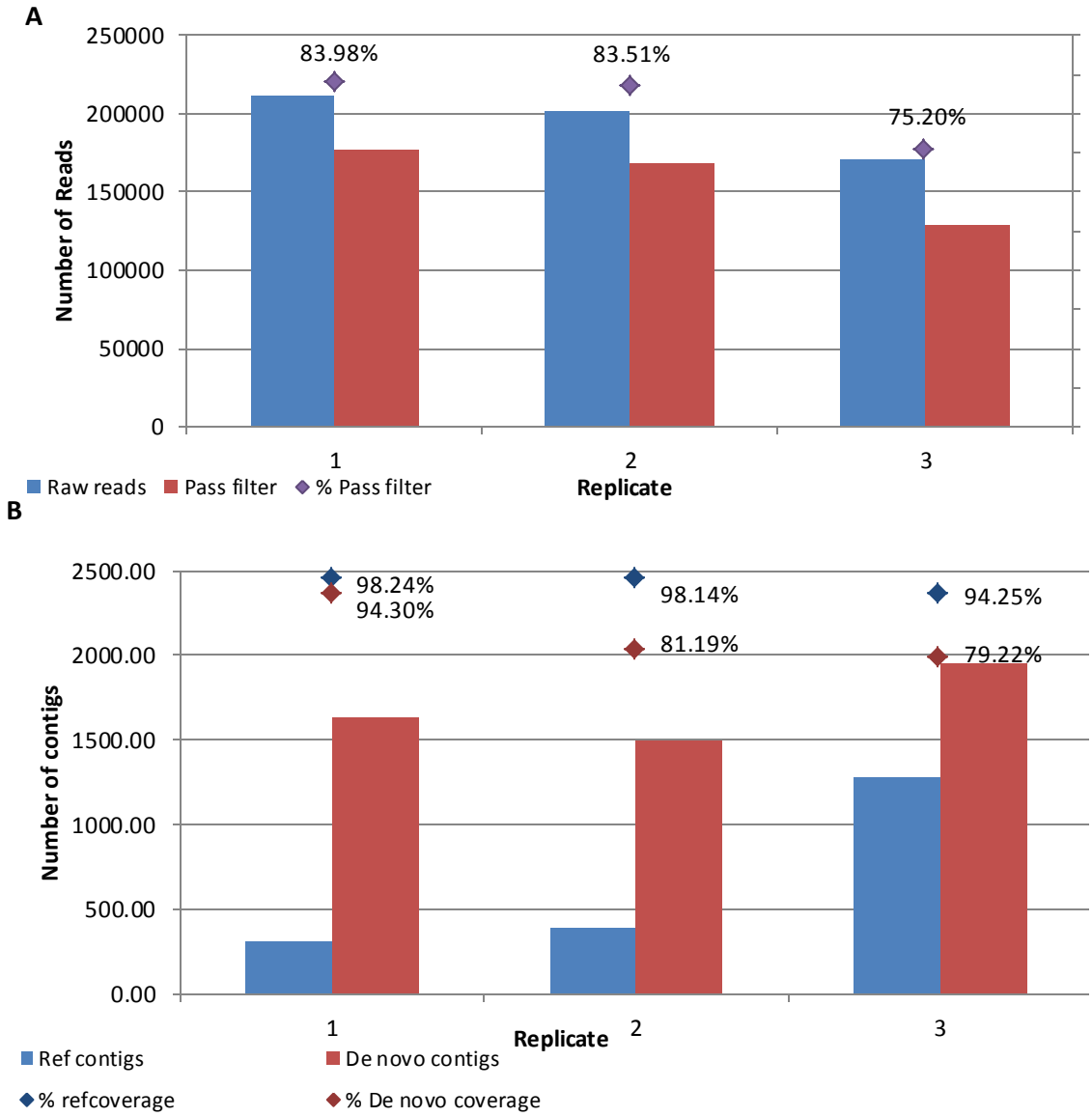


Figure 3-6 Analysis of sequencing results from non-amplification control *E. coli* K12, (A) Raw and Filtered Reads with % reads passing filter and (B) the number of contigs in the reference and *de novo* assemblies along with the genome coverage of each assembly.

3.6.2 Reproducibility of ϕ29MDA Reactions of E. coli K12

In addition to the previously sequenced ϕ29 MDA *E. coli* K12, two further replicates were amplified and sequenced. The amount of DNA produced in the ϕ29 MDA reactions for the two additional biological replicates performed for *E. coli* was 998 ng/µl and $1.24x10^3$ ng/µl producing 235,215 and 278,531 sequencing reads **(Figure 3-7(A)).** The mean read lengths produced were 447.62 and 384.45 bases, for each run. After reference assembly the genome coverage for each run was 97.18% and 98%. The total number of contigs in the reference assemblies was 394 for both runs. *De novo* assemblies were performed for the three replicates covered which 94.12%, 90.4% and 91.6% of the reference genome across 1568, 1345 and 1032 contigs **(Figure 3-7 (B))**. The number of misassemblies in the *de novo* assemblies was 21, 22 and 23 for each run. The largest *de novo* contigs were 32639, 38791 and 50561 bps, with the N50s for the runs being 455, 5520 and 7897. A summary of the results obtained for the three replicates (including the original sequencing run) can be found in **Table 3-12.**

	Replicate 1	**Replicate 2**	**Replicate 3**
DNA produced (ng/µl)	$1.12x10^3$	998	$1.24x10^3$
No: Raw reads	238, 827	235,215	278,531
Reads passing filter	180,144	178,378	222,066
% reads passing filter	75.43	76.03	79.72
Mean read length (bps)	439.16	447.62	384.45
	Reference assembly		
% ref coverage	98.44	97.18	98
% reads mapped	91%	90.16	94.06
No: Contigs	242	394	394
	***De novo* assembly**		
% ref coverage	94.12	90.4	91.6
No: contigs	1568	1345	1032
largest contig (bps)	32,639	38,791	50,561
N50 (bps)	455	5520	7897
#misassemblies	21	22	23

Table 3-12 Sequencing results of the three *E. coli* K12 ϕ29 MDA reaction replicates, including reference and *de novo* assembly results

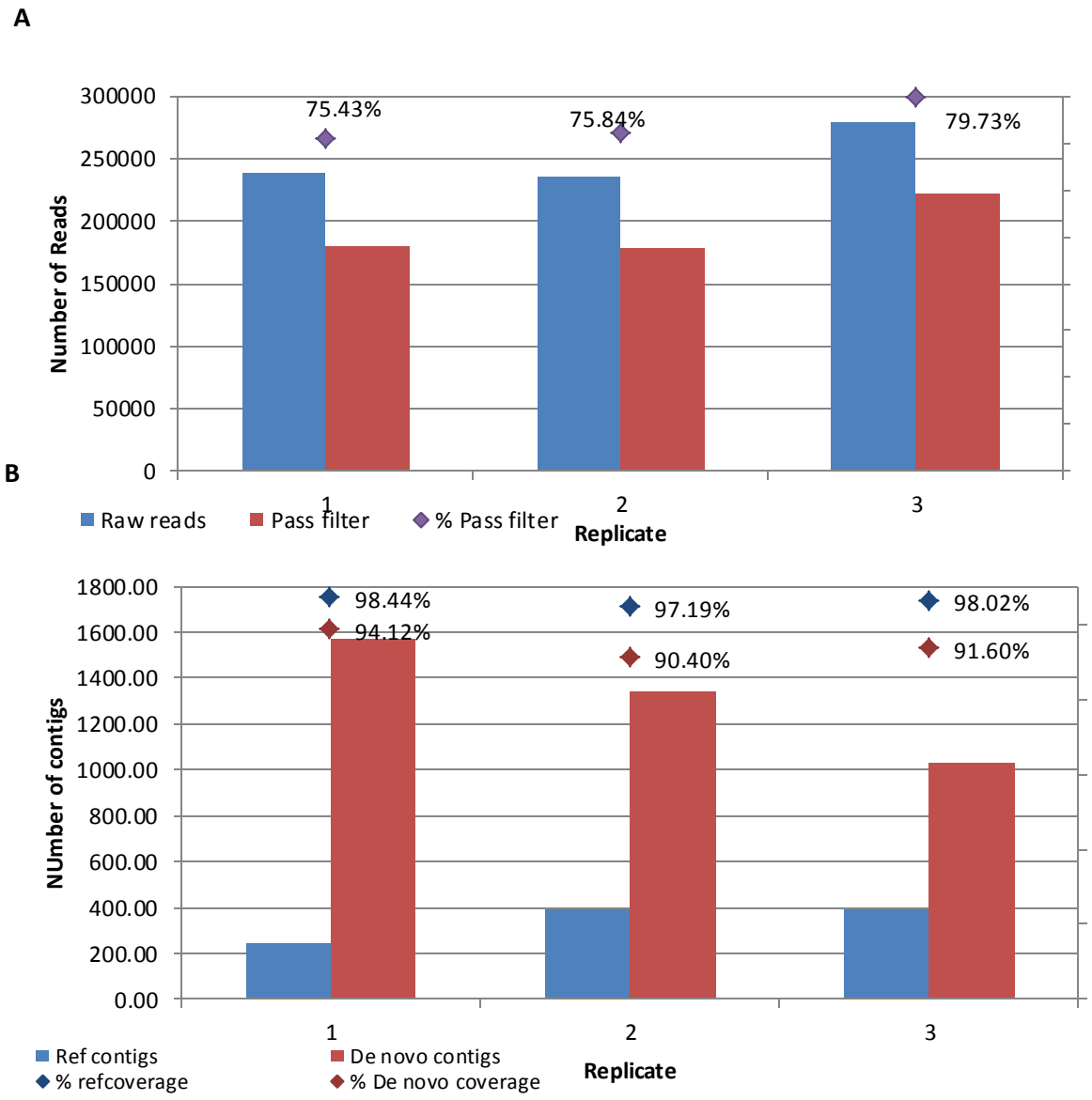


Figure 3-7 Sequencing results of the *E. coli* K12 ϕ29 MDA reactions, (A) Raw and Filtered Reads including minimum reads required to pass run(dotted line) and (B) the number of contigs in the reference and *de novo* assemblies including the proportion of the genome covered.

3.6.3 Comparison of ϕ29 MDA and Culture Control in Sequencing of E. coli k12

The average results of the non-amplification control and ϕ29 MDA are summarised in **Table 3-14,** along with the F test results (equal or unequal variance) and the T test results, including interpretation with a 0.05 value.

The average number of raw reads produced by ϕ29 MDA amplified libraries was 250,858 compared to 194,646 produced using the non-amplification method. The difference between the two approaches being statistically significant when tested with a two-sample T-Test, assuming equal variances. However the number of pass filter reads was not significantly different between the two methods the average percentage of reads passing filters was 77% for ϕ29 MDA produced libraries and 80.9% for non-amplification libraries, again this wasn't shown to be a significant difference (**Figure 3-8(A)**). There was no significant difference in length of sequencing reads produced, with the average read length being 423 bases for ϕ29 MDA produced libraries and 466 bases for non-amplification libraries. The proportion of the reference that was covered was 97.88% and 96.87%, for ϕ29 MDA and control libraries respectively. The number of reads used to construct the reference assembly was slightly higher in the non-amplification libraries (94.18%) compared to ϕ29 MDA libraries (91.75%) however this was not found to be significant. The average number of contigs produced was lower in the ϕ29 MDA library (343.33) than the non-amplification library (660.67), but this was not found to be significant (**Figure 3-8(B)**).

The largest contig produced in the reference assembly was significantly larger (176638.33 bases) in the ϕ29 MDA library compared to the non-amplification library (81579.00 bases), but the average size on the contigs was not significantly different. The number and size of contigs generated by the *de novo* assembly was not statistically significantly different between the ϕ29 MDA and non-amplification libraries (**Figure 3-8(C)**). The number of misassemblies in the ϕ29 MDA libraries was significantly less (22) than the non-amplification libraries (27).

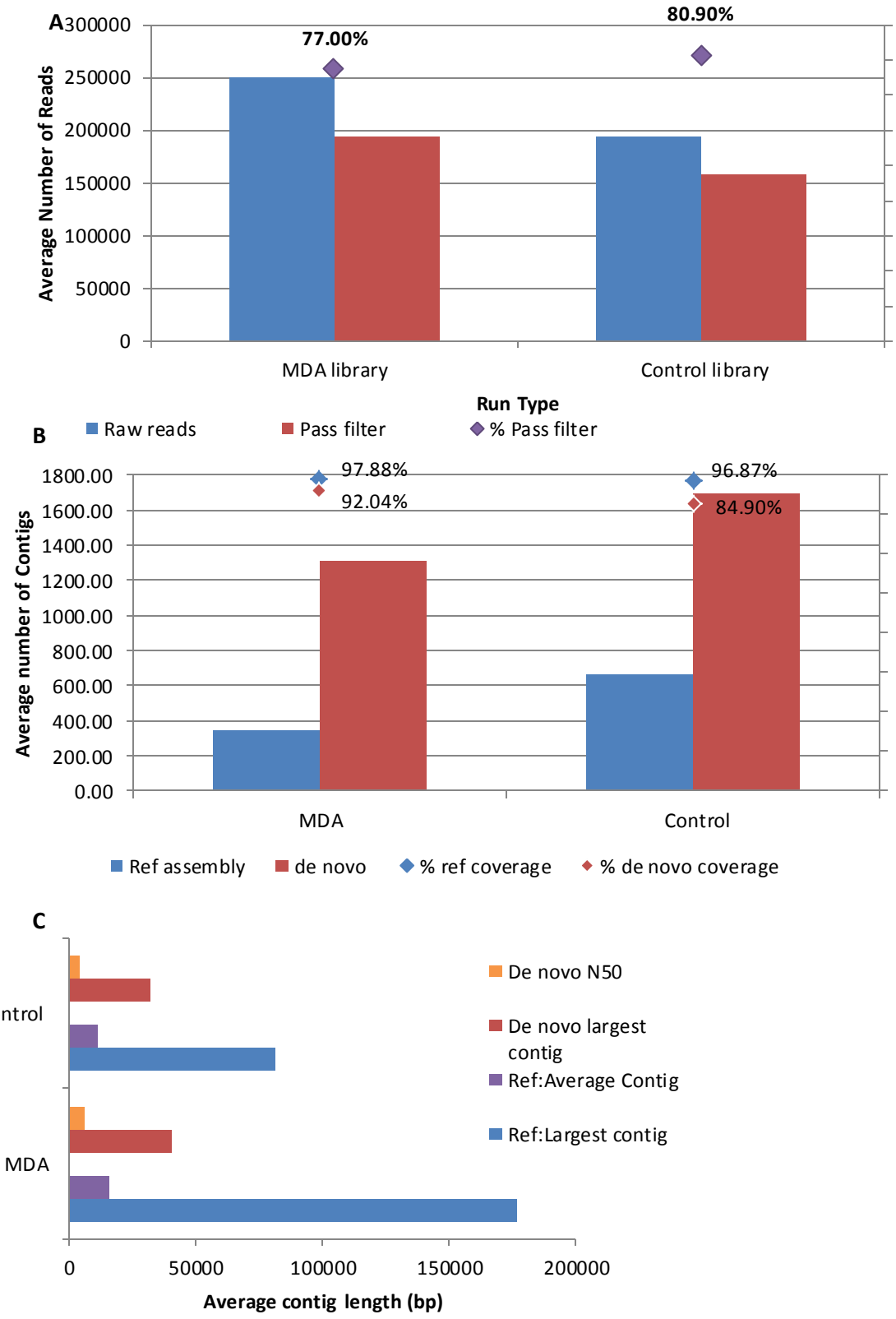


Figure 3-8 Comparison of ϕ29 MDA and non-amplification preparation of sequencing of *E. Coli* K12, (A) number of raw and filter passed reads, including % reads passing filter (B) number of contigs in the reference and *de novo* assemblies and the proportion of genome covered and (C) contig sizes produced by reference and *de novo* assemblies

3.6.4 Extra Chromosomal Reads of E. coli

After reference assembly of the reads against the plasmid F reference sequence, **Table 3-14**, the coverage for ϕ29 MDA amplified libraries was 99.90%, 99.89% and 99.93%, using 8.59%, 8.54% and 5.67% of the passed filter reads. Non-amplification libraries covered 98.49%, 97.97% and 94.79%, using 2.20%, 3.46% and 3.42% of reads that passed filter **Figure 3-9-A**. This was found to be significantly less than the ϕ29 MDA produced libraries.

The number of contigs in the ϕ29 MDA reference assembly against the F plasmid was 3, 3 and 2 with the largest contigs being 46869, 46903 and 74509 bases long, with average contig lengths of 33021, 33017 and 49543 bases. The non-amplification libraries were assembled into 12, 8 and 24 contigs with the largest being 31216, 40380 and 20366 bases and the average contig size being 8138, 13807 and 4233 bases, which was significantly shorter that the ϕ29 MDA produced libraries.

For the three sets of data produced by ϕ29 MDA amplification of *E. coli* K12, the number of reads mapping to lambda phage were 143, 128 and 100, which covered 67.85%, 52.93% and 47.94% of the lambda genome. The non-amplification libraries covered 39.69%, 40.88% and 21.17% of the lambda genome **Figure 3-9-B**.

Additionally, 374, 394 and 5 reads from the ϕ29 MDA produced library were associated with human genome elements, and 162, 158 and 81 were unidentified. Of the remaining reads 0, 1638 and 414 were not mapped above the Genus level, but were mainly within the Primate Order, or the Rosales (order of flowing plants) order. For the non-amplification libraries, 175, 226 and 298 reads mapped to human, and 2802, 572 and 512 reads were unmapped. The number of reads classed as other was, 4140, 1968 and 2610, which fell in the Orders of either Primate or Rosales.

	All Reads	K12 Genome	F Plasmid	Lambda	Human	Not Mapped	Other
				MDA			
1	180144	91.04%	8.59%	0.08%	0.21%	0.09%	0.00%
2	178378	90.16%	8.54%	0.07%	0.22%	0.09%	0.92%
3	222066	94.06%	5.67%	0.05%	0.00%	0.04%	0.19%
				Non-amplification			
1	177442	93.76%	2.20%	0.03%	0.10%	1.58%	2.33%
2	168106	94.87%	3.46%	0.02%	0.13%	0.34%	1.17%
3	128849	93.91%	3.42%	0.02%	0.23%	0.40%	2.03%

Table 3-13 proportional read identification in *E. coli* sequences for ϕ29 MDA and non-amplification replicates (reads identified as Rosales (order of flowing plants) have been grouped with reads classed as 'other')

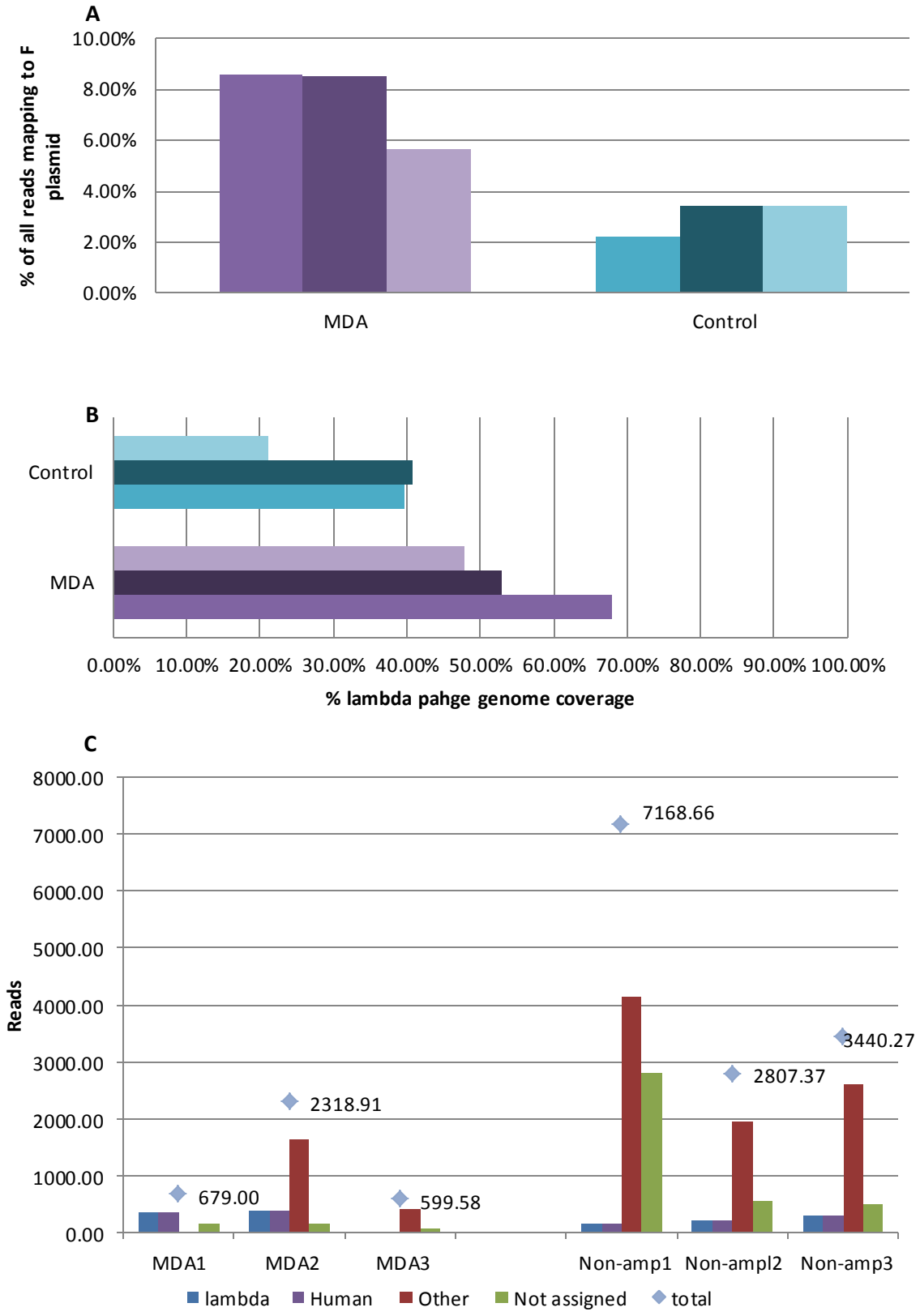


Figure 3-9 comparison of *E. coli* K12 none chromosomal elements of ϕ29 MDA and non-amplification prepared sequencing. (A) Percentage of reads mapping to plasmid F and (B) percentage coverage of Lambda phage genome and (C) identification of the remaining reads in the three replicates of the ϕ29 MDA and control libraries

	ϕ29 MDA amplified	Non-amplification control	F test	T test value	Significant (0.05)
No: Raw reads	250858	194646	equal	0.0375	**Yes**
No: Pass filter reads	193529	158132	unequal	0.1614	No
% Pass filter	77.00%	80.90%	equal	0.2851	No
Mean read length (bp)	423.74	466.84	equal	0.0954	No
Median read length (bp)	440	496	equal	0.0545	No
Modal read length (bp)	458	503	equal	0.1290	No
Reference mapping					
% ref coverage	97.87%	96.88%	unequal	0.5273	No
% reads mapped	91.75%	94.18%	equal	0.1200	No
No: Contigs	343	661	unequal	0.3709	No
Largest contig (bp)	176638	81579	equal	0.0105	**Yes**
Average contig (bp)	15804.67	11441.67	equal	0.3888	No
***De novo* assembly**					
% ref coverage	92.04%	84.90%	equal	0.2158	No
No: Contigs	1315	1697	equal	0.1360	No
Largest contig (bp)	40664	32327	equal	0.4916	No
N50	5990	4506	equal	0.4574	No
#Misassemblies	22.00	27.00	equal	0.0179	**Yes**
Plasmid F Reference Mapping					
% ref coverage	99.91%	97.09%	unequal	0.1347	No
% reads mapped	7.60%	3.03%	equal	0.0121	**Yes**
No: contigs	2.67	14.67	unequal	0.0337	**Yes**
largest contig (bp)	56093.67	30654.00	equal	0.0794	No
average contig (bp)	38527.00	8726.00	equal	0.0085	**Yes**

Table 3-14 Summary of Average Sequencing Results for *E. coli* K12 ϕ29 MDA and culture control including results of reference and *de novo* assemblies and plasmid reference assembly, showing results F test and T test for statistical significance

3.7 Determining Sensitivity and Processivity of ϕ29 MDA in the Context of Bacterial Genomes

After establishing that ϕ29 MDA was suitable for whole genome sequencing, the method was optimised to provide reliable and rapid results. For the optimisation study the *E. coli* K12 previously sequenced was used. The aim was to have method that provided results are as quick as possible, whilst not compromising genome coverage or data quality. Initially the conditions were 50 µl reactions incubated for 16 hours, as instructed by the manufacturer. Smaller volumes were investigated (25 µl and 12.5 µl), as well as reduced incubation times, scaling down from 8 hours to 1 hour incubations. The sensitivity of the assay was also tested by reducing the input to a single bacterial cell.

3.7.1 Reaction Volume Reduction

Reaction volumes were scaled down from 50 µl to 25 µl and 12.5 µl, and two biological replicates of each reaction volume were sequenced, results are shown **Table 3-15.**

There was a reduction in the amount of DNA produced by the 25 µl ϕ29 MDA reaction (average 825.5 ng/ µl) compared to the 50 µl reaction (average 1240 ng/ µl), however this was not significant. The 12.5 µl reaction produced significantly less DNA (average 633.5 ng/ µl) than the 50 µl reaction (**Figure 3-10 (A)**).

There was no significant difference in the number of raw reads produced when sequencing the DNA produced, (average values of 250,858 in 50 µl, 219,657 in 25 µl and 277,902.5 in 12.5 µl ϕ29 MDA reactions.). The proportion passing filter was very similar in the 50 µl and 25 µl reactions (77% and 70.4%), however significantly less passed filter for the 12.5 µl reaction (47.36%) shown in **Figure 3-10 (B)**

After reference assembly, the amount of genome coverage was significantly lower in both the 25 µl (61.97%) and 12.5 µl (55.84%) reactions compared to the 50 µl (97.88%) reaction **Figure 3-10 (C)**. The percentage of reads used to create the reference assembly from the 50 µl (91.75%) and 25µl (85.45%) reactions were similar; however, was significantly lower in the 12.5 µl (42.62%) reaction.

The number of contigs produced in the reference assembly was significantly higher in both the 25 µl (1149) and 12.5 µl (1332.5) reactions, compared to the 50 µl reaction (343). The largest contigs produced were significantly shorter (51195, 25 µl and 51204, 12.5 µl compared to 176638, 50 µl). The average contig length were also shorter (3589, 25 µl and 3409, 12.5 µl compared to

15805, 50 µl). There was also significantly less of the genome covered after *de novo* assembly (45.4%, 25 µl and 41.22%, 12.5 µl compared to 92.04%, 50 µl). However, there was little difference in the contig number (1446 and 1843 compared to 1315) there was significantly more misassemblies in both the 25 µl (33) and 12 .5 µl, (44) reactions compared to the 50 µl, reaction (12). Results are summarised in **Table 3-15**

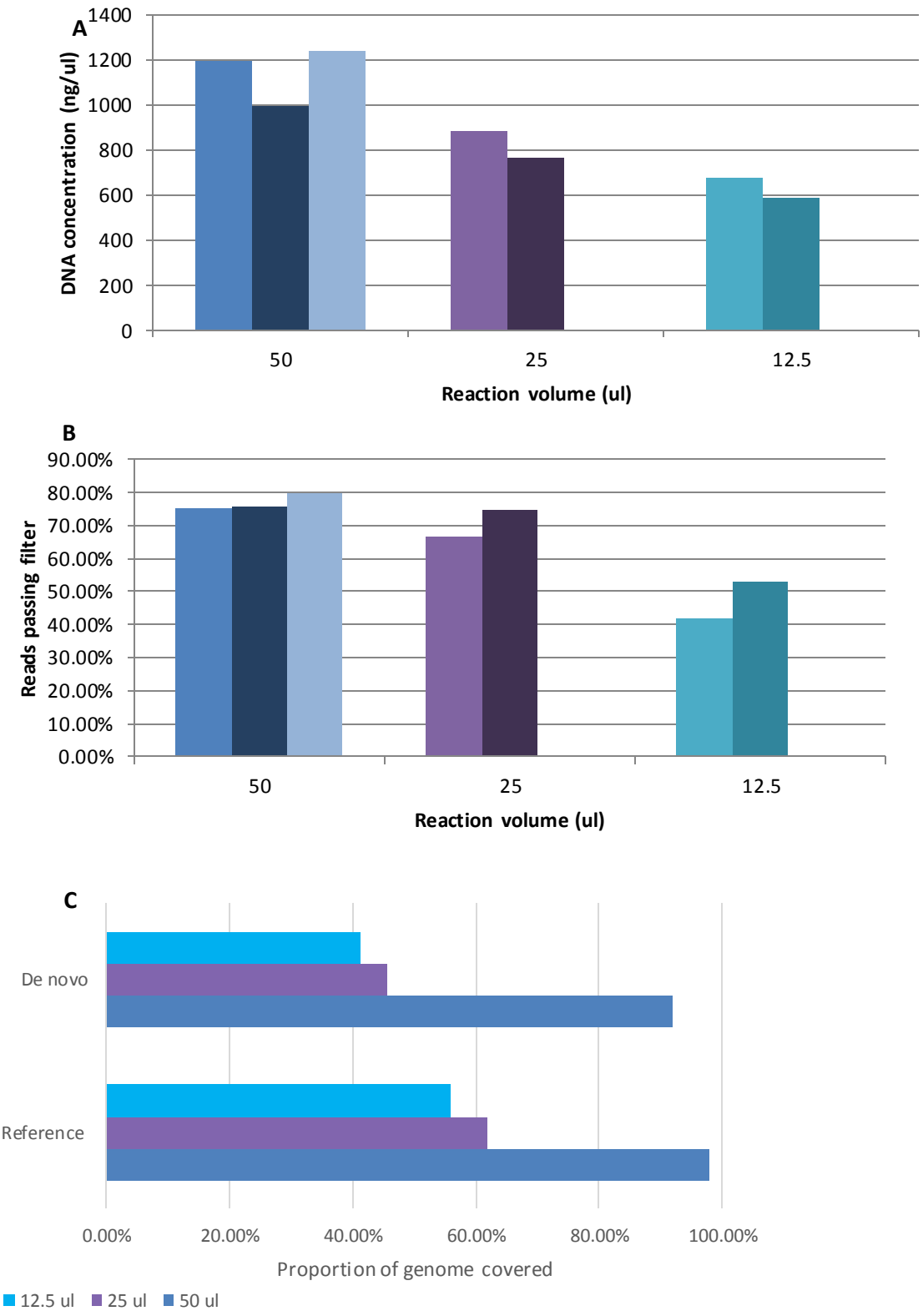


Figure 3-10 Results of reducing the ϕ29 MDA reaction volume of *E. coli* K12, (A) DNA concentration produced, (B) % of reads passing filter and (C) % reference coverage using reference and *de novo* methods

	50 µl *vs.* 25 µl			Average results			50 µl *vs.* 12.5 µl		
	F test	T test value	Sig (0.05)	50 µl	25 µl	12.5 µl	F test	T test value	Sig (0.05)
DNA produced (ng/µl)	E	0.0588	No	1240	825.5	633.5	E	0.0153	Yes
No: Raw reads	UE	0.3948	No	250858	219657	277903	E	0.2294	No
No: Pass filter reads	UE	0.3521	No	193529	155928	131728	E	0.0685	No
% Pass filter reads	E	0.1590	No	77.00%	70.5%	47.36%	E	0.0073	Yes
Mean read length (bp)	E	0.4002	No	424	449.07	393.95	E	0.4332	No
Reference assembly									
% ref coverage	E	0.0000	Yes	97.88%	61.97%	55.84%	E	0.0009	Yes
% reads mapped	E	0.2172	No	91.75%	85.45%	42.62%	E	0.0001	Yes
No: Contigs	E	0.0060	Yes	343.33	1149	1333	E	0.0111	Yes
Largest contig (bp)	UE	0.1212	No	176638	51195	51204	E	0.0264	Yes
Average contig (bp)	E	0.0232	Yes	15805	3589	3410	E	0.0222	Yes
***De novo* assembly**									
% ref coverage	E	0.0001	Yes	92.04%	45.50%	41.22%	E	0.0002	Yes
No: Contigs	E	0.7566	No	1315	1446	1843	E	0.0818	No
Largest contig (bp)	E	0.0108	Yes	40664	90015	40265	UE	0.9928	No
N50	E	0.3250	No	5990	4423	3611.5	UE	0.1886	No
#Misassemblies	E	0.0071	Yes	22	33	44	E	0.0001	Yes

Table 3-15 Summary of comparison between 50µl, 25µl, and 12.5µl, ϕ29 MDA reactions volumes to amplify *E. coli* K12 Including statistical significance testing of DNA production, sequencing output and assembly results.

F test E=Equal UE = unequal

3.7.2 Determining the Sensitivity of ϕ29 MDA

Single cells of *E. coli* K12 were prepared and sequenced in triplicate as described in **2.3.4.1** and sequenced **Table 3-16**.

There was no significant different between the amount of DNA produced from a colony and single cell, the average being 1146 ng/µl for colony input and 1268 ng/µl for the single cell input. There was also no significant difference between the number of reads produced when sequenced, 250,858 and 217,141 for the colony and single cell input respectively. The percentage of reads that passed filter also showed no significant difference being 77% (colony) and 78.4% (single cell). The length of reads showed little variation with the average read lengths being 423.75 (colony) and 428.77 (single cell) bases. The proportion of the genome covered with the reference assembly method showed little variation between the two sets of data (97.87% and 97.03%). Slightly fewer reads were used to construct the assembly using the single cell library (86.49% compared to 91.75%) but this was not found to be a significant difference. Further analysis of the non-chromosomally mapped reads revealed most of the colony input reads mapped to the F plasmid, whereas the single cell input showed a greater variety of mapping hits and a greater number that could not be identified. As shown below in **Figure 3-11 .**The number of contigs produced in the chromosome reference assembly was on average 343 and 316, with the average contig length being 15805 and 18818 bp. The *de novo* assembly covered 92.05% and 92.00% of the genome, in an average of 1315 and 1297 contigs. The N50s of the two sets of data were 5990 and 5168, showing no significant difference. The average number of misassemblies was 22 and 21. The results are summarised in **Table 3-17**

	Replicate 1	**Replicate 2**	**Replicate 3**
DNA produced (ng/µl)	1214 ng/µl	1450 ng/µl	1140 ng/µl
No: Raw reads	238836	224702	187886
No: Reads passing filter	189211	171250	150240
% reads passing filter	79.22%	76 .21%	79.96%
	Reference assembly		
% ref coverage	95.50%	96.99%	98.59%
% reads mapped	90.84%	82.74%	85.89%
No: Contigs	313	254	380
	***De novo* assembly**		
% ref coverage	91.79%,	89.42%	94.78%
No: Contigs	1036	1522	132
N50	4276	5031	6196
#misassemblies	22	20	21

Table 3-16 summary table of single cell *E. coli* K12 ϕ29 MDA reactions, performed in triplicate. Including sequencing output and assembly results.

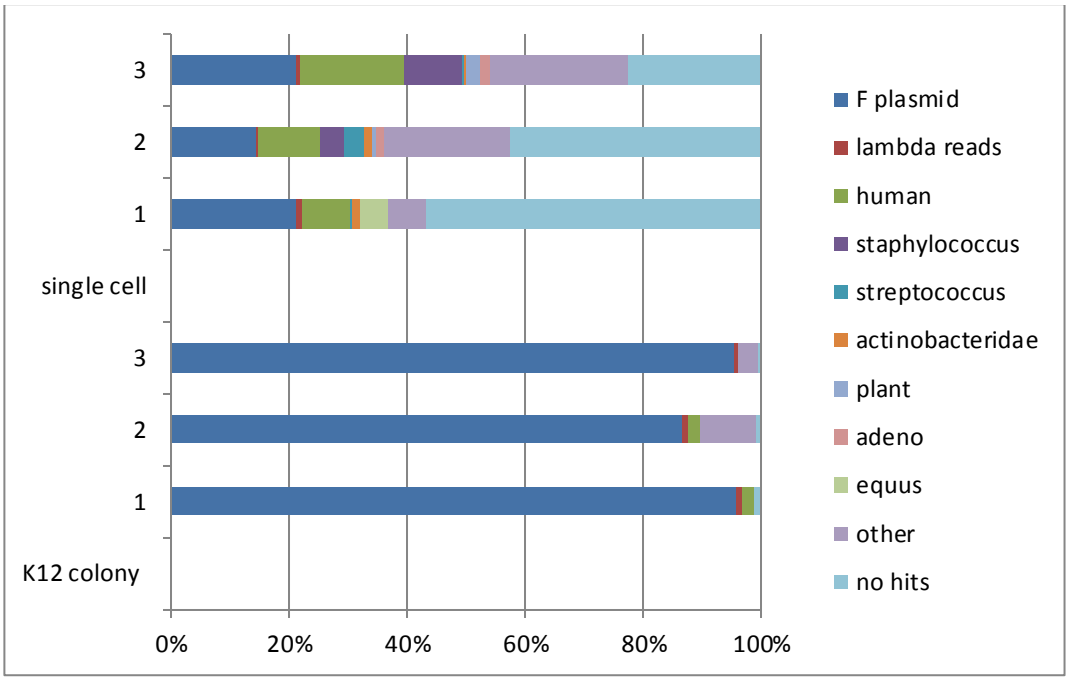


Figure 3-11 proportional read identity of reads not mapping to the *E. coli* K12 chromosome after ϕ29 MDA of single colony or single cells

	Colony ϕ29 MDA amplified	Single cellϕ29 MDA amplified	F test	T test value	Sig (0.05)
DNA produced (ng/µl)	1146	1268	Equal	0.3660	No
No: Raw reads	250858	217141	Equal	0.1765	No
No: Pass filter reads	193529	170233	Equal	0.2694	No
% Pass filter	77.00%	78.4%	Equal	0.4574	No
Mean read length(bp)	423.74	428.77	Equal	0.8383	No
Median read length (bp)	440	476	Equal	0.2019	No
Modal read length(bp)	458	494	Equal	0.2459	No
		Reference mapping			
% ref coverage	97.87%	97.03%	Equal	0.4223	No
% reads mapped	91.75%	86.49%	Equal	0.1166	No
No: Contigs	343	316	Equal	0.6803	No
Largest contig (bp)	176638	178657	Equal	0.9030	No
Average contig (bp)	15804.67	18817.67	Equal	0.2723	No
		***De novo* assembly**			
% ref coverage	92.04%	91.99%	Equal	0.9825	No
No: Contigs	1315	1297	Equal	0.9347	No
Largest contig (bp)	40664	35872	Equal	0.5632	No
N50	5990	5167	Equal	0.5103	No
#Misassemblies	22	21	unequal	0.2879	No

Table 3-17 Comparison of average sequencing results produced by colony and single cell input of ϕ29 MDA reactions of l *E. coli* K12, including results of statistical test (T test)

3.7.3 Incubation Time Reduction

Single cell input ϕ29 MDA reactions were incubated at 30°C for 8, 4, 2 or 1 hour, and if sufficient DNA was produced, were sequenced **Table 3-18.** Overall the concentration of DNA produced from the ϕ29 MDA reactions decreased with the decreasing incubation time. The average amount of DNA produced being 1268, 1211.3, 1105.3, 778 and 184 ng/µl, after 16, 8, 4, 2 and 1 hour incubations **Figure 3-12 (A).** The percentage of the reference genome covered was consistently high for the 16, 8, 4 and 2 hours' incubation, always being above 97%. However, this dropped to 78.58% in the 1-hour incubation sample. A similar pattern was found with the *de novo* assemblies. The number of contigs constructed remained similar across the first four incubation times (316, 285, 301 and 262) but rose to 2889.5 in the 1-hour incubation sample. From the *de novo* assemblies a similar pattern was observed. With the number of contigs decreasing from 1297, 1295, 859 and 786 from 16, 8, 4 and 2 hours, but then increasing to 1581 for the 1 hour incubation **Figure 3-12 (B))**

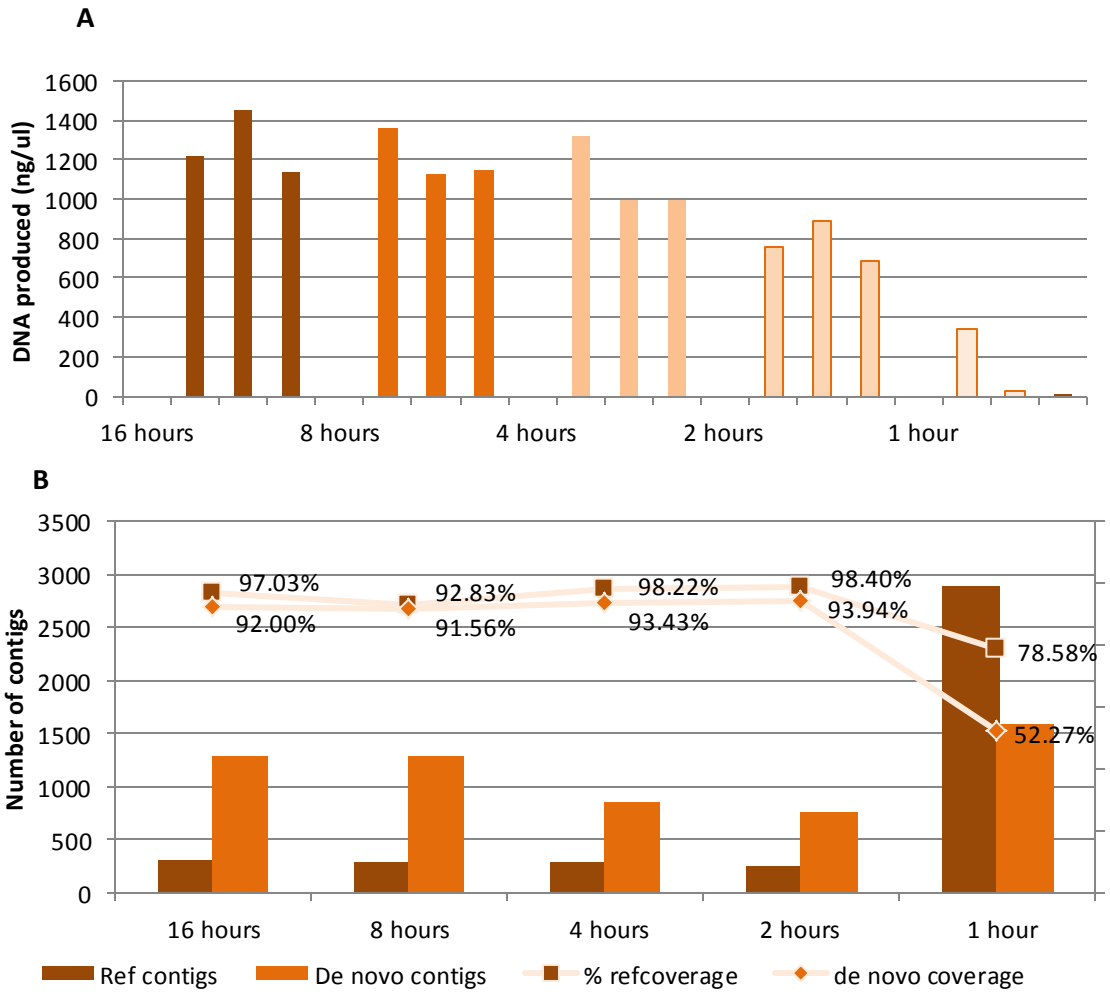


Figure 3-12 Impact of reducing the ϕ29 MDA incubation times for amplifying single *E. coli* cells , (A), DNA concentration production (B)number of contigs in the reference and *de novo* assembly along with proportion of the genome covered

	8 hours	4 hours	2 hours	1 hour
DNA Produced (ng/µl)	1. 1,364, ng/ µl 2. 1,125 ng/ µl 3. 1,145 ng/ µl	1. 1,364 ng/ µl 2. 1,125 ng/ µl 3. 1,145 ng/ µl	1. 756 ng/ µl 2. 889 ng/ µl 3. 689 ng/ µl	1. 345 ng/ µl 2. 23 ng/ µl 3. 2 ng/ µl
		Reference assembly		
% ref coverage	1. 96.63%, 2. 83.99% 3. 97.88%.	1. 97.59 % 2. 98.56% 3. 98.52%.	1. 98.41 % 2. 98.46% 3. 98.22%	1. 71.69% 2. 80.67%
No: Contigs	1. 268 2. 297 3. 290	1. 352 2. 320 3. 230	1. 254 2. 243 3. 190	1. 3567 2. 2212
Average contig size (bp)	1. 12283 2. 16684 3. 17414	1. 31051 2. 39410 3. 36847	1. 31445 2. 32395 3. 3754	1. 9036 2. 9445
Largest contig (bp)	1. 190728 2. 132053 3. 141624	1. 189335 2. 289014 3. 286883	1. 519189 2. 680322 3. 408750	1. 19102 2. 71919
		***De novo* assembly**		
% ref coverage	1. 91.29% 2. 90.80% 3. 92.6%	1. 92.43% 2. 91.89% 3. 95.95%	1. 93.4% 2. 93.53% 3. 94.81%	1. 58.52 % 2. 46.02%
No: Contigs	1. 1343 2. 1214 3. 1329	1. 948 2. 648 3. 982	1. 708 2. 809 3. 787	1. 1622 2. 1540
Largest contig (bp)	1. 35379 2. 24145 3. 42959	1. 31453 2. 30713 3. 92122	1. 61435 2. 89754 3. 75063	1. 13416 2. 8565
N50	1. 5178 2. 5066 3. 5012	1. 5685 2. 5690 3. 7790	1. 5671 2. 7416 3. 6347	1. 1001 2. 2157
#Misassemblies	1. 20 2. 22 3. 23	1. 22 2. 22 3. 21	1. 21 2. 23 3. 24	1. 33 2. 38

Table 3-18 Results of time reduction of ϕ29 MDA reactions to eight, four, two and one hour incubations. Results include the sequencing outputs and assembly results.

3.8 Assessing ϕ29 MDA for Whole Genome Amplification of Varying GC Contents

To check the applicability of amplification using ϕ29 MDA to multiple pathogens two bacteria with extreme GC contents were investigated. The results were then compared to *E. coli* K12 (50.75% GC) previously sequenced. The two bacteria were *Clostridium difficile* (29% GC) and *Actinomyces naeslundii* (69%GC). Additionally, these bacteria were selected as they had Gram positive cell walls, to ensure the extraction technique used was suitable.

3.8.1 Clostridium difficile

C. difficile 630 is a well characterised pathogenic strain of *C. difficile*, which is a Gram positive, spore forming anaerobic bacteria, with a GC content of 29.06% and a total genome size of 4290252 bases. It also known to carry a plasmid, pCD630 which is also well characterised with a GC content of 27.9 and total length 7881 bases. The bacterium has been well studied and its genome was fully characterised in 2006[136], the isolate has multiple drug resistances and a highly mobile genome.

Two replicates were performed for non-amplification controls of *C. difficile* and three single cell ϕ29 MDA reactions were sequenced. The concentration of the three ϕ29 MDA reactions was $1.25x10^3$ ng/µl, $1.07x10^3$ ng/µl and $1.02x10^3$ ng/µl.

When the reads from the ϕ29 MDA and culture control were mapped against the reference for plasmid_pCD630, no reads mapped to the reference in any file. A specific PCR targeting the plasmid_pCD630 was negative for the ϕ29 MDA and control. A fresh culture from an earlier sub cultured set of storages beads was also negative. When whole cell lysis, shown in **Figure 3-13** was visualised on the tape station (Agilent) using genomic tapes, no products were found below 10K base pairs (plasmid size 7881 bases).

The average number of reads produced by the ϕ29 MDA amplified methods was 225,451 and for the non-amplification control was 227,593, with the number of reads passing filter being 174,907 and 167,093, which was 77.6% and 72.75% of the total reads. None of these variables were statistically significant from each other. There was no significant difference found between the mean read length (460 vs. 444 bases) for the ϕ29 MDA and none-amplification. The proportion of the reference genome covered after reference assembly was higher in the ϕ29 MDA libraries 95.73% compared to 88.66% for the control libraries, however this was not found to be a significant difference. This is shown below in **Figure 3-14 (A)**

The average number of contigs in the reference assemblies was 430 and 850 for the ϕ29 MDA and control libraries respectively. The average contig size was larger in the ϕ29MDA library being 11357 compared to 5655 bp in the control library; once again this was not found to be a significantly different **Figure 3-14 (B)**. The size of the largest contig was shown to be significantly bigger in the ϕ29 MDA library, 146,811 bases compared to 57823 bases in the non-amplification libraries. The average amount of the genome covered by the *de novo* assembly was 92.88% in the ϕ29MDA libraries and 83.16% in the non-amplification libraries, with the N50 values being very similar in the two assemblies (5194 and 4688 bp respectively).

However, the number of misassemblies was significantly higher in the control libraries with 34 misassemblies compared to 24 in the ϕ29 MDA libraries. **Table 3-20** summarises the average results for sequencing parameters, including the T test values and interpretations.

The reads not mapping to the *C. difficile* chromosome were investigated. In the non-amplification sample, the number of reads which mapped to the human genome was 51 and 212 for the two replicates. The number of reads mapping to the order Clostridiales was 196 and 116, with 32 and 11 mapping to Clostridium. The numbers of unassigned reads were 173 and 438. In the ϕ29 MDA libraries, the number of reads mapping to the human genome was 12, 150 and 2098 for the three replicates. The number of reads mapping to the Order Clostridiales was 2150, 87 and 0, with a further 292, 25 and 0 reads mapping to the Genus Clostridium. The number of unassigned reads was 117, 1022 and 4925. Other reads accounted for 0, 503 and 3 reads, with the majority being in the Primate Order. The third replicate of the ϕ29 MDA library had 4072 reads that were identified as Enterobacteriaceae.

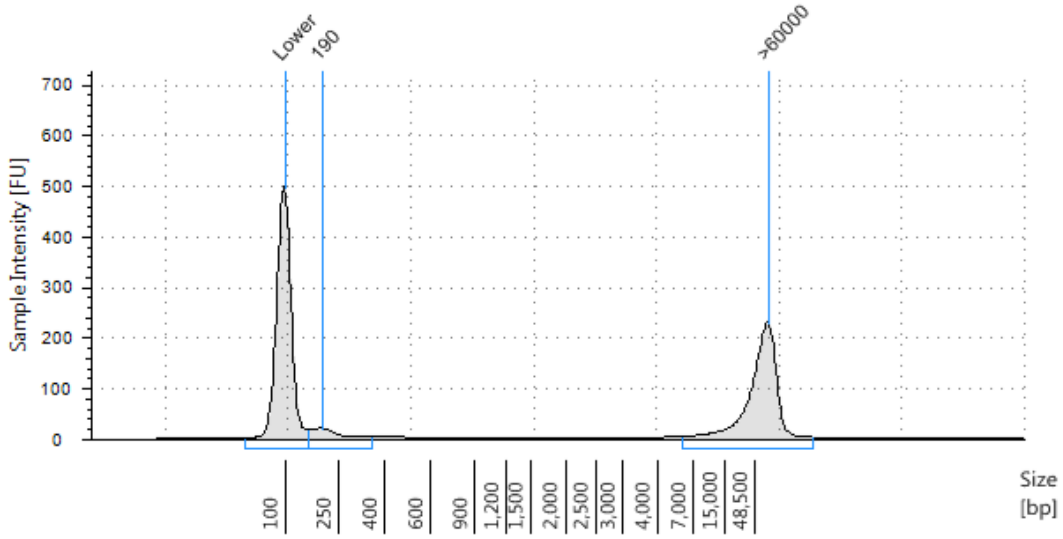


Figure 3-13 Tape station Output for Whole Cell Lysis of *C. difficile* showing the absence of plasmid DNA

	Culture Control		ϕ29 MDA single cell reactions		
	Control 1	**Control 2**	**Reaction 1**	**Reaction 2**	**Reaction 3**
DNA produced (ng/µl)			$1.25x10^3$ ng/µl	$1.07x10^3$ ng/µl	$1.02x10^3$ ng/µl
No: Raw reads	205220	249966	216897	238758	220698
No: Reads passing filter	135341	198844	179767	171830	173123
% reads passing filter	65.95%	79.55%	82.88%	71.97%	78.44%
Mean read length (bp)	452.04	436.33	452.67	470.01	457.47
Reference assembly					
% ref coverage	93.69%	83.62%	96.25%	94.23%	96.62%
% reads mapped	99.67%	99.29%	98.57%	98.96%	93.59%
No: Contigs	548	1152	479.00	523.00	311.00
Average contig size (bp)	7902	3408	9440	10738	13893
***De novo* assembly**					
% ref coverage	89.59%	76.73%	90.63%	98.38%	89.62%
No: contigs	654	1145	1397	1200	1108
N50	4941	4435	4380	3839	7364
#misassemblies	32	35	26	27	22

Table 3-19 Raw sequencing data from *C. difficile* non-amplification and single cell ϕ29 MDA prepared sequencing including raw output and results of reference and *de novo* assembly

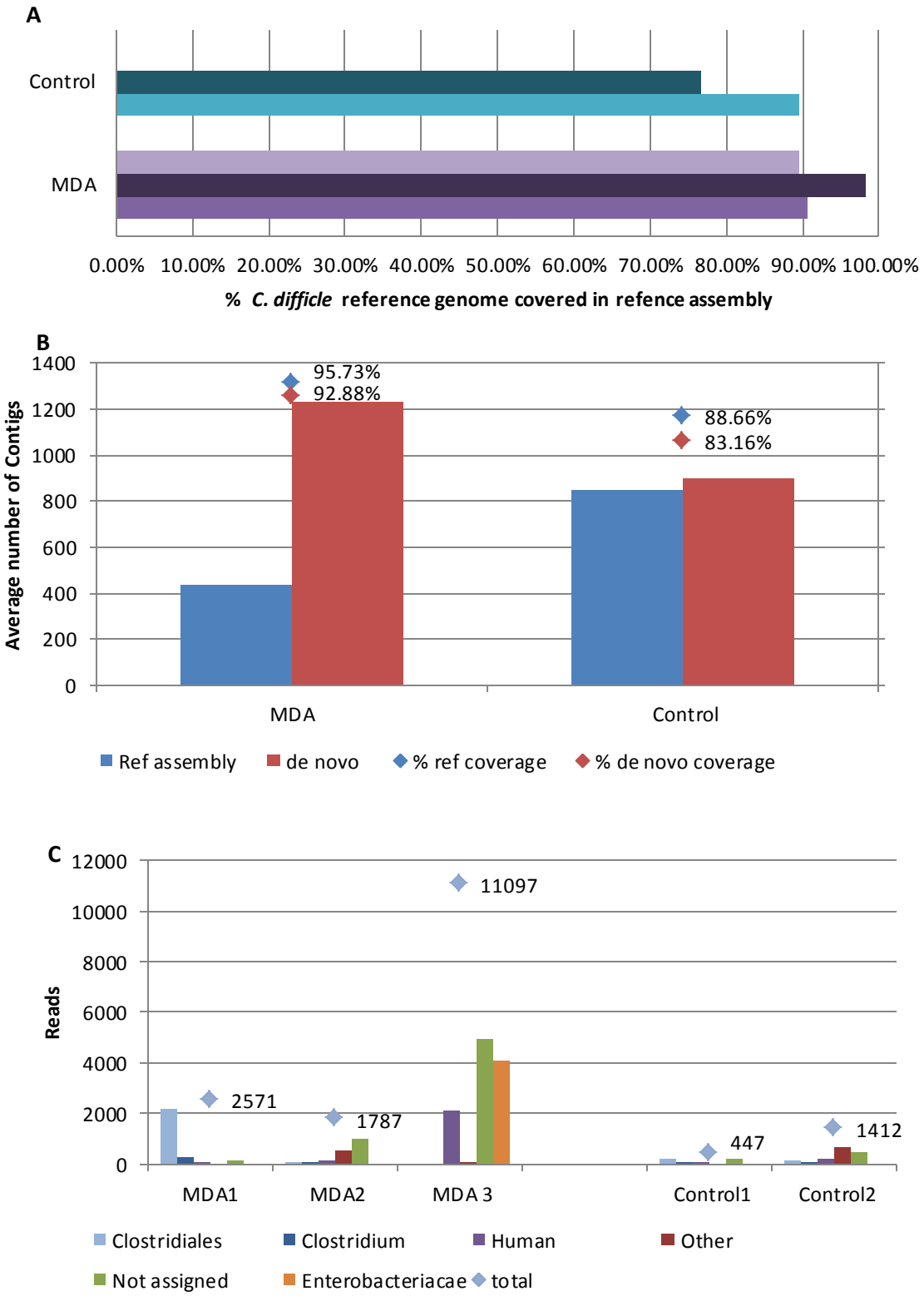


Figure 3-14 Comparison of single cell ϕ29 MDA and non-amplification preparation of sequencing on (A) *C. difficile* reference coverage and (B) average number of contigs rpoduced in the reference and *de novo* assembllies along with average % genome coverage. (C) the indentification of none *C. difficile* DNA in non-amplification and ϕ29 MDA samples

	ϕ29 MDA amplified	None Amplification control	F test	T test value	Sig (0.05)
No: Raw reads	225451	227593	equal	0.9165	No
No: Pass filter reads	174907	167093	unequal	0.8465	No
% Pass filter	77.76%	72.75%	equal	0.4973	No
Mean read length (bp)	460	444	unequal	0.1717	No
		Reference mapping			
% ref coverage	95.73%	88.66%	unequal	0.3905	No
% reads mapped	97.04%	99.48%	equal	0.3550	No
No: contigs	438	850	equal	0.1843	No
largest contig (bp)	146811	57823	equal	0.0146	Yes
average contig (bp)	11357	5655	equal	0.0972	No
		***De novo* assembly**			
% ref coverage	92.88%	83.16%	equal	0.2025	No
No: contigs	1235	900	equal	0.2142	No
largest contig (bp)	29806	30559	equal	0.8957	No
N50	5194	4688	equal	0.7463	No
#misassemblies	24	34	equal	0.0331	Yes

Table 3-20 comparison of sequencing results obtained from single cell ϕ29 MDA and non-amplification for library preparation of C. *difficile* 630 including raw output and results of reference and *de novo* assemblies

3.8.2 Actinomyces naeslundii

Actinomyces naeslundii is a Gram positive pathogen with a GC content of 68.5% and a total genome size of 3,042,856 bases.

Two sets of non-amplification control libraries were sequenced and three replicate using single cell ϕ29 MDA reactions were sequenced **Table 3-22.** The total DNA produced by the ϕ29 MDA reaction was $1.46x10^3$ ng/µl, 997 ng/µl and $1.24x10^3$ ng/µl.

The average number of raw reads produced by the ϕ29 MDA and non-amplification libraries was very similar, 206578 and 201128. The percentage that passed filter being slightly higher in the ϕ29 MDA library, 78.14% compare to 76.87% in the non-amplification library, but this was not statistically significant. There was very little difference in the length of reads produced by the sequencing, with the average read lengths being 429 and 433 bases. The proportion of the reference that was covered after reference assembly was higher in the ϕ29 MDA runs 76.30% compared to 63.39% in the non-amplification library, but this wasn't significant. The number of reads used to construct the reference mapping was 81.62% in the ϕ29 MDA and 79.43% in the non-amplification libraries. The number of contigs produced was similar in both sets of data 1624 and 2018, and the largest contigs were 11,628 and 10,312 bases, with average lengths of 1116 and 1216 bases, none of which were significant. *De novo* reference coverage was 64.95% in the ϕ29 MDA and 60.94% in the non-amplification library, and the number of contigs was slightly higher in the ϕ29 MDA data at 1663 compared to 1583 in the control library, but again this was not significant. The N50 scores were also similar between the two data sets with the ϕ29 MDA being 3835 bp and the non-amplification being 4229 bp. The number of misassemblies was significantly higher in the non-amplification at 33 and only 22 in the ϕ29 MDA libraries. The comparison results are summarised in **Table 3-23** showing the average values for each variable, the F test results and the T test values and interpretation.

	Non-amplification		ϕ29 MDA single cell reaction		
	Control 1	Control 2	Reaction 1	Reaction 2	Reaction 3
DNA produced (ng/µl)			1.46x10^3 ng/µl	997 ng/µl	1.24x10^3 ng/µl
No: Raw reads	187569	214686	186891	242620	190222
No: Reads passing filter	149569	158825	153174	124301	95534
% reads passing filter	79.74%	79.98%	81.92%	51.23%	50.22%
Mean read length (bp)	463	430	457.25	419.49	409.23
Reference assembly					
% ref coverage	63.22%	63.56%	78.52%	80.49%	69.90%
% reads mapped	81.35%	77.51%.	83.45%	81.77%	79.63%
No: Contigs	2136	1900	1465.00	1755.00	1651.00
Average contig length (bp)	1076	1355	1198	1104	1047
***De novo* assembly**					
% ref coverage	61.78%	60.09%	67.89%	65.59%	61.37%
No: Contigs	1868	1297	1560	1760	1668
N50	4226	4231	4132	4035	3339
#misassemblies	33	33	22	22	23

Table 3-21 results of sequencing single cell ϕ29 MDA and non-amplification control replicates of *A. naeslundii* including raw output and reference and *de novo* assembly results

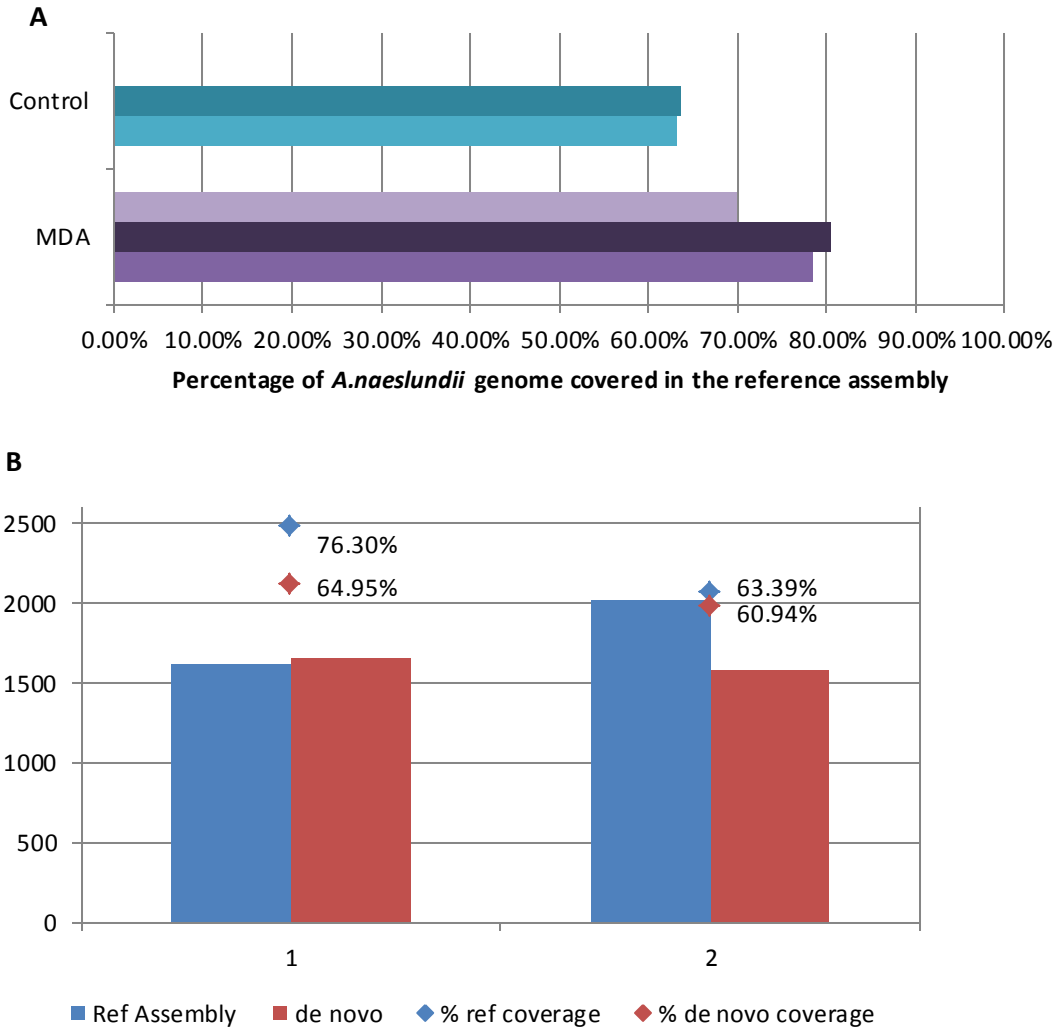


Table 3-22 Comparison of single cell ϕ29 MDA and culture control on *A.naeslundii* (A) % reference coverage and (B) number of contigs in the reference and *de novo* assemblies, including % genome coverage

	ϕ29 MDA amplified	Non-Amplification control	F test	T test value	Sig (0.05)
No: Raw reads	206578	201128	Equal	0.8438	No
No: Pass filter reads	161003	154215	Equal	0.6944	No
% Pass filter reads	78.14%	76.87%	Unequal	0.7505	No
Mean read length (bp)	429	433	Equal	0.8337	No
Median read length (bp)	489	461	Unequal	0.2091	No
Modal read length (bp)	508	496	Equal	0.2564	No
Reference mapping					
% ref coverage	76.30%	63.39%	Unequal	0.0576	No
% reads mapped	81.62%	79.43%	Unequal	0.4441	No
No: Contigs	1624	1818	Equal	0.2190	No
Largest contig (bp)	11628	10312	Unequal	0.4321	No
Average contig (bp)	1116	1216	Equal	0.4641	No
***De novo* assembly**					
% ref coverage	64.95%	60.94%	Equal	0.2126	No
No: Contigs	1663	1583	Equal	0.7457	No
Largest contig (bp)	21743	22587	Equal	0.9282	No
N50	3835	4229	Unequal	0.2561	No
#misassemblies	22	33	Unequal	0.0010	Yes

Table 3-23 Comparison of sequencing results from single cell ϕ29 MDA and culture control replicates of *A. naeslundii,* including raw output and results of reference and *de novo* assemblies

3.8.3 Analysis of the Effect of GC Content on DNA Amplification, Sequencing and Assembly

The ϕ29 MDA single cell reactions for the three bacteria covering a wide range of GC contents, *C. difficile* (29.06%), *E. coli* (50.75%) and *A. naeslundii* (68.5%), were investigated to determine the impact of GC content on whole genome amplification using ϕ29 MDA. Further analysis was performed on sequencing results across the range of GC content, looking at sequencing and assembly quality. Culture control sequencing was used to control for instrument and programmes to allow differentiation between ϕ29 MDA bias and inherent issues in sequencing and analysis techniques.

The average concentration of DNA produced by the *C. difficile*, *E. coli* and *A. naeslundii* ϕ29 MDA reactions were 1113, 1268 and 1232 ng/µl the three sets of data demonstrated no significant difference between them. **Figure 3-15** illustrates the amount of DNA produced in each replicate along with the average of each replicate.

For the non-amplification control the average number of reads for the three bacteria *C. difficile*, *E. coli* and *A. naeslundii)* was 227,593, 250,858 and 201,128 and the number of reads that passed filter was 167,093, 193,529 and 154,215. There was no significant difference between bacteria. There was also no significant difference between read lengths produced, with the average mean read lengths being 444, 423.74 and 433 bp.

The average number of reads for the three bacteria *C. difficile*, *E. coli* and *A. naeslundii* amplified by ϕ29 MDA was 225,451, 217,141 and 206,578 and the number of reads that passed filter was 174,907, 170,233 and 161,003. There was no significant difference between bacteria. There was also no significant difference between read lengths produced, with the average mean read lengths being 460, 428.77 and 429 bases. **Table 3-24** summarises the average results and statistical results for sequencing data produced.

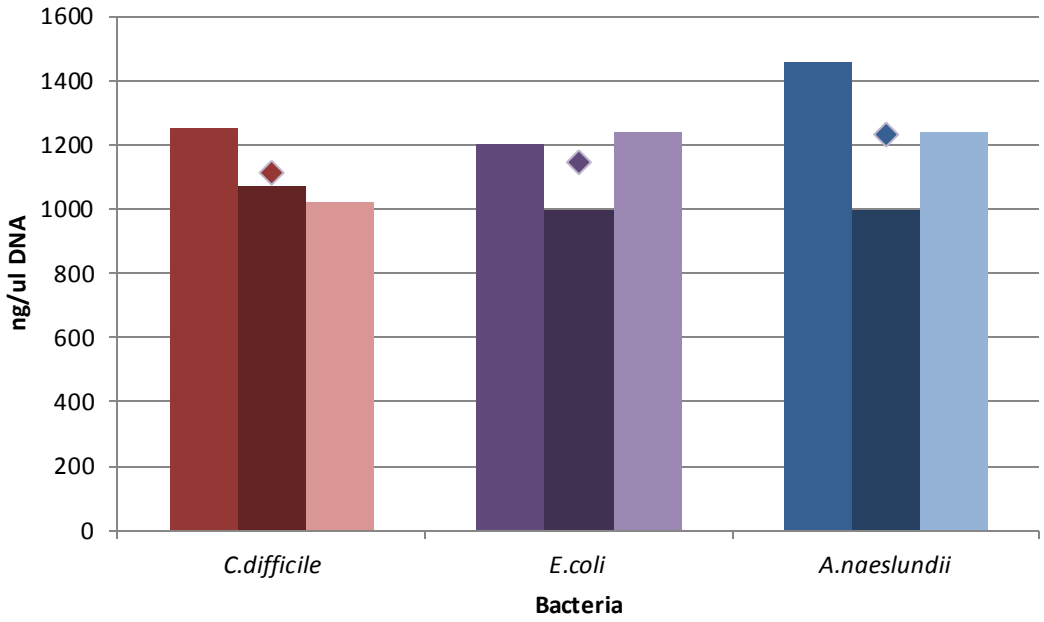


Figure 3-15 A Graph Showing the concentration of DNA produced by each replicate in the three bacteria studied, (*C. difficile*, *E. coli* and *A. naeslundii*), with averages concentration shown as diamonds

	C. difficile vs. *E. coli*			Average results			*A. naeslundii* vs. *E. coli*		
Non-amplification libraries									
	F test	T test value	Sig (0.05)	*C. difficile*	*E. coli*	*A. naeslundii*	F test	T test value	Sig (0.05)
No: Raw reads	UE	0.3509	No	227593	250858	201128	E	0.7493	No
No: Pass filter reads	E	0.7880	No	167093	193529	154215	E	0.8538	No
% Pass filter	UE	0.4270	No	72.75%	77.00%	76.87%	E	0.4146	No
Mean read length (bp)	E	0.0330	Yes	444	423.74	433	UE	0.0011	Yes
ϕ29 MDA single cell libraries									
No: Raw reads	E	0.1749	No	225451	217141	206578	UE	0.1283	No
No: Pass filter reads	UE	0.3212	No	174907	170233	161003	E	0.1545	No
% Pass filter	E	0.8351	No	77.76%	78.4%	78.14%	UE	0.6544	No
Mean read length bp)	E	0.1506	No	460	428.77	429	UE	0.8522	No

Table 3-24 table summarising the average results for sequencing output of ϕ29 MDA for the three bacteria with differing GC contents (C. *difficile, E. coli and A. naeslundii*). Additionally, the results of a statistical test comparing results of E. *coli* (50% GC) sequencing to the extreme GC content bacteria. (F tests E= equal UE=unequal)

After reference assembly the average coverage of the reference genomes for the non-amplification controls were 88.66% for *C. difficile*, 96.87% for *E. coli* and 63.39% for *A. naeslundii.* There was no significant difference between the *C. difficile* and *E. coli,* but the reference coverage of the *A. naeslundii* was significantly lower. The percentage of reads used to create the reference map was significantly higher in the *C. difficile (99.84%)* assembly than the *E. coli* (94.18%), but in the *A. naeslundii* reference assembly the percentage was significant lower (79.43%). There was no significant difference in the number of contigs produced for the *C. difficile* and *E. coli* reference assembly *(850 C. difficile* and 661 *E. coli)*; however, there was a significantly higher number of contigs produced in the *A. naeslundii* assembly (1216)**.** The size of contigs didn't significantly differ in the *C. difficile* and *E. coli,* the average size of the largest contigs was 57823 bp for *C. difficile* and 81579 bp for *E. coli.* The average contig size for *C. difficile* was 5655 and 22,442 for *E. coli.* The contigs produced by the *A. naeslundii* assembly were significantly shorter (average size 1216 compared to 11442 bp).

When examining the ϕ29 MDA library data the average genome coverage of the *C. difficile*, *E. coli* and *A. naeslundii* reference sequences were 95.73%, 97.03% and 76.30%, with no significant difference between *C. difficile* and *E. coli,* but the *A. naeslundii* genome coverage was significantly lower. The number of reads used to generate the reference assembly was significantly higher for the *C. difficile* (97.04%), assembly when compared to *E. coli* (86.49%) assembly. The number of contigs produced by the reference assembly was 438, 316 and 1624, with no significant difference between the *C. difficile* and *E. coli* assemblies. However, there was a significant increase in the number of contigs produced by the *A. naeslundii* assembly. The largest contig and average contig size of the *C. difficile* and *E. coli* assemblies showed no significant difference, but again there was a significant difference between the *A. naeslundii* and *E. coli* assemblies, with the *A. naeslundii* assembly producing much smaller contigs. **Figure 3-16** shows the average and largest contig size produced by each bacterial non-amplification control and ϕ29 MDA libraries along with the genome coverage and reads mapping.

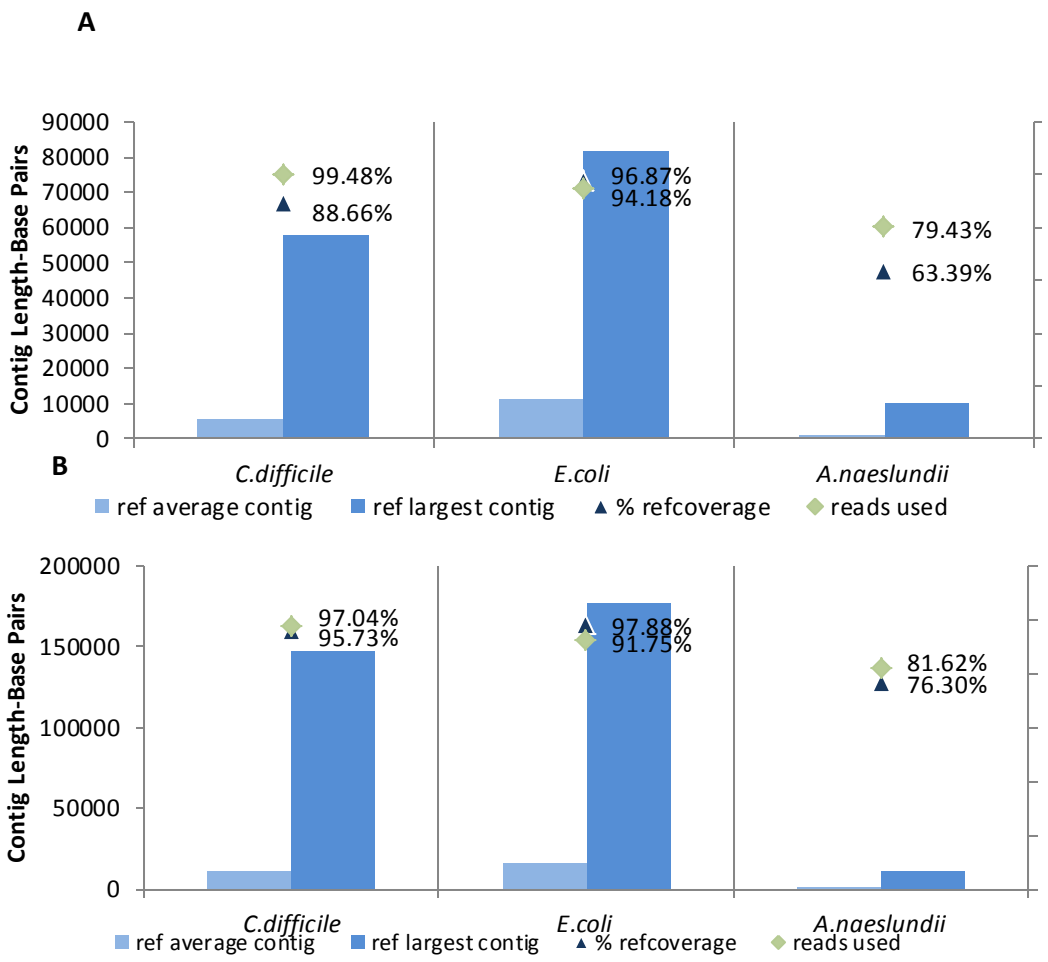


Figure 3-16 comparison of average and longest contigs produced from reference assembly of the (A) non-amplification control and (B) ϕ29 MDA libraries. Including reference coverage and % of reads used to create reference assembly

	C. difficile vs. *E. coli*			Average results			*A. naeslundii* vs. *E. coli*		
Culture control libraries									
	F test	T test value	Sig (0.05)	*C. difficile*	*E. coli*	*A. naeslundii*	F test	T test value	Sig (0.05)
% ref coverage	UE	0.3369	No	88.66%	96.87%	63.39%	E	0.0003	Yes
% reads mapped	E	0.0015	Yes	99.48%	94.18%	79.43%	E	0.0022	Yes
No: Contigs	E	0.7085	No	850	660.67	2018	E	0.0457	Yes
Largest contig (bp)	E	0.4385	No	57823	81579	10312	UE	0.0433	Yes
Average contig (bp)	E	0.3598	No	5655	11442	1216	UE	0.0357	Yes
ϕ29 MDA single cells									
% ref coverage	UE	0.0690	No	95.73%	97.03%	76.30%	E	0.0027	Yes
% reads mapped	UE	0.0439	Yes	97.04%	86.49%	81.62%	E	0.0533	No
No: Contigs	UE	0.3179	No	438	316	1624	UE	0.0006	Yes
Largest contig (bp)	UE	0.1216	No	146811	178657	11628	UE	0.0011	Yes
Average contig (bp)	E	0.1575	No	11357	15805	1116	UE	0.0215	Yes

Table 3-25 average values for reference assembly of non-amplification and ϕ29 MDA sequencing for *C. difficile*, *E. coli* and *A.naeslundii* including F and T test results comparing extreme GC content with *E. coli* (F test E=equal and UE=unequal)

When the non-amplification libraries of the three bacteria were *de novo* assembled, the average genome coverage was 83.16% for the *C. difficile* library, 84.90% for the *E. coli* library and 60.94% for the *A. naeslundii,* with the *A. naeslundii* coverage being significantly lower than the *E. coli*. The number of contigs, and contig sizes produced were similar between the three bacteria. The average number of contigs for *C. difficile*, *E. coli* and *A. naeslundii* was 900, 1697 and 1583. The average largest contig of the replicate assemblies was 30559, 32327 and 22587 bps for each bacterium. The number of misassemblies when comparing *C. difficile* and *E. coli* showed significant difference (34 *C. difficile* vs. 27 *E. coli*); the number of misassemblies in the *A. naeslundii de novo* assembly was also significantly higher at 33. Further investigation of the QUAST analysis showed the total lengths of the *de novo* assemblies for *E. coli* and *C. difficile* were similar to the reference sequence length. The average total assembly length for *E. coli* was 4,444,129 bases (reference length 4,639,675 bp) and the average total length assembly length of the *C. difficile* was 3,908,710 bp (reference length 4,290,252 bases). However, the total length of the *A. naeslundii* assembly was only 81% of the reference length (2,464,713 bp, with a reference length of 3,042,856 bp).

When inspecting the data produced from ϕ29 MDA libraries after the reads were assembled using a *de novo* approach the proportion of the genome that was covered was 92.88% for the *C. difficile* library, 92.21% for the *E. coli* library and 64.95% for the *A. naeslundii.* With the coverage being significantly lower in the *A. naeslundii*. The number of contigs produced and the contigs sizes showed no significant difference between the three bacterium (1235, 1230 and 1663). The number of misassemblies also showed no significant differences (25, 21 and 22). Further investigation of the QUAST analysis showed similar results to the non-amplification sample, with the *E. coli* and *C. difficile* assembly lengths being similar to the total reference length, and the *A. naeslundii* total reference length being shorter.

	C. difficile vs. *E. coli*			Average results			*A. naeslundii* vs. *E. coli*		
Non-amplification control									
	F test	T test value	Sig (0.05)	*C. difficile*	*E. coli*	*A. naeslundii*	F test	T test value	Sig (0.05)
% ref coverage	UE	0.8467	No	83.16%	84.90%	60.94%	UE	0.0335	Yes
No: contigs	UE	0.1321	No	900	1697	1583	UE	0.7617	No
largest contig (bp)	E	0.9029	No	30559	32327	22587	E	0.5278	No
N50	UE	0.9317	No	4688	4506	4229	UE	0.5134	No
#misassemblies	UE	0.0398	Yes	34	27.00	33	E	0.0351	Yes
Total assembly length (bp)				3908710	4444129	2464713			
Reference sequence length (bp)				4290252	4639675	3042856			
ϕ29 MDA single cell									
% ref coverage	UE	0.7932	No	92.88%	92.21%	64.95%	E	0.0003	Yes
No: contigs	E	0.6752	No	1235	1230	1663	E	0.1041	No
largest contig (bp)	E	0.1256	No	29806	35872	21743	UE	0.0641	No
N50	UE	0.6193	No	5194	5168	3835	E	0.1032	No
#misassemblies	UE	0.0677	No	25	21	22	E	0.6433	No
Total assembly length (bp)				4,023,458	4,453,125	3,025,154			
Reference sequence length (bp)				4,290,252	4,639,675	3,042,856			

Table 3-26 average values for *de novo* assembly of culture control and ϕ29 MDA sequencing for *C. difficile, E. coli and A.naeslundii* including F and T test results for comparison of extreme GC genomes with *E. coli*

3.9 Application to Low Level Mixed Bacteria

To assess the utility of the approach developed in this study for the analysis of low number mixed infection two bacteria were mixed, amplified and sequenced. *E. faecalis (*3.2MB*)* and *H. influenzae* (1.8MB) were mixed in PBS in different ratios, *E. faecalis: H. influenzae* was 1:1, 1:10, 1:100 and 1:1000 **Table 3-27.**

The 1:1 ratio mix reference mapping results showed 93% coverage of the *E. faecalis* genome using 62% of the reads. The amount of the *H. influenzae* genome covered was 92% using 25% of the reads. In the 1:10 ratio, 64% of the *E. faecalis* genome was covered using 13% of the reads, and 94% of the *H. influenzae* using 85% of the reads. In the 1:100 ratio 15% of the *E. faecalis* genome was covered using 2% of the reads, and 94% of the *H. influenzae* genome was covered using 83% of the reads. In the 1:1000 genomes no reads mapped to *E. faecalis* and 93% of the *H. influenzae* genome covered was using 95% of the reads.

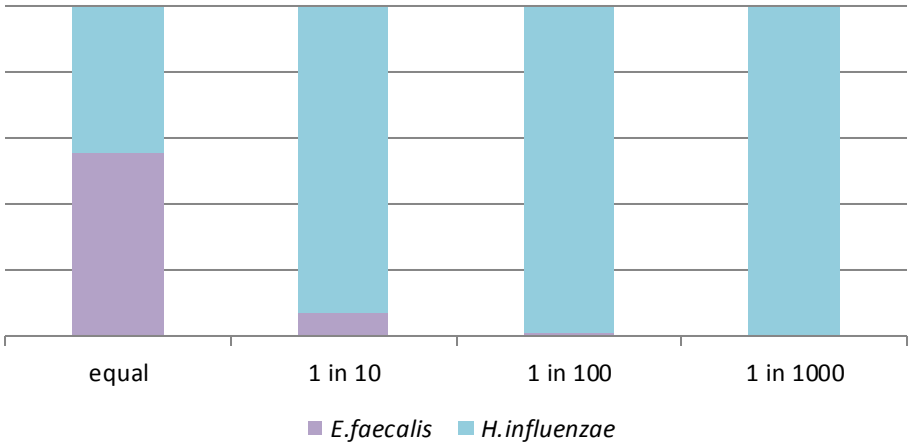


Figure 3-17 Proportion of reads mapping to *E. faecalis* and *H. influenzae* in different ratio mixes, normalised for genome size

Ratio	Total reads	Reads *E. faecalis: H. influenzae*	% reads *E. faecalis: H. influenzae*	% reads Normalised for genome size *E. faecalis: H. influenzae*	%genome coverage *E. faecalis*	%genome coverage *H. influenzae*
1:1	133320	83286:33254	62:25	31:25	93%	92%
1:10	229667	30086:195217	13:85	6.5:83	64%	94%
1:100	148139	3169:122308	2:83	1:83	15%	94%
1:1000	152879	0:145235		(:95)	0%	93%

Table 3-27 Results of reference mapping to *E. faecalis* and *H. influenzae* in different ratio mixes, including proportion of reads before and after normalised for genome size and genome coverage.

3.10 Amplification of Viral Genomes

To determine whether ϕ29 MDA can be applied to viral genomes adenovirus 40 (Ad40) and adenovirus 41 (Ad41) were amplified and sequenced. Adenoviruses are none enveloped icosahedral viruses of between 90-100nm which have a double stranded DNA genome of 35-36kb. The group F adenoviruses were cultured as described in **section 2.1.3.1** extracted and amplified using ϕ29 MDA as previously described.

Three replicates of Ad40 were sequenced producing 123070, 194887 and 140934 reads, with 73.32%, 73.79% and 79.68% of these reads passing filters. When these were assembled against the reference genome they formed a single contig that covered 99.98%, 99.83% and 100% of the genome. The percentage of reads used to form these assemblies were 20.38%, 27.21% (**Figure 3-18**) and 22.95% of all the reads. The *de novo* assemblies formed 261, 108 and 204 contigs respectively covering 96.68%, 96.4% and 96.87% of the genome. The numbers of misassemblies were 3, 0 and 0 for the three runs. When the reads were examined using Blastn and LCA analysis on average 52.8% of reads were identifies as the family Homininae, and in one sample there was a contamination of *E. coli* representing 4.7% of the reads **Figure 3-18-A**. The remaining reads were unassigned.

Three replicates of Ad41 were sequenced. After sequencing 68225, 128802 and 147597 reads were produced with 64.4%, 65.05% and 76.7% of these reads passing filters. When these were assembled against the reference genome they formed a single contig that covered 99.69%, 99.93% and 99.96% of the genome. The percentage of reads used to form these assemblies 40.75%,27.17% and 22.95% of all the reads. The *de novo* assemblies formed 28, 102 and 41 contigs respectively covering 98.47%, 99.71% and 98.75% of the genome. The number of misassemblies was 0, 0 and 2 for the three runs. On average 24.6% of reads mapped to the family Homininae and the remaining reads were unassigned **Figure 3-18-B**.

To lower the presence of host commination, before extraction the sample was treated with DNase1 to remove any host DNA that was reaming in the supernatant. After sequencing 311767, 219881 and 168407 reads were produced with 45.79%, 74.92% and 80.36% of these reads passing filter. The reads were reference assembled into a single contig covering 100%, 99.96% and 99.88% of the reference genome. The number of reads used was 75.98%, 82.92% and 66.14%. The number of *de novo* contigs was 98, 106 and 211 covering 99.04%, 98.55% and 90.12% of the reference with 1, 0 and 0 misassemblies. In two out of three replicates no reads were identified as the family Homininae and in the final sample 3.6% of reads were assigned to the family

Homininae. In one sample there was a small proportion of reads (1.3%) identified as *H. influenzae* **Figure 3-18-C**.

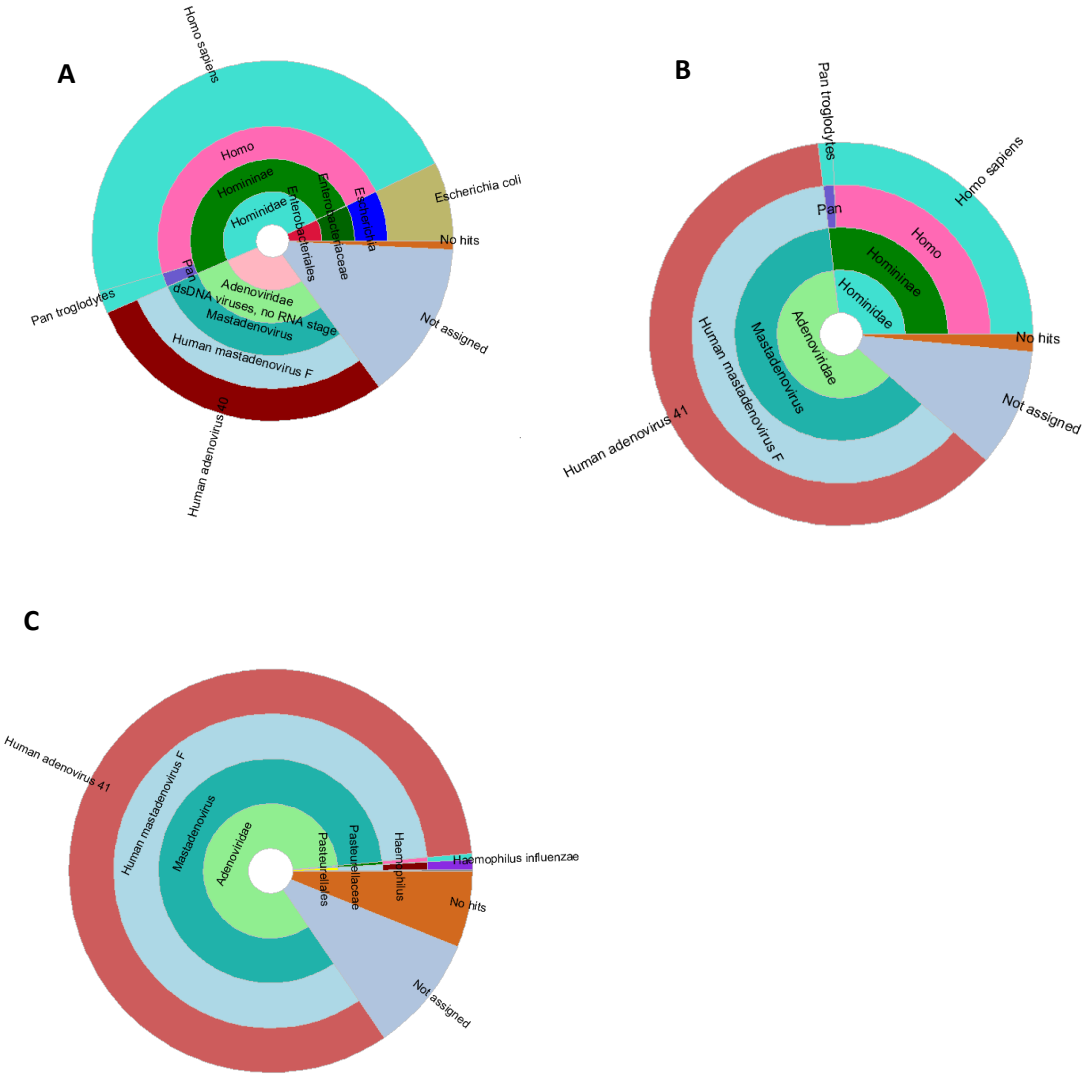


Figure 3-18 LCA analysis demonstrating proportion read identification after ϕ29 MDA and sequencing of (A) Adenovirus 40, (B) Adenovirus 41 and (C) post DNase treatment Adeno41

3.10.1 Application to Mixed Adenovirus and HIV

3.10.1.1 Extracted Nucleic acid

Initially extracted nucleic acid, from a DNA (Ad41) and a RNA (HIV) virus were used to establish the applicability of reverse transcription and ϕ29 MDA to both RNA and DNA pathogens simultaneously.

The number of particles of the adenovirus tissue culture was quantified as $2.3x10^8$ virions per ml. 500 µl of adenovirus tissue culture supernatant was concentrated and extracted (including a DNase and RNase stage) as described in section **2.5.1.1,** the DNA was diluted to contain the equivalent of 25 viral copies per µl (0.001 pg/µl). This was added to the equivalent of 25 virions of HIV (0.0002 pg/ µl). (Using End Memo online tool for copy number calculator, based on HIV having two copies of the RNA genome per Virion.) The mixture of the extracted RNA and DNA was added to the reverse transcription reaction with SuperScript IV as described in **2.2.2.5** . The resulting DNA was then amplified using ϕ29 MDA for two hours. The DNA output was 584 ng/ µl, which was then sequenced using the 454 Junior as previously described. In total 103115 sequencing reads were produced, with 72143 reads passing filter, of these 6954 (9.64%) mapped to the human genome. 66.2% of the remaining reads were identified as Adenovirus, and 20.76% as HIV and 12.6 % of reads were not identified. The remaining reads were identified as Proteobacteria. When the reads were assembled using *de novo* assembly 92.3% of the adenovirus genome and 91.4% of the HIV genome were covered.

3.10.1.2 Adenovirus and HIV-Viral Particles

Dilutions of the supernatant of HIV and Adenovirus tissue cultures were made so as to contain an estimated 25 viral particles. These were pooled and then the total volume made up to 10 ml. This was then concentrated with PEG 20,000 incubated on ice for one hour (**2.5.1.1**), host nucleic acid removed (**2.5.1.2**) and the sample was extracted using column extraction (**2.5.1.3).** The resulting eluted nucleic acid was then added to a reaction with SuperScript IV using random primers (**2.5.1.5**) before being heated to 95^oC for two minutes followed by 2 minutes on ice. The sample was then amplification in a two hour ϕ29 MDA reaction, resulting in a DNA concentration of 491 ng/ µl. Which was fragmented and sequenced on the 454 junior as previously described. After sequencing 126391 reads were produced with 91001 passing filter. When removing human signals, 6.8% of reads were found to be human, of the remaining reads 62.7% mapped to

adenovirus and 28.4% mapped to HIV, 8.8% of reads had no identity. The remaining reads were once again Proteobacteria.

The adenovirus *de novo* assemble covered 97.1% of the genome in 392 contigs, with three misassemblies. The HIV *de novo* assembly covered 99.9% of the genome in three contigs, with two misassemblies. When the *de novo* assembly results were analysed using an automated genotyper, REGA version 3 (Stanford University)[137] (**Figure 3-19**) the results were sub-type B, which correlated with the input.

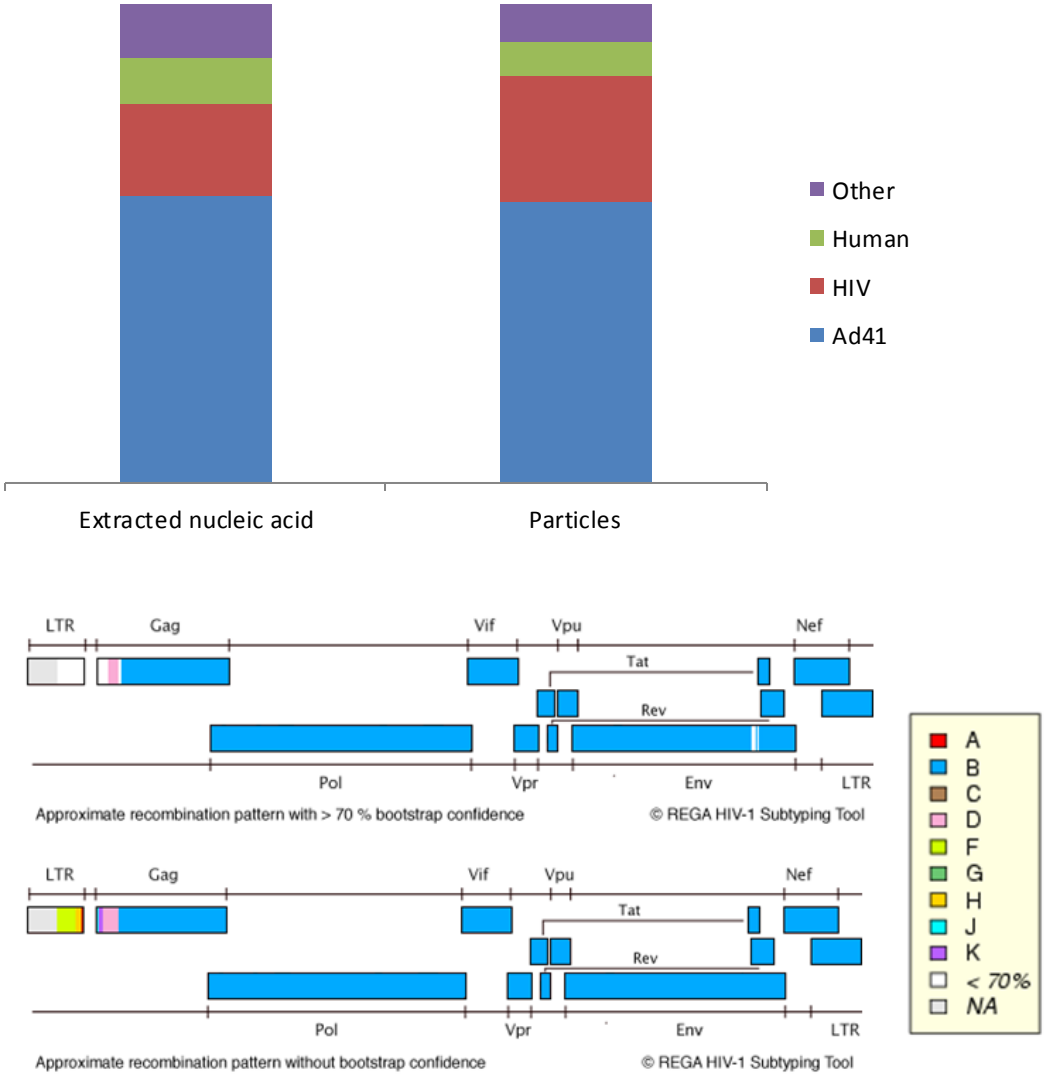


Figure 3-19 HIV and DNA viral mix ϕ29 MDA sample (A) proportional read identification of amplification of extracted nucleic acid and viral particles of Ad41 and HIV mix. (B) HIV genotyping using *de novo* assembly of ϕ29 MDA prepared Ad41 and HIV mix using REGA version 3 (B)

3.10.2 Application to Multiple Viral Pathogens

3.10.2.1 Preparation of Initial Viral Mix

A viral mixture was prepared as described in **2.5.3.1**, briefly; this included the same volume of HIV and Adenovirus tissue culture supernatant as in **3.10,** which aimed at 25 virus particles. Additionally an estimated 25 viral particles of HBV, Parvovirus 19, HAV, and HBV were added, volumes were calculated using the IFU/ml given by NIBSC **Table 2-3**. For the VZV, Norovirus, Measles and Influenza A only Ct values were provided and so 20 µl of each of these were added. The total volume was then made up to 10 ml and the viruses were concentrated and extracted as previously described.

The number of reads assigned to Ad41 was 338, which when assembled using *de novo* assembly covered 32.7% of the genome in 36 contigs. The number of reads assigned to VZV was 1111481, which equated to 2309 viral copies, when these reads were assembled using Spades 94.1% of the genome was covered in 39 contigs. 98 reads were assigned to HBV, which equated to 95 viral particles; the *de novo* assembly covered 42.3% of the genome in 29 contigs. Parvovirus 19 had 255 reads assigned, equating to and input of 264 virus particles, when assembled 86.0% of the genome was covered in 19 contigs **Table 3-28.**

Using LCA analysis 164 reads were assigned to HIV, and when assembled into 34 contigs were formed which covered 40.8% of the genome. Norovirus had 11136 reads assigned, which was calculated to represent 4117 virus particles, when these reads were assembled 97.5% of the genome was covered in 23 contigs. HAV had 40 reads assigned which was equated to 16 viral particles, when *de novo* assembled 21.9% of the genome was covered in 19 contigs. HCV had 451 reads assigned; equal to 133 viral particles, when this was assembled 51.3% of the genome was covered in 21 contigs. No reads were assigned to Measles virus. Influenza A had 12960 reads assigned, with 2948 viral copies, when these reads were assembled 93.5% of the genome was covered in 25 contigs **Figure 3-20** and **Figure 3-21.**

Virus	Genome type	Genome size (kb)	GCPC	Reads	Normalised Reads	Relative viral input	Genome Cov %	Contigs
Ad41	DS DNA	35	1	338		25	32.7	36
VZV	DS DNA	125	1	111481	9300	2309	94.1	93
HBV	Partial DS DNA	3.2	1	98	6969	95	42.3	29
Parvovirus B19	SS DNA	5	1	255	18284	264	86	19
HIV	SS (+) RNA	9.7	2	164		25	40.8	34
Norovirus	SS (+) RNA	8	1	11136	13311	4117	97.5	23
HAV	SS (+) RNA	7.5	1	40	20435	16	21.9	19
HCV	SS (+) RNA	10	1	451	13894	133	51.3	21
Measles	SS (-) RNA	15.5	1	0	118	0		
Influenza A	SS (-) RNA	13	1	12960	5956	2948	93.5	25
Other				1575				
Total				138498				

Table 3-28 results of sequencing of mixed viral input using reverse transcription with SuperScript III and DNA amplification using ϕ29 MDA, including -reads assigned, calculated viral particle number and assembly

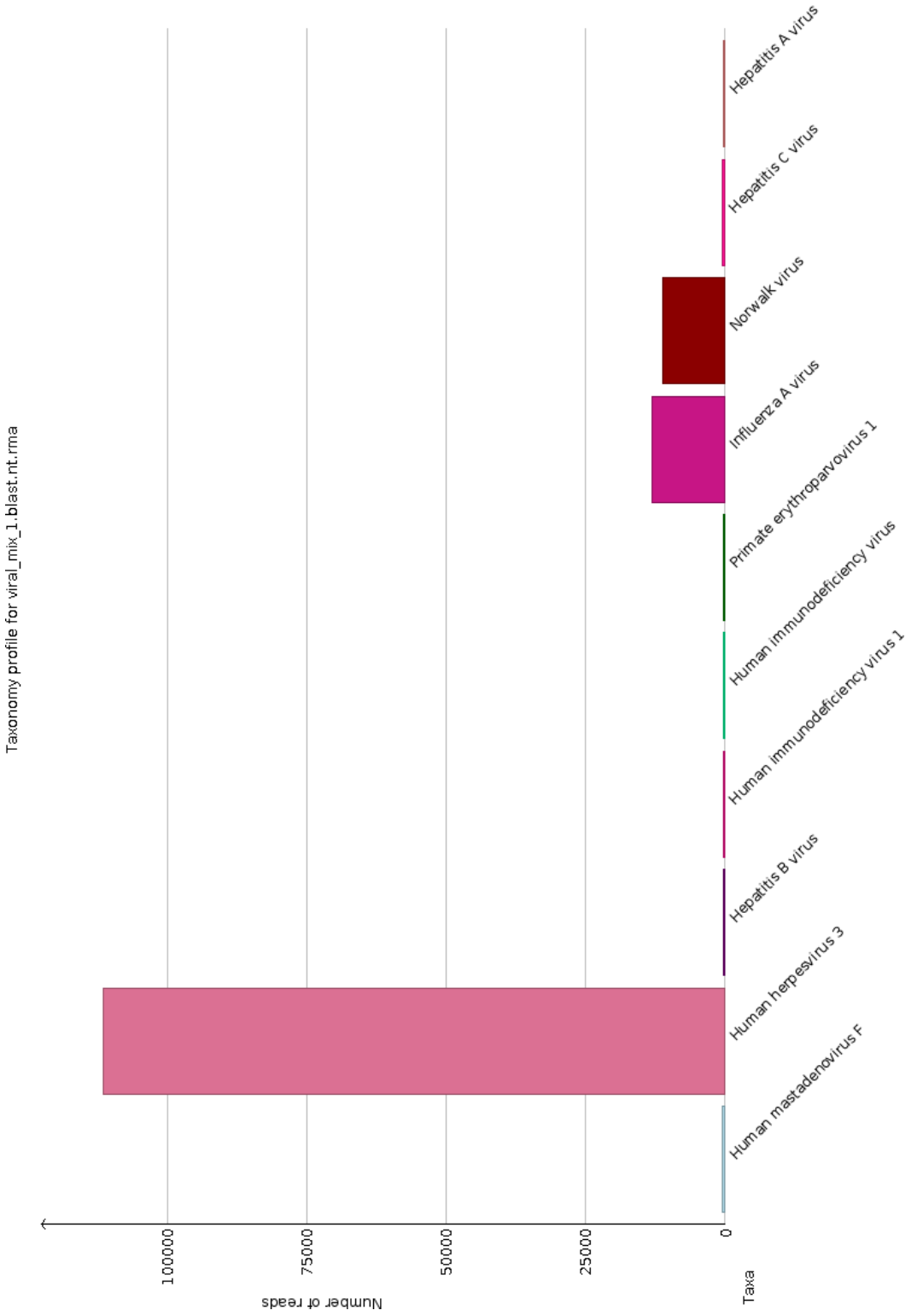


Figure 3-20 visualisation of proportional read identification using LCA analysis for initial mixed virus reaction amplified using ϕ29 MDA**.**

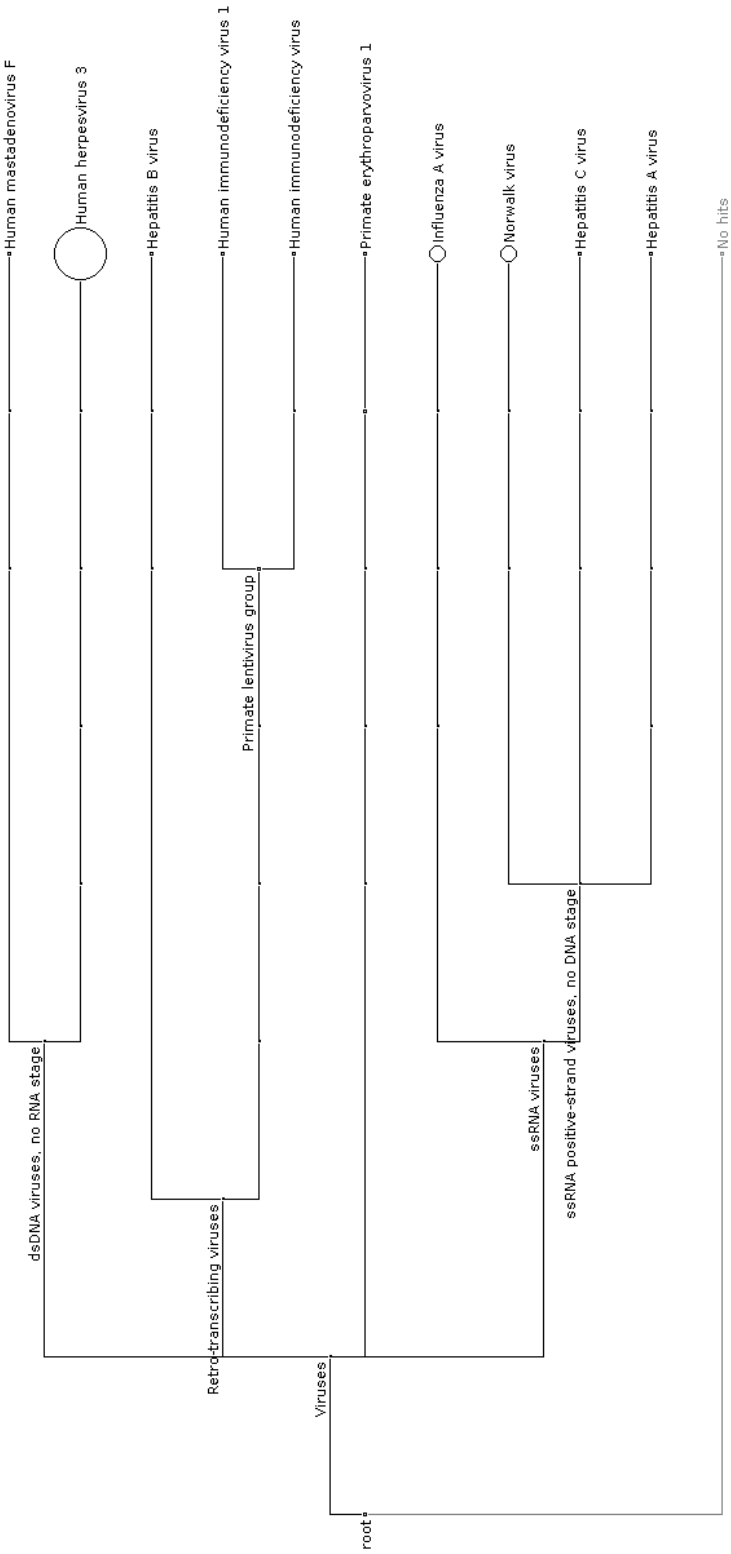


Figure 3-21 LCA analysis of initial mixed viruses amplified with ϕ29 MDA

3.10.2.2 Calculating Viral Input

An equation was developed (Equation 2) to which allowed the number of viral particles present in a sample to be calculated by using a known quantity viral spike. This equation takes into account the genome size, the genome type and the number of copies of the genome the virus carries. The use of a known spike will allow more accurate quantification of the viral particles identified, as opposed to the relative abundance. In a high through put assay the values for the reference would be fixed. In this case the number of control particles was 25. The assigned reads refers to the identification given to reads using Blastn and LCA analysis using Megan, and so represents the number of reads identified as each virus. The reads are normali sed for genome size to allow for the different genome sizes, due to the assumption that, a genome that was twice the size would be expected to produce twice the number of reads. The strand designation value allows the nature of the genome to be taken into account, such that a virus with a double stranded genome would start off with twice the amount of reads as a virus with the same sized genome but a single stranded genome. Also in this case a different RNA and DNA reference spike was included to account for inefficiencies in RNA conversion, and any RNA degradation that may have occurred during processing.

A

$$P\begin{pmatrix} A\binom{r}{a}\binom{c}{d} \\ R\binom{s}{t} \end{pmatrix} = V$$

V=viral number
P=Number control particles
A=assigned read number
R=Spike read number
r=Spike genome size
a=assigned genome size
c=genome copies per virus, assigned
d=genome copies per virus reference
s= spike strand designation
t= assigned strand designation
Strand type designations= double stranded=1, single stranded = 2,
Partial double stranded = the number of bases double stranded divided by number of bases single stranded.

B

$$25\begin{pmatrix} A\binom{35}{a}\binom{c}{1} \\ 338\binom{1}{t} \end{pmatrix} = V$$

C

$$25\begin{pmatrix} A\binom{9.7}{a}\binom{c}{2} \\ 164\binom{2}{t} \end{pmatrix} = V$$

Equation 2 (A) Equation for calculating number of viral partides based on a known spike (B) worked example for Ad41 spike from **3.10.2** and (C) worked example for HIV spike from **3.10.2**

3.10.2.3 Sequencing Output Viral Mix after Adjustment of Input

The input volumes of the viral mix were adjusted based on the previous result to an expected outcome of 25 viral particles in total. The resulting number of reads for adenovirus was 5960, which was used as the reference for 25 viral particle input. When these reads were *de novo* assembled they covered 96.1% in 48 contigs. The number of reads assigned to VZV was 33215, which equated to an input number of 39. When the VZV reads were assembled they covered 94.2% of the genome in 115 contigs. The reads assigned to HBV totalled 531, which was equivalent to 29 input particles. When assembled they covered 98.7% of the genome in three contigs. The number of reads assigned to Parvovirus was 1306, equating to 77 viral particles. When the Parvovirus19 reads were assembled 96.3% of the genome was covered in six contigs.

The number of reads assigned to HIV was 3240, which was used as the reference to adjust the RNA virus numbers. The *de novo* assembly covered 99.2% of the genome in four contigs. The number of reads assigned to Norovirus was 5489 which equated to 103 viral particles. When the assigned reads were assembled 98.7% of the genome was covered in three contigs. The number of reads assigned to HAV was 7900 which was the equivalent of 158 virus particles. The *de novo* assembly covered 97.9% of the genome in 9 contigs. The total number of reads assigned to HCV was 7162, which was comparable to 107 viral particles. The *de novo* assembly of the assigned reads consisted of 11 contigs, which covered 99.1% of the genome. The number of reads identified as Measles was 94, equivalent to 1 viral particle. 11.2% of the measles genome was covered using *de novo* method, in seven contigs. The number of reads assigned to influenza A was 3991, equating to 46 copies. These reads were assembled into 32 contigs covering 88.1% of the genome.

3.10.2.4 Detection of Viral Genotypes

Using the Blastn identification and LCA analysis of the viruses' genotypes were identified correctly for most of the viruses in the mixture. The adenovirus was correctly identified as Human adenovirus 41, the VZV was subtyped as B, the HBV was identified as genotype A (defined by S-gene type), and also subtype adr (surface antigen typing). Using Blastn the HIV was only identified as HIV-1, but when an online tool was used, it was correctly identified as subtype B. The human parvovirus 19 was not further identified. The measles virus was also not further identified. The influenza was correctly identified as being H3N2, and the Norovirus was correctly identified as genotype II but not further subtyped. The HCV was identified as genotype 3, and the HAV was identified as group 1.

Virus	Genome type	Genome size (kb)	GCPC	Reads	Normalised Reads	Relative viral input	Genome Cov %	Contigs
Ad41	DS DNA	35	1	5960		**25**	96.1	48
VZV	DS DNA	125	1	33215	9300	39	94.2	115
HBV	Partial DS DNA	3.2	1	531	6969	29	98.7	3
Parvovirus B19	SS DNA	5	1	1306	18284	77	96.3	6
HIV	SS (+) RNA	9.7	2	3240		**25**	99.2	4
Norovirus	SS (+) RNA	8	1	5489	13311	103	98.7	3
HAV	SS (+) RNA	7.5	1	7900	20435	158	97.9	9
HCV	SS (+) RNA	10	1	7162	13894	107	99.1	11
Measles	SS (-) RNA	15.5	1	94	118	1	11.2	7
Influenza A	SS (-) RNA	13	1	3991	5956	46	88.1	32
Other				1375				
Total				70263				

Table 3-29 results of sequencing of mixed viral input after adjustment using reverse transcription with SuperScript III and DNA amplification using ϕ29 MDA, including -reads assigned, calculated viral partide number and assembly

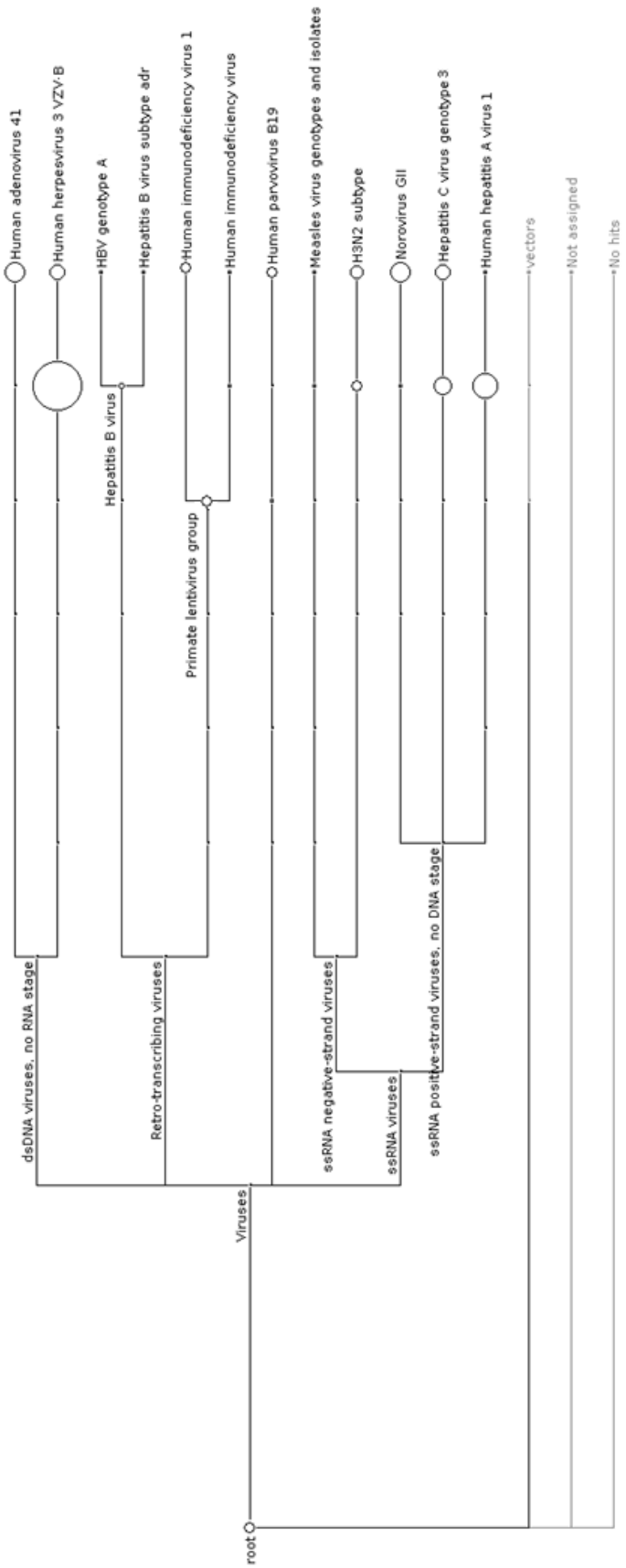


Figure 3-22 LCA Output of viral mix Blastn using MEGAN after adjustment showing viral genotype identification

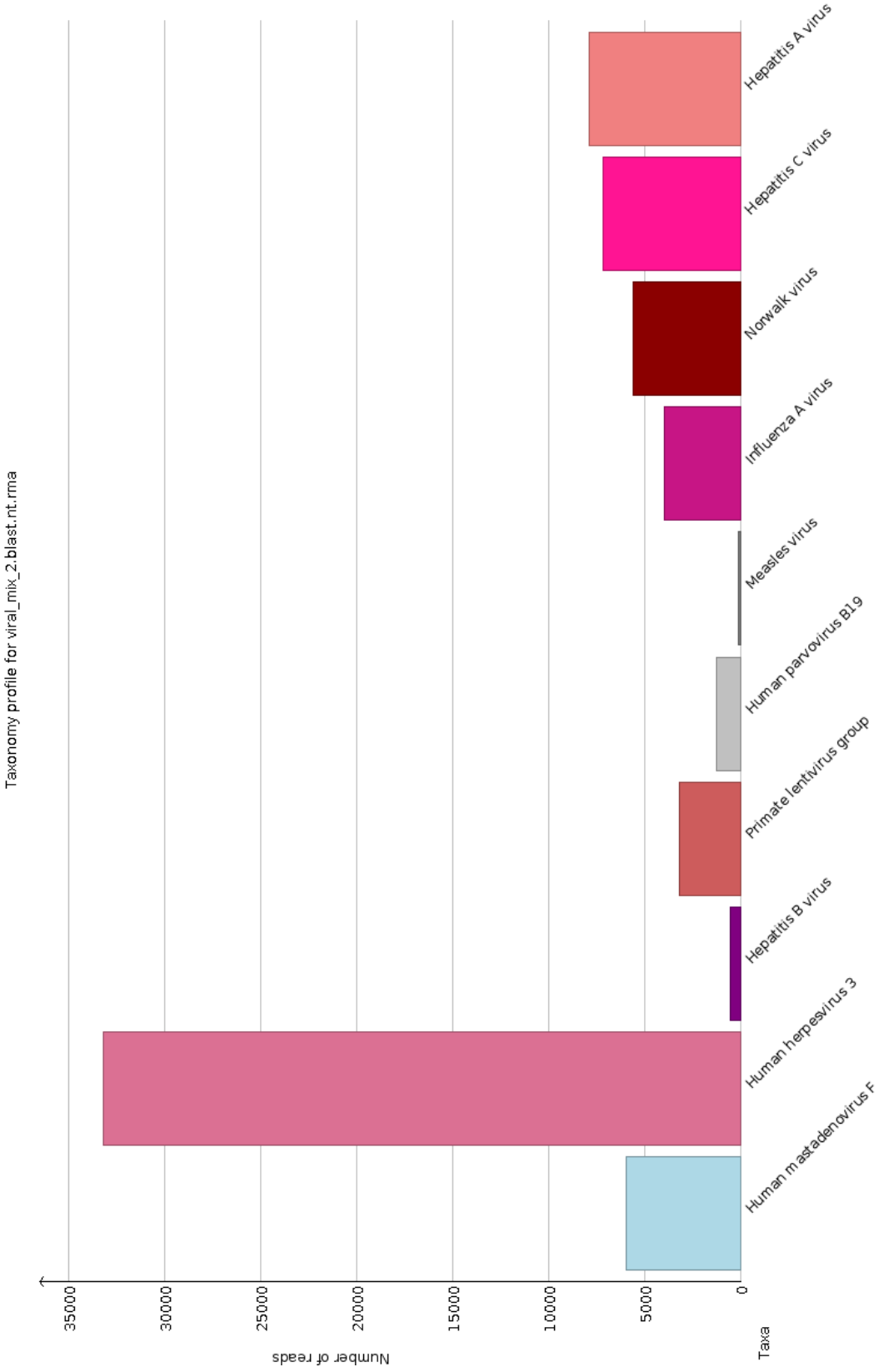


Figure 3-23 proportional read identification using LCA analysis of Blastn results of the adjusted viral mix after amplification with ϕ29 MDA

3.11 Chapter Discussion

3.11.1 Assessment of RNA Conversion Enzymes

RNA viruses are a major cause of local and global outbreaks, such as seasonal Norovirus outbreaks. One study suggested that two Norovirus seasons cost one hospital £1.2 million due lost bed days and staff absences[138]. Rapid and accurate whole genome sequencing will inform epidemiological studies, helping to interrupt transmissions.

The first aim was to efficiently concentrate and extract viruses, which may be circulating in low numbers. The initial PEG concentration was based on Kohno et al's [121] study, which used an overnight incubation at 4°C. This will be a major bottle neck in sample processing and so shorter incubation times were investigated. One-hour incubation at 4°C had a fairly poor performance compared to overnight (RNA concentration of 0.76 vs. 1.14 ng/µl), however by incubating the sample on ice for one hour, comparable results were achieved (1.07 ng/µl). Incubating viral tissue culture supernatant with PEG for one hour on ice recovers 94% of the viral RNA compared to overnight incubations at 4°C in 6.5% of the time (**3.2.1**). For this study, which aims at rapid sequencing from samples one-hour incubation on ice was seen as a better option.

After initial concentration and nucleic acid quantification, high levels of DNA in the sample (1.87 ng/µl) indicated the presence of host nucleic acid, indicating some of the RNA detected would likely be the presence of host RNA, particularly ribosomal RNA. Whole cells should already have been removed by centrifuging at lower speeds without PEG, and so the source of contaminating nucleic acids should be free nucleic acid. The treatment with RNAse and DNAse left no detectable DNA and very little RNA, presumably belonging to HIV. Use of RNase and DNAse treatment to remove host signals is important when using non-specific amplification (3.2.1), as all DNA including host will be amplified.

For RNA viral detection there needs to be a rapid and reliable method for producing long length transcripts from a variety of inputs. Initially specific amplicon targets were used to assess the sensitivity and compatibility of three reverse transcription enzymes. Amplicons of varying lengths aided in indicating the ability to produce long length transcripts. On initial testing after PEG concentration and alkali extraction SuperScript III and IV were both able to produce long transcript (2283 bases) at 0.1 pg. SuperScript III had a lower sensitivity of 0.1 pg compared to SuperScript IV, 0.01pg; however SuperScript IV failed to produce products over 1233 bases at this lower concentration. SuperScript IV is known to be less affected by inhibitors which may have

been present in the sample explaining the slightly higher sensitivity. However, both SuperScript enzymes performed more poorly than expected. PyroPhage was unable to produce transcripts for any size amplicon at all input concentration inputs. When the SuperScript enzymes were combined with ϕ29 MDA, the results were very poor, with only inputs of 1000 pg producing products. The source of these PCR products were most likely the original SuperScript products and not DNA amplified by ϕ29 MDA. Suggesting that ϕ29 was not able to produce DNA from these transcripts.

The poor performance of all enzymes could be caused by inhibition of enzymes; due to the presence of PEG or lack of viral extraction due to PEG interfering with the alkali extraction method. Also PEG allows concentration of small molecules by removing space in the solution; this action may prevent enzymes accessing the nucleic acids as freely. A more efficient extraction method which included the removal of PEG was investigated to purify the HIV extracts after PEG concentration. A spin column, PureLink Viral, was selected, with the substitution of carrier RNA with LPA, as carrier RNA would be converted and sequenced. The small filter allows elution of nucleic acids in small volumes, allowing additional concentration of the nucleic acids.

After column extraction the performance of SuperScript II and IV improved, and when ϕ29 was added, full transcripts were produced. The sensitivity of SuperScript IV was found to be 0.00001 pg, which was better than SuperScript III. PyroPhage 3137 was able to successfully amplify targets at 422 bases, the exonuclease negative enzyme had a higher sensitivity, (0.001 pg vs. 1000 pg) than the wild type enzyme. Although the exonuclease negative enzyme showed higher sensitivity, the enzyme would be less suitable for highly sensitive conversion due to the higher error rate. The inability of the enzyme to produce products greater than 400 bases has previously been reported [78], this combined with the high error rate and lower sensitivity means that PyroPhage 3137 is unsuitably for high sensitivity whole genome amplification.

Overall the use of column extraction improved the performance of all enzymes, probably because traces of PEG were removed. SuperScript IV had the highest sensitivity, and has the advantage of more rapid reaction times (10 minutes, compared to 1 hour with SuperScript III) and so was selected for further work. Previous work coupling ϕ29 MDA with randomly primed reverse transcription used SuperScript III, and cDNA production took 100 minutes [139]. By using a more recently engineered enzyme the reaction time was reduced, allowing results to be obtained quicker.

3.11.2 Non-PCR Based Amplification of Bacterial Genomes

Strand displacement enzymes are able to continually produce DNA, even through regions of double stranded DNA. This has the potential to produce reads of a long length which cover whole pathogen genomes without the need for specific priming.

Escherichia coli K12 substr. MG1655 was used for initial experiments, due to its 50% GC content, ease of growth and fully characterised genome. Initially a PCR with targets across the genome of *E. coli* K12 was used to indicate the success of amplifying the whole genome. Although the ten points chosen only represent a very tiny portion of the genome, it allowed quick and cheap indication of the success of the ϕ29 MDA amplification. *E. coli* was extracted using the alkali method and amplified using ϕ29 MDA producing a high concentration of DNA, which with repeats was shown to be reproducible. The ten-target PCR was positive in all replicates for all targets, indicating successful genome amplification. This demonstrated good DNA production across the *E. coli* genome. Additionally, DNA of high concentration was produced from a negative sample; however, this produced negative results when tested in the *E. coli* PCR, ruling out contamination. The most likely cause is unspecific amplification due to the high processivity of ϕ29 which could be using the primers present to start random replication in the absence of a template to copy. This work combining ϕ29 MDA with targeted PCR amplification, is similar to earlier work, such as the amplification of the intracellular pathogen *Coxiella burnetti*[140]. The aim is to develop the method to rapidly produce whole genome data suitable for sequencing

3.11.3 DNA Amplification from Nicks

Potentially by removing primers from amplification reactions the introduction of primer bias could be prevented allowing a more even coverage of genomes. Additionally, removal of primers could prevent the production of non-specific DNA in low input samples. DNA has previously been successfully amplified from nicks [134] using other polymerase enzymes, so this section aimed to investigate the potential of ϕ29 MDA to produce full length amplification from nicks in DNA. Firstly, bioinformatic predictions of frequency of the nicking sites predicted that the number of nicking sites varied with the GC content of pathogen, with the high GC genomes being nicked more frequently. If the method was successful, an alternative enzyme or a combination of nicking enzymes may have been considered to even out the cutting frequency.

After incubation of nicked DNA with ϕ29, all libraries failed to amplify DNA (3.3.2). The control with nicked DNA and primers also failed to produce DNA. This may be due to enzyme

inhibition from the nicking buffer, or enzyme competition. A clean up using isopropanol was trialled in an effort to remove the inhibitor. There was still no DNA production from the nicked samples, however the control with nicked DNA and primers did produce DNA but at a lower level than the non-nicked sample with primers. Potential explanations include damage to the DNA or inhibition due to ethanol carry over from the clean-up. Another reason for failure could be that the nicking points are repaired before ϕ29 has a chance to bind. To overcome potential nick repairing co-nicking and amplification was investigated. The use of co-nicking and amplification would also have the added advantage of continually producing nick points for DNA production, thereby increasing amplification efficiency. The addition of a single stranded binding protein (SSBP) was also trialled in an effort to stabilise the nicked single stranded DNA. These reactions also failed to produce any DNA, potentially because ϕ29 is too large to bind in at the point of nicking, or ϕ29 actually starts replication upstream of the primer site. Previous investigation into the ability of ϕ29 to amplify from nicks also failed to produce DNA[134], strengthening the argument that the enzyme is not capable of producing DNA in this manner. Other strand displacement enzymes could theoretically begin the process of amplification from nicks, increasing the size of the single stranded DNA section allowing ϕ29 to bind and continue amplification

The first enzyme investigated was Klenow fragment (3.3.3.1), which has been previously used to produced DNA from nicks[97]. When incubated with nicked DNA, Klenow did produce low levels of DNA, with product sizes 15-300 bases, with the sizes of the DNA fragments produced slightly increasing with the addition of the SSBP. Only 24% of reads produced by Klenow fragment were above 300 bps and none were over 2000 bp (3.3.3.1). this was an improvement on previous uses of Klenow Fragment[135], where when it was used for TB amplification products of only 200bp were achieved. However, the addition of ϕ29 failed to produce any extra DNA. The control sample with primers was successful with ϕ 29, but at lower concentration than previous amplifications. This indicated that whilst the enzyme was able to produce DNA under these conditions they were not optimal for enzyme function.

The next enzyme investigated was Vent, which has known strand displacement activity, however it is normally used for thermocycling production of DNA, where there are high levels of secondary structure. Firstly the ability of the enzyme to perform isothermal amplification was investigated (3.3.3.2). No DNA was produced below 60^{O}C and at 60 ^{O}C only very low concentrations of DNA was detected. There was an increase in DNA production above 70^{o}C, suggesting that the enzymes strand displacement activity is only active at high temperatures. Even at the higher temperature less DNA was produced, (2.98 ng/µl) compared to the thermocycling control (12.3 ng/µl). When investigating the extension from nicks a small amount

of DNA was produced above 80°C, however this is probably down to the DNA denaturing at nicking points. Inclusion of SSBP had little effect on DNA production using $Vent_R$. There was a small increase in the concentration of DNA when the DNA was incubated with ϕ29 for six hours; however, this was most likely low level background amplification. Although this enzyme's strand displacement ability was previously known[141], this study provides the first evidence of isothermal amplification with this enzyme. The increased ability to produce DNA through the stand displacement mechanism is in line with the findings of Kong et al [141], who described increasing strand displacement activity at higher temperatures.

The final enzyme investigated was *Bst*, which is commonly used in stand displacement amplification in DNA production using Multiple Annealing and Looping Based Amplification Cycles (MALBAC) approach. *Bst* was able to produce DNA from nicks in isothermal conditions at 50°C, 60°C, and 70°C, and to a lesser efficiency at 40°C (3.3.3.3). Combing *Bst* with SSBP had little impact on DNA input. Addition of ϕ29 to *Bst* reactions produced less DNA than *Bst* used alone, suggesting that the working conditions of the two enzymes are not compatible. Similarly to Klenow Limitations in the length of DNA produced by *Bst* from nicks have previously been described, with evidence of unspecific amplification when products of greater than 600 bp are amplified[142]

Additionally previous work suggested that amplification from nicks using Klenow and *Bst* was very sensitive to nicking enzyme concentration, success here suggests that balance of nicking enzyme and amplification enzymes are near optimal and limited improvements could be made[134].

In conclusion, ϕ29 is not able to produced DNA from nicks in DNA, even in combination with other strand displacement enzymes. One explanation for the failure of ϕ29 to amplify from nicks is the re-sealing of nicks when the reaction is cooled to 30°C. However, the most likely explanation is the failure of ϕ29 at the nicking point, either as it is too large, or its initiation point is downstream of the primer point. Another enzyme known as Sequenase, has been shown to produce product lengths of up to 5000bp[134]. However this enzyme is a modified T7 polymerase which has a higher error rate and a lower processivity than ϕ29[141] and the aim of this study was to rapidly produce long lengths of DNA with high fidelity

3.11.4 DNA Production from RNA and DNA using DNA Tagging

Addition of tags could aid in stabilisation of the 3' ends of DNA and RNA, allowing the production of more full length transcripts. There is the additional potential to include unique identifiers in the tag to allow identification of initial mixes in samples.

Initially small (330 base) amplicons were used to check for tag attachment (3.4.2), which resulted in an increase in the peak size when visualised on the Bioanalyser. When this tagged DNA amplicon was incubated for four hours with varying size tags attached, the tag with a five base pair single stranded section failed to produce DNA. The tag with a 10 base pair section produced DNA (0.497 ng/µl), and the 15 base pair tag produced less (0.311 ng/µl) showing there was successful amplification from these tags. The ten base overhang DNA tags were the most efficient at producing DNA, indicating ϕ29 needs between 6 and 10 bases to bind and initiate replication. The 15 base pair tag was less successful suggesting that the tag was less stable. There were small increases in the concentration of DNA in all reactions (including negative control) after six hours of incubation, suggesting background non-specific amplification. When tags with RNA segments were used the results were similar, but with less DNA production in all cases, implying the secondary structure is less stable than with DNA.

When the tags were attached to longer DNA fragments (3.4.2.2) similar results were found, with failure to amplify from 5 bases, and 10 base tags giving the best outcome. When the fragments were investigated the products were longer than the input sizes, suggesting either chimera formation or the secondary structure from replication caused altered DNA movement through the Bioanalyser chip. There was less DNA produced using tags than primers, as there is only a single starting point in each fragment as opposed to many random primer binding points. From these experiments, it is possible to infer that, for optimal initiation of replication, ϕ29 needs a gap greater than 5 bases and less than 15 bases. There are some similarities between this method and Loop mediated isothermal amplification (LAMP), which uses a stem-loop feature for DNA extension[92]. However, LAMP requires primer targets and limits the size of DNA produced, and so is not suitable for the rapid production of untargeted DNA.

When tagged RNA was incubated with reverse transcriptase it failed to produce any cDNA. Reasons for this include failure to attach the DNA tag to the RNA target, however T4 ligase has better activity to join RNA and DNA [143]. Alternatively, the RNA target may not have been dephosphorylated, or SuperScript may have failed to initiate reverse transcription as the tag is made from DNA. A reverse transcription LAMP protocol exists, suggesting the possibility of reverse transcription from stem-loops[144], however this technique is based on using MMLV reverse

transcribing enzyme which has a higher error rate and a lower working temperature. Additionally, as reverse transcription is the first stage, it mainly occurs prior to stem-loop formation.

Overall there was no cDNA production using the tagging methods, and RNA genomes had the most to gain from the stabilisation of 3' ends, which are often missing from sequencing data. The aim of this project was to develop a rapid method of producing sequencing data from a variety of inputs with a single method. The DNA production is also less efficient as there are less starting points, and the DNA needs to be fragmented prior to tag attachment. However, these tags could be used to provide accurate quantification of the starting mixes in complex samples, by adding an index to each original DNA fragment before a longer amplification time was used.

Overall use of random primers was the most efficient way to produce DNA. When low starting material is used multiple priming points across the genome will allow rapid production of DNA without prior knowledge of the pathogen. However, this will also produce non-specific DNA without the presence of a template and so use of quantification of DNA as a measure of positivity would not be suitable.

3.11.5 Development of ϕ29 MDA for Whole Genome Sequencing

Often ϕ29 MDA is used alongside targeted techniques to increase the sensitivity of PCR[140,145] or micro-arrays[146]. Here the aim was to produce a library suitable for whole genome sequencing.

The first attempt to sequence a library prepared using ϕ29 MDA **(3.5)** failed, producing no sequencing beads after enrichment. The enrichment process removes beads which have no amplified target bound, leaving only those beads which have successfully amplified the library attached. There are three potential reasons for no beads remaining after enrichment, either the DNA failed to attach to the beads, the library was inefficiently amplified once on the beads, or the library completely failed to amplify during the EmPCR stage. In order for the DNA to attach to the beads it must have the correct adapters attached to the DNA fragments. During the library quantification it is these ligated adapters that are used to quantify the library. Any unligated adapters should be removed during the library size selection and clean up so only successfully ligated adapters would be visualised during quantification. Therefore, any unattached adapters will not be quantified, thus the most likely cause was problems during amplification once the DNA was attached to the beads. The lengths of the DNA strands after library preparation were slightly larger than the recommended maximum length, which could have potentially caused poor amplification of the product (due to a finite amount of reagents such as dNTPS and magnesium ions). However, there was a variety of lengths of DNA produced, with some of them being within the recommended length, and so this would have produced a poor yield of enrichment beads

rather than a total lack of enrichment beads. Therefore, the most likely reason for failure was lack of amplification of library once on the beads. The amplification reaction on the be ads is performed using a *Taq* polymerase, which has no strand displacement ability. The highly complex DNA structure produced by the ϕ29 MDA amplification enzyme may interfere with the ability of the *Taq* polymerase to amplify the target DNA. There is a denaturation stage prior to addition of the DNA to beads, but this may be insufficient to overcome all of the secondary structure of the branched DNA.

To attempt to overcome the issue of failure to produce libraries, two additional steps were incorporated into the library preparation. The first additional stage was to add an S1 nuclease reaction (3.5.1), which cuts DNA at points where it is single stranded; this will remove the branches from the DNA product and lower the overall complexity of the DNA product. Removing branches also makes DNA quantification using probes more accurate because of the lowered complexity of the DNA structure. There have also been suggestions in the literature that use of S1 nuclease lowers the number of chimeric reads produced[147,148].

The second step to ensure the success of sequencing runs was, to increase fragmentation of the DNA produced by ϕ29 MDA after S1 nuclease digestion. The reads produced using ϕ29 MDA amplification are very long, and so additional nebulisation was needed to reduce the DNA fragments to a suitable size. Two fragmentation methods were investigated, the use of the Fragmentase enzyme, and physical shearing via nebulisation.

Use of the Fragmentase enzyme after S1 nuclease treatment found that both the 10 and 15 minute incubations produced DNA that was too large to quantify on the Bioanalyser (>10380 bp) and even the 20-minute incubation produced most fragments above this point. When the two enzymes were combined in a reaction no fragmentation occurred. This inefficiency of the Fragmentase may be because of the very long lengths of the ϕ29 MDA, or suboptimal enzyme conditions due to carry over of ϕ29 MDA buffer. One possible solution would be to clean-up to DNA between S1 treatment and incubation with Fragmentase. However due to the large size of products produced during ϕ29 MDA most DNA clean-up techniques aren't suitable, column based clean ups would fail to allow large DNA fragments elute through the column, and so to clean-up a method such as isopropanol would have to be used. The additional reaction time required for the enzyme fragmentation compared to the nebulisation (>30minutes compared to 2 minutes) plus the extra clean-up stage required, makes the physical shearing a better option. Although enzymatic methods are preferred for some applications, as less DNA is lost during the process, this is not a problem when using ϕ29 MDA as so much DNA is produced. Nebulisation produced

fragments of a suitable size for sequence library preparation and also has the advantage that there is less bias in the points at which the DNA is fragmented as the process is random and enzymes have a target. This is particularly important when looking at *de novo* assemblies as DNA will be fragmented at different places. When looking at pathogens at the extremes of GC content, cutting bias will be increased. In conclusion the best library preparation for ϕ29 MDA prepared DNA is a combination if S1 nuclease to remove braches and reduce chimers, and extended physical shearing.

3.11.5.1 Reproducibility of ϕ29 MDA in Whole Genome Amplification

After successfully sequencing libraries produced using ϕ29 MDA, the reproducibility of the methods was investigated in terms of the amount of DNA produced during the ϕ29 MDA reaction, the amount of reads produced by the sequencing and the amount of the reference that was covered showed no significant difference. This confirmed that the method is robust for DNA amplification and sequencing and is suitable for whole genome amplification and analysis

3.11.6 Comparison of ϕ29 MDA to Current Methods of Bacterial Sequencing

The sequence data produced by the ϕ29 MDA amplification method was compared to the most commonly used (culture based) non-amplification method in section 3.6.1. This showed very similar sequencing outputs, verifying the good quality of DNA produced. There were small differences in the sequencing output which were most likely due to sequencer variation and problems with equipment overheating. This had no impact on the amount of the genome which was covered with both methods covering the same percentage of the reference genome, with the average ϕ29 MDA reference assembly coverage of 97.88% vs. average non-amplification reference coverage of 96.87% **Figure 3-8.** When the reads were assembled both the reference and *de novo* assembly produced fewer and longer contigs using ϕ29 MDA libraries. The number of misassemblies in the *de novo* assemblies was also lower in the ϕ29 MDA libraries. This could be down to better quality DNA input, as the extraction method used for the control libraries used columns and may have caused shearing and DNA damage. The ϕ29 MDA library included more reads that mapped to the plasmid than the culture control (Figure 3-9). This could either be caused by over amplification by the ϕ29 MDA of the plasmid DNA or poor recovery of the plasmid DNA during the control extraction. The results of the ϕ29 MDA assembly for the plasmid created fewer and longer contigs, likely due to the additional reads that were available to use.

Detection of the lambda phage was lower in the control, most likely due to the same reasons as the lower plasmid representation. The amount of human contamination was similar in both

sets of libraries (0.14% average in ϕ29 MDA and 0.15% in the control libraries) suggesting the contamination was post amplification. There was a higher number of reads classified as 'other' or not mapped in the non-amplification library, which could be due to the samples being extracted in a none 'PCR clean room', low level contamination of the extraction kit, or water used during the extraction process, meaning residual low quality DNA was present.

Over all ϕ29 MDA amplification was shown to be a robust method to amplify the whole genome of *E. coli* (with a GC content ~50%), the results were reproducible and comparable to the current method. The method was also found to be suitable for detection of extra chromosomal elements including phages and plasmids.

Most uses of ϕ29 MDA have been for high sensitivity detection of un-culturable bacteria[145,147], or used in combination with targeted techniques[140149]. This is the first time that the method has been investigated thoroughly and compared to traditional culture based techniques for the production of sequencing libraries.

3.11.7 Determining the Sensitivity and Processivity of ϕ29 MDA amplification in the Context of Bacterial Genomes

When the volume of the reaction was lowered in section **3.7.1** the amount of DNA produced in a ϕ29 MDA reaction was also reduced, probably due to less input reagents especially dNTPs causing premature termination of the reaction. However, in all cases there was still sufficient DNA produced to sequence the products.

The sequencing output in terms of reads produced was similar in the reduced volumes compared to the full reaction volumes. However significantly fewer of the reads passed filter (47% in the 12.5 µl reaction compared to 77% in the 50 µl reaction), perhaps because the DNA produced was of poorer quality in the low volume reaction. After reference assembly significantly less of the reference genome was covered in both reduced volume reactions (61.97% and 47.36% compared to 97.88%) this could be because of incomplete ϕ29 MDA reactions, producing shorter DNA reads. The *de novo* assemblies also performed poorly with data resulting from reduced volumes.

Overall the lowering of the reaction volumes produced poor results, suggesting that the reaction volume was already optimal. The poor amplification would also have been exaggerated when the starting input was lowered. Low volume pipetting is difficult and with the input volumes

being just 1.5 µl or 0.75 µl, this may have been a factor. When working from clinical samples it would be technically difficult to concentrate the input to such low volumes. Therefore, it would not be appropriate to lower the volumes below the original 50 µl.

The amount of DNA produced (3.7.2) when the input was lowered to a single cell was equivalent to that produced when an entire colony was input to the reaction. There were also very similar results after sequencing, assembly and mapping. Confirming that the ϕ29 MDA method is extremely sensitive and suitable for use on single bacterial cells. Application of this method to clinical samples has the potential to detect pathogens in samples at lower levels than some currently used molecular methods, in a way that doesn't require prior knowledge of the target.

In section 3.7.3 ϕ29 MDA incubation times were reduced from 16 hours to a lowest incubation time of 1 hour. As shown in figure 4.4.3.5 the average amount of DNA produced by the ϕ29 MDA reaction fell as incubation times fell, which would be expected with reduced time for the reaction. However, until 1-hour incubation, all reduced time incubations consistently produced sufficient DNA to sequence. When using the Junior 454 system high amounts of input DNA are required (500 ng) compared to other platforms such as Illumina where as little as 1 ng is required. So if this method is sufficient for Junior library preparation it will be sufficient for other, more sensitive platforms.

The proportion of the reference covered was also consistently high until the 1-hour reaction library. When assembling the data, the number of contigs in both reference and *de novo* assemblies were lower as the incubation time decreased, until 1-hour incubation when the number of contigs increases significantly. This lowering of contig number could suggest that the assembler finds the data easier to deal with, perhaps due to lower complexity of the DNA input. The poor performance of the one-hour incubation could be because the reaction has had insufficient time to copy all of the genome.

Overall two hours' incubation time seems to be the optimum time to produce whole genome amplification using ϕ29 MDA from a bacterial input. It provided consistently high DNA concentrations with good coverage of the genome. This rapid time to whole genome production is comparable to turn around times of commonly used PCRs in the routine laboratory. Most studies utilising ϕ29 MDA have incubated samples for a least ten hours [140,149,150], here the incubation time has been successfully reduced to allow application in a clinically relevant time.

When using the term single cell in this thesis, it is an estimate based upon dilutions and CFU counting of bacteria, and so more bacteria than a single cell may be present.

3.11.8 ***Assessing ϕ29 MDA for Whole Genome Amplification of Varying GC Contents***

C. difficile has a very low GC content of 29%, and so was used an example of extreme GC content. Reproducibility was good across the three ϕ29 MDA replicates **(3.8.1),** with the amount of DNA and sequencing reads produced being constant. When comparing the use of ϕ29 MDA and non-amplification control in *C. difficile* no significant difference was found between the sequencing output, however there was slightly higher % genome coverage resulting from the ϕ29 MDA library (95.73% compared to 88.66%), and the assemblies contained fewer and longer contigs than the control assemblies. Similarly, to the *E. coli* this could be due to better quality of DNA input due to DNA shearing during spin-column extraction. However, one of the control runs did produce significantly more contigs in the reference assembly than the other, with lower genome coverage after reference assembly. This may have been down to the sequencer over heating during the sequencing run.

There were significantly fewer assembly errors in the *de novo* assembly of the ϕ29 MDA library. These factors would suggest that for extremely low GC pathogens ϕ29 MDA amplification provides better assembly results than non-amplification methods. No extra chromosomal elements were detected either in the non-amplification or ϕ29 MDA libraries, with either *de novo* or reference assembly techniques. The results of whole cell lysis and specific PCR were also negative. These results suggest that the expected plasmid was not present in the subculture of *C. difficile* used in this study. The plasmid in this strain of *C. difficile* 630, has no functions assigned to any of its CDSs [136], and so the plasmid could have been lost during sub cultures without any obvious loss to functionality.

When looking at the reads that did not map to *C. difficile* chromosome, there was a higher proportion of reads not mapping in the ϕ29 MDA samples. In two runs this only accounted for less than 1.5% of the reads, compared to an average of 0.52% in the control libraries, with these reads mainly mapping to either the genus Clostridiales (MDA1) or non-assigned reads. The third sample had a higher proportion of unmapped reads (6.41%). With 2.35% of these mapping to Enterobacteriaceae and 1.21% mapping to human, suggesting cross run contamination, most likely after amplification. There were also 2.84% of reads that were unmapped, suggesting degraded DNA contamination.

Overall the ϕ29 MDA approach seems appropriate for a low GC content genome, with better results being produced than the non-amplification control.

Actinomyces naeslundii has a high GC content at 68.5% and was selected to represent the other end of the GC content scale to *C. difficile.* Firstly the amount of DNA produced by the ϕ29 MDA for the *A. naeslundii* reaction was the same as that produced by the other reactions **(**3.8.2). The sequencing output was also similar both between runs of *A. naeslundii ϕ29* MDA and *A. naeslundii* non-MDA control libraries. The ϕ29 MDA sequencing performed better in the reference assembly, there was low reference coverage for both the ϕ29 MDA and the control for *A. naeslundii.* The ϕ29 MDA covered 76% of the genome and the control covering only 63.39% of the genome, additionally the total *de novo* assembly length was less than the reference sequence length. Both sample preparation methods performed worse than expected suggesting that the problem doesn't lie with the amplification but perhaps with either the sequencing technology or the assembly, either due to a poor reference or the programme struggling with the high GC content. Poor assembly could be down to the assembly algorithm used or the repetitive nature of the genome or because significant differences in genome content between the strain sequenced in this study and the published reference genome.

Overall the fact that both the control and ϕ29 MDA libraries both performed similarly, any problems that are associated with the high GC content of the pathogen lie after the amplification and not with the ϕ29 MDA process itself. Significant work has been carried out using ϕ29 MDA to investigate organisms, and complex ecological and evolutionary questions [151]. Here, the goal was to determine the robustness of the method for a variety of clinical organisms. Previous work was expanded to determine the robustness of the method for a variety of clinical organisms

3.11.8.1 Impact of GC on MDA and Sequencing

When comparing the performance of reference assemblies across the range of GC studied the first difference to note is the percentage of reads assembled against the reference was higher in *C. difficile* than *E. coli*, however this was mainly accounted for when considering the extra chromosomal elements detected. In contrast the number of reads mapping to the reference sequence of *A. naeslundii* was lower than *E. coli*, in both the control and ϕ29 MDA library, again suggesting that this was down to post amplification problems. However overall GC content doesn't appear to affect the ability of ϕ29 MDA to amplify whole genomes.

3.11.9 *Application to Low Level Mixed Bacteria*

To further investigate the scope of ϕ29 MDA for pathogen sequencing mixed bacterial cells and a viral pathogen were amplified and sequenced.

The results from the even ratio of cells show that outcome genome coverage is equal for both of the starting bacteria, however the percentage of reads mapping is different, with more reads mapping to the *E. faecalis* genome. When the total genome sizes of the bacteria (*E. faecalis* 3.2MB, *H. influenzae* 1.8MB) were factored in, the number of reads indicated an equal split per base of input genome. A low level mix of two bacteria were successfully amplified and sequenced, producing reads roughly equal to input ratios. Detection limit was found to be at 1:1000 ratio, however this was more likely down to the sequencing platform, with higher sequencing depth, a lower limit of detection is expected. Our findings indicate that ϕ29 MDA may be suitable for relatively simple communities, while some studies have reported that ϕ29 MDA is suitable for investigating complex communities like the cystic fibrosis lung[152]. On the other hand, it has also been reported that sequencing resulting from ϕ29 MDA prepared libraries does not reflect the initial abundance[153]. This application would require further testing before it could approach the clinic.

3.11.10 Amplification of Viral Genomes

Good results were achieved when sequencing Adenovirus 40 and 41 (3.10), with complete coverage of the genome after reference assembly being achieved in a single contig even when using a low percentage of the reads. Improvements in the proportion of reads used for the reference assembly were achieved through the use of DNAse on the sample prior to extraction. Adenovirus is a double stranded DNA virus, and was a starting point for the application of ϕ29 MDA to viral genomes.

To investigate the applicability of ϕ29 MDA as a sensitive method across multiple pathogen types, mixes of RNA and DNA viruses were investigated. This was used to explore the ability to perform reverse transcription and ϕ29 MDA on the same extract to produce good quality genome data. Initially viral extracts were used, with roughly the equivalent of 25 virus particle input; both viruses were successfully detected from this sample. The results showed that 66.2% of the reads were assigned to adeno virus and 20.76% assigned to HIV. When these numbers are adjusted for input genome size, the ratio of reads is 1:1.19 (adeno: HIV). HIV carries two copies of its genome in each virion, suggesting that there is inefficiency in the RNA conversion, or that some RNA

degrades during the process. When whole viruses were concentrated, extracted and amplified, the number of reads assigned to adenovirus was 62.7% and HIV had 20.76%. When this was normalised for genome size, the ratio was 1:1.7, which was an improvement from the extracted sample. However, it was still not at 1:2, suggesting there is some loss during the reverse transcription process, or inaccuracy in quantifying the input. Improvement may be down to low stability of extracted RNA, and show that extraction and RNA conversion should be performed in quick succession to provide highly sensitive outputs. When *de novo* assembled, high genome coverage was achieved for both viruses (97.1% and 99.90%). Annotations from both of the low input viruses were equivalent to the reference genome. The data produced from the HIV was of good enough quality to use a publically available tool (REGA version 3) to allow accurate tying of the virus.

A mixed viral sample was then developed using a variety of viruses, including multiple viral particle types and assortment of viral genome types. This allowed the testing of the concentration, extraction and amplification techniques. An equation was developed that allows the use of a known copy number spike to quantify viral mixes within a sample. This equation takes into account the large variety of genome architecture among viruses, including the nucleic acid type, genome size and genome copy number. A spike for an RNA and DNA virus was used to account for inefficiencies in the reverse transcription stage and the more fragile nature of RNA.

Much of the initial mixture had to be estimated, as only the Ct values were supplied which are not a direct quantification of the viral particles. A volume in scale with that used for other viruses provided was used to provide a starting point. After sequencing and reads assignment, VZV had a very high input estimation (2309), Norovirus and Influenza also had higher numbers all of which had been estimated from Ct values. Parvovirus 19 was also in higher numbers, the provided IFU range for this virus was 569-1058, which is a large range, showing a possible inaccuracy of the initial numeration method.

When the sample was recreated taking into account quantification results from the first sequencing run, all calculated viral numbers were closer to the 25 viral aim. Other than measles, which had a very low read assignment. Measles was the only virus that was provided inactivated; the inactivation method may have damaged the virus's viability. Alternatively, the inactivation method may have acted by 'fixing' the virus, making extraction more difficult. After the sample input was adjusted, the RNA viruses were quantified at a higher level than expected (Norovirus, HAV, HCV). Possibly suggesting HIV is a poor spiking standard, which would account for the higher numbers of other RNA viruses.

Overall the method was successful at sequencing all viruses (other than measles). *De novo* assemblies for most viruses recovered more that 94% of the reference genome. Influenza virus *de novo* assembly performed slightly worse (88%), this is perhaps down the poor performance of the assembler on segmented genomes. Also 3' ends of genomes are often missing as DNA will only be produced towards the 5' end of primer. In addition, an equation was developed to allow quantification of mixed viral input from a known spike. If this was used in clinical samples, a non-human pathogen for a DNA and RNA virus should be used.

3.11.11 Extraction using Alkali Method

The extraction method used uses a combination of SDS (Sodium dodecyl sulphate) which as a detergent solubilises the cell membrane and disrupts the cellular proteins, and NaOH which causes dissociation of double stranded DNA allowing simultaneous extraction and denaturation of nucleic acids. After extraction there is a neutralisation stage, which raises the pH of the solution using potassium acetate, which may allow some double bonds to reform[154]. However as genomic DNA is so large it would be unlikely that the entire genomic DNA re-annealed completely allowing binding of the small hexamer primers. Due to the strand displacement ability of the ϕ29 enzyme sections of reformed double bonds would still be copied. The selection of bacteria and viruses studied covered many cell wall and virion types, providing evidence that this extraction method is suitable for bacterial and viral pathogens.

3.11.12 Conclusion: Development of Non-PCR Amplification Techniques for Rapid and Sensitive Whole Genome Amplification of Pathogens

The use of ϕ29 MDA on bacterial and viral input is a robust and reproducible method for the production of whole genome amplification from as little as a single bacterial cell. Before sequencing can be performed the secondary structure of the amplified DNA needs to be reduced though the use of S1 nuclease, and extended physical fragmentation. The amplification method copes well with varied GC content input and in mixed cultures results in sequencing data that reflects the initial mix. The starting volume of 50 µl appears to be the optimum volume for bacterial amplification and the method is successful with only two hours of incubation at 30°C.

However, there were some issues with the use of *de novo* methods to assemble genomes with the genome covered by all *de novo* assemblies in all bacteria tested being lower than the reference mapping method. The aim of the project is to be able to use *de novo* methods to

identify and characterise unknown infections. There was also significant problem with the assembly of data produced from the high GC pathogen *A. naeslundii*, which requires further investigation, but the fact that both the *de novo* and non-amplification control performed poorly would suggest that the problem is external to the amplification process.

4. Results: Development of an Analysis Pipeline for data Generated by NGS of ϕ29 MDA Prepared Libraries Focussing on de novo Methods

In parallel with the development of a highly sensitive laboratory method, a robust analytical pipeline was developed to profile and characterise clinically relevant isolates. Analytical pipelines allow interpretation of NGS outputs converting data to information. Data produced by NGS technologies vary depending on the input genome type, library preparation techniques and the sequencing technologies themselves. There are multiple algorithms designed for the analysis of NGS data, each with different strengths. Using ϕ29 MDA to prepare libraries brings unique data characteristics, which need to be considered and evaluated when analysing the data, meaning default parameters may not be suitable. Therefore, to ensure that the best quality data outcomes are achieved, a specific pipeline will be tested and developed. The aim is to provide an automated analysis method to allow the identification and characterisation of pathogens that can be applied to unknown pathogens.

The first stage developed will be data preparation including removal of host and environmental contaminants, followed by investigation of the best error trimming algorithms for this data. This will include optimisation of quality cut offs, defining the best parameters for maintaining data whilst removing low quality data. A comparison will be made between two trimming algorithms, a window based and running sum algorithm. As well as comparing the impact of end and average trimming techniques

The average peak depth of ϕ29 MDA produced libraries was 129 whereas the non-amplification peak depth of 61. Reducing the peak depth of the sequencing data will reduce the amount of data the assembler needs to process and so will potentially speed up the assembly. A process known as digital normalisation[107] allows peak depth reduction without the need for a reference.

Multiple *de novo* assembly algorithms will then be tested to find the most accurate one for this data set, assessment will include the impact of GC content on assembly quality and investigation of assembly errors. The primary focus will be on using *de novo* assembly techniques which would allow a single pipeline to be used for unknown pathogens. The data will then be investigated for the accuracy of characterisation of genomes in this dataset, including virulence determinants and resistance genes.

4.1 Removal of Contaminants from Sequencing Data

Host and environmental contamination can be problematic for assembly, with the potential for false genome annotations, and so these were removed prior to error trimming and assembly.

4.1.1 Host Contamination

The adenovirus assemblies in **section 3.10** had low number of reads that mapped to the relevant adenovirus genome (average 23.51% for Ad40). Most of the unmapped reads belong to the tissue culture (393 monkey kidney) that the virus was cultured in. In one of these samples only 22.95% of reads mapped to the target Ad40 genome, when mapped against the host genome 69.24% of the reads mapped. After removal of unmapped reads, 229956 reads were put into a new SFF file, as described in **2.3.5.5.** When this new SFF was mapped against the Adenovirus 40 reference, 100% was covered using 72.35% of the reads. The remaining unmapped reads were extracted and compared to the Blastn nr library. When this was investigated, the remaining reads were either identified as Enterobacteriaceae reads (5634), *E. coli* (918) or unidentified (734). **Figure 4-1** shows the proportional mapping of reads before and after removal of host signals.

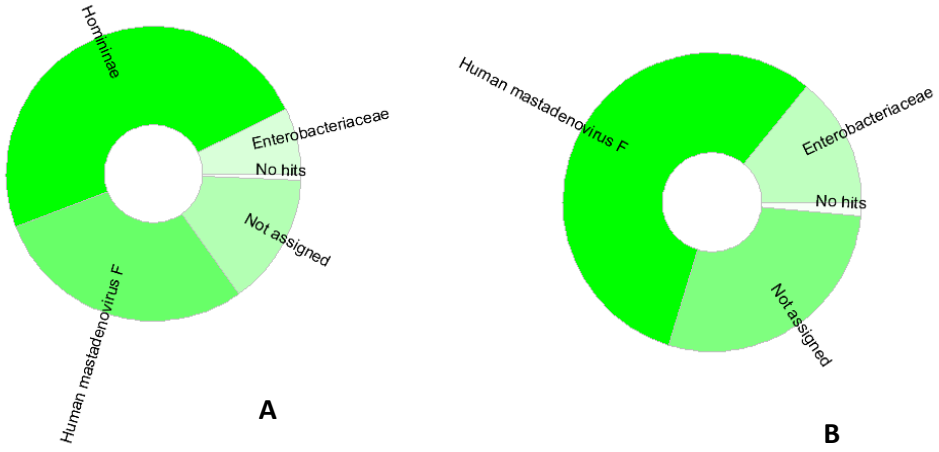


Figure 4-1 proportional read assignment of ϕ29 MDA prepared Adenovirus 40 sequence data before (A) and after (B) removal of host reads

4.1.2 Removal of Kit Contaminants

In addition to the expected host contamination, the lack of specificity and high sensitivity of the ϕ29 enzyme enables the amplification of any kit or environmental contaminants. Failure to acknowledge and remove these contaminants can lead to erroneous data interpretation. Negative libraries were frequently amplified and analysed to investigate the presence of contaminants.

4.2 Negative Library

The output files from negative runs were concatenated into a single SFF file, which was then converted to a fasta and a Blastn search performed using the non-redundant nucleotide database. The results were viewed in MEGAN5.

The majority of reads (78.97%) had no hits to the nucleotide library. However several bacterial species were identified**, Figure 4-2**. One major represented Genus was the Mycobacterium, with 13% of all reads mapping within this Genus.

A small number of reads mapped to significant human pathogens including *Haemophilus influenzae* (0.7%), *Salmonella enterica* (0.06), *E. coli* (0.03%), *Enterococcus faecalis* (0.04%), and *Staphylococcus* sp (0.24%). These reads were removed along with reads mapping to *Mycobacterium abscessus* (1.13%), other Mycobacterium sp. reads were left in the negative library. The resulting library was used to as a reference to remove contamination from subsequent sequencing runs

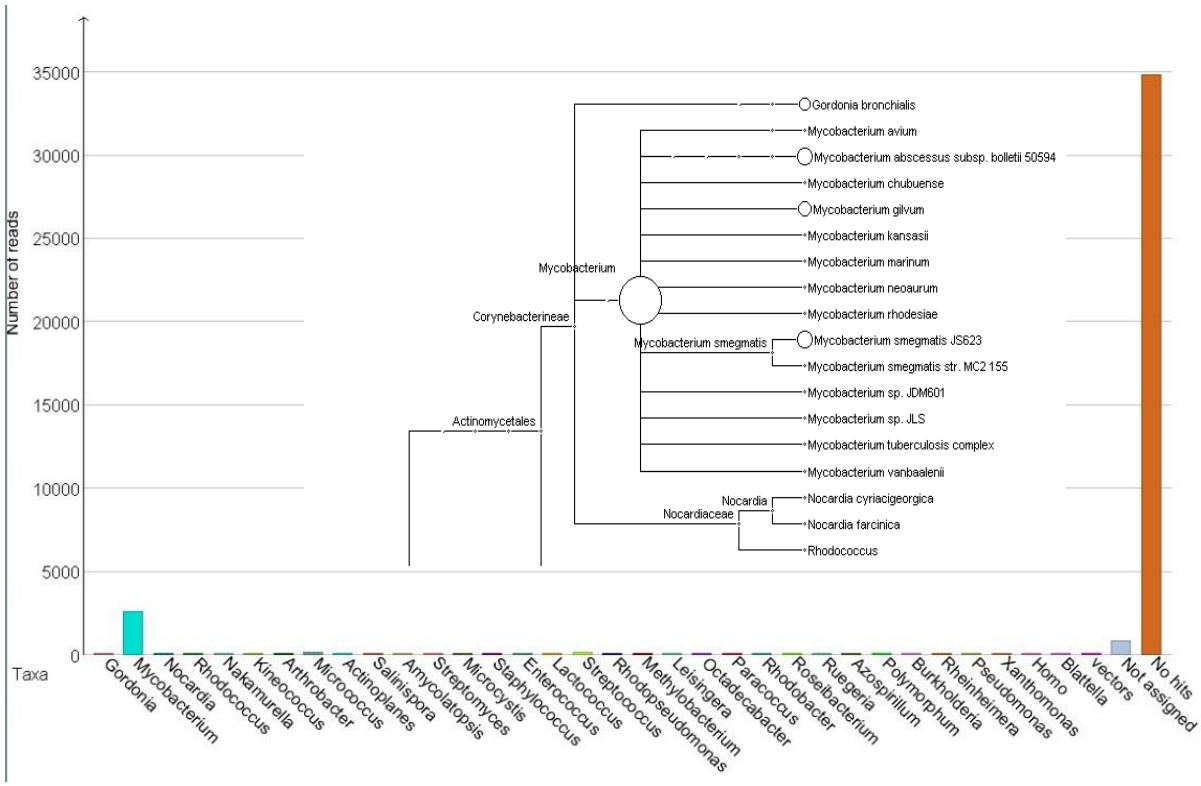


Figure 4-2 Species identification within negative ϕ 29 MDA sequence data using Blastn and LCA analysis, the insert shows Mycobacterium species identified within this library

4.2.1.1 Removal of Contamination from Sequencing Run

The negative library was tested by mapping the sequencing file produced by 16 hour ϕ29 MDA of a single cell, as it had a high number of unidentified reads (3.34% of all reads). The SFF file was mapped against the negative library, 0.98% of all the reads mapped to the negative library. With most of these mapping to reads that had no hits or were unassigned (92%) and a small number mapping to the Order Actinomycetales (1%), with a subsection identified as Mycobacterium sp **Figure 4-3**.

4.2.1.2 Improvement of Contamination Library

All unidentified reads from previous sequencing runs were added to the negative library. When this improved library was used as a reference in the same sequencing file, 1.26% of reads mapped to the negative library.

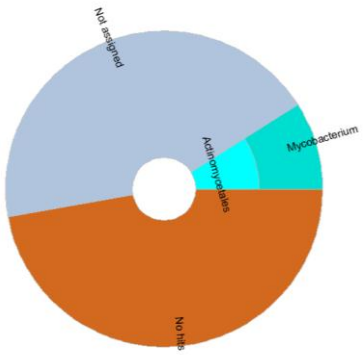


Figure 4-3 proportional identification of reads that map to the negative library

4.3 Quality Trimming of Raw Reads

One *E. coli* k12 single cell sequencing run was randomly selected to investigate the impact of different trimming metrics using two algorithms, a window based and a running sum error trimming programme.

4.3.1 Raw Data

The number of raw reads that passed standard on-instrument filtering was 150,144 with the total number of bases of 79,359,696. The range of length of reads was 48 to 1200 bases with an average read length 529 bases. When this was assembled against the reference 98.49% of the genome was covered using 91.28% of the total reads into 242 contigs. The average contig size was 19,514 bases with the largest contig being 220,401 bases. When the reads were *de novo* assembled a total of 1692 contigs with 1239 (73.23%) of these being over 1000bp. The largest contig was 32,639 bases long with an N50 value of 4554, which covered 94.12% of the genome with a total of 20 misassemblies.

4.3.2 Tag Removal

After tag removal there was no impact on the read number, but as expected a small decrease in the total number of bases (99.23% compared to the raw data) **Table 4-1**. The lengths of the shortest, longest and average reads decreased by the expected four bases, confirming removal of only the tag sequence. The reference assembly genome coverage dropped slightly to 98.47% and the contig number increased marginally to 245. Conversely the genome coverage of the *de novo* assembly slightly increased, as did the N50 whilst the number of contigs increased. There was no impact on the number of misassemblies in the *de novo* assembly. The same results were obtained using both trimming algorithms.

4.3.3 End quality Trimming

4.3.3.1 Prinseq

After removal of sequencing tags the 3' and 5'ends of the reads were trimmed using Q35, Q30, Q25 and Q20 as cut offs, trimming outputs along with assembly information can be found in **Table 4-1.**

Quality trimming the ends of reads had little impact on the total number of reads; however, there were reductions in the total number of bases remaining after trimming. The strictest trimming parameter (Q35) removed 38% of all bases, as the trimming parameters became less strict fewer bases were removed, with the Q20 data only removing 9% of the bases. After end quality trimming the minimum read length was reduced to a single base pair in all conditions other than Q20 trimming. The longest read length was considerably reduced in the Q35 trimmed data, where the longest read was reduced to just 39% of the raw data read length. The impact was lessened as the strictness of trimming was reduced, with the Q20 trimmed data having a longest read length at 96% of the raw data **Figure 4-4-A**. This method of trimming also had an impact on the average read length, again with a similar pattern of Q35 end trimming reducing the mean read length to most (to 65% of raw) and Q20 reducing it the least (91% of raw).

End trimming had very little impact on the genome coverage following reference assembly. However the contig number increased to more than twice the number following Q35 end trimming **Figure 4-4-B**. Again this affect was lessened as the strictness of trimming was reduced, with the Q20 actually having one fewer contig than in the data following only tag removal. Contig sizes reflected the changes in the contig number, with higher contig numbers leading to shorter contig lengths.

Following *de novo* assembly the reference coverage was reduced, with the Q35 sample coverage falling to 90.93%. As the trimming strictness was reduced the reference coverage increased with the Q20 trimmed sample showing similar reference coverage to the raw sample (Q20 94.85% vs. raw 94.91%). Contig number was lower in all the end trimmed samples with fewer contigs being produced when the trimming parameters were less strict, following this as the contig numbers fell, the contig length increased **Figure 4-4-C**. There was a slight declined in the misassemblies in all trimmed samples.

	Raw	tag	Q35	Q30	Q25	Q20
No: Reads	150,144	150,144	149,589	150,098	150,144	150,144
No: Bases	79,359,696	78,750,120	49,426,937	58,191,182	65,017,487	72,283,905
Shortest read (bps)	48	44	1	1	1	45
Longest read (bps)	1200	1196	479	577	607	1156
Average read length (bps)	528.5	524.5	330.42	387.69	433.03	481.43
Reference assembly						
Ref coverage (%)	98.49	98.47	97.5	98.11	98.4	98.47
No: Contigs	242	245	584	375	264	244
Average contig size (bps)	19,514	19,345	8,464	12,802	17,959	19,427
Largest contig size (bps)	220,401	220,385	72,143	93,890	184,905	220,387
***De novo* assembly**						
Ref coverage (%)	94.12	94.91	90.93	93.30	94.57	94.85
No: Contigs	1692	1600	1312	983	834	804
N50 (bps)	4554	11356	7043	10787	12422	13863
Largest contig (bps)	32,639	53191	41488	46088	71924	72324
misassemblies	20	20	17	15	14	15

Table 4-1 Output of end quality trimming using Prinseq at four different cut offs. Including basic read information and results of reference and *de novo* assemblies. tag=tag removed data, Q number refers to the quality cut off used to trim the ends of data

4.3.2.1. Cutadapt

The sequencing tags were removed and the 3' and 5' ends of the reads were trimmed to Q35, Q30, Q25 and Q20, the samples were then reference and *de novo* assembled **Table 4-2**

Post end trimming using Cutadapt, the number of reads remaining was reduced, with only 88% of the raw read remaining in the Q35 data. As the trimming became less strict fewer reads were removed with 99.9% remaining after Q20 trimming. The number of bases was reduced to 39% after Q35 trimming, and again as trimming strictness was reduced the proportion of bases remaining increase, with 78% of bases retained after Q20 trimming. The shortest read length in all trimmed samples was one base. The longest read length was reduced to 479 bps (39% of raw length), in the Q35 trimmed sample. The longest read length increased as trimming strictness was reduced, with the Q20 sample's longest read being 877 bps (73% of the raw length) **Figure 4-4-A**. A similar pattern was observed in the average read lengths with the Q35 sample showing the shortest average read length at 237.43 bps and the Q20 having the longest at 412.71 bps (compared to the raw average read length of 528.4 bps).

Following reference assembly of the data the genome coverage was reduced, with the most reduction seen following the strictest trimming (94.47% genome coverage), and genome coverage increased as trimming decreased. Contig number increased in all samples, with the highest number found in the Q35 trimmed data, which had more than seven times the number of contigs compared to the raw data. Contig number declined as the trimming strictness decreased with the assembled Q20 data consisting of 307 contigs (242 contigs in raw data) **Figure 4-4-B**. As expected the higher the number of contigs produced in an assembly the shorter the contigs lengths became.

Once the reads were *de novo* assembled, reference coverage was lower in all samples when compared to the tag trimmed data. The Q35 data showed the lowest reference coverage at 75.24% and coverage increased as trimming decreased. The Q20 data having comparable genome coverage (94.29%) to the tag trimmed data (94.91%). The number of contigs in the assembled data was highest in the samples with the strictest end trimming. The Q35 and Q30 data had more contigs with shorter lengths than in the raw data. However the Q25 and Q30 trimmed data had fewer longer contigs than the raw data **Figure 4-4-C**. Misassemblies were lower in all trimmed data with the Q25 trimmed data having the least misassemblies.

	Raw	tag	Q35	Q30	Q25	Q20
No: Reads	150,144	150,144	132,763	143,601	148,920	150,023
No: Bases	79,359,696	78,750,120	31,521,918	42,227,367	51,823,657	61,823,657
Shortest read (bps)	48	44	1	1	1	1
Longest read (bps)	1200	1196	479	563	607	877
Average read length (bps)	528.5	524.5	237.43	298.06	348	412.71
Reference assembly						
Ref coverage (%)	98.49	98.47	94.47	96.81	97.82	98.30
No: Contigs	242	245	1737	874	501	307
Average contig size (bps)	19,514	19,345	3137	5743	9753	15549
Largest contig size (bps)	220,401	220,385	44108	72085	89217	93807
***De novo* assembly**						
Ref coverage (%)	94.12	94.91	75.24	87.29	91.70	94.29
No: Contigs	1692	1600	2671	1729	1213	896
N50 (bps)	4554	11356	1535	3754	7212	11087
Largest contig (bps)	32,639	53191	21254	29605	57573	71935
misassemblies	20	20	18	16	13	17

Table 4-2 Output of end quality trimming using Cutadapt at four different cut offs. Including basic read information and results of reference and *de novo* assemblies. tag=tag removed data, Q number refers to the quality cut off used to trim the ends of data

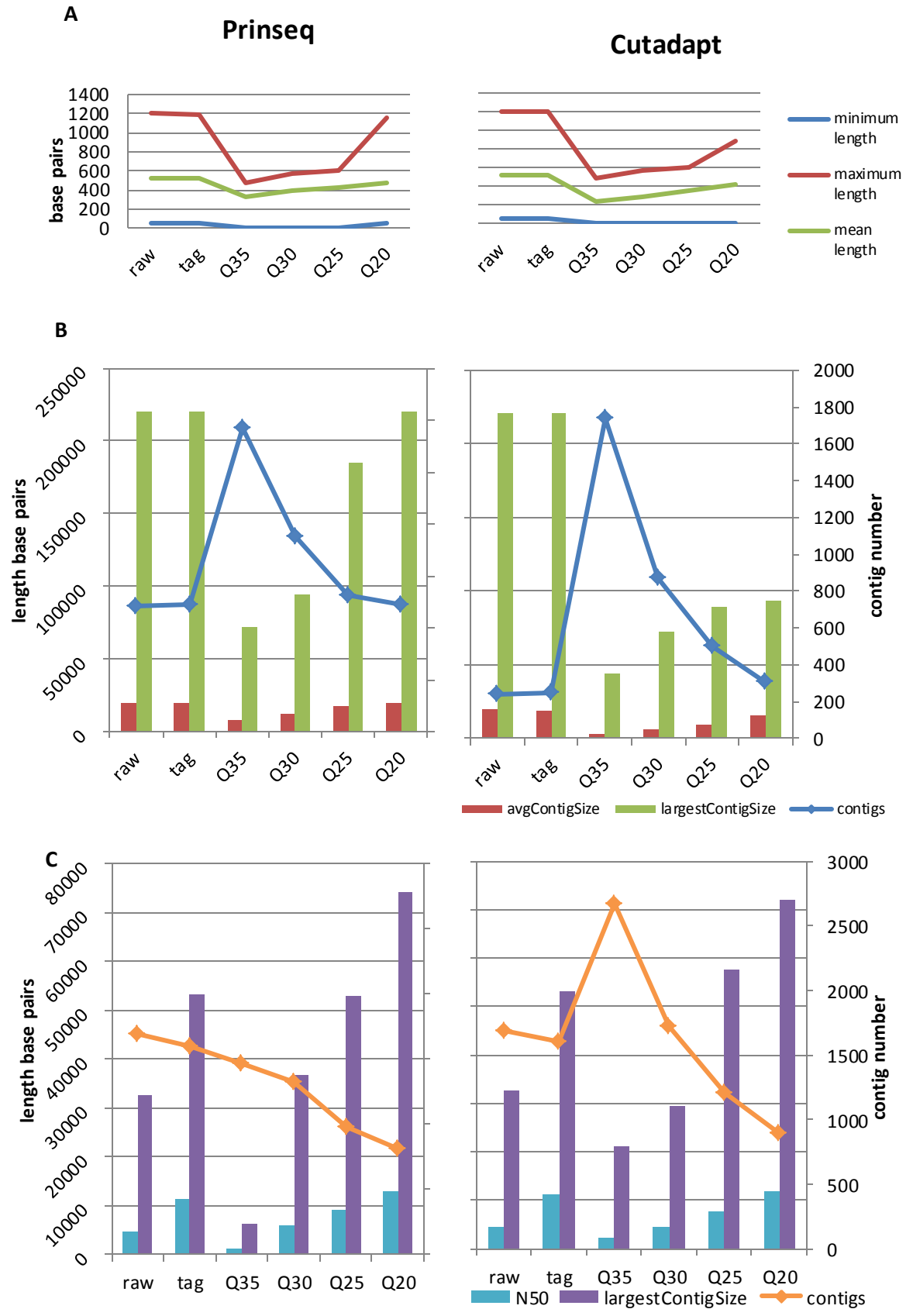


Figure 4-4 impact of end quality trimming at varying quality cut-offs using Prinseq and Cutadapt on (A) read length (B) reference assembly (number of contigs, average and longest contig length) and (C) *de novo assembly* (number of contigs, average and longest contig length)

4.3.2.2. Comparison of Widow Based and Running Average Trimming Techniques

Similar trends were observed in data following trimming using both algorithms, with a general trend showing more bases removed when a stricter cut off is used. Overall Cutadapt consistently removed more bases during trimming than Prinseq.

After data was reference assembled a similar trend was again observed in both algorithms, with strict trimming leading to the formation of more contigs with shorter lengths on average, and longer fewer contigs produced as the trimming became less stringent. When comparing the two algorithms Prinseq consistently produced fewer longer contigs than Cutadapt. Prinseq trimmed data also covered a higher proportion of the reference genome, with the affect being amplified as trimming became stricter.

Following *de novo* assembly very similar patterns were seen, with more contigs produced with stricter trimming. The major difference between the two algorithms was that all trimmed data produced fewer contigs when Prinseq was used, whereas Q35 and Q30 trimmed data using Cutadapt produced more contigs. Similarly, to the reference mapping a result, higher genome coverage was achieved when Prinseq trimmed data was used, with the affect being more obvious as trimming strictness increased.

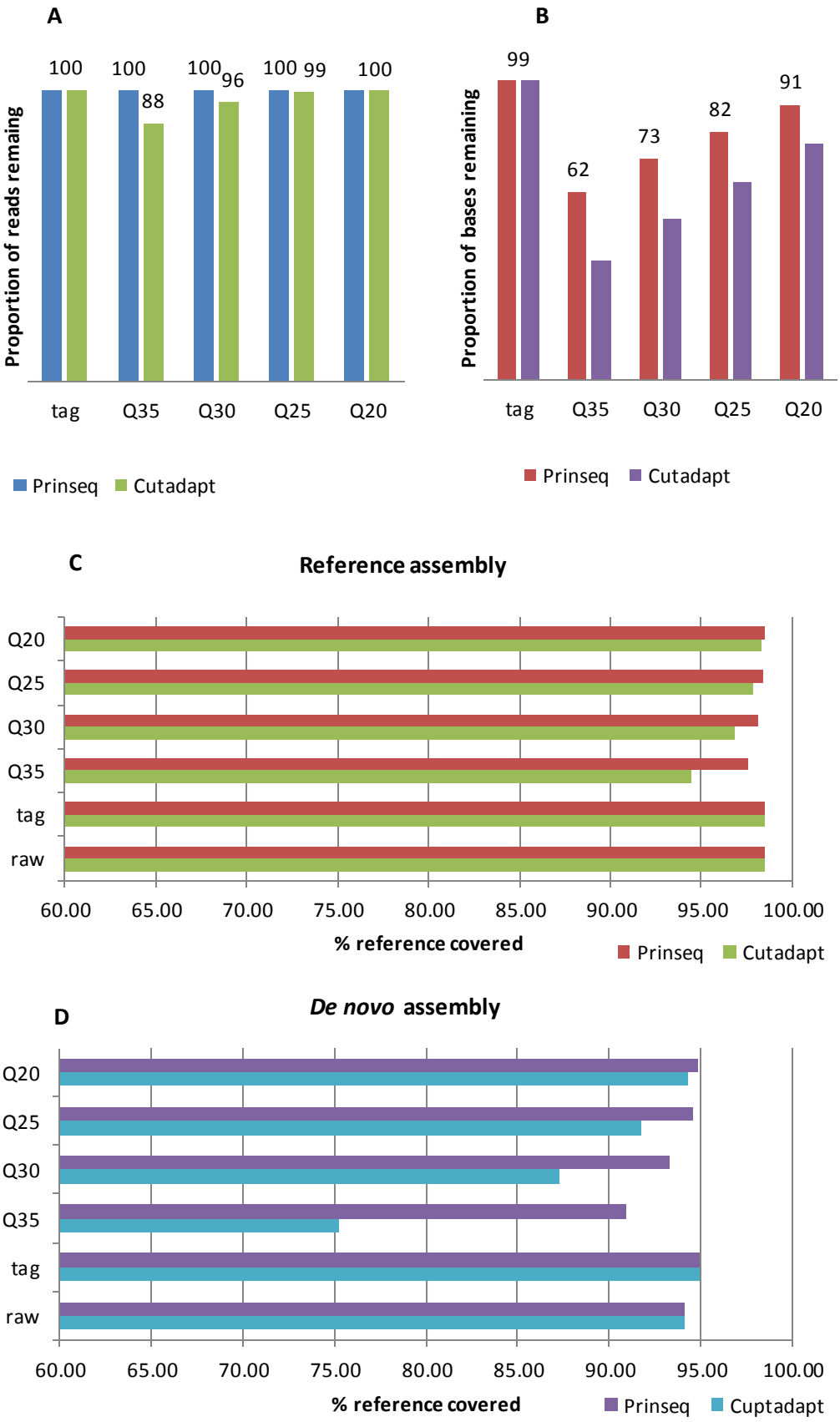


Figure 4-5 comparison of end quality trimming at varying quality cut-offs using Prinseq and Cutadapt on (A) proportion of reads remaining compared to raw data (B) proportion of bases remaining compared to raw data number (C) % reference assembly genome coverage (D) *de novo* assembly % reference coverage

4.3.3. Average Read Quality Trimming-Prinseq

Four different parameters were tested for trimming the data based on average quality of the reads, Q35, Q30, Q25 and Q20.

Q35 trimming led to a large reduction in read number with only 18% of reads remaining, as trimming became less strict fewer reads were removed with 91% of reads remaining after Q20 trimming. The same pattern was observed with the number of bases remaining after trimming, with only 16% left after Q35 trimming and 89% remaining after Q20 trimming. Longest read lengths were reduced in all trimmed data with the Q20 data having the longest read length at 778 bps (1200 bps in raw), as trimming became stricter the longest read length was reduced. There was less impact on the average read length with the Q35 sample having an average read length of 486 bps (vs. 528 bps in raw data). Again the average read length increased as trimming strictness decreased.

Once the data was reference assembled the genome coverage of the Q35 sample was reduced compared to the raw data (78.23% vs. 98.49%). There was less impact on reference coverage in the other trimming parameters. The Q35 sample was assembled into a high number of contigs (>11x the raw data), with the average contig size dropping from 19514 in the raw data to 1443 in the Q35 data. Again these changes lessened as the trimming became less strict.

After *de novo* assembly the proportion of the reference genome covered decreased as trimming became stricter, (94.57%, 92.87%, 88.05% and 33.62% for Q20. Q25, Q30 and Q35 trimmed data respectively). The number of contigs in all trimmed datasets was fewer than in the raw dataset, again numbers decreased as trimming strictness decreased. The number of misassemblies was at its lowest in the Q20 trimmed data, with the Q35 trimmed data having more misassemblies than the raw dataset.

	Raw	Q35	Q30	Q25	Q20
No: Reads	150,144	26,783	83,110	112,229	149,589
No: Bases	79,359,696	13,017,445	42,113,413	57,406,212	70,940,838
Shortest read (bps)	48	51	51	51	51
Longest read (bps)	1200	586	609	674	778
Average read length (bps)	528.5	486.03	506.72	511.51	516.91
			Reference assembly		
Ref coverage (%)	98.49	78.23	97.07	98.02	98.33
No: Contigs	242	2769	617	367	271
Average contig size (bps)	19,514	1443	7458	12657	17079
Largest contig size (bps)	220,401	9755	54997	97404	142037
			***De novo* assembly**		
Ref coverage (%)	94.12	33.62	88.05	92.87	94.57
No: Contigs	1692	1473	1327	983	818
N50 (bps)	4554	1277	5916	9081	13103
Largest contig (bps)	32,639	6203	36642	52917	74191
misassemblies	20	22	20	18	15

Table 4-3 Output of average read quality trimming using Prinseq at four different cut offs. Including basic read information and results of reference and *de novo* assemblies

4.3.4. Combination Average and End Trimming-Prinseq

The overall aim of trimming is to retain as much information as possible whilst improving the reliability of the assembly, and so a combination of end trimming and average read trimming was investigated. Q20 average read trimming had shown the best output and so this was used as the cut off for average trimming. Four different end trimming cut offs were investigated, Q30, Q28, Q25 and Q20. Q35 wasn't included as previously it had shown to be very detrimental to the assemblies, Q28 was included to provide an extra data point for optimising the trimming parameters. As before the first stage was to remove the sequencing tags from the sequences, following this ends were trimmed to the varying cut offs. Once this was complete the reads were average quality trimmed to remove reads at less than Q20 average quality. Finally, all reads less than 99 bps were removed, as these complicate assemblies whilst providing little additional information. An additional data set was included which only had reads less than 99 bp removed **Table 4-4**.

Once trimming was complete, more than 99% of reads remained in all trimmed samples except in the Q20 end trimmed sample, which had 96% of the reads remaining. The number of bases removed from the data increased as the end trimming cut off increased, with 73% of bases remaining in the Q30 end trimmed sample and 88% remaining in the Q20 end trimmed sample. Both the longest and average read lengths increased as trimming became less strict.

There was very little impact on the reference coverage after reference assembly of all the data sets, and more than 98% of the genome was covered in all samples. More contigs were produced in all the trimmed samples compared to the raw data. The highest number of contigs was found in the most strictly trimmed data set.

After *de novo* assembly the number of contigs produced in all samples was lower. This included the sample with no error trimming but with reads less than 99 bp removed. The lowest number of contigs were formed in the Q20 data set with the highest being found in the Q30 dataset. Again as contig number increased contig length decreased. Genome coverage was good in all samples with >93% covered in all samples, the highest genome coverage was found in the Q20 sample.

	Raw	Min_99	Q30	Q28	Q25	Q20
No: Reads	150,144	150,062	148,638	148,940	148,591	144,730
No: Bases	79,359,696	78,743,117	58,043,544	60,571,470	64,594,719	69,689,814
Shortest read (bps)	48	99	99	99	99	99
Longest read (bps)	1200	1196	577	606	607	759
Average read length (bps)	528.5	524.74	390.5	406.68	434.71	481.52
			Reference assembly			
Ref coverage (%)	98.49	98.47	98.10	98.23	98.39	98.42
No: Contigs	242	245	376	336	268	258
Average contig size (bps)	19,514	19345	12765	14183	17679	18253
Largest contig size (bps)	220,401	220385.	93890	93889	156112	163822
			***De novo* assembly**			
Ref coverage (%)	94.12	94.91	93.30	93.78	94.49	94.73
No: Contigs	1692	787	983	919	825	799
N50 (bps)	4554	14744	10787	11306	12644	14347
Largest contig (bps)	32,639	74191	60036	60038	71926	74189
misassemblies	20	19	15	14	14	14

Table 4-4 Output of end quality combined with Q20 average read quality trimming using Prinseq at four different cut offs. Including basic read information and results of reference and *de novo* assemblies **Min**_99 refers to raw data with reads shorter than 99 bases removed

4.3.5. Final Trimming Algorithm

The final trimming algorithm used in further work was using Prinseq, tag removal, Q20 end trimming, and Q20 overall quality trimming with a minimum of 99 base pairs read length.

4.4 Impact of Abundance Trimming in de novo assembly

Digital normalisation was used to investigate potential improvements in *de novo* assembly due to lower data complexity, the first stage was to optimise the parameters for this particular dataset. As stated in Brown et AL's [107] paper there is a relationship between the accuracy of the approach and the read length. Maximum lengths of K can be calculated using the following formula where L is read length and K is Kmer size used.

Equation 3

$$L > 3K - 1$$

If this formula is used a single substitution error will not affect the median kmer abundance, which the programme uses as a base for normalisation. The average read length for single cell two hour ϕ29 MDA reactions was 432, so according to the formula above 143 would be the maximum kmer size. A range of kmer sizes were tested from 20 (suggested by manual, based on short read data sets) to 140. Normalisation was performed on both raw and error trimmed data. The resulting reads were then assembled using reference and *de novo* methods and assessed for quality. The aim was to reduce the coverage to a similar level of the none amplification control assembly results (peak depth of 60)

4.4.1 *Abundance Trimming on Raw Reads*

Digital normalisation was applied to the sequencing results from the single cell ϕ29 MDA two hours incubation with the highest peak depth (160).

After abundance trimming using the K value 20 only 59.9% of reads remained, which was lower than K values 50-140 where 81.8-85.5% of reads remained **Figure 4-6-A**. After reference assembling the data genome coverage above 98.4% was achieved in all samples other than the K20 data which gave slightly lower genome coverage of 98.0%. Abundance trimming had very little impact on the number of contigs in the reference assembly, with a slight increase in number observed in the K20 sample. The peak depth of the assembly when K values 50-140 was used gave a range from 52-56, whereas the K20 assembly was reduced to 26 **Figure 4-6-B**. A similar pattern was observed in the average depth with abundance trimming using K50-140 giving an average depth the range of 14-14.6 and the K20 sample having an average depth of 9.7 (**Figure 4-6-C)**. When investigating the *de novo* assembly the genome coverage was greater than 95% in all abundance trimmed samples. Contig numbers were higher in all abundance trimmed assemblies, with the highest number being produced in the K20 sample. The largest contig size and N50 values were reduced in all abundance trimmed assemblies. However the number of misassemblies was lower than the raw data set **Table 4-5** and **Figure 4-6-D**.

	raw	k20	k50	k80	k110	k140
No: Reads	181,058	108,474	152,805	154,035	154,854	148,009
		59.91%	84.40%	85.07%	85.53%	81.75%
No: Bases	81064083	46,422,541	68,195,117	69,394,995	70,237,617	67,356,833
		57.27%	84.12%	85.61%	86.64%	83.09%
			Ref Assembly			
% Covered	98.46%	97.97%	98.45%	98.45%	98.46%	98.47%
Contig Number	143.00	153	140	141	145	143
Peak Depth	160	26	52	56	57	54
Average Depth	16.5489	9.68678	14.2109	14.461	14.6325	14.0223
			***De Novo* Assembly**			
% Covered	95.53%	95.156%	95.297%	95.269%	95.245%	95.256%
Contig Number	538	614	567	558	567	559
Largest Contig (bp)	87216	72500	68435	68436	68436	80293
N50	20471	18648	20346	20596	20470	20418
Misassemblies	23	18	17	18	19	18

Table 4-5 Results of abundance trimming of raw reads using varying K values, including read and base number, reference assembly and *de novo* assembly data including misassemblies. % of bases and reads compared to raw data

4.4.2 Abundance Trimming on Error Trimmed Reads

The raw reads were error trimmed using the optimised parameters described in 4.3.5**,** the resulting reads were then abundance trimmed using K values from 20-140.

The K20 data again removed more reads than the other values, only leaving 60.4%, with the other K values giving values 84.1-87.7% **Figure 4-6-A**. This had no impact on the proportion of the genome covered in the reference assembly with all assemblies covering more than 97.4% of the genome. The peak depths from K values 50-140 was 49-58, whereas the peak depth of the K20 data was 26 **Figure 4-6-B**. The average depth of the K20 sample was also lower (9.7) compared with the other K values (14.0-14.6) **Figure 4-6-C**. After *de novo* assembly there was genome coverage greater than 95% in all samples. The number of contigs produced was equivalent to the raw data, with a slight increase in the K20 data. The number of misassemblies was lower in the abundance trimming samples, other than the K20 sample, with K150 abundance trimming giving the lowest number **Figure 4-6-D**.

	trimmed	**trimmedk 20**	**trimmed k50**	**trimmed k80**	**trimmed k110**	**trimmed k140**
Total reads	177,737	107,312	151,596	154,273	155,900	149,382
	98.17%	60.38%	85.29%	86.80%	87.71%	84.05%
Total bases	79,427,727	46,076,860	67,428,717	68,670,284	69,598,188	66,912,909
		58.01%	84.89%	86.46%	87.62%	84.24%
Ref assembly						
% covered	98.42%	97.94%	98.43%	98.43%	98.46%	98.46%
Contig number	155	156	145	145	143	147
Peak depth	155	26	53	56	58	49
Average depth	16.3425	9.67412	14.1478	14.4025	14.5887	14.0126
***De novo* assembly**						
% covered	95.495%	95.179%	95.296%	95.266%	95.292%	95.286%
Contig number	557	607	576	572	574	566
Misassemblies	14	14	13	12	13	11

Table 4-6 Results of abundance trimming of error trimmed reads using varying K values, including read and base number, reference assembly and *de novo* assembly data including misassemblies

4.4.3 Abundance Followed by Error Trimming

To investigate the impact of abundance trimming prior to error trimming the raw data was first abundance trimmed using K values 20-140 and then error trimmed using the previously optimised parameters.

The K20 sample once again removed more bases than the other K values **Figure 4-6-A**; however this had no impact on the proportion of the genome covered after reference assembly. The number of contigs was lower in all the abundance trimmed samples. The peak depth of the K50-140 samples ranged from 51-57, whereas the peal depth of the K20 sample was once again 26 (**Figure 4-6-B)**. Genome coverage after *de novo* assembly was slightly lower in all abundance trimmed data compared to the error trimmed sample. However the number of misassemblies was lower, with the lowest number (9) being found in the K150 sample **Figure 4-6-D**.

	trimmed	**k20 trimmed**	**k50 trimmed**	**k80 trimmed**	**k110 trimmed**	**k140 trimmed**
Total reads	177737	108296	152675	153955	154793	147959
	98.17%	60.93%	85.90%	86.62%	87.09%	83.25%
Total bases	79427727	45646768	67031069	68220160	69054559	66228323
		56.31%	82.69%	84.16%	85.19%	81.70%
			Ref assembly			
% covered	98.42%	97.97	98.43	98.43	98.45	98.46
Contig number	155	150	137	137	140	140
Peak depth	155	26	51	56	57	53
Average depth	16.3425	9.5781	14.2769	14.3079	14.4782	13.8758
			***De novo* assembly**			
% covered	95.495	95.175	95.238	95.277	95.259	95.252
Contig number	557	620	578	579	575	569
Misassemblies	14	12	12	11	11	9

Table 4-7 Results of abundance trimming of raw reads using varying K values followed by error trimming, including read and base number, reference assembly and *de novo* assembly data including misassemblies

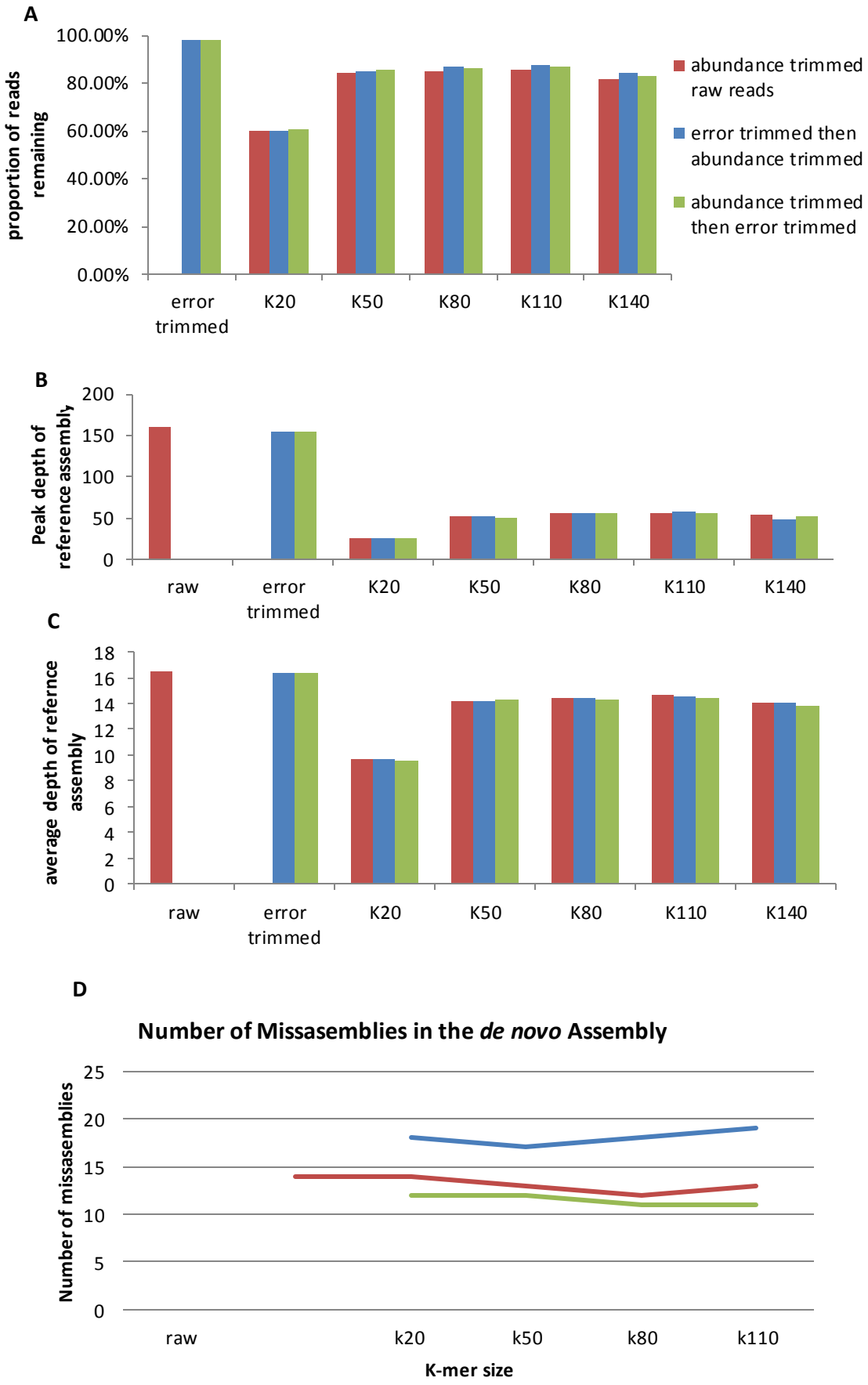


Figure 4-6 Impact of abundance trimming using different K values on (A) read removal compared to raw data, (B) peak reference assembly depth (C) average reference assembly depth and (D) number of misassemblies

4.5 Comparison of Genome Assembly quality using Varying De Novo Assemblers

The performances of six different assemblers on sequenced single cell ϕ29 MDA samples were investigated. Four sequencing files, *C. difficile*, *E. coli* and *A. naeslundii* and Adenovirus 41, representing a range of GC contents as well as a viral genome were abundance and error trimmed as described previously before being used in each assembler **Table 4-8**.

Initial assessment of the assembler was based on the ability of the assemblers to give good genome coverage. SSAKE performed poorly on all the genomes tested particularly at the GC extremes with only 12% of the *A. naeslundii* and 22% of the *C. difficile* genomes covered **Figure 4-7-A**. Due to the low genome coverage the assembler gave misleadingly low numbers of contigs and misassemblies and so was excluded from all other analysis. The next stage was to consider the assemblers performances across the selected genomes, particularly across the GC range. The remaining five assemblers (Newbler, Abyss, SPAdes, Ray and MIRA) had comparable genome coverage for the two pathogens with GC at 48.5% and 50.38% (Adenovirus and *E. coli)* **Figure 4-7-B**. However, Ray and Abyss were outliers in coverage of the two bacteria with extreme GC content. This was particularly obvious in the low GC pathogen *C. difficile*.

Assembly accuracy was the next consideration, as interpretation of data is reliant on good quality assembly. Data assembled using MIRA had considerably more misassemblies in every data set and so was not selected as the assembler **Figure 4-7-C**. Newbler and Spades had comparably low numbers of misassemblies with no clear advantage to either.

The final stage of selection was to consider the number of contigs the genome was assembled into, with a lower number desirable. In all cases Newbler produced more contigs than SPAdes, and so SPAdes was considered the most appropriate assembler for this dataset **Figure 4-7-D**.

	% cov	contigs	largest	N50	misassemblies
SSAKE					
Ad41	65.095	563	9543	908	5
C. dif	21.908	729	3520	705	0
E. coli	58.816	2195	11533	1503	8
A. naes	11.838	547	1768	781	1
Newbler					
Ad41	96.873	62	16453	6901	2
C. dif	89.075	1046	27971	6348	11
E. coli	95.252	569	78005	6074	9
A. naes	65.59	1660	22726	4406	8
Abyss					
Ad41	90.687	634	11007	1516	32
C. dif	72.228	2651	6218	1366	59
E. coli	88.567	1621	22949	4008	52
A. naes	56.325	1780	20141	2283	42
Spades					
Ad41	96.994	53	14532	4560.00	2
C. dif	92.899	469	72426	8273.00	8
E. coli	95.963	242	129944	12170.00	11
A. naes	63.417	1300	36873	7437.00	5
Ray					
Ad41	92.355	54	14671	1515	32
C. dif	54.625	1685	13803	1811	31
E. coli	89.232	1388	55257	5860	38
A. naes	45.165	1049	17938	1918	21
Mira					
Ad41	96.991	87	16978	1835	18
C. dif	90.812	1385	33820	5517	35
E. coli	97.226	954	79580	16280	20
A. naes	68.579	1076	32784	3294	17

Table 4-8 results of the *de novo* assembly of *C. difficile*, *E. coli* and *A. naeslundi*i single cell ϕ29 MDA data and ϕ29 MDA of Adenovirus 40 using six different assemblers.

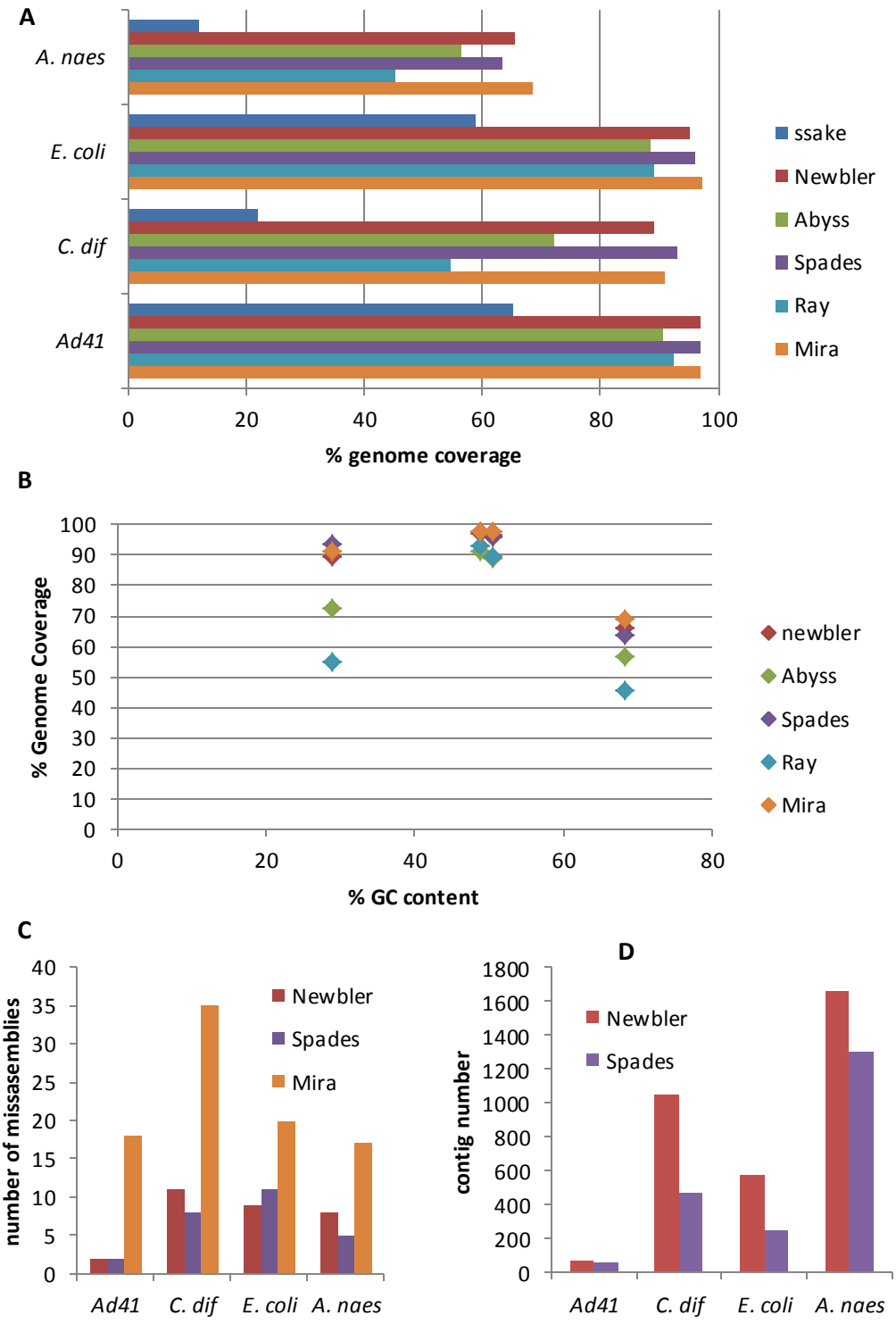


Figure 4-7 results of the *de novo* assembly of *C. difficile*, *E. coli* and *A. naeslundi* i single cell ϕ29 MDA data and ϕ29 MDA of Adenovirus 40 using six different assemblers on (A) genome coverage, (B) genome coverage at different GC contents, (C) number of misassemblies and (D) the number of contigs

4.6 Characterisation of Pathogens using de novo Assembled Genome Data

4.6.1 Annotation of the Genome

To assess the potential for *de novo* assemblies to provide detailed genome data, annotations of completed reference genomes were compared to *de novo* assembled datasets. *De novo* assemblies produced by none amplification controls and single cell ϕ29 MDA amplifications for *C. difficile, E. coli* and *A. naeslundii* were investigated alongside ϕ29 MDA amplification of adenovirus. Files for *de novo* assembly were error and abundance trimmed before assembly using Spades. The result assemblies and reference files were then annotated using Prokka[115], which is a tool specifically designed for rapid annotation of prokaryotic genomes. After annotations were completed the number of hypothetical proteins was calculated before they were removed from analysis along with any repeated proteins, **Table 4-9.**

Overall for all bacterial annotations the majority of predicted proteins were found i n all three annotations **Figure 4-8.** In the *E. coli* annotations, a similar number of proteins were predicted in all three annotations (3102, 3103 and 3094), with 2934 of these being identified in all three annotations. A small number of proteins (<1%) were found in only a single annotation, and a slightly larger proportion (2%) were found in only two annotations. In the *C. difficile* annotations, the reference and single cell ϕ29 MDA annotations identified a similar number of proteins (2413 and 2356), whereas fewer proteins were identified in the non-amplification control (2123). The non-amplification control also contained more hypothetical proteins than the other two annotations. Almost 5% of the genes predicted in the *C. difficile* non-amplification control annotations were only identified in that annotation, whereas 14% of proteins were identified in the reference and the ϕ29 MDA sample. In the *A. naeslundii* annotations a similar number of proteins were identified in the non-amplification control and ϕ29 MDA sample (1123 and 1121), however this was less than identified in the reference (1420). 30% of the genes identified in the reference annotation weren't found in the non-amplification or ϕ29 MDA sample.

After annotation of the adenovirus reference and ϕ29 MDA assembly a similar number of proteins were identified in both (22 and 28), but only one of these was identified in both. However when the two annotation results were arranged by transcription unit (supplementary Table 7-1) 13 out of 14 units were found in both annotations.

	Reference	Control	Single cell MDA
		E. coli	
Genes	4552	4735	4826
Hypothetical	764	815	647
Unique hits	3102	3103	3094
		C. difficile	
Genes	4113	3982	4745
Hypothetical	877	1126	964
Unique hits	2413	2123	2356
		A.naeslundii	
Genes	2520	2141	2028
Hypothetical	852	915	615
Unique hits	1420	1129	1121
		Adenovirus	
Genes	33		53
Hypothetical	10		32
Unique hits	22		28

Table 4-9 genome annotations for reference genome, non-amplification control and single cell ϕ29 MDA *de novo* assemblies for *E. coli*, *C. difficile* and *A. naeslundii* alongside reference and ϕ29MDA *de novo* assembly annotations of Adenovirus.

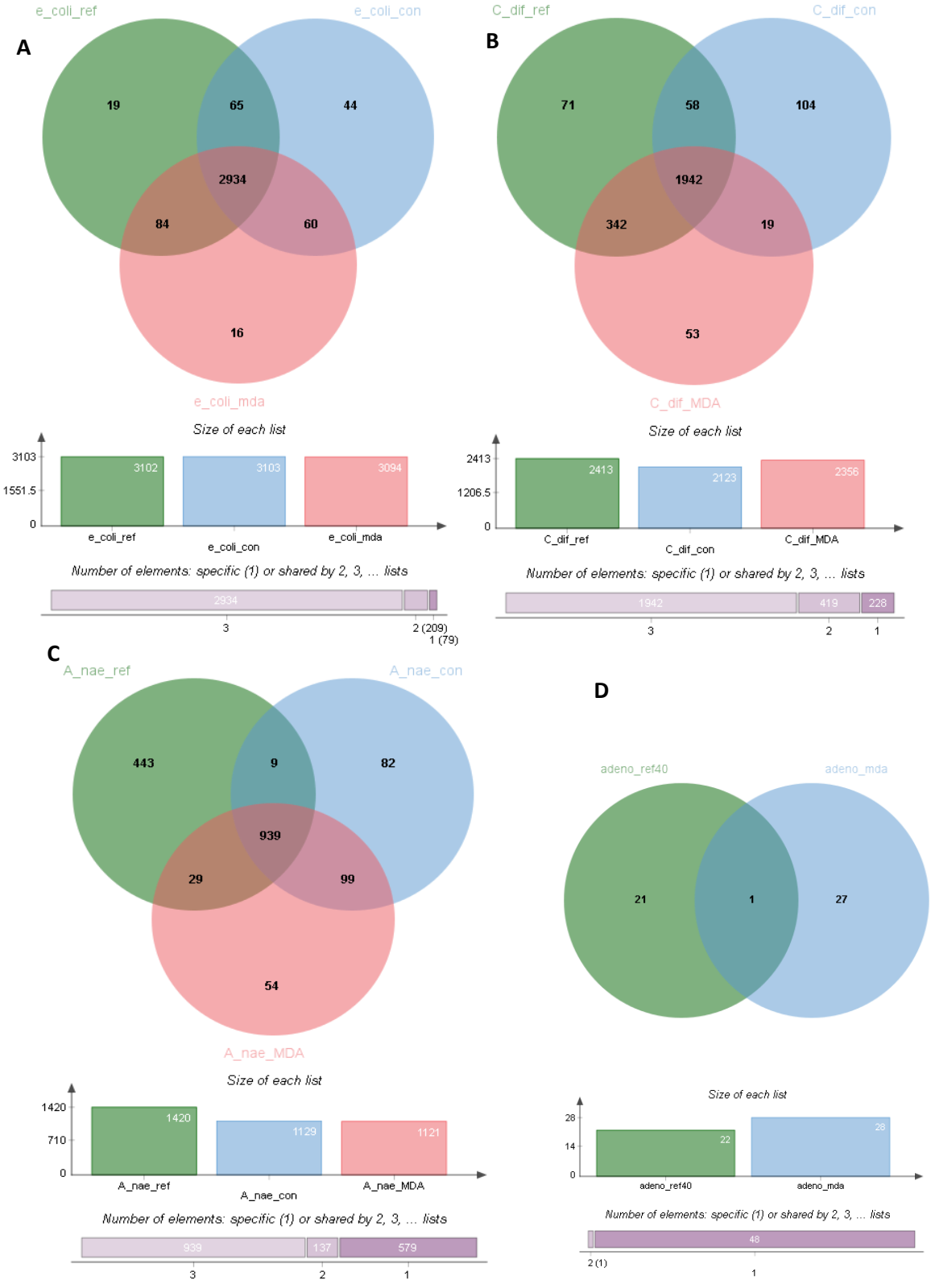


Figure 4-8 Venn diagrams for genome annotations results for unique hits for reference genome, non-amplification control and single cell de novo assemblies for (A) *E. coli,* (B) *C. difficile* (C) *A. naeslundii* and (D) Adenovirus

4.6.2 Virulence Determinants

4.6.2.1 Key Word Search

To further investigate the Prokka outputs, a key word search was performed to identify potential virulence factors identified. A list of the virulence key words can be found in **2.4.10.1**

The annotated reference sequence for C. *difficile* predicted 15 virulence factors, 13 of which were found in all annotations and included six phage associated genes, three transposons and four toxins including toxin A and B associated with *C. difficile* virulence. Two additional genes were identified in the ϕ29 MDA and the reference annotations, a conjugative transposon and a pathogenicity locus. In the non-amplification control a plasmid associated gene was identified, which was not found in the other two annotations. The overlapping virulence factors are summarised in **Figure 4-9-B** and the details of all factors identified are found in **Table 4-11.**

After filtering for repetition the *E. coli* reference sequence contained 19 possible virulence determinants, 12 of which were phage associated, five were toxins, and two were virulence associated. In total the non-amplification control had 47 genes predicted and the single cell ϕ29 MDA had 45 genes identified. 27 of these were not identified in the reference, **Figure 4-9**, of these 24 were phage associated, one was a toxin, one was a plasmid element and one was a conjugative transposon. All factors identified are shown in **Table 4-10**

In the reference annotation of *A. naeslundii*, four factors were identified, two plasmid genes, one toxin and one transposon. The control annotation included two factors, plasmid maintenance and a toxin. The ϕ29 MDA annotation contained four factors, two toxins, one plasmid factor and one transposon **Table 4-12**

Common elements All	Common elements in E_coli_con E_coli_MDA :
Phage antitermination protein Q	Conjugative relaxosome accessory transposon
phage exclusion protein Lit	Phage portal protein, lambda family
Phage Terminase	Phage terminase large subunit (GpA)
Phage portal protein	Phage tail fibre repeat
Phage Tail Collar Domain protein	Phage NinH protein
Phage shock protein A	Phage minor tail protein L
Phage shock protein B	Phage minor tail protein
Phage shock protein C	Phage minor tail protein U
Phage shock protein D	Phage Head-Tail Attachment
Phage shock protein G	Phage major capsid protein E
phage T7 F exclusion suppressor FxsA	Plasmid-derived single-stranded DNA-binding
Toxin YkfI	Toxin CcdB
Toxin YoeB	Transposon gamma-delta resolvase
Toxin CptA	Phage lysozyme
Toxin YhaV	Phage holin family (Lysis protein S)
Toxin-antitoxin biofilm protein TabA	Phage integrase family protein
Virulence regulon transcriptional activator	Lambda phage tail tape-measure protein
Virulence factors putative positive	Bacteriophage CII protein
	Bacteriophage Lambda NinG protein
Common elements in E_coli_ref E_coli_con :	Prophage tail length tape measure protein
Phage DNA packaging protein Nu1	Bacteriophage lambda minor tail protein (GpG)
	Prophage minor tail protein Z (GPZ)
	Bacteriophage lambda head decoration protein D
Elements only in E_coli_con :	Bacteriophage lambda Kil protein
Plasmid SOS inhibition protein (PsiB)	Lambda Phage CIII
	Bacteriophage replication protein O
	Bacteriophage lambda tail assembly protein I

Table 4-10 -all virulence factors identified in the *E. coli* genomes using Prokka output keyword search

Common elements in All :	Common elements in C_dif_ref C_dif_MDA :
"Phage terminase large subunit"	"Conjugative transposon protein TcpC"
"Phage portal protein, SPP1 Gp6-like"	"Pathogenicity locus"
"Phage minor structural protein GP20"	
"Phage tail sheath protein"	
"Phage-like element PBSX protein XkdM"	**Elements only in C_dif_con :**
"Phage XkdN-like protein"	"Plasmid maintenance system killer protein"
"Toxin B"	
"Toxin A"	**Elements in C_dif_con and C_Dif_MDA :**
"Toxin-antitoxin biofilm protein TabA"	"Phage tail fibre repeat"
"Transposon Tn10 TetD protein"	
"Transposon gamma-delta resolvase"	
"Transposon Tn3 resolvase"	
"Virulence protein"	
"Toxin HigB-1"	

Table 4-11 all virulence factors identified in the *C. difficile* genomes using Prokka output keyword search

Elements only in A_nae_ref :	Common elements in A_nae_con A_nae_MDA :
"Plasmid stabilisation system protein"	"Plasmid maintenance system killer protein"
"Plasmid encoded toxin Txe"	"Toxin HigB-1"
Common elements in A_nae_ref A_nae_MDA :	
"Toxin Doc"	
"Transposon Tn10 TetC protein"	

Table 4-12 all virulence factors identified in the *A. naeslundii* genomes using Prokka output keyword search

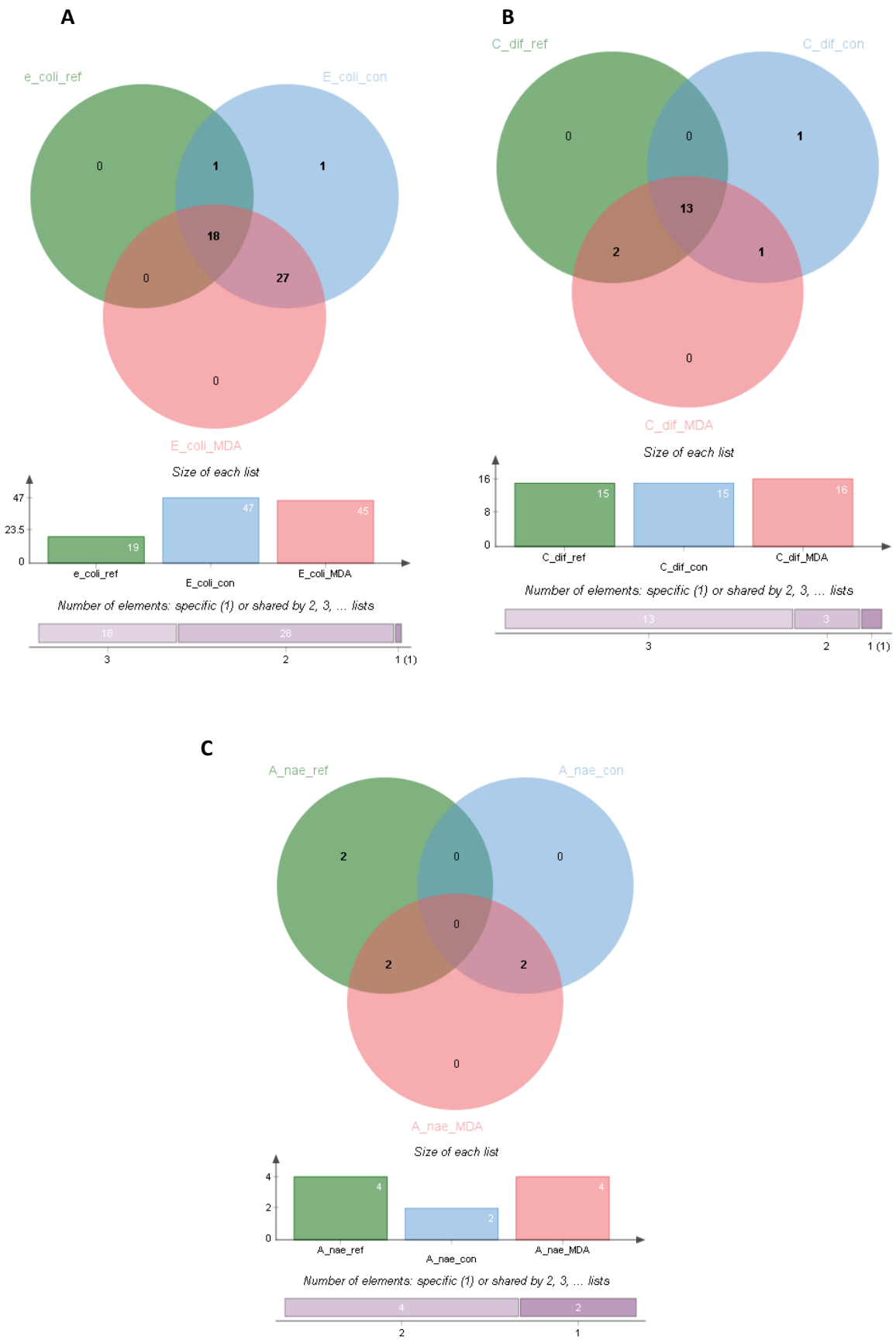


Figure 4-9 Venn diagrams for virulence factors identified using Prokka output keyword search for reference genome, non-amplification control and single cell de novo assemblies for (A) *E. coli*, (B) *C. difficile* and (C) *A. naeslundii*

4.6.2.2 Database search

The same files used in the Prokka annotations were used to predict virulence factors using the VFDB. The database identified 79 hits in the *E. coli* reference, and 82 in the non-amplification control and single cell ϕ29 MDA *de novo* assemblies. Of these identified factors 28 were associated with the genes were *E. coli* CFT073 and 15 were associated with *E. coli* 536 (uropathogenic strains). Additionally, seven factors were associated with E. *coli* 0157:H7 and three unspecified *E. coli* strains. There were also 25 hits against *Shigella flexerni* serotype 2a, four against *Shigella dysenteraie.* When the proteins were grouped into classes, 43 were unknown, hypothetical or had unknown functions. 20 were involved in ferric export, 9 were fimbri ae associated, three were involved in efflux or export, and two were insertion elements found in all files, two additional insertion factors were found only in the ϕ29 MDA and control samples. Two phage proteins were also identified. A full list of the ide ntified factors can be found in the supplementary material.

When using the VFDB to predict resistance in *C. difficile* the reference, non-amplification control and ϕ29 MDA all had four hits identified, toxin A and B, along with cdtA and cdtB. Shown in **Table 4-13**

No virulence determinants were identified in any *A. naeslundii* assemblies.

Gene	ID	Bacterial origin
tcdA	toxin_A	Clostridium_difficile_630
tcdB	toxin_B	Clostridium_difficile_630
cdtB	CdtB	Clostridium_difficile
cdtA	CdtA	Clostridium_difficile

Table 4-13 output of VFDB for *C. difficile,* the same output was achieved using completed reference and *de novo* assemblies of the non-amplification and ϕ29 MDA samples

4.6.3 Antibiotic Resistance Prediction

4.6.3.1 Key Word Search

An additional key word search was performed to investigate the potential of the Prokka annotation to provided antibiotic resistance prediction. A list of terms included can be found in **2.4.10.1**

When the *E. coli* reference was investigated for the word resistance there were 33 hits, seven (21%) of which were environmental resistance factors, all other hits also contained another word in the search the full list can been seen in the supplementary material **Table 7-5.**When the *C. difficile* was investigated for the work resistance 26 hits were found, six of which were environmental resistance factors, all others but one contained other words in the search. One hit conferred resistance to phages which would have been detected in the previous search . When *A. naeslundii* was investigated for the word resistance 13 hits were found, five (38%) of which were environment associated, all others had another searched for string.

When the *E. coli* reference was investigated for efflux pumps, 24 were identified, 20 of which were housekeeping associated, all others were found to have other search terms in. When the *C. difficile* reference was searched for efflux there were 11 results seven of which were housekeeping, one had an unknown function and one was only identified as a putative component. Two were associated with resistance and had other search strings in their name. In the *A. naeslundii* reference four efflux pumps were identified, all of which were housekeeping efflux pumps.

Searches were repeated without the words resistance or efflux. When the *E. coli* reference genome was investigated for resistance factors there were 53 identified, the control had 51 identified proteins and the ϕ29 MDA sample had 53. 51 of these genes were associated will all three searches; two appeared only in the reference and the ϕ29 MDA. In all three datasets 27 efflux pumps and export systems were identified associated with drug resistance, with an additional two found in the ϕ29 MDA and reference datasets **Table 4-10.** Seven penicillin binding proteins were identified, and two repressor or activator genes were detected. Six of the identified genes had an unknown or undefined role and five identified genes weren't associated with antibiotic resistance.

When the *C. difficile* reference genome was investigated for resistance factors 39 genes were identified, **Table 4-10** the control found 33 and the ϕ29 MDA identified 36. 28 of these genes were found in all three, seven were found only in the reference and ϕ29 MDA sample, four were found in the control and ϕ29 MDA and one gene was found in the control, and one found only in the ϕ29 MDA sample. Four beta-lactamases were identified in all datasets, 12 efflux or export systems were found in all samples, along with an additional two in the reference and ϕ29 MDA. One penicillin binding protein was found in the ϕ29 MDA and control and one in only the ϕ29 MDA. Four activator or repressor were identified in all datasets, and an additional one was identified in the reference and the ϕ29 MDA. Two vancomycin specific resistance factors were found in the all samples and an additional one in the control sample. One tetracycline resistance associated gene was identified in all samples and two additional genes were found in the ϕ29 MDA and reference. Three proteins had no defined function and four weren't associated with antibiotic resistance.

In the *A. naeslundii* reference 17 factors were identified in the search for antibiotic resistance **Table 4-10**, 14 factors were identified in the control and 18 were identified in the ϕ29 MDA sample. Ten of these factors were common to all, five export or efflux systems, three without defined function, one penicillin binding protein and one factor not associated with antibiotic resistance. Three factors were only found in the reference sequence, one with an unknown function and two penicillin binding proteins. Four factors were found in the ϕ29 MDA and reference sequence, two efflux pumps, one penicillin binding protein and a vancomycin resistance associated gene. Four genes were only identified in the ϕ29 MDA and control, two beta lactamase associated gens, a resistance operon repressor and an export protein.

4.6.3.2 Database Search

When the ardbAnno script was used to detect resistance factors in *E. coli* none were identified any of the datasets. In all the *C. difficile* the ermB resistance gene was detected, which is associated with resistance to macrolides, Lincosamides and Streptogramin B. No resistance factors were identified in the *A. naeslundii* genomes.

When the Resfinder database was used to find resistance factors in *A. naeslundii*, no matches were found. In all the *C. difficile* genomes ermB and tetM gens were identified. In the *E. coli* genomes, the oqxB gene and the blaCMY operon were identified.

	Reference	Control	MDA
A. naeslundii	None detected	None detected	None detected
C. difficile	erm(B)_10_U86375	erm(B)_10_U86375	erm(B)_10_U86375
	erm(B)_11_M19270	erm(B)_11_M19270	erm(B)_11_M19270
	erm(B)_12_U18931	erm(B)_12_U18931	erm(B)_12_U18931
	erm(B)_15_U48430	erm(B)_15_U48430	erm(B)_15_U48430
	erm(B)_18_X66468		
	erm(B)_1_JN899585	erm(B)_1_JN899585	erm(B)_1_JN899585
	erm(B)_20_AF109075	erm(B)_20_AF109075	erm(B)_20_AF109075
	erm(B)_21_U35228	erm(B)_21_U35228	erm(B)_21_U35228
	erm(B)_6_AF242872	erm(B)_6_AF242872	erm(B)_6_AF242872
	erm(B)_7_AF368302	erm(B)_7_AF368302	erm(B)_7_AF368302
	erm(B)_9_AF299292	erm(B)_9_AF299292	erm(B)_9_AF299292
	tet(M)_10_EU182585	tet(M)_10_EU182585	tet(M)_10_EU182585
	tet(M)_11_JN846696	tet(M)_11_JN846696	tet(M)_11_JN846696
E. coli	blaCMY-12_1_Y16785	blaCMY-12_1_Y16785	blaCMY-12_1_Y16785
	blaCMY-15_1_AJ555823	blaCMY-15_1_AJ555823	blaCMY-15_1_AJ555823
	blaCMY-38_1_AM931008	blaCMY-38_1_AM931008	blaCMY-38_1_AM931008
	oqxB_1_EU370913	oqxB_1_EU370913	oqxB_1_EU370913

Table 4-14 Results from Resfinder for predicting resistance in *A. naeslundii*, *C. difficile* and *E. coli* for completed references and *de novo* assemblies of non-amplification control and ϕ29 MDA samples

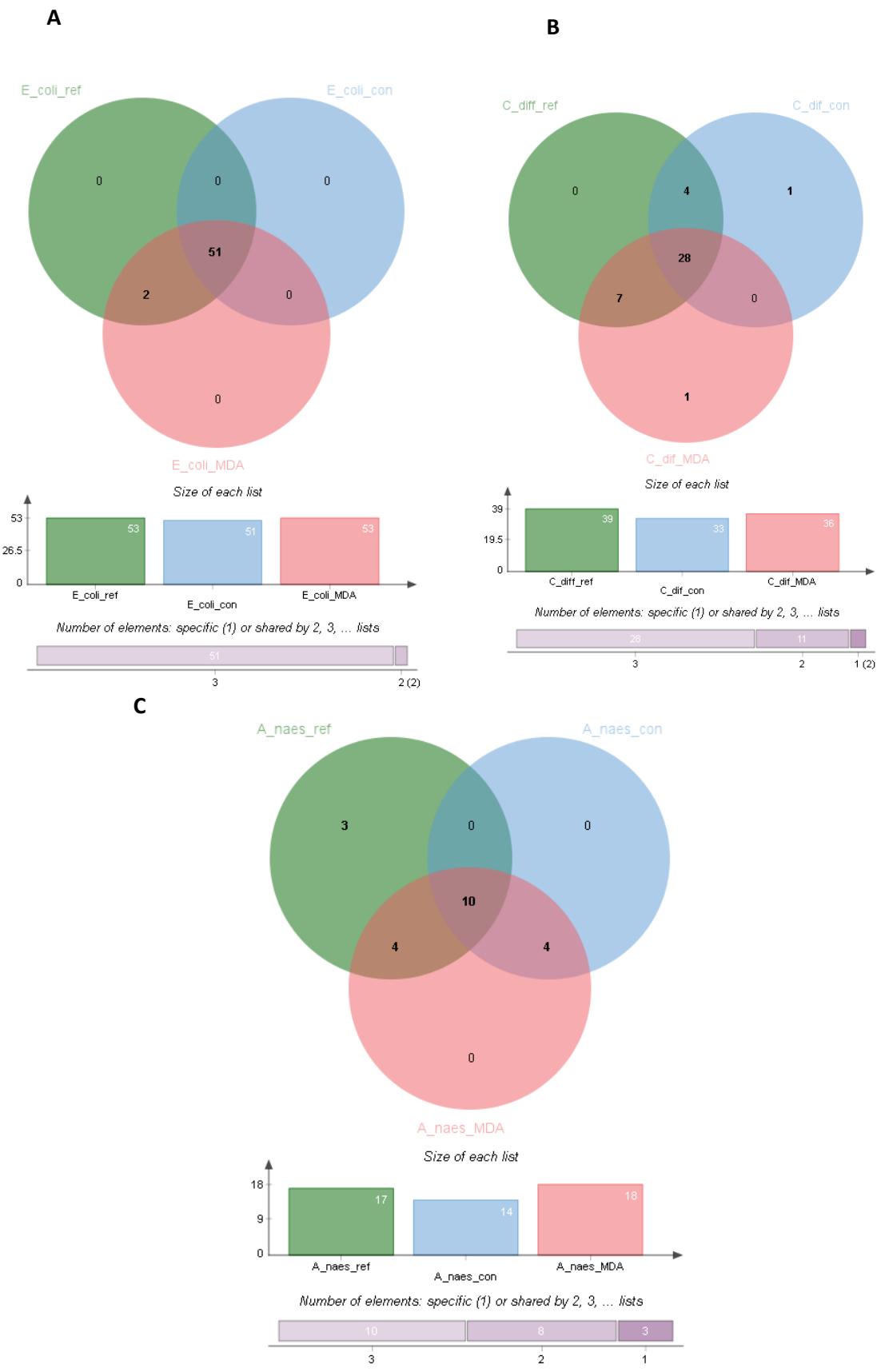


Figure 4-10 Venn diagrams for resistance predictions for reference genome, non-amplification control and single cell ϕ29 MDA *de novo* assemblies for *E. coli, C. difficile* and *A. naeslundii*

4.7 Analysis of Mixed Cell Input Sequencing Data

To investigate the impact of abundance trimming on mixed samples, the four samples consisting of mixed bacteria at different ratios (3.9) were error and abundance trimmed. The resulting reads were then identified using Blastn and LCA analysis, with reads belonging to each bacteria being extracted into separate fastq files. These fastq files were then *de novo* assembled using Spades, before being annotated using Prokka.

After completion of the trimming stage the 1:1 reads had more reads remaining (54%) than the other ratio samples. As the ratio between the bacteria increased the proportion of reads remaining after trimming decreased **Figure 4-11-A**. The *E. faecalis genome (*3,218,030 base pairs) is twice the size of the *H. influenzae* genome (1,830,140 base pairs), and so the proportion of reads mapping to the *E. faecalis* genome was halved to normalised the read proportion for genome size. After completion of trimming the proportion of reads identified as each bacterium was slightly closer to the original cell mix ratio **Figure 4-11-B**. As the ratio of *E. faecalis* reads decline the proportion of the genome covered dropped, with no genome coverage achieved in the 1:1000 ratio **Figure 4-11-C**. However, Enterococcus phages BE2 and EF62phi were identified in all samples. Genome coverage of the *H. influenzae* remained high in all samples after abundance trimming (92-94%) **Figure 4-11-D**

	1:1		**1:10**		**1:100**		**1:1000**	
	Before Pipeline	**After Pipeline**	Before Pipeline	**After Pipeline**	Before Pipeline	**After Pipeline**	Before Pipeline	**After Pipeline**
Total reads	133320	**72583**	167837	**80562**	148139	**61361**	152879	**59208**
Reads *Enterococcus*	83286	**43710**	30086	**13293**	3169	**2002**	0	**0**
% reads *Enterococcus*	62%	**60%**	13%	**17%**	2%	**3%**	0%	**0%**
% reads *Enterococcus* normalised	31%	**30%**	7%	**8%**	1%	**1.6%**	0%	**0%**
%genome coverage *Enterococcus*	93%	**93%**	64%	**65%**	15%	**15%**	0%	**0%**
Reads Haemophilus	33254	**22098**	195217	**66866**	122308	**58331**	145235	**57211**
%reads *Haemophilus*	25%	**30%**	85%	**83%**	83%	**95%**	95%	**97%**
% genome coverage *Haemophilus*	92%	**92%**	94%	**94%**	94%	**93%**	93%	**92%**

Table 4-15 summary of results of sequencing of different starting ratios of *E. Faecalis* and *H. Influenzae* before and after application of the developed analysis pipeline

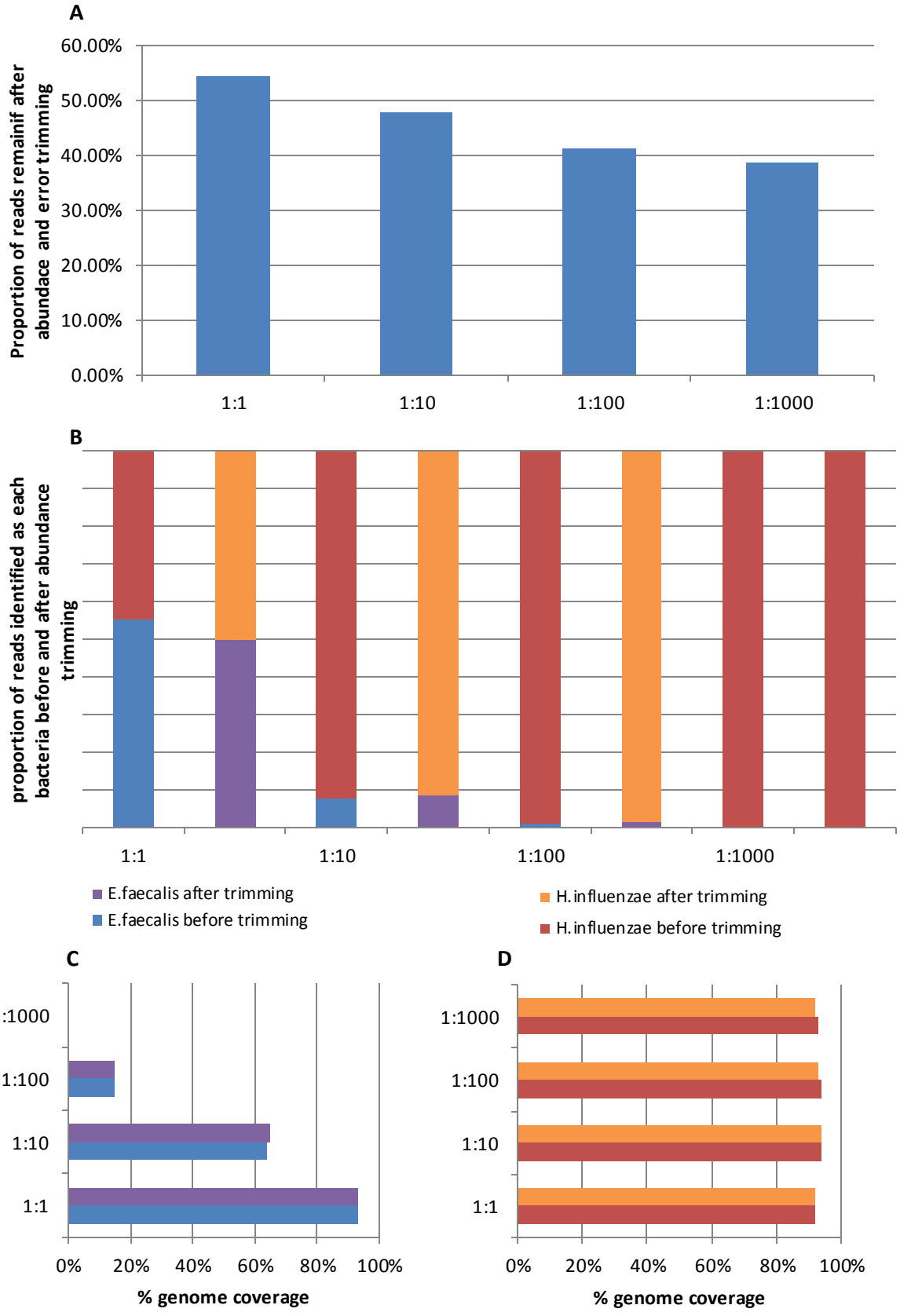


Figure 4-11 Impact of abundance and error trimming on different mixed cell rations, (A) proportion of total reads remaining (B) normalised proportion of reads identified as each bacterium before and after abundance trimming (C) % genome coverage of *E. faecalis* at different mix ratios before and after abundance trimming (D) % genome coverage of *H. influenzae* at different mix ratios before and after abundance trimming

4.7.1 Characterisation of Mixed Samples

4.7.1.1 Mixed bacterial Input

As the ratio of *E. faecalis* declined the number of predicted proteins also declined with 2476 unique proteins predicted in the 1:10 ratio sample and only 716 predicted in the 1:100 ratio. A total of 650 of these products were identified in all ratios, 1232 were found in the 1:1 and 1:10 ratio, and 571 were only found in the 1:1 ratio sample.

When the *H. influenzae* samples were annotated the number of unique proteins identified was fairly constant (2392-2428)**Figure 4-12-B**. The total number of proteins predicted and the number of hypothetical proteins were lower in samples with a lower ratio of E. *faecalis* **Figure 4-12-A** A total of 2290 of these products were identified in all ratios, with 62 being identified in three datasets, 104 were identified in two samples and 105 were identified in only a single sample.

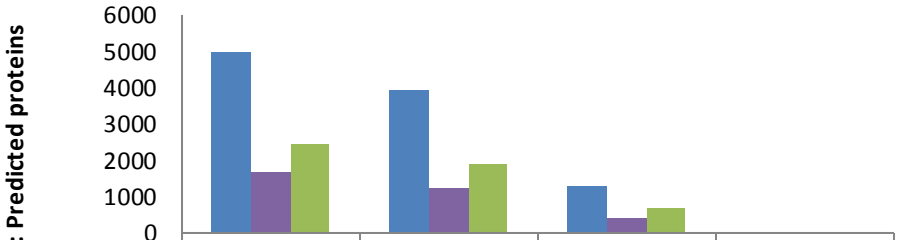
The annotation data from the *H. influenzae* also gave an insight into the reproducibility of the amplification and annotation method **Figure 4-13**. Over 94% of predicted proteins were found in all four *H. influenzae* amplifications, 2.5% of proteins were found in three annotations, 4% in two annotations and 4% were identified only in one annotation.

4.7.1.2 Mixed Viral Input

When the sequencing results from the amplification of mixed HIV and Adenovirus (**3.10.1.2**) was annotated using Prokka, the adenovirus results were comparable with those found in **4.6.1,** with a total of 52 genes predicted with 22 of these being hypothetical and 28 remained after filtering for repetition.

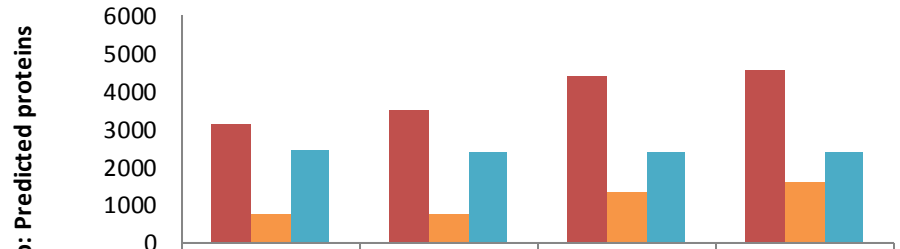
The HIV *de novo* assembly covered 99.9% of the genome in three contigs, with two misassemblies. The output from the Prokka annotation was nine products, one of which was hypothetical and after filtering for repeats, four proteins remained. When the reference was annotated using Prokka, eight products were identified, three of which were hypothetical, and after filtering for repeats four remained. The four remaining gene annotations were the same in both the reference and *de novo* assembly which were Gag polyprotein, Gag-Pol polyprotein, Virion infectivity factor and Envelope glycoprotein gp160 precursor.

A



	1:1	1:10	1:100	1:1000
all	5023	3928	1292	0
hypothetical	1675	1262	397	0
unique	2476	1903	716	0

B



	1:1	1:10	1:100	1:1000
all	3153	3486	4390	4566
hypothetical	750	753	1326	1618
unique	2428	2412	2421	2392

Figure 4-12 Results of Prokka annotation on different mixed call ratios of (A) *E .faecalis* and (B) *H. influenzae* showing total number of predicted genes, number of hypothetical genes and number of unique genes

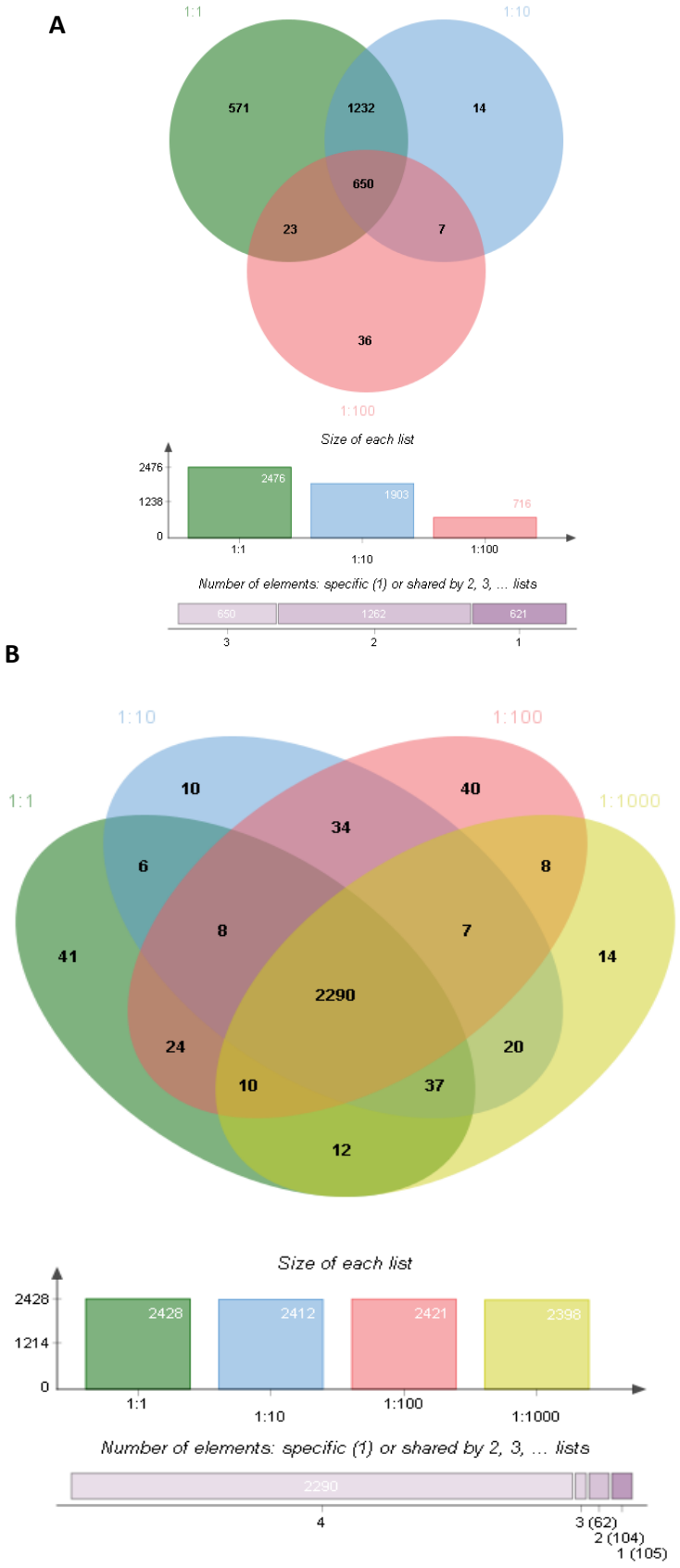


Figure 4-13 Venn diagram of unique genes identified in the (A) *E. faecalis* and (B) *H. influenzae* in each mix ratio using Prokka

4.7.1.3 *Resistance in Enterococcus*

The number of resistance genes predicted dropped as the ratio of *E. faecalis* declined. When the 1:1 enterococcus annotation was interrogated for resistance factors 47 genes were identified, 12 were penicillin binding proteins and four weren't associated with antibiotic resistance. 19 were efflux pumps associated with drug resistance, four were specifically tetracycline resistance genes, two were beta-lactamases, and two were associated with streptomycin resistance, one specifically with vancomycin/teicoplanin resistance and three with other antibiotic breakdown. When the 1:10 enterococcus annotation was investigated for resistance factors 39 genes were identified, nine were penicillin binding proteins and four weren't associated with antibiotic resistance. 16 were efflux pumps associated with drug resistance, four were specifically tetracycline resistance genes, and two were beta-lactamases, one specifically with vancomycin/teicoplanin resistance and three with other antibiotic breakdown. In the 1:100 annotation of the enterococcus genome, 14 resistance genes were identified, seven were penicillin binding proteins, one wasn't associated with antibiotic resistance, five were drug efflux pumps and one was a beta-lactamase. This is summarised in **Table 4-16**

	All ratios	1:1 only	1:1 and 1:10	1:10 only	1:100 only
None antibiotic resistance	1	1	2	1	
PBP	6	4	2	1	1
Efflux	5	6	8	3	
Tetracycline resistance			4		
Beta-lactamase	1		1		
Streptomycin resistance		2			
Vanc/Teic resistance			1		
Other ABx breakdown			3		

Table 4-16 summary of antibiotic resistance factors identified in *E. faecalis* in a mixed bacteria samples at different input ratios

4.7.1.4 *Haemophilus Resistance*

In the 1:1 and 1:10 sample 52 resistance genes were identified, 50 were identified in the 1:100 and 52 were identified in the 1:1000. 47 genes were found in all samples, including 11 PBP and two not associated with antibiotic resistance. 21 drug efflux pumps were identified, five genes specifically associated with tetracycline were identified, two genes associated with streptothricin were identified, along with two genes associated with vancomycin and teicoplanin resistance, and four genes which broke down other antibiotics. Three genes were found in three datasets, two efflux pumps and one PBP, three genes were only found in two datasets, one efflux, one PBP and one gene not associated with antibiotic resistance. An additional PBP was only found in one dataset.

4.8 Final Assembly Pipeline

A final analysis pipeline was designed using the methods optimist in this chapter to provide robust data analysis of sequencing produced from ϕ29 MDA on the 454 Junior, **Figure 4-14.**

The first stages of the pipeline include removal of host and possible environmental and kit contamination using reads from negative amplification samples. This is then followed by digital normalisation and error trimming. Read identification is then performed using Blastn and LCA analysis. At this point the results are visualised to allow assessment of mixed samples, and provide microbial species identification. Species specific reads are then extracted from the trimmed fastq, and *de novo* assembled. Finally, annotation is performed alongside use of database searches for virulence and resistance factors.

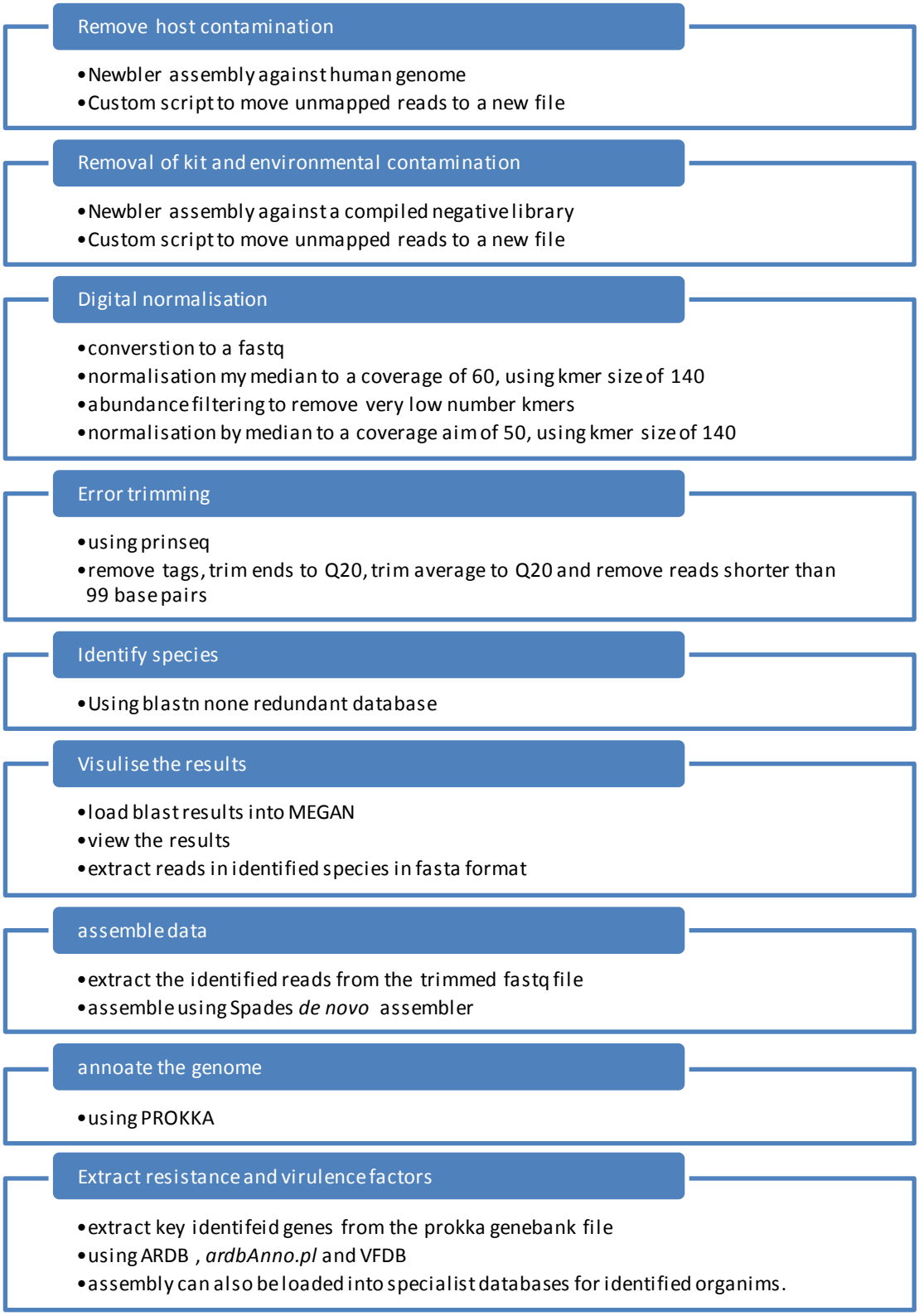


Figure 4-14 summary of bioinformatic analysis of sequencing produced using ϕ29 MDA on the 454 Junior

4.9 Chapter Discussion

4.9.1 Removal of Contaminants from Sequencing Data

Host signals increase data size and make *de novo* assemblies more complex and so a simple mapping method was used to remove them. The adenovirus sequencing in 3.10 which had the highest level of host contamination was used to model the bioinformatic stage of the procedure to remove host signals. After the human signals were removed from the file, the reads belonging to the adenovirus and Enterobacteriaceae were left intact; the only information removed was the host contamination. This method allows simple and rapid removal of host signals without affecting the pathogen signals, thereby lowering the overall complexity of the downstream assembly.

Kit and environmental contaminants are important factors to consider when using highly sensitive molecular methods [103]. Throughout this study negative samples were amplified and sequenced. The sequencing was then investigated using LCA analysis to provide a way of monitoring possible contaminations. Within the negative samples in **4.1**, the majority of the reads present were either classed as unmapped or unidentified. The possible source of these reads could be polymerisation and joining of very degraded DNA, which was introduced at some point in the sample processing. Also the ϕ29 could be randomly joining nucleotides together or could be using primers as sites of replication because of its high processivity. These reads make up a much smaller proportion in positive runs, suggesting ϕ29 will preferentially copy good quality DNA.

There was a significant presence of Mycobacterium sp which have previously been identified as a laboratory contaminant problem with the 16S sequencing service run in the same laboratory, with the most probable source being water. In this case the contamination appears to be post amplification; otherwise the Mycobacterium would be present in much higher numbers. Most Mycobacterium sp reads remained in the negative library which was later used to remove mapping reads. If these reads had remained they may lead to misassemblies and mis-annotations, appearing as gene variants or inserted genes. If there was a true infection with mycobacterium, there would be a higher coverage of the genome, and so this would still be identifiable. This highlights the need for manual visual monitoring of samples and contamination libraries side by side to monitor contaminants, false positives and true pathogen presence.

Other contaminating species were present in very low numbers, again suggesting post amplification contamination. All of these were species that had previously been sequenced in the

laboratory. The most likely source of these would be when loading the PTP, as the same loading apparatus is used for each sequencing run.

In conclusion removing contaminants, amplification and sequencing artefacts improves the reliability of the data. It also lowers the overall size of the files which need to be analysed, speeding up processing. However negative libraries need to be manually monitored and not just added into a negative reference file. Negative extractions and amplification will allow monitoring of all processes in the laboratory. Other studies have begun to recognise the importance of monitoring results for contamination, when 57 sequencing runs (across six sites) from the 1000 genome project were analysed reads classed as 'contaminating sequences' were present in all runs, representing a range of 0.000007% to 0.015% of all reads[155]. The main source to which these contaminants were attributed was an UltraPure water system, and a high level of *Bradyrhizobium* Spp was identified. A whole genome sequencing study of a blood sample from a febrile patient identified *Bradyrhizobium sp* as the possible cause citing that this was the first known infection with this bacterial species,[156] however this conclusion is quite possibly erroneous in light of the findings of Laurence et al [155]. No negative samples or environmental testing were undertaken in this study identifying *Bradyrhizobium sp* as a pathogen, this highlights the need for caution and the use of negative controls.

4.9.2 Quality Trimming of Raw Reads

To improve reliability of assemblies and data interpretation, poor quality data needs to be removed from the raw reads. The trimming method needs to be balanced to improve data quality whilst retaining as much information as possible. Two different trimming algorithms were investigated, running sum and window based, with the specific programmes selected based on the ability to trim 454 data.

There was no real difference observed after the control tags were removed (4.3.2), however Newbler is designed for 454 data and has inbuilt detection of these tags, other assemblers may not recognise these. No information is found in these tags and so they may as well be removed as part of the trimming process.

Ends of reads often have the lowest quality bases, removing poor quality bases at the end of reads allows better overlaps with higher confidence. After end quality trimming there was very little impact on the number of reads using Prinseq; however strict trimming (Q35) using Cutadapt

reduced the read length so far that some reads were removed entirely. Most of the impact on data was through the removal of bases, leading to shorter reads.

There was very little impact on the reference assembly after end error trimming with Prinseq, the main difference was more and shorter contigs, probably owing to the reduced read lengths, leading to less overlaps made. Both the Q30 and Q35 trimming using Cutadapt were actually detrimental to the reference assembly, lowering the reference coverage and increasing the number of contigs, as more information was removed with this trimming algorithm.

Strict (Q35) end error trimming was detrimental to the *de novo* assemblies both Prinseq and Cutadapt trimmed samples. Reference coverage was lowered to 90.93% and 75.2% from 94.12%. However, the Prinseq assembly actually had fewer and longer contigs, with the N50 increasing 155% and the longest contig increasing 127%, misassemblies dropped from 20 to 17. Cutadapt assembly had a shorter N50 (34%) and longest contig (65%), but the misassemblies did decrease to 18. Lower genome coverage is probably down to loss of information due to over trimming, especially using Cutadapt, which only left 40% of the bases.

Less strict (Q20) end trimming improved the *de novo,* with the reference coverage increasing to 94.85% using Prinseq and 94.29% using Cutadapt. Contig number was reduced, and the N50 length increased compared to the raw data to 222% using Prinseq and 220% using Cutadapt, and the largest contigs increased to 304% and 259%. Misassemblies dropped to 15 using Prinseq and 17 using Cutadapt. Fewer contigs were produced because the overall quality of overlaps increases allowing more overlaps to be accepted by the assembler. This leads to longer contigs being constructed because there are more high quality overlaps. There is particular improvement in the Q20 trimmed assemblies, as these have maintained the longest read lengths.

Cutadapt removes more bases in every quality cut off compared to Prinseq. Window based trimmers (Prinseq), make allowances for poor quality base that are surrounded by high quality bases, however running sum algorithms (Cutadapt) will take averages of sections of reads. Poor quality bases essentially lower the quality of the surrounding bases, and so high quality bases may be removed. Removal of more bases predictably leads to shorter reads, lowering overlaps causing higher contig numbers and shorter lengths of the resulting contigs. This was apparent in both the reference and *de novo* assembly. Whilst there was no improvement in reference assembly results using Q20 Prinseq, there was improvements in the *de novo* assembly, showing there is an advantage to end trimming.

In addition to end quality trimming, Prinseq has the ability to trim reads based on overall read quality. As expected the higher the quality metric the more reads were removed. The longest and average read lengths were reduced in all metrics, due to poor quality bases at the end of long reads, which lower the overall read quality. Very strict average read trimming (Q35) led to a drop in the genome covered to 78.3% after reference assembly; there was also an increase in contig number and a decrease in contig size, with the average contig size being only 7% of the original. This is probably due to the large amount of data lost to trimming, coupled with the removal of long reads, Q20 trimming also lead to a poorer performance of the reference assembly, although to a lesser scale. The Q35 trimming reference assembly was able to still cover 78% of the genome using only 18% of the original reads, suggesting the occurrence of poor quality reads was random and spread across the genome.

The *de novo* assembly of the strictly trimmed file only covered 34% of the genome, suggesting difficulty in matching and overlapping reads, probably due to the large data loss especially in read length. The *de novo* assembly of the Q20 average trimmed file was actually improved, with a slight increase in genome coverage (94.12% to 94.57%). The number of contigs was halved, and larger contigs were produced, both the average contig size and the N50 doubled in length. The *de novo* assembly has benefited from removal of erroneous reads, allowing more high quality overlaps to be made.

Overall there is very poor assembly of reads when very strict quality metrics are used, over trimming causes loss of data, particularly those reads with the longest length. Less strict error trimming had a slightly detrimental effect on the reference assembly, but actually improved the performance of the *de novo* assembly. Reference assembly metrics are better adapted to interpret erroneous reads as they have a something to compare the reads to and allow better interpretation of reads. *De novo* metrics must rely completely on the read information given, and removing erroneous reads leads to a less complex data set with only high quality reads, improving the performance of *de novo* assemblers. For rapid genome coverage assessment where a similar reference is available untrimmed data can be used to reference assemble, saving time and processing power needed by trimmers, however the improvements in the *de novo* assemblers is such that error trimming should be performed.

Very short reads (some as little as one base) left behind after error trimming contain little or no information, and increase the complexity of the assembly, and often will be ignored by assemblers. On its own it has negligible in its impact on assembly, but when combined with error

trimming parameters it will lower the number of reads the assemblers have to process without loss of information.

The impact of preceding average read trimming with end read trimming was investigated with the reasoning that by removing the poor quality ends from reads before average quality trimming, more reads and therefore more data should remain. As very high quality trimming has proved detrimental to assembly, Q20 was chosen as the average quality cut off and combined with various end quality trimming varying from Q20 to Q30.

Once again the reference assembly was slightly negatively impacted with all metrics, with the reads being assembled into more and shorter contigs, but there was no impact on the coverage of the genome. And in contrast the *de novo* assemblies were once again improved across all in terms of contig number and length, however the strictest end trimming did lead to a small decrease the reference coverage (94.12% to 93.3%), probably down to loss of data. The Q20 end trimming combined with Q20 average trimming led to the best performance with a small increase in reference coverage (94.73%), less than half the number of contigs, and the largest contig more than double in length compared to the raw data. Of particular note was the increase in the N50 value which more than trebled. The number of misassemblies was only 14 compared to the original 20. Again this is most likely down to the removal of poor quality overlaps which interfere with the assembly algorithm, good quality overlaps allow more to be accepted and larger contigs to be built.

Overall for rapid genome coverage assessment using a close reference, error trimming appears to be slightly detrimental and so may be contraindicated. However, for *de novo* assembly, the aim of this thesis, the quality of the output is improved. Additionally, in this data set window based trimming algorithms provide a better performance than running sum, allowing more data to be retained, whilst still improving the data quality. The combination of end quality trimming prior to average quality trimming allows a better outcome, by allowing more reads and therefore more data to be retained. Over trimming is detrimental to the assembly because a large amount of data is lost. The findings reported here are in line with those of a larger trimming study performed by Del Fabbro et al [104], who concluded that when trimming strictness was too high there was a significant reduction in the dataset lowering genome coverage, whereas too little trimming increased the data complexity and lowers overall data quality. By optimising the trimming parameters reliability of downstream analysis increases, Del Fabbro et al showed that the same quality cut-offs have different outcomes with different trimming algorithms. Ultimately

the data analysis pipeline has to reflect the biological question and the dataset being analysed, which has been achieved here.

4.9.3 Impact of Abundance Trimming on de novo Assembly

Data was digitally normalised to reduce the dynamic range of the coverage depth, using Khmer[107]. The aim of digital normalisation was to decrease the data size by discarding redundant reads, whist maintaining the information within the file. This software also has the additional benefit of removing those reads that appear at very low numbers, which are most likely due to random sequencing errors. This approach has been used previously to aid in the assembly of metagenomes[157], transcriptomes[158], as well as genomes[159].

The programme uses a reference free algorithm to estimate the genome coverage of the data by looking at the abundance distribution of kmers on the assumption that kmers tend to have similar abundances within a read. (The more times a piece of DNA is copied the higher the kmer abundance will be in that region). In the absence of errors, the average kmer number can be used to estimate the depth of coverage. Additionally, any kmers that overlap errors will have low abundance, which is the basis for the additional error trimming.

The initial stage focused on optimising the method for this data set by finding the value of K that gave the best assembly outcome. The overall aim was to reduce the peak depth to a depth similar to that achieved in the non-amplification assemblies (50), whilst not reducing the genome coverage achieved. A Kmer size of 20 removed more reads than any other value of K, with only 58% of reads remaining after normalisation of the raw data. The use of a short K value appears to have overestimated the read depth at each point, perhaps because more kmers will be present in each read. The peak depth was lowered to 26, much lower than the targeted peak depth, removal of so many reads also had a large impact on the average depth which was reduced to 9.6 from 16.5. However, there was only a very small reduction in the genome coverage from 98.46% to 97.97%, showing that the reads removed were from areas of high coverage, and so not losing much information.

When larger values of K (50-140) were used, more accurate estimation of the depth was achieved, with the resulting peak depth being reduced to a range of 49-58 across these kmer values. These kmer values also predictably had less impact on the average depth of the reads,

ranging from 14-14.6. The proportion of reads retained varied between 84.2% and 87.6%. Reducing the peak depth had very little impact on assembly, both reference and *de novo,* with the only real difference being a reduction in the misassemblies from 23 to 17-19 across the kmer range, probably mostly attributable to the removal of erroneous kmers.

The percentage reads removed from the post trimmed data was very similar to that from the non-trimmed dataset (84%-87%), lowering the peak depth by the same amounts. Again there was little impact on the assemblies, other than to lower the misassemblies, which were as low as 11 when using a kmer value of 140. When performing abundance trimming prior to error trimming, the amount of reads removed was again similar. The main difference was the further lowing of the misassembly rates to 9 in the k140 sample.

In conclusion, the removal of overrepresented reads lowers the amount of data to be processed without removal of information, and has very little impact on the assembly statistics. However, when combined with error trimming it does have a positive impact on the number of misassemblies present in the *de novo* assembly. One of the factors to bear in mind is this algorithm will remove repetitive sequences, which in some cases may contain additional information and may be used for typing (e.g. MIRU typing), however in most cases the read length will be longer than the repetitive element which will prevent their removal. The removal of low abundance kmers is a good additional error check, removing errors that other trimmers may miss. The error trimming is also performed in the context of the depth at that point, a fact that other error trimmers cannot account for, leading to less data loss. The Newbler assembler used does not use a kmer based algorithm, and so the impact on the assembly time is negligible, however other assemblers will be trialled for this pipeline and so the impact there might be greater.

4.9.4 Comparison of Genome Assembly quality using varying De Novo Assembler Algorithms

Robust *de novo* assembly will allow the data to be interpreted dependably; the best assembly for a data set will depend on the type of nucleic acid sequenced, the amplification method used and the platform upon which the DNA was sequenced. Numerous assemblers are publically available, utilising a number of different algorithms, several of these were investigated for their suitability to assemble single cell data from three bacteria with differing GC contents, as well as viral sequencing.

SSAKE (Greedy graph algorithm) was one of the very first assemblers to be designed for short read assemblers, for this data set the assembler performed poorly, with all the genomes having

low coverage, 21.9%, 58.8%, 11.8% and 65.1% for the *C. difficile, E. coli, A. naeslundii* and Adenovirus respectively. Short contigs and N50s suggest that the reads aren't successfully being joined. The method was optimised for read lengths of 50-300 base pairs in lengths, which are shorter than the reads produced on the 454 Junior.

Newbler (overlap consensus algorithm) is the recommended assembler for using with 454 data and was specifically designed for data produced by this sequencing chemistry. The assembler performed well across all genomes, with coverage being 89.1%, 95.3%, 65.59% and 96.9%. Contig numbers varied from 247 for the Adenovirus, and 569-1660 for the bacterial genome assemblies. Misassemblies were low across all of the assemblies, from 2-11, with the lowest being for the adenovirus, which is probably attributable to its small genome size. The highest was for the *C. difficile*, which may be down to the mosaic nature of the genome, with many insertion and repetitive elements which are difficult to assemble.

Three *De Bruijn* Graph Algorithms were tested with the four datasets. The first of these types of assembly algorithms to be trialled was Abyss, which had the highest kmer setting of 96, a short size considering the reads lengths produced by 454 sequencing chemistry. There was a decent coverage of the pathogens, 72.2%, 88.6%, 62.8% and 90.7% for the Adenovirus. There was a high number of contigs produced by this assembler (1621-2651), and the number of misassemblies was also high, between 32 and 52. The poor performance was perhaps due to the short kmer sizes, and the loss of the advantage of long reads.

The second of the *de Bruijn* graph algorithms to be tested was Spades; the algorithm is able to use several kmer sizes and the relationship between these to produce a distance histogram, allowing the route across the graph to be resolved using relative distances. This assembler provided very good coverage, 92.9%, 96.0% and 62.4% for the bacterial genomes and 97% for the adenovirus genome. This was the best resolution of the *C. difficile* genome using a *de novo* approach. The number of contigs produced was also the lowest; however, there was a still large number of contigs for *the A. naeslundii* genome, and the coverage was still lower than in the other bacteria.

The final *de Bruijn* graph algorithm to be tested was Ray; here the maximum kmer size available was 140. The genome coverage was quite low, especially for the *C. difficile* genome, which only covered 54.6%, suggesting either the assembler struggles to adapt to GC content, or has difficulty resolving repetitive elements. There misassemblies across all genomes were high, 21-38.

De Bruijn graph algorithms are increasing popular to assemble short read data sets, as they are rapid and fairly computationally light. It is important to note that the three *De Bruijn* assemblers didn't perform equally and so optimisation of assembly parameters for individual datasets is important, as resolution of GC content and repetition in the genome may be handled differently. We hypothesise that the main cause for poor performance on this dataset is the limitations caused by small kmer sizes, which would suit other data types such as those produced on the Illumina system.

Mira-4 is a Weighted Graph Algorithm assembler and produced good coverage of the genomes, 90.8%, 97.2%, 68.6% and 96.99%. This assembly shows good results across all the GC contents studied, however the contig numbers are still high, with *C. difficile* in 1385 contigs.

Overall Spades showed the best assembly performance on the dataset, in terms of genome coverage across GC contents, it also had the lowest contig number and most importantly the lowest misassembly number. Other assemblies studied here mainly struggled because of the limitation in kmer sizes, which was poor for this data type. Spades is specifically designed to use a variety of kmer sizes, allowing a more dynamic *de Bruijn* graph approach. Newbler also performed well on this data set, which is unsurprising as it specifically designed for 454 data. The preferred data entry for Newbler is the original SFF produced by the sequencer, however the additional benefits of error and abundance filtering of this dataset outweighs any additional performance using the SFF file.

4.9.5 Characterisation of Pathogens using de novo Assembled Genome Data

Preliminary investigation was carried out to compare annotation outputs across three datasets, closed complete 'reference' genomes, and two draft genomes consisting of *de novo* assemblies of culture controls and single cell ϕ29 MDA sequencing.

The published, annotated *E. coli* K-12[160] was reported to have 4288 genes, with 1632 being hypothetical, leaving 2656 remaining with putative functions. Use of Prokka in this study identified 446 more genes than previously stated, perhaps attributable to improvements in databases since this publication in 1997, and improvements in open reading frame calling. This also demonstrates the importance of controls when implementing new techniques, and not relying solely on references. Both the ϕ29 MDA and non-amplification control annotations produced a higher number of gene predictions; however, the difference was greatly reduced when the genes were filtered for repetition, suggesting that the cause behind the higher number

was the same genome elements being assembled into multiple contigs. 95% of all the genes identified were found in all three annotations, showing high concordance suggesting that the annotation of 'draft' genomes is comparable to polished, closed genomes. Additionally, the control and single cell data were comparable showing no increase in potential false genes introduced from single cells, or missing genes. The small number of differences could be accounted for due to differences in naming nomenclature, or with very similar genes being identified, (redundancy in the database), which is a downside to comparing exact strings of data. The ϕ29 MDA and non-amplification control samples both had a number of additional genes which were associated with phages, either suggesting the phage is extra chromosomal, and therefore not present in the reference genome, or the phage has been acquired during sub culturing events since the publication of the reference. Overall the *E. coli* annotation showed that the data produced from the single cell reaction was equivalent to the non-amplification control and comparable to the polished control.

This process was repeated with the *C. difficile* data. The publication associated with the reference genome[136] for this strain stated that the genome had 3776 genes, with 250 having an unknown function and 370 being classed as conserved unknown, leaving 3156 predicted genes. After repeat filtering the reference was predicted to have 2413 genes, so either there are genes missing from the database, or there was no repeat filtering performed. Many repetitive elements were identified in the *C. difficile* genome, such as clustered regulatory interspersed palindromic repeats (CRISPRs) and insertion elements so if these weren't filtered to singularity it would explain the disparity between results. This again highlights the need for controls when establishing new techniques for analysis.

The non-amplification sample had fewer genes identified than both the ϕ29 MDA and the reference, this may relate to the lower coverage found in the control, possibly due to the poorer DNA quality that was input into the sequencer due to the extraction technique. Overall 80.5% of genes were found in all three annotations, and 95% were shared between the reference and the single cell ϕ29 MDA. With the small discrepancies again probably down to different naming, or multiple database entry for the same or similar gene products. In this case the ϕ29 MDA from single cells actually outperformed the non-amplification and was comparable to the reference.

For *A. naeslundii* the reference paper suggested that there were 2761 [161] genes present in *A. naeslundii*, however there was no information regarding the number of hypothetical genes. The reference annotation in this studied gave 1420 genes after removal of hypothetical and repeat genes. 66% of the genes identified were found in all three annotations, with the non-amplification

and ϕ29 MDA sharing 92% of genes. The reference coverage of the *A. naeslundii* never exceeded 68%, which is in line with the number of annotated genes shared with the reference. The possible explanations for this include that the reference and the bacteria sequence are not the same strain, most probably due sample mislabelling when the bacterium was archived. Alternatives could be that genes have been lost, or the reference could be incorrectly joined creating a larger genome, or there could be considerable diversity within *A. naeslundii* and the reference genome and the genome analysed here could be at opposite ends of that diversity. Alternatively, these missing sections could be down to the sequencing errors, and the inability of the platform to cope with very high GC regions. The most probable explanation is that the reference genome does not match the bacteria sequenced in this study, and a smaller genome is present in this strain. The non-amplification and the ϕ29 MDA matched, and so any errors will be down to either the reference mismatching or the sequencing platform, and not the amplification of the single cell to sequencing quantities of DNA.

When the adenovirus sequencing was annotated, initial investigation suggested a poor outcome with only one gene found in both the reference and the ϕ29 MDA sample. However, upon closer inspection the discrepancies were resolved by comparison of function and in the context of viral expression. This was possible, although time consuming in this example where only 28 genes were identified in a single comparison, this however would be impractical at a larger scale. This highlights the problem of comparative genomics using a generic database, especially for viral genomes, where there are multiple naming nomenclatures and complex expression systems in viruses allowing multiple expression products from coding regions. Also the Adenovirus used here was grown in a tissue culture, but was originally isolated from a clinical isolate, and may have slight genome differences to the reference.

Overall the annotation of single cell ϕ29 MDA reactions has been comparable or better than multicellular inputs. The use of *de novo* multi contig inputs have performed well and given comparable results to the polished reference genomes, but have highlighted the need for filtering of repeat genes which may appear on more than one contig. Limitations of using a reference set to annotate data includes, inconsistency of the database due to multiple strain entries, spelling errors in the database, different naming nomenclature for the same gene and the misrepresentation of the output 'hypothetical protein'[162].

4.9.6 Virulence Prediction Using Genome Data

Further investigation into characterisation of pathogens using genomic data was performed. Firstly, methods for prediction of virulence factors were tested.

When using Prokka on the completed reference genome of *E. coli* K12 only 19 virulence factors were identified using Prokka, compared to 47 and 45 in the control and ϕ29 MDA sample. The majority of these were phage associated, suggesting phage acquisition since the reference was sequenced, or an extra chromosomal phage. There was also a conjugation factor identified, which is associated with the F plasmid.

Using the VFDB a large number of factors were identified, mostly originating from *Shigella* sp. *E. coli* and *Shigella* are genetically similar, with many systems failing to differentiate them, including 16S studies, multi-locus enzyme electrophoresis (MLEE) and a housekeeping gene sequence study[163] and so it is no surprise to find *Shigella* associated genes in the *E. coli* strain studied. However, many of the genes identified actually have no known function, and only appear in the database as they were identified in pathogenic *Shigella*. This highlights a weakness in this database, particularly with the inclusion of genes without solid functional evidence of being a virulence gene. It also highlights issues of using genomic data for *E. coli*, and there is a huge genetic and disease type variation of this organism. This comparative method seems unsuitable for *E. coli* and highlights that careful choices of database are required for virulence prediction, and expert interpretation is still required.

Using Prokka for the three *C. difficile* assemblies, overall 13 out of 16 factors were identified in all the samples, with 15 out of 16 being identified in both the ϕ29 MDA and the reference. Importantly toxin A and B were identified in all samples, and additional factors identified included transposons and phage elements. The pathogenicity island was identified in the reference and ϕ29 MDA but not culture control. Once again the ϕ29 MDA performed better than the control and equal to the reference in prediction of factors. Using VFDB, a very clear prediction of the toxin A and B was identified in all samples. This database worked well in the case of toxin detection for *C. difficile*, although no additional data was found, here selection of Prokka elements performed better. Use of databases compiled by others restricts what factors will be identified.

When using Prokka for *A. naeslundii* only four factors were identified in the reference and ϕ29 MDA, and only two were found in the control. The reference has phage elements not found in the other two sequence files, probably explained by previous mis-matches identified in the genomes. The VFDB produced no hits, these results either demonstrate the low pathogenicity of the isolate or the lack of study means there are few genes identified as causing virulence.

The use of databases to predict virulence has had mixed results on this dataset. In the case of *E. coli* use of VFDB gave a huge number of hits, with very little application to this strain of *E. coli*, over estimating the potential virulence of the bacteria. In the case of *A. naeslundii*, no factors were identified, despite evidence of its ability to cause infections[164]. However, *A. naeslundii* is considered an opportunistic pathogen, which is associated with less traditional virulence factors, and so more investigation of factors which cause this bacterium to become pathogenic would be required. Although there are demonstrable problems with using comparative genomics for prediction of virulence, for rapid *de novo* characterisation use of data bases is the best current option. Outputs need to be intelligently interpreted to prevent over estimation of virulence, such as in the case of *E. coli* in this study. In general, when using external databases, the quality of the output is dependant of the quality of the database including contributors and curators. As whole genome sequencing becomes more widely used, these databases will grow and their duration will become increasingly important, possibly too important to be left to purely voluntary curators.

4.9.7 Resistance Prediction using Genome Data

A significant proportion of the genes on bacterial genomes are responsible for environmental survival, which includes resistance to toxic substances. When searching for the word 'resistance' within the Prokka output files, all relevant genes that were identified were also identified using another term in the search, such as 'drug', making this search term redundant. Efflux pumps are increasingly being appreciated as important resistance mechanisms, especially where multi factor resistance is present; however, efflux pumps are also essential for bacterial survival. Similarly, to resistance, searching for the term efflux produces a large number of irrelevant results. When an efflux pump is associated with drug resistance they are mostly renamed accordingly and so the word efflux was removed. However, if a study was to look for difference between two similar strains, but one had an increased antibiotic resistance, the term could once again be included, which would potentially prove useful in comparative studies.

Using the Prokka output from *E. coli*, there was a large and complex output, which was difficult to interpret. Many drug efflux systems were identified, as well as many PBPs, which would need to be investigated to distinguish housekeeping genotypes from those which are associated with drug tolerance or resistance. Potentially a second database would need to be built to remove commonly found elements (such as housekeeping PBP), which don't confer resistance. In total 51 out of 53 elements were found in all three datasets, showing good performance in draft genomes compared to the completed reference. Using *ardbAnno.pl* no resistance factors were identified, suggesting this strain has no acquired resistance factors. Using Resfinder the bacteria was identified as having the blaCMY genotype, which is an ampC that

confers resistance to ceftriaxone[165]. It also identified oqxB which is part of the oqxB efflux system that is a multidrug efflux pump[166], however when oqxB is present on its own it has been shown that it is simply part of the membrane make-up. This shows that using more targeted tools, such as ResFinder, can provide improved results compared with general tools such as Prokka. However, it also emphasises the need for context and expert interpretation of results, as there is not always a simple relationship between genotype and phenotype when it comes to complex drug resistance.

In the case of *E. coli,* the three databases all produced very different results, either an overwhelming number of results with difficult biological interpretation, no results, or a simple output with limitations.

Using the Prokka output to predict resistance in *C. difficile* 39 factors were identified in the reference, 33 in the control and 36 in the ϕ29 MDA. Several resistances were identified in all, including beta-lactams, metallo-beta-lactamase, resistance to flouroquinolones including ciprofloxacin, resistance to tetracycline and streptomycin. VanB was also identified as being present, although only *VanW* was identified through similarity, which is only one gene in the seven gene operon, and isn't even essential for production of resistance. *ArdbAnno.pl* identified ermB which induces resistances to erythromycin, clindamycin and macrolides. Resfinder identified ermB and tetM (tetracycline resistance). This demonstrates the current need to look at several sources to gain a full picture of the resistance factors that are present. It also highlights again the need to match prediction with the biological output, particularly when partial operons are shown. Overall the ϕ29 MDA showed good concordance with the reference.

Most of the factors identified by Prokka as resistance factors in the *A. naeslundii* assemblies were efflux pumps and PBP, which are commonly found in all bacteria. This is probably attributable to the lack of studies focussing on the genome of this organism. In the culture control and ϕ29 MDA a beta-lactamase was identified that wasn't present in the reference which has either been acquired since the reference sequence production, or is additional evidence to the sequenced strain being different to the reference. *ArdbAnno.pl* and Resfinder didn't identify any resistance factors.

It is beyond the scope of this thesis to write new databases for antibiotic resistance predictions, several new species specific databases are available to use, which are useful if one of these species is identified. However, this did demonstrate that the ϕ29 MDA sample is as reliable as the reference for prediction of antibiotic resistance from genome data.

It is also worth noting that through this study it is obvious that some bacteria such as *E. coli* have been extensively studied, and so the databases are almost overpopulated with predictions for genome function. Whereas other bacteria such as *A. naeslundii* have been almost neglected in studies making prediction from genomes very limited. Prokka often outputs complex results which are hard to interpret on their own, making targeted tools like ResFinder more appropriate, however when comparing two similar bacteria comparison of the resistance output might prove valuable.

4.9.8 Analysis of Mixed Cell input sequencing data

Use of Blastn and LCA analysis allows the identification and separation of mixed bacterial species allowing *de novo* assembly. The limit of detection for mixed cells lies between the 1:100 and 1:1000 dilution factor as in the 1:1000 sample no chromosomal Enterococcus genes are detected, however phages, which may be present at more than one per cell, were detected confirming the presence of the Enterococcus in the sample. After abundance trimming the relative abundance of the two bacteria was closer to the original mix ratio. One possible explanation for this could be removal of repetitive genome elements, as some genomes will have a higher proportion of repetitive elements than other, and removing them will help normalise the data. This assay would not be able to exactly quantify the input cells, but it would be able to give an estimation of the relative abundance of the bacteria. When characterising the Enterococcus at lowering relative abundance, 77% of the gene products were identified in the 1:10 sample and 29% were identified in the 1:100 sample. Demonstrating that even at very low numbers without whole genome coverage it is still possible to gain insight into the genome.

5. Application of ϕ29 MDA to Real and Modelled Clinical Samples

In order to take advantage of the single cell amplification, sequencing and interpretation methods developed in this thesis, a process for isolating pathogens from clinical samples was developed. Initial development work focused on direct pathogen sequencing from blood. Bacteraemia is a major cause of morbidity and mortality, and is a particular problem in healthcare settings. Current methods for diagnosis of bacteraemia involves specialist equipment in the form of blood culture machines, which constantly monitor blood samples for possible bacterial growth. Once growth is detected these blood cultures must be sub cultured in order to isolate and identify the bacterial pathogen present. Applying whole genome sequencing to blood samples would allow rapid pathogen diagnosis, along with simultaneous pathogen typing. The application of unbiased pathogen detection would also allow a combined workflow for viral and bacterial pathogens.

A bacteraemia model was developed, using defibrinated horse blood to model whole human blood infection additionally selective host cell lysis was explored using Saponin. Separation of RBCs from other cellular components in blood was investigated using both density gradients and specific RBC aggregation. HetaSep® induces aggregation of red blood cells through the formation of stacks of erythrocytes which leads to a rapid sedimentation rate. And has previously been used to isolate nucleated cells in the blood particularly granulocytes.

Once a specific workflow for isolating pathogens from whole blood is developed it will then be applied to multiple sample types, both real and modelled. This will allow testing of the isolation and amplification of pathogens to multiple sample types. Additional bacteraemia models will be performed to evaluate the use of ϕ29 MDA to more challenging pathogens. Also tissue models will be investigated to investigate the potential for isolating and sequencing of pathogens from tissue infections. Various clinical samples were obtained from multiple sources and tested, including urines, CSF and STI swabs.

5.1 Blood Culture Model

5.1.1 Survival of Bacterial in Horse Blood and 2% Saponin

Initial investigations were undertaken to examine bacterial survival in horse blood and saponin. A panel of 27 bacterial isolates were collected from the Royal Free Hospital Hampstead, from clinical cases, with the details **Table 2-2**. These bacteria represented a variety of cell wall types, morphology and growth requirements. For these 27 isolates approximately 500 bacterial cells were spiked into 1 ml of horse blood, 2% saponin, PBS or 1ml of 2% saponin in horse blood. The samples were then incubated at room temperature for 2 hours, and the number of bacterial cells was calculated by culturing three sets of 100 µl in each condition and counting the CFU, survival percentage was calculated using the PBS as the control. Survival of isolates in all conditions is shown in **Table 5-1** and **Figure 5-1**. All of the isolates showed better that 95% survival in 2% saponin, except *Paenibacillus anaericanus* (87.2%) and *Streptococcus pneumoniae* (94%). 20 of the isolates showed greater than 95% survival in horse blood. Seven of the isolates had poorer survival in the horse blood, *Staphylococcus epidermidis* (73%), *Streptococcus pneumoniae* (6%), *Proteus vulgaris* (60.9%), *Streptococcus agalactia*e (45%), *Listeria monocytogenes* (50.3%), *Proprionibacterium acnes* (23.4%) and *Shigella sonnei* had a 33.3% survival rate. Nine of the isolates showed altered colony morphology after incubation in horse blood, with the colonies appearing smaller and with a lobate margin, marked with a * in **Table 5-1**(*Escherichia coli, Staphylococcus aureus* two isolates, *Klebsiella pneumoniae, Pseudomonas aurginosa* one isolate, *Enterobacter cloacae, Salmonella Typhimurium, Serratia marcescens* and *Listeria monocytogenes)*.

Isolate number	Species	PBS			2% Saponin				Horse blood				Horseblood and 2%saponin			
		1	2	**av**	1	2	av	%	1	2	av	%	1	2	av	%
1	*Escherichia coli*	58	49	**53.5**	56	54	**55**	102.8	53	54	**53.5**	100.0	58	51	**54.5**	101.9
2	*Staphylococcus epidermidis*	36	38	**37**	34	38	**36**	97.3	26	28	**27**	73.0	31	35	**33**	89.2
3	*Staphylococcus aureus*	68	62	**65**	72	69	**70.5**	108.5	71	65	**68**	104.6	69	61	**65**	100.0
4	*Staphylococcus aureus*	75	67	**71**	69	75	**72**	101.4	73	71	**72**	101.4	68	79	**73.5**	103.5
5	*Enterococcus faecium*	56	49	**52.5**	52	53	**52.5**	100.0	48	58	**53**	101.0	53	49	**51**	97.1
6	*Klebsiella pneumoniae*	44	52	**48**	57	51	**54**	112.5	51	48	**49.5**	103.1	51	54	**52.5**	109.4
7	*Streptococcus pneumoniae*	25	25	**25**	24	23	**23.5**	94.0	1	2	**1.5**	6.0	3	0	**1.5**	6.0
8	*Pseudomonas aeruginosa*	62	65	**63.5**	68	62	**65**	102.4	75	72	**73.5**	115.7	65	70	**67.5**	106.3
9	*Pseudomonas aeruginosa*	72	75	**73.5**	75	69	**72**	98.0	78	72	**75**	102.0	71	84	**77.5**	105.4
10	*Proteus vulgaris*	93	99	**96**	101	87	**94**	97.9	10	107	**58.5**	60.9	112	106	**109**	113.5
11	*Enterobacter cloacae*	29	52	**40.5**	31	50	**40.5**	100.0	44	46	**45**	111.1	43	39	**41**	101.2
12	*Streptococcus agalactiae*	19	21	**20**	18	20	**19**	95.0	8	10	**9**	45.0	6	11	**8.5**	42.5
13	*Bacteroides vulgatus*	58	61	**59.5**	57	59	**58**	97.5	62	57	**59.5**	100.0	63	57	**60**	100.8
14	*Streptococcus oralis*	51	48	**49.5**	55	51	**53**	107.1	51	49	**50**	101.0	55	49	**52**	105.1
15	*Streptococcus mitis*	38	36	**37**	42	35	**38.5**	104.1	36	35	**35.5**	95.9	41	45	**43**	116.2
16	*Streptococcus anginosus*	72	71	**71.5**	78	75	**76.5**	107.0	69	81	**75**	104.9	75	76	**75.5**	105.6
17	*Haemophilus influenzae*	15	14	**14.5**	15	15	**15**	103.4	16	12	**14**	96.6	15	13	**14**	96.6
18	*Streptococcus pyogenes*	52	51	**51.5**	49	55	**52**	101.0	58	52	**55**	106.8	61	49	**55**	106.8
19	*Salmonella Typhimurium*	87	88	**87.5**	91	86	**88.5**	101.1	96	91	**93.5**	106.9	90	94	**92**	105.1
20	*Serratia marcescens*	102	110	**106**	108	100	**104**	98.1	111	111	**111**	104.7	123	114	**118.5**	111.8
21	*Fusobacterium necrophorum*	102	109	**105.5**	111	99	**105**	99.5	112	103	**107.5**	101.9	121	99	**110**	104.3
22	*Listeria monocytogenes*	90	75	**82.5**	95	80	**87.5**	106.1	42	41	**41.5**	50.3	43	39	**41**	49.7
23	*Actinomyces naeslundii*	99	112	**105.5**	101	107	**104**	98.6	125	132	**128.5**	121.8	89	145	**117**	110.9
24	*Proprionibacterium acens*	46	47	**46.5**	49	52	**50.5**	108.6	51	48	**49.5**	106.5	55	47	**51**	109.7
25	*Paenibacillus anaericanus*	25	22	**23.5**	22	19	**20.5**	87.2	2	9	**5.5**	23.4	4	8	**6**	25.5
26	*Clostridium butyricum*	52	51	**51.5**	48	53	**50.5**	98.1	42	56	**49**	95.1	42	54	**48**	93.2
27	*Shigella sonnei*	37	26	**31.5**	32	33	**32.5**	103.2	11	10	**10.5**	33.3	13	12	**12.5**	39.7

Table 5-1 survival rates of 27 clinically isolated bacteria in PBS, horse blood and saponin, numbers represent average CFU of each bacterium calculated from three replicates

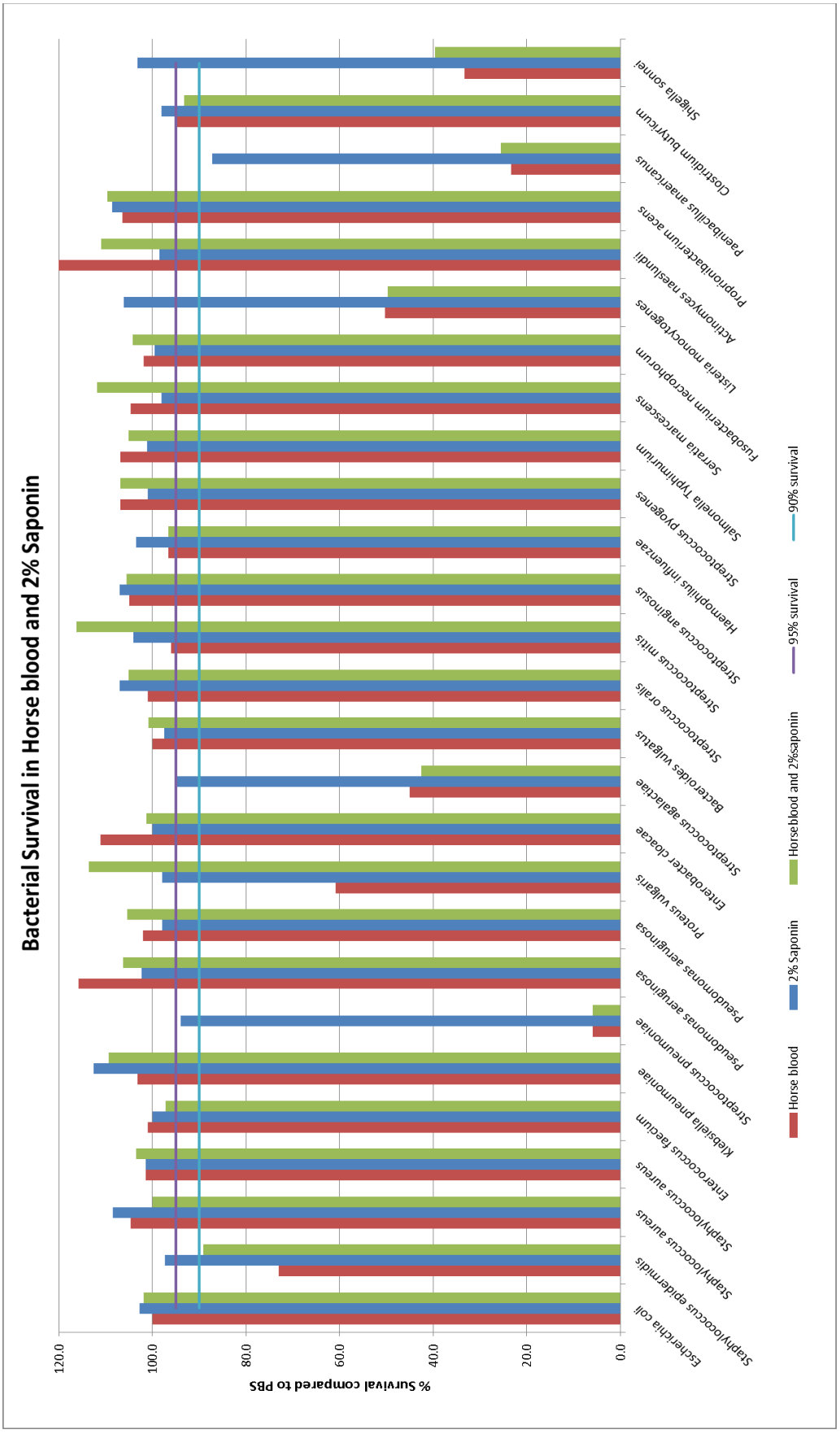


Figure 5-1 Survival rates of 27 clinically isolated bacteria in horse blood, 2% saponin and combined horse blood and saponin. Survival rates are compared to the same incubation in PBS

5.1.2 Bacterial Isolation using Density Gradients

To assess the applicability of using density gradients to isolate bacteria from whole blood, *E. coli* and *S. aureus* were spiked into two separate 1:1 mixture of horse blood and PBS. This was then slowly added to a Percoll density gradient, described in **2.6.2.** After centrifuging for 25 minutes there were three visible bands, the top segment being a pale red colour of about 3 ml in volume. The middle layer was very dense and red, being about 2.5 ml in volume. The bottom layer was clear and about 500 µl in volume.

The positive control (same inoculum cultured) indicted that the starting numbers of bacteria cells were 150 CFU for *E. coli* and 123 CFU for *S. aureus*. After the spiked blood was centrifuged with the Percoll to create a gradient, culture was performed at several points down the gradient (Table 5-2). Each bacterial spiking experiment was performed in triplicate and averages reported. When the *E. coli* spiked sample was investigated, bacterial cells were recovered at most points in the gradient. The three samples from the top segment gave average results of 1, 35 and 40 CFU. The middle segment gave average results of 20, 30, 26 and 4 CFU at the four sampling points. No *E. coli* cells were recovered from the bottom segment. When *S. aureus* spiked blood was centrifuged into a Percoll gradient no *S. aureus* was recovered from the top segment. From the middle segment 30, 35, 41 and 12 CFU of *S. aureus* was recovered on average at each of the four sampling points. Four addition CFUs of *S. aureus* were recovered from the very bottom segment of the gradient. The number of CFUs of each bacteria recovered is shown in **Table 5-2** alongside the volume of sample cultured at each point, the sedimentation of *E. coli* and *S. aureus* compared to RBCs is visualised in **Figure 5-2.**

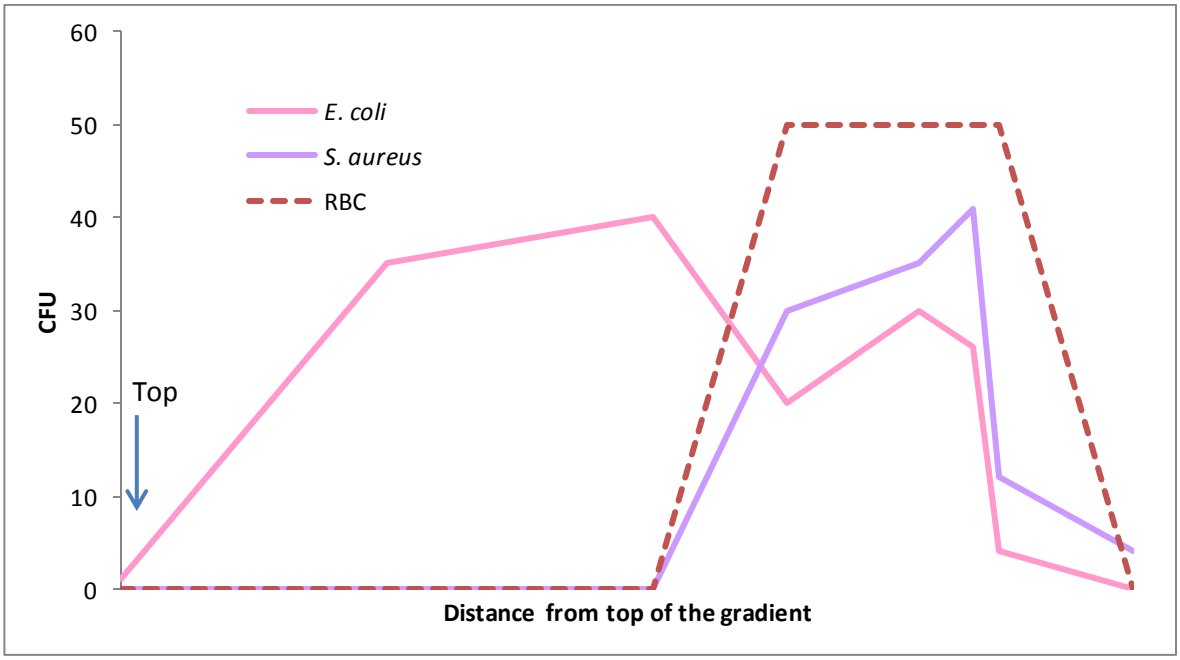


Figure 5-2 graphical representation of the sedimentation of *E. coli* and *S. aureus* compared to RBCs in a Percoll gradient (RBCs numbers for illustration purposes only, to demonstrate RBC sedimentation position in the Percoll gradient)

Layer	Volume	*E. coli* (CFU)	*S. aureus* (CFU)
Top	1ml	1	0
Top	1ml	35	0
Top	1ml	40	0
RBC	500µl	20	30
RBC	500µl	30	35
RBC	200 µl	26	41
RBC	100 µl	4	12
Clear	500 µl	0	4

Table 5-2 sample volumes and CFU recovery of sampling points in a Percoll gradient to show sedimentation of *E. coli* and *S. aureus* in whole horse blood

5.1.3 Erythrocyte Depletion using HetaSep®

5.1.3.1 Centrifugation vs. Gravity Sedimentation

HetaSep® can either be incubated at 37°C to allow sedimentation of RBCs, or can be centrifuged to facilitate sedimentation. Bacterial cells (*E. coli* and *S. aureus*) were spiked into either neat blood or blood diluted 1:1 in PBS. HetaSep was added and the samples either centrifuged or incubated at 37°C. Three replicates for each condition were performed and averages calculated. After both centrifugation and incubation at 37°C there was separation of the sample with RBCs pelleted at the bottom of the tube and a red coloured clear liquid at the top of the tube. There was also a small joining phase was visible between the two sections.

The control samples (bacterial cells incubated in PBS) contained on average 18 CFUs *S. aureus* and 27 CFUs *E. coli*. In the centrifuged sample with *E. coli* the top segment contained and average of 20 CFUs, the middle segment contained 5 CFUs and the bottom contained no bacterial cells. When the blood was dilution 1:1 in PBS the top segment contained 15 CFU, the middle segment had 2 CFUs and the bottom contained 10 CFUs. In the centrifuged sample with *S. aureus* 12 CFUs were found in the top segment 1 CFU was isolated from the middle segment and 4 CFU in the bottom segment. With the added PBS 15 CFU were found in the top segment, 1 CFU in the middle segment and 2 CFUs were found in the bottom segment **Table 5-3.**

In the gravity sedimented samples, no bacterial cells were found in the bottom segment of any samples. In the *E. coli* sample 21 CFUs were found in the top and 2 CFU in the middle segment. When samples were diluted with PBS 25 CFUs was recovered from the top segment and 5 CFUs were recovered from the middle segment. In the *S. aureus* sample 18 CFU were found in the top segment and 1 CFU was found in the middle segment. With the additional PBS 15 CFUs were found in the top segment and 2 CFUS were recovered from the middle segment.

	Upper Segment (CFU)	Middle segment (CFU)	Lower Segment (CFU)
	Centrifuge		
E. coli	20	5	0
***E. coli* with PBS**	15	2	8
S. aureus	12	1	4
***S. aureus* with PBS**	15	1	2
	Gravity sedimentation		
E. coli	21	2	0
***E. coli* with PBS**	25	5	0
S. aureus	18	1	0
***S. aureus* with PBS**	15	2	0

Table 5-3 average number of bacterial CFU (*E. coli* and *S. aureus*) recovered from the three different HetaSep segments after centrifugation or incubation at 37°C with and without PBS dilution.

5.1.3.2 Single vs. Double Incubation with HetaSep

In order to remove more of the RBCs from the middle segment an additional incubation stage with HetaSep was investigated. The single incubation was performed as before with the samples being incubated at 37°C for 20 minutes and culture performed from each segment. For samples having a double incubation, after 10 minutes' incubation at 37°C the top and middle segment was removed and more HetaSep was added and the sample was again incubated at 37°C for 15 minutes.

From the *S. aureus* control 89 CFU were isolated and from the *E. coli* control 125 CFU were isolated. The one step separation showed a clear separation with a top clear (red coloured) segment and a dark bottom layer, with a small visible stage in between. With the single incubation samples, no bacteria cells were isolated from the bottom segment. The majority of bacterial cells were isolated from the top layer, 115 CFUs of *E. coli* and 75 CFUS of *S. aureus*. The middle segment also contained some bacterial cells, 12 CFUs of *E. coli* and 6 CFU of *S. aureus* **(Table 5-4)**

The separation after the second incubation seemed to be less defined that previously. The top segment still had the majority of bacterial cells, 103 CFUs and 70 CFUs for *E. coli* and *S. aureus* respectively. The middle segments had slightly higher numbers than in the single incubation sample 26 CFUs for *E. coli* and 14 CFUs *S. aureus*. The bottom segment of the *E. coli* sample had no bacterial cells, but 2 CFUs were found in the *S. aureus* sample. No bacteria were isolated from the pellet left after the initial incubation of 10 minutes **(Table 5-4)**.

	Upper Segment (CFU)	Middle segment (CFU)	Bottom segment (CFU)	
	Single Incubation			
E. coli	115	12	0	
S. aureus	75	6	0	
	Double Incubation			Bottom segment (first incubation)
E. coli	103	26	0	0
S. aureus	70	14	2	0

Table 5-4 number of bacterial CFUs present at each point in the HetaSep gradient. Comparing recovery after either a single or double incubation with HetaSep

5.1.4 Low Level Bacterial Isolation from Horse blood

Approximately ten bacterial cells of either *E. coli* or *S. aureus* were spiked into 1 ml of horse blood and processed as described in **2.6.4.** After the completion of each stage of the process, the sample was cultured to identify bacterial survival **Figure 5-3** and **Table 5-5.**

The initial starting CFUs were obtained by culturing the same volume of spiking material as was used to spike the blood samples. This yielded 9 CFUs of *S. aureus* and 12 CFUs for *E. coli*. After incubating the sample with HetaSep to pellet the RBCs, the 500 µl supernatant contained 8 CFUs for the *S. aureus* sample with 1 CFU isolated from the pellet. A similar pattern was observed in the *E. coli* sample with 9 CFUs isolated from the supernatant and 2 CFUs from the pellet. Once the saponin treatment was completed 12 CFUs of *S. aureus* and 10 CFUs of *E. coli* were recovered. Water shock treatment didn't affect the *S. aureus* recovery (9 CFUs) but had an impact on *E. coli* CFU numbers which decreased to 5 CFU. Post-salt restoration samples showed similar numbers to those recovered from the water shock samples (8 CFUs for *S. aureus* and 5 CFUs for *E. coli*). All *E. coli* cells pelleted during the centrifugation step at 3000xg whereas only 2 CFUs of *S. aureus* were recovered from the pellet. For all washes at 6000xg both bacteria were only recovered from the pellet. When the DNA in the final supernatant was quantified, 0.35 ng/ µl was detected from the *S. aureus* sample and 0.76 ng/ µl was detected in the *E. coli* sample.

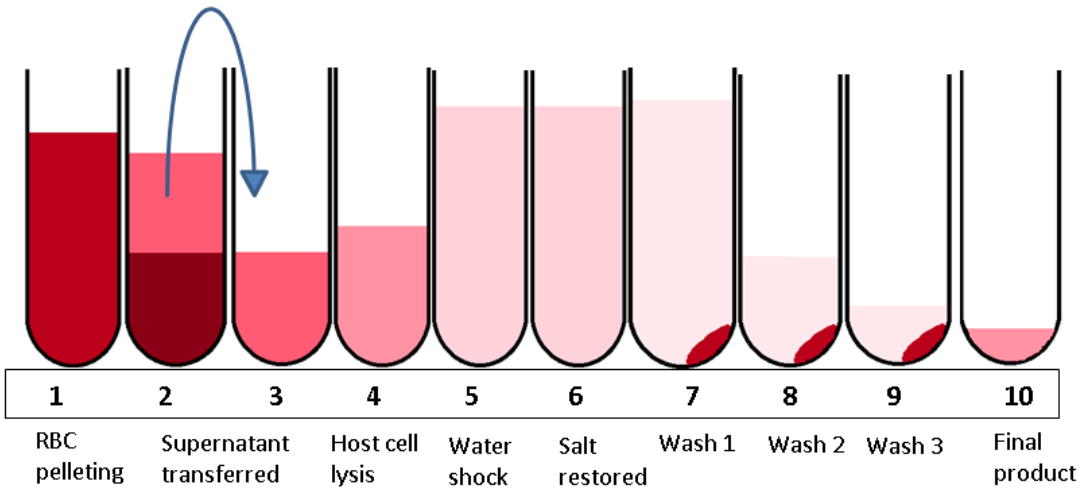


Figure 5-3 Illustration of the process for bacterial isolation from whole blood using HetaSep and selective lysis with Saponin. Numbers refer to sampling points where bacterial recovery was investigated.

Stage		*S. aureus*	*E. coli*
1	Spiked cells	9	12
2	RBC separation-bottom	1	2
3	RBC separation-top	8	9
4	Addition of Saponin	12	10
5	Water shock	9	5
6	Salt restoration	8	5
7	Cell pelleting-top Cell pelleting-bottom	2 11	0 5
8	Wash 1-top Wash 1-bottom	0 8	0 4
9	Wash 2-top Wash 2–bottom	0 9	0 5
10	Final pellet	10	5

Table 5-5 the number of CFU isolated at each processing stage for bacterial isolation from whole blood using HetaSep and selective lysis with Saponin. Numbers refer to sampling points where bacterial recovery was investigated as illustrated in **Figure 5-3**.

5.1.4.1 Assessment of DNase inactivation

To remove host DNA in the supernatant following cell lysis a DNase treatment was used, however heat inactivation of the enzyme caused remaining RBCs and other blood components to coagulate. Therefore, the removal of the DNase activity through chelation and multiple washes was investigated. DNase was added to either PBS or horse blood post RBC aggregation and saponin treatment. Positive controls were set up alongside, to which DNA was added to a concentration of 50 ng /µl. The samples were then incubated at 37°C for 15 minutes, after incubation the concentration of DNA in the positive controls were 0.06 ng / µl in the PBS and 0.08 ng / µl in the processed horse blood sample. For the test samples after incubation with DNase EDTA was added and the samples were washed three times with PBS. DNA to a final concentration of 50 ng / µl was added to both and the samples incubated at 37°C for one hour. DNA concentrations after incubation were 49.2 ng / µl in the PBS sample and 50.3 ng / µl in the processed horse blood.

5.1.4.2 Improvements to Work Flow

Based on the results from **5.1.4** several improvements were made to the work flow. This included increasing the volume of supernatant kept after RBC pelleting, reducing the water shock treatment time and increasing the centrifugation speed from 3000xg to 4000xg. Additionally, a DNase stage was included and an extra pellet wash was added. The full details of the improved work flow are described in **2.6.4.1.**

The number of cells initially spiked into the blood was quantified as 10 CFU for *S. aureus* and 11 CFU for *E. coli*. Similarly, to the previous work flow samples were cultured at several points along the process to ascertain bacterial survival rates. After aggregation of the RBC using HetaSep no bacteria were isolated from the RBC pellet. (10 CFU *S. aureus* and 9 CFU *E. coli* was cultured from the supernatant). Incubation with saponin caused no decrease in the number of CFU recovered for either bacterium. The decreased water shock improved the recovery of *E. coli* with 12 CFU cultured from this stage. After initial pelleting of the bacterial cells at the higher speed of 4000xg no bacterium were found in the supernatant (9 CFU *S. aureus* and 9 CFU *E. coli* was cultured from the pellet.) DNase treatment and inactivation had no impact on bacterial survival and all three wash stages only recovered bacteria in the pellets. The final number of bacteria recovered on completion of the process was 11 CFU for *S. aureus* and 9 CFU for *E. coli*.

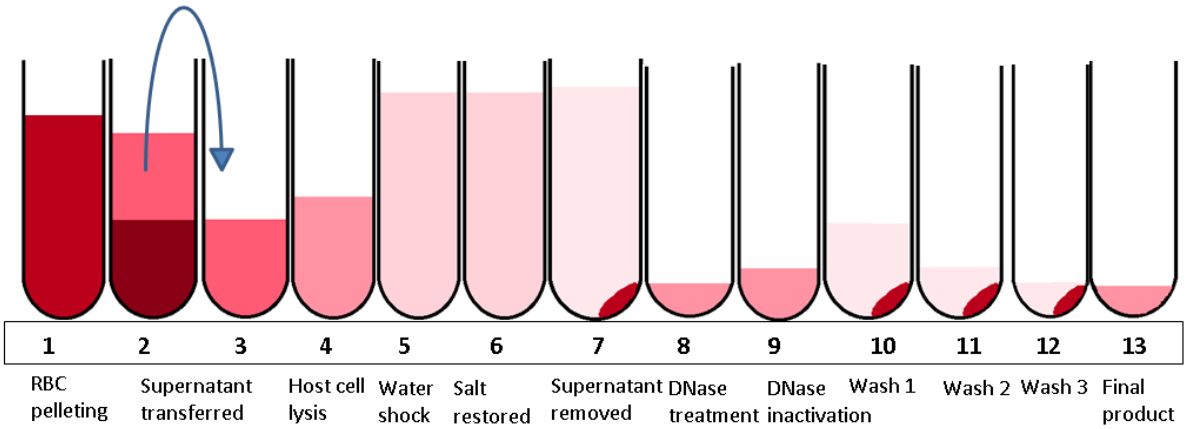


Figure 5-4 Illustration of the process for bacterial isolation from whole blood using HetaSep and selective lysis with Saponin after workflow improvements. Numbers refer to sampling points where bacterial recovery was investigated.

Stage		*S. aureus*	*E. coli*
1	Spiked cells	10	11
2	RBC separation-bottom	0	0
3	RBC separation-top	10	9
4	Addition of Saponin	12	9
5	Water shock	12	12
6	Salt restoration	10	10
7	Cell pelleting-top Cell pelleting-bottom	0 9	0 9
8	DNAse treatment	11	12
9	EDTA and PBS	10	12
10	Wash 1-top Wash 1-bottom	0 9	0 11
11	Wash 2-top Wash 2–bottom	0 10	0 10
12	Wash 3-top Wash 3–bottom	0 8	0 9
13	Final pellet	11	9

Table 5-6 the number of CFU isolated at each processing stage for bacterial isolation from whole blood using HetaSep and selective lysis with Saponin after workflow improvements. Numbers refer to sampling points where bacterial recovery was investigated as illustrated in **Figure 5-4**

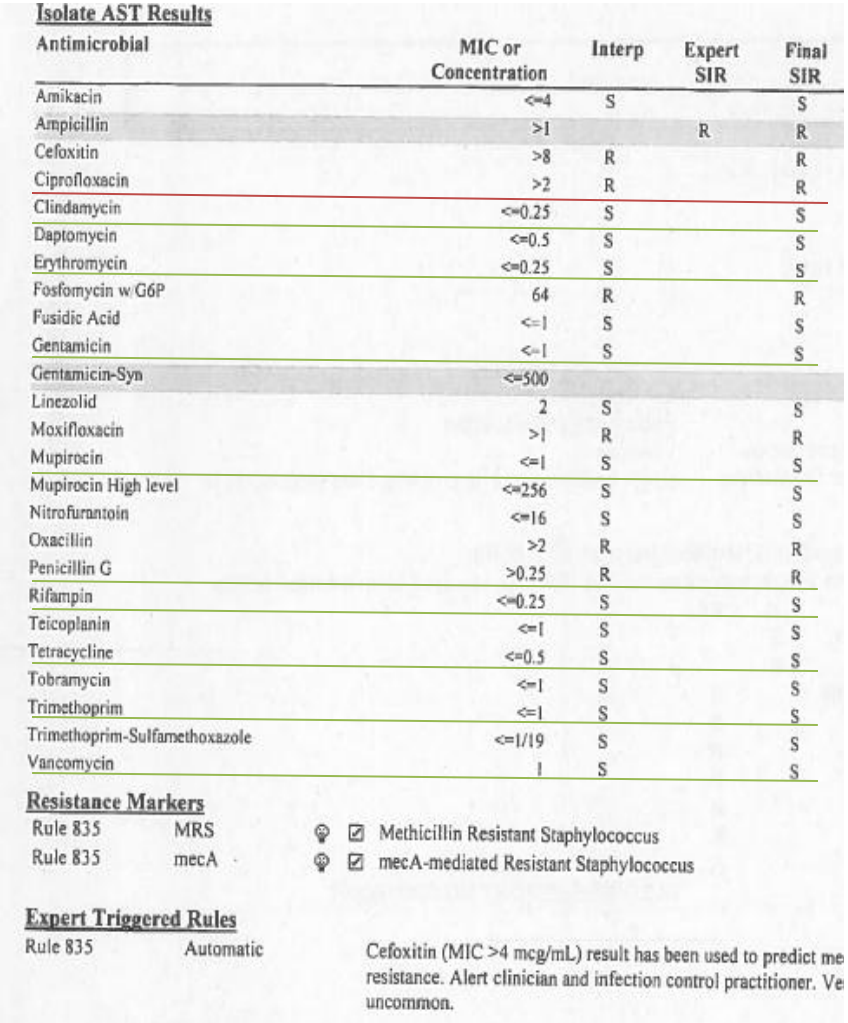
5.1.5 Sequencing from Low Levels in Blood

Single cells of *E. coli* and *S. aureus* were spiked into 1ml of horse blood to simulate low level bacteraemia. This was then processed as shown in **Figure 5-4,** with the exception of the final pellet being suspended in 4 µl PBS. The resulting sample was then extracted and amplified for two hours. The resulting DNA was then sequenced on the 454 Junior and the resulting file analysed using the pipeline described in **section 2.7.3**

5.1.5.1 S. aureus

After processing and sequencing the horse blood spiked with *S. aureus* the number of reads passing filter was 128500.Once the processing pipeline described in 2.4.13 was complete, 124,145 reads remained. Using Blastn and Megan, 77131 (62.1%) reads were identified as *S. aureus*. 8,395 (6.76%) of the reads were identified as the genus Equus, and 2,138 (1.72%) were identified as *Parascaris equorum*. When examining the *S. aureus* reads closer 4254 reads were identified the subspecies level, (*Staphylococcus aureus subsp. aureus* HO 5096 0412), this subspecies has a complete genome available, and so the fasta file was downloaded from NCBI and used as a reference for reference mapping. The reference mapped reads were assembled into 451 contigs and covered 92% of the reference genome. When the reads identified as *Staphylococcus* using the LCA analysis were extracted and the reads de *novo* assembled 1212 contigs were produced. When this was compared to the same reference as the reference assembly 83% of the genome was covered with 10 misassemblies and a N50 of 3882. The *de novo* assembly was then investigated using a programme specifically written to predict resistance in *S. aureus*, Mykrobe-predictor [167]. Using Mykrobe a genotypic result was produced for 12 antibiotics, which were compared to the phenotypic results (produced using the BD Phoenix™). The genotypic and phenotypic results matched in 11 of the 12 antibiotics with the output of both methods shown in **Figure 5-5**. The results for ciprofloxacin were inconclusive in genotypic tests, but resistant by phenotypic methods. Additionally, Mykrobe ruled out the presence of the PVL gene in this isolate.

A



Isolate AST Results

Antimicrobial	MIC or Concentration	Interp	Expert SIR	Final SIR
Amikacin	<=4	S		S
Ampicillin	>1		R	R
Cefoxitin	>8	R		R
Ciprofloxacin	>2	R		R
Clindamycin	<=0.25	S		S
Daptomycin	<=0.5	S		S
Erythromycin	<=0.25	S		S
Fosfomycin w/G6P	64	R		R
Fusidic Acid	<=1	S		S
Gentamicin	<=1	S		S
Gentamicin-Syn	<=500			
Linezolid	2	S		S
Moxifloxacin	>1	R		R
Mupirocin	<=1	S		S
Mupirocin High level	<=256	S		S
Nitrofurantoin	<=16	S		S
Oxacillin	>2	R		R
Penicillin G	>0.25	R		R
Rifampin	<=0.25	S		S
Teicoplanin	<=1	S		S
Tetracycline	<=0.5	S		S
Tobramycin	<=1	S		S
Trimethoprim	<=1	S		S
Trimethoprim-Sulfamethoxazole	<=1/19	S		S
Vancomycin	1	S		S

Resistance Markers

Rule 835	MRS	☑ Methicillin Resistant Staphylococcus
Rule 835	mecA	☑ mecA-mediated Resistant Staphylococcus

Expert Triggered Rules

Rule 835	Automatic	Cefoxitin (MIC >4 mcg/mL) result has been used to predict me resistance. Alert clinician and infection control practitioner. Ve uncommon.

B


```
"phylogenetics": {
        "phylo_group": {
                "Staphylococcus aureus": "4"
        },
        "species": {
                "S. aureus": "4"
        },
        "lineage": {
                "N/A": "-1"
        }
},
"susceptibility" :{
        "Gentamicin": "S",
        "Penicillin": "R",
        "Methicillin": "R",
        "Trimethoprim": "S",
        "Erythromycin": "S",
        "FusidicAcid": "S",
        "Ciprofloxacin": "Inconclusive",
        "Rifampicin": "S",
        "Tetracycline": "S",
        "Vancomycin": "S",
        "Mupirocin": "S",
        "Clindamycin": "S"
},
"called_variants" :{
},
"called_genes" :{
        "blaZ" :{
                "per_cov": "63",
                "median_cov": "4",
                "conf": "11",
        "induced_resistance": "Penicillin"
        },
        "mecA" :{
                "per_cov": "99",
                "median_cov": "10",
                "conf": "29",
        "induced_resistance": "Methicillin"
        }
},
"virulence_toxins" :{
        "PVL": "negative"
}
```


Figure 5-5 Results of the resistance profile prediction of the clinical isolate of *S. aureus* using genotypic and phenotypic tools. The phenotypic (A) results output by the BD Phoenix instrument and the genomic (B) results output from Mykrobe after single cell isolation from horse blood. The 12 antibiotics predicted by both methods are underlined.

5.1.5.2 E. coli

After completion of the sequencing 173597 reads passed the initial filter, once the analysis pipeline was complete 170243 reads remained. 69538 of these were identified by LCA as *Enterobacteriaceae*, 2529 as *Escherichia*, and 51479 were identified to the species level as *E. coli*. Overall 73% of all reads remaining after the analysis pipeline were identified as *E. coli*. 5661 reads were identified as the genus *Equus* and 1927 were identified as *Equorum*. 31959 (18%) reads had no identity. Unlike *S. aureus* it was not possible to sub-type the *E. coli,* 150 reads were assigned to O7:K1 and 211 reads were assigned to JJ1886, **Figure 5-6**. Reads which were identified as *Enterobacteriaceae, Escherichia* and *E. coli* were extracted from the fastq produced after pipeline completion. The genome of JJ1886 was available on Integrated Microbial Genomes (ID 2558309052), and the chromosomal sequence was used as the reference against which the extracted reads were assembled. After reference assembly 93.5% of the genome was covered in 548 contigs. When the reads were *de novo* assembled, 89% of the same reference was covered in 1334 contigs. The number of misassemblies was 23. The *de novo* assembly was used to identify several resistance markers using a combination of ardbAnno and ResFinder **(Table 5-7).** The resistance markers included dfrA, conferring trimethoprim resistance, gyrA conferring resistance to fluoroquinolones. mdtK an efflux pump conferring resistance to norfloxacin and blaCMY-2 which is an AmpC, conferring resistance to beta-lactams including cephalosporins. Additionally, eight drug efflux systems were identified. These results are comparable to the phenotypic data (supplementary Figure 7-1)

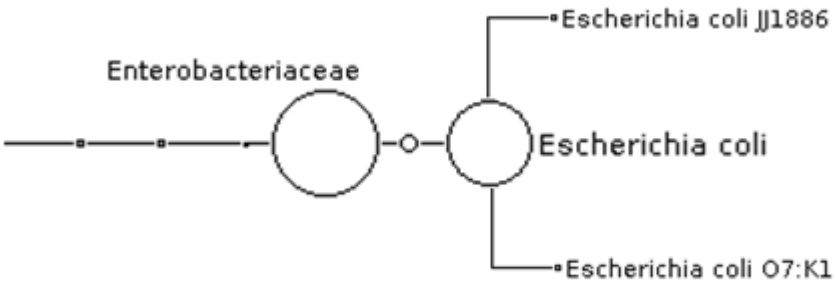


Figure 5-6 LCA analysis of *E. coli* isolated from blood showing reads identified as *Enterobacteriaceae* or higher, with the circle size being proportional to the number of reads identified at each taxonomic point, demonstrating that it was not possible to identify a single sub-species

Gene name	Gene	Gene function
dfrA	D-arabinitol 4-dehydrogenase	dfrA The DHFRs in this family are 64–88% identical in amino acids and mediate high-level resistance to trimethoprim
mdtB	multidrug efflux system, subunit B	MdtB is a transporter that forms a heteromultimer complex with MdtC to form a multidrug transporter. MdtBC is part of the MdtABC-TolC efflux complex.
gyrA	DNA gyrase (type II topoisomerase), subunit A.	Point mutation of Escherichia coli gyrA resulted in the lowered affinity between fluoroquinolones and gyrA. Thus, conferring resistance
emrB	multidrug resistance protein B	emrB is a translocase in the emrB -TolC efflux protein in *E. coli*. It recognizes substrates including carbonyl cyanide m-chlorophenylhydrazone (CCCP), nalidixic acid, and thioloactomycin.
emrY	putative multidrug efflux system	emrY is a multidrug transport that moves substrates across the inner membrane of the Gram-negative *E. coli*. It is a homolog of emrB.
tolC	Multi-drug efflux pump	TolC is a protein subunit of many multidrug efflux complexes in Gram negative bacteria. It is an outer membrane efflux protein and is constitutively open. Regulation of efflux activity is often at its periplasmic entrance by other components of the efflux complex
mdtK	multidrug efflux system transporter	A multidrug and toxic compound extrusions (MATE) transporter conferring resistance to norfloxacin, doxorubicin and acriflavine
mdtG	drug efflux system	Efflux protien mdtG
emrA	multidrug efflux system	EmrA is a membrane fusion protein, providing an efflux pathway with EmrB and TolC between the inner and outer membranes of *E. coli*, a Gram-negative bacterium.
CMY-2	blaCMY-2 protein	CMY-2 is an AmpC type beta-lactamase (resistance to amoxicillin, amoxicillin plus clavulanic acid, cephalothin, cefoxitin, ceftazidime and cefotaxime, but susceptible to cefepime and imipenem)
mdtE	multidrug resistance efflux transporter	MdtE is the membrane fusion protein of the MdtEF multidrug efflux complex. It shares 70% sequence similarity with AcrA

Table 5-7 Resistance indicators predicted using of ardbAnno and ResFinder on *de novo* assembly of single cell sequencing of *E. coli* isolated from blood. Including the gene ID, gene name and gene function according to UniProt[168]

5.1.6 *Final Sample Processing for Isolation of Pathogens from Human Samples*

Using the methods developed previously in **2.5.3** for viral concentration and extraction alongside those methods for isolating bacteria from whole blood in this chapter a sample processing workflow was developed to allow detection of pathogens from multiple clinical sample types. This sample process was then combined with the amplification and data analysis developed in previous chapters. An overview of the work flow is shown in **Figure 5-7** with full processing details shown in **Figure 2-4.**

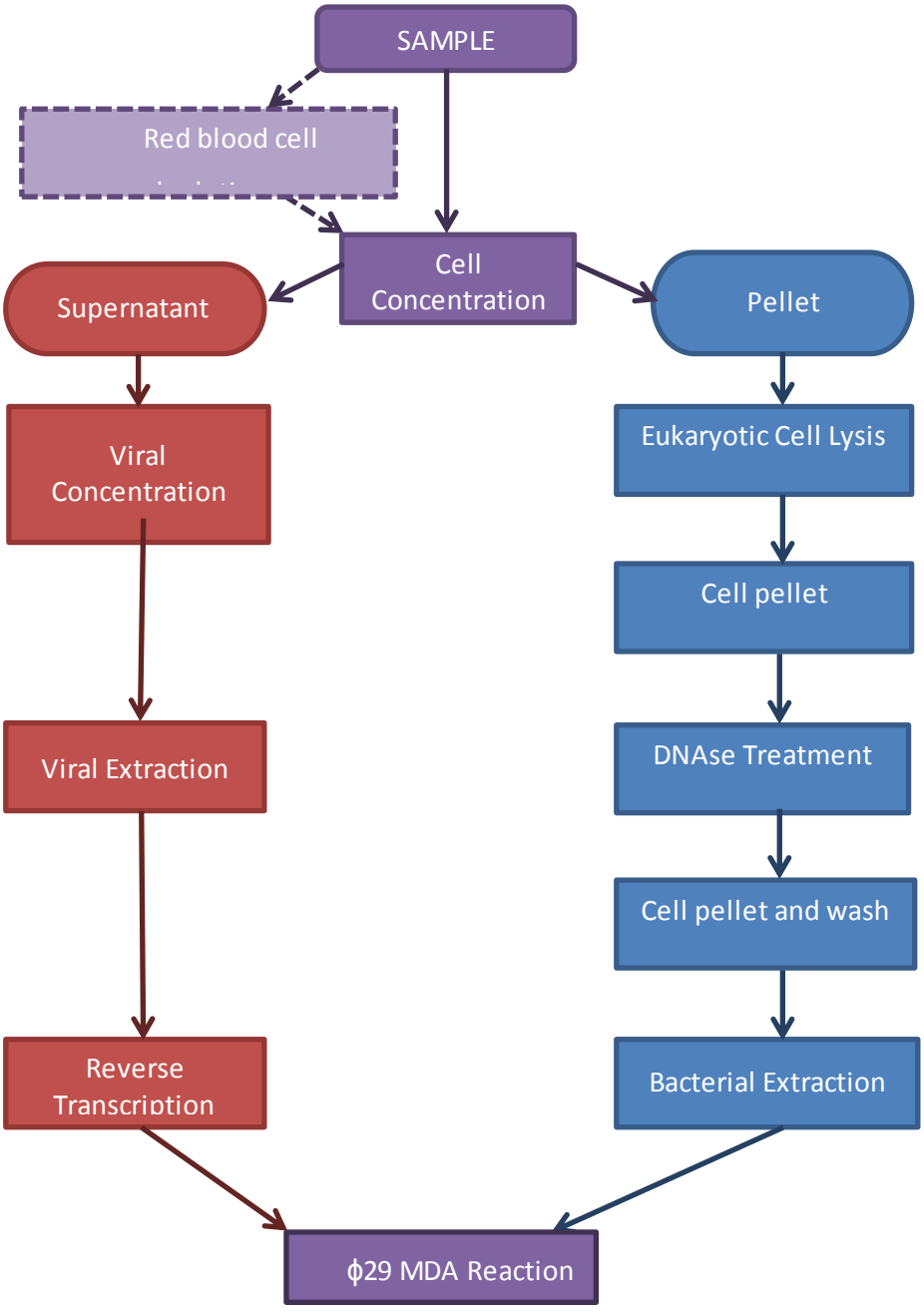


Figure 5-7 overview of sample processing for isolation and amplification of low level pathogens from clinical samples

5.2 Application to Multiple Sample Types

Several sample types were extracted, amplified and sequenced (on Illumina HiSeq platform) in parallel to test the method for suitability for application across sterile site infections. Summaries of the output from all sample types are shown in **Figure 5-15** and **Table 5-11.**

5.2.1 Negative Amplifications

Two negative controls were amplified and sequenced, the first was a negative extract starting with 1ml of sterile PBS and following the full extraction process. The second was a negative amplification sample, which was sterile PBS which underwent the reverse transcription and ϕ29 MDA amplification. After amplification the negative extract had a DNA concentration of 126 ng/µl and in the negative amplification the DNA concentration was 68 ng/µl. The number of raw reads and number of reads remaining after each pipeline element is shown in **Table 5-11** and **Table 5-11**. The number of reads that were removed using the existing contamination database was 127 for read one and 157 for read two in the negative extraction and 108 for read one and 144 for read two in the negative amplification.

After completion of the Blastn and LCA analysis, 99.55% of the negative extract reads and 99.61% of the negative amplification reads had no hits **Table 5-8**. Of the reads identified a similar pattern was observed for both samples. 29.66% and 27.26% of the identified reads for the negative extract and negative amplification respectively were identified as bacteria **(Figure 5-8)**. Within these reads the two major Phylums were Firmicutes (2.24% and 4.30%) and Proteobacteria (27.34% and 22.41%). When looking closer at the Firmiculates the major genus identified was *Alicyclobacilli,* with 2.28% and 3.57% of the identified reads for each sample. Within the Proteobacteria, those which were assigned a higher level fell within the class *Alphaproteobacteria*. With the majority of these falling the in *Bradyrhizobium* genus, with 17.91% of the identified reads in the negative extract sample being identified within this genus, and 12.77% of the negative amplification. Other genera represented in this class were *Nitrobacter* and *Rhodopseudomonas,* which were identified in both samples.

Within the Eukaryotes, the reads were identified as fungal, with 59.78% of the identified reads from the negative extraction, and 71.90% of the reads from the negative amplification being identified as fungal. Two division of the fungal phylum were identified in both samples the *Ascomycota* and *Basidiomycota.* Within the *Ascomycota* phylum the main represented class was the *Leotiomycetes*, with 4.25% of the negative extract reads and 3.83% of the negative

amplification reads. Within the *Basidiomycota* phylum (which represented 28.27% and 34.60% of the two samples), the main class represented was the *Agaricostilbales*, which had 9.76% and 12.49% of the identified reads. No reads within the fungal family were identified beyond the class level. Additionally, reads identified as *Horduem vulgare* were only identified in the negative extract, which represented 10.93 % of all identified reads. This information is summarised in **Figure 5-8** and **Table 5-8**

All reads from the negative runs were added to the contamination library, and this updated contamination library was used in the pipeline for all other sequencing runs.

	Negative extract			**Negative amplification**		
	reads	% all reads	%identified reads	reads	% all reads	%identified reads
Bacteria	3010	0.13%	29.66%	2382	0.10%	27.26%
P:Firmicutes	227	0.01%	2.24%	376	0.02%	4.30%
G:Alicyclobacilli	231	0.01%	2.28%	312	0.01%	3.57%
P:Proteobacteria	2775	0.12%	27.34%	1958	0.09%	22.41%
G:*bradyrhizobium*	1818	0.08%	17.91%	1116	0.05%	12.77%
G:*Nitrobacter*	67	0.00%	0.66%	20	0.00%	0.23%
G:Rhodopseudomonas	308	0.01%	3.03%	194	0.01%	2.22%
Fungi	6067	0.27%	59.78%	6283	0.28%	71.90%
P:Ascomycota	622	0.03%	6.13%	629	0.03%	7.20%
C:*Leotiomyces*	431	0.02%	4.25%	335	0.01%	3.83%
P:Basidiomycota	2869	0.13%	28.27%	3023	0.13%	34.60%
C:*Agaricostilbales*	991	0.04%	9.76%	1091	0.05%	12.49%
Horduem vulgare	1109	0.05%	10.93%		0.00%	0.00%
total	2280503	100.00%		2269242	100.00%	
no hits/not assigned	2270354	99.55%		2260504	99.61%	
total identified	10149	0.45%		8738	0.39%	

Table 5-8read identification in the two negative samples using Illumina sequencing , including read number and proportion of all reads.

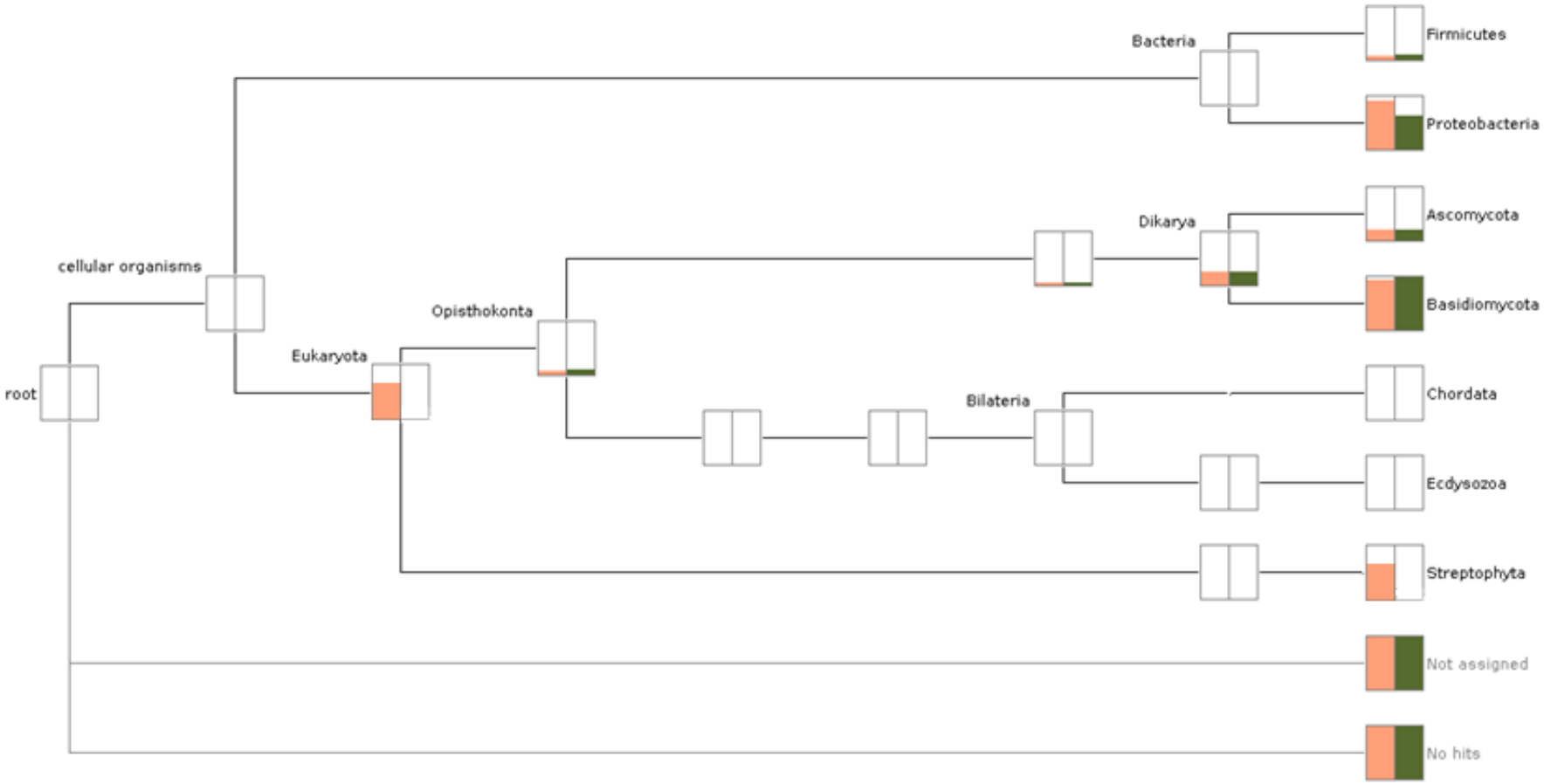


Figure 5-8 comparison of read identification using Blastn shown as LCA analysis, with pink representing the negative extraction samples and green showing reads identified in the negative amplification sample

5.2.2 Blood Model

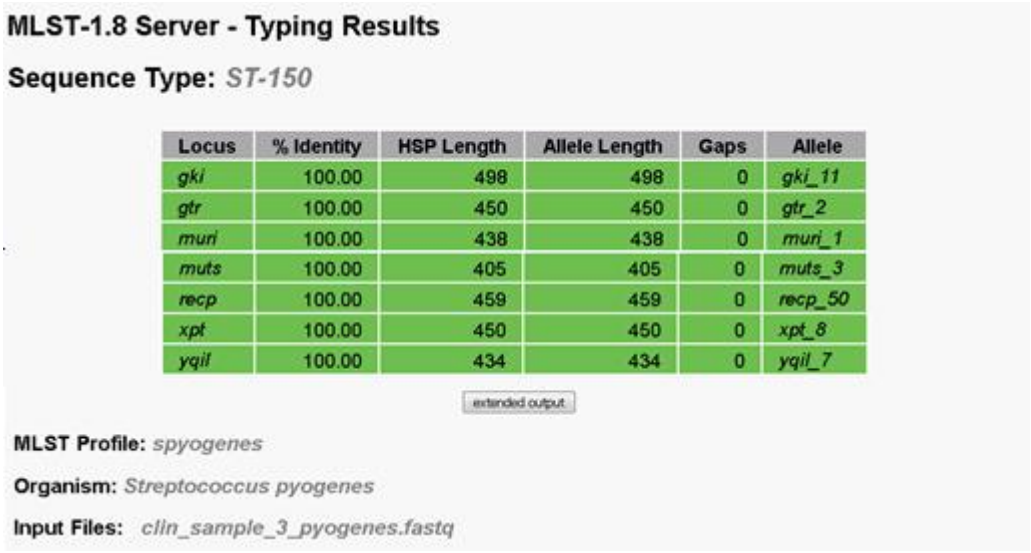
Further spiked blood models were used to test more challenging scenarios, one sample consisted of a mixed bacteria and virus spike. The second was spiked with *Shigella sonnei* to test the ability to differentiate *E. coli* and *Shigella* directly from sample. The third was a clinically invasive non typhoidal *Salmonella*, which had previously been whole genome sequenced, allowing comparison of direct from sample to culture workflows. The final spike was a carbapenem resistant *Pseudomonas*. One ml aliquots of horse blood were spiked with bacteria and or viruses and processed for isolation of bacteria and viruses as described in **2.6.5.**

5.2.2.1 Mixed Bacterial and Viral

Quantification of DNA after amplification using ϕ29 MDA gave a concentration of 139 ng/µl and after abundance trimming, removal of orphan reads and error trimming was complete 1436158 reads remained.

Investigation of the LCA analysis allowed identification of both Influenza and *S. Pyogenes.* When the reads identified as influenza were examined, 45,897 reads were identified as Influenza, and 997 were identified as H3N2. *De novo* assembly of these reads covered 91% of the genome in 51 contigs.

When investigating the reads using LCA analysis, 959,760 reads were identified as *Streptococcus* or higher, 544,384 reads were identified to species level as *S. pyogenes* and 4135 reads were identified as type M2, (MGAS10270). When these reads were extracted from the file, converted to a fastq and mapped against the MGAS10270 reference 93.4% of the reference was covered in 432 contigs, using 96.2% of the reads. The *de novo* assembly covered 89.7% of the genome and when this assembly was uploaded to the online tool 'Multilocus Sequence Typing of Total Genome Sequenced Bacterial' [169], it was able to give a serotype of ST150, with all the targeted regions covered **Figure 5-9**. When the *de novo* assembly was annotated, 1581 proteins were identified, 313 of these were hypothetical, and when repeats were removed 1203 remained. When the reference was annotated 1781 proteins were identified, 461 of these were hypothetical. When repeated proteins were removed 1206 proteins remained. When these two annotations were compared there was an overlap of 1174 proteins (98%), with 29 only identified in the *de novo* assembly, and 32 only identified in the reference genome.



MLST-1.8 Server - Typing Results

Sequence Type: *ST-150*

Locus	% Identity	HSP Length	Allele Length	Gaps	Allele
gki	100.00	498	498	0	*gki_11*
gtr	100.00	450	450	0	*gtr_2*
muri	100.00	438	438	0	*muri_1*
muts	100.00	405	405	0	*muts_3*
recp	100.00	459	459	0	*recp_50*
xpt	100.00	450	450	0	*xpt_8*
yqil	100.00	434	434	0	*yqil_7*

extended output

MLST Profile: *spyogenes*

Organism: *Streptococcus pyogenes*

Input Files: *clin_sample_3_pyogenes.fastq*

Figure 5-9 output from online MLST analysis of *de novo* assembly of *S.Pyogenes* reads showing good coverage of all genes used for MLST analysis.

5.2.2.2 Shigella

Quantification of DNA after amplification using ϕ29 MDA gave a concentration of 143 ng/µl. After abundance trimming, removal of orphan reads and error trimming 242940 reads remained.

When inspecting the LCA output, 800,165 reads were assigned to the family Enterobacteriaceae or above **(Figure 5-10)** with 131,799 of these assigned to the species *E. coli*. 57,266 reads were assigned to the genus *Shigella* or higher, 55,374 to the species *Shigella sonnei*, and 16,229 reads were identified to the strain level as Ss046. When the reads were mapped against the IpaH gene (gene for invasive toxin), 100% of the gene was covered.

When the reads were extracted and assembled against the reference Ss046 92.4% of the genome was covered in 212 contigs. When the reads were *de novo* assembled 85.5% of the same reference was covered in 1134 contigs. The *de novo* assembly was used to predict resistance to antibiotics using Resfinder[71] one resistance gene was identified, dfrA1 which confers resistance to trimethoprim. According to the phenotypic data (supplementary Figure 7-2), the only antibiotic to which the isolate was phenotypically resistant was trimethoprim. There were additional antibiotics changed to resistant based on clinically ineffectiveness of the antibiotics for treatment of *Shigella* (first and second generation cephalosporins and primary aminoglycosides). When the *de novo* assembly was used for MLST typing the result was ST-152 with full coverage of all genes used for typing.

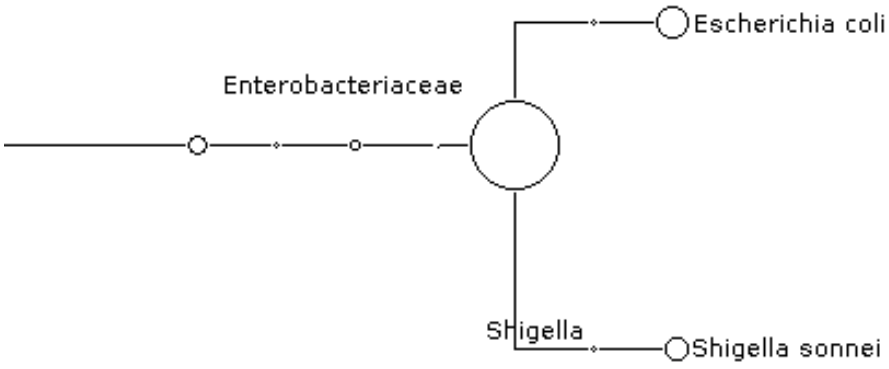


Figure 5-10 LCA analysis of *Shigella* sonnei isolated from horse blood, showing reads identified to *Enterobacteriaceae* or higher**.**

5.2.2.3 Salmonella

The *Salmonella* reference laboratory routinely uses whole genome sequence analysis (from cultured bacteria) to characterise and investigate *Salmonella* sp. In the hospital lab the isolate was identified as *Salmonella* sp. Via API20E, (2704542) but VI negative ruling out Typhi. Sensitivities by disc diffusion found no resistance markers.

When the DNA produced by ϕ29 MDA was quantified the concentration was 239 ng/µl. After sequencing 17732898 reads were produced, after completion of the trimming pipeline 2117443 reads remained. When the Megan LCA analysis was investigated 1,922,946 reads were identified as the family *Enterobacteriaceae* or higher. 1,870,222 reads were identified as the genus *Salmonella* or above. 1,815,371 reads were identified to the species level as *Salmonella* enterica, and 1,698,669 to the subspecies level as *Salmonella* enterica subsp. Enterica. And 12,975 of these reads were identified to the serovar level of Typhimurium. After *de novo* assembly when comparing the assembly to a closely related *Salmonella* whole genome sequence 95.4% of the genome was covered in 389 contigs, when this assembly was used to sequence type, the result was ST19 with full coverage of all the required genes. This matched the reference laboratory ID of *Salmonella* enterica serovar Typhimurium ST19. No resistance markers were detected using the *de novo* assembly file which was consistent with the phenotypic results (supplementary **Figure 7-3**).

5.2.2.4 Highly Resistant P. aeruginosa

When the Megan LCA analysis was investigated 1,142,554 reads were identified as the family *Pseudomonadaceae* or higher and 1,099,834 reads were identified as the genus *Pseudomonas* or above. 1,091,384 reads were identified to the species level as P. *aeruginosa*, and 298,669 to the strain level of MTB-1. Additionally, two phages were identified, Pseudomonas phage PAJU2 and Pseudomonas phage vB_Paes_PGM1.

When interpreting resistance markers for *P. aeruginosa* it is important to remember the high number of antibiotics that *P. aeruginosa* are intrinsically resistant to. (Glucose none fermenting Gram-negative bacteria are intrinsically resistant to penicillin, cefoxitin, cefamandole, cefuroxime, glycopeptides, fusidic acid, macrolides, Lincosamides, Streptogramin, rifampicin, daptomycin and linezolid. *P. aeruginosa* is intrinsically resistant to ampicillin, Augmentin, cefazolin, cefotaxime, ertapenem, chloramphenicol, trimethoprim, trimethoprim-sulfamethoxazole, tetracyclines and tigeclyine. Additionally, nitrofurantoin is not clinically effective for the treatment of Glucose none fermenting Gram-negative bacteria.) In addition to the intrinsic resistances phenotypic methods also identified resistance to cefepime, ciprofloxacin, Gentamicin, Levofloxacin, meropenem, piperacillin, piperacillin-tazobactam and tobramycin. The isolate was also found to have an intermediate MIC for aztreonam (supplementary Figure 7-**4**). These phenotypic resistances were confirmed at the reference laboratory.

When ResFinder was used on the *de novo* assembly of the *P. aeruginosa* four resistance genes were identified. The resistance markers were blaOXA-*50 (*carbapenemase-hydrolysing *oxacillinase), blaPAO (ampC), fosA and aph (3')-IIb (*aminoglycoside phosphotransferase) **Table 5-9**. The presence of both an ampC and a carbapenase would confer resistance to all beta-lactam antibiotics. Presence of an aminoglycoside phosphotransferase would confer resistance to gentamicin and tobramycin; fosA would confer resistance to fosfomycin

The reference laboratory for this isolate reported the presence of blaVIM metallo-carbapenemase gene, which was not identified in the chromosomal *de novo* assembly. *P. aeruginosa* has been reported to carry resistance markers on plasmids, which may not have been assigned to *P. aeruginosa* using LCA analysis. The unassigned reads were extracted from the trimmed fastq file and *de novo* assembled and this was then used in ResFinder. This identified the *blaVIM-7* gene.

Resistance gene	%Identity	Query/HSP length	Predicted phenotype	Accession number
Chromosomally encoded				
Beta-lactam				
blaOXA-50	100.00	789 / 789	Beta-lactam resistance	AY306132
blaPAO	99.25	1194 / 1194	Beta-lactam resistance	AY083592
Fosfomycin				
fosA	99.26	408 / 408	Fosfomycin resistance	NZ_ACWU01000146
Aminoglycoside				
aph(3')-IIb	99.01	807 / 807	Aminoglycoside resistance	X90856
Probable Plasmid Encoded				
blaVIM-7	100.00	798 / 798	Beta-lactam resistance	AJ536835

Table 5-9 Resistance genes detected using ResFinder on a *de novo* assembly of *P. aeruginosa* isolated from blood
HSP=High-scoring Segment Pairs

5.2.3 Tissue Model

Four tissue infection models were created and testing using rhesus macaque tissue. The four models were bacterial endocarditis, viral hepatitis, mixed bacterial brain abscess and viral pyelonephritis.

5.2.3.1 Endocarditis Model

When the DNA from the ϕ29 MDA was quantified a concentration of 846 ng/µl was found. After sequencing 1709690 reads were produced and after trimming pipeline completion 868754 reads remained, with the average read lengths being 93.02 and 85.64 bases for the first and second read. The majority of the reads (794,782) were removed during the abundance trimming stage. 3161 reads were assigned to Gammaproteobacteria, 8911 reads were assigned to Pasteurellacaeae, 29,870 were assigned to Haemophilus and 786252 reads were assigned to *H. influenzae*. Further identification based on LCA analysis showed six possible sub-types, 2019 (1366 reads), 86-028NP (2,459 reads), CGShiCZ412602 (2,554 reads), KR494 (3,781 reads), pittEE (7.733 reads) and R2846 (5,236 reads) **Figure 5-11**. The result of the MLST typing was ST-165 with all required genes fully covered. When the Resfinder data base was used to detect resistance markers one gene was identified, *blaTEM-1B (*100% gene identity, 861/861 bases) which confers

resistance to beta-lactam antibiotics. The phenotypic results for this isolate can be found in the supplementary material (Figure 7-5)

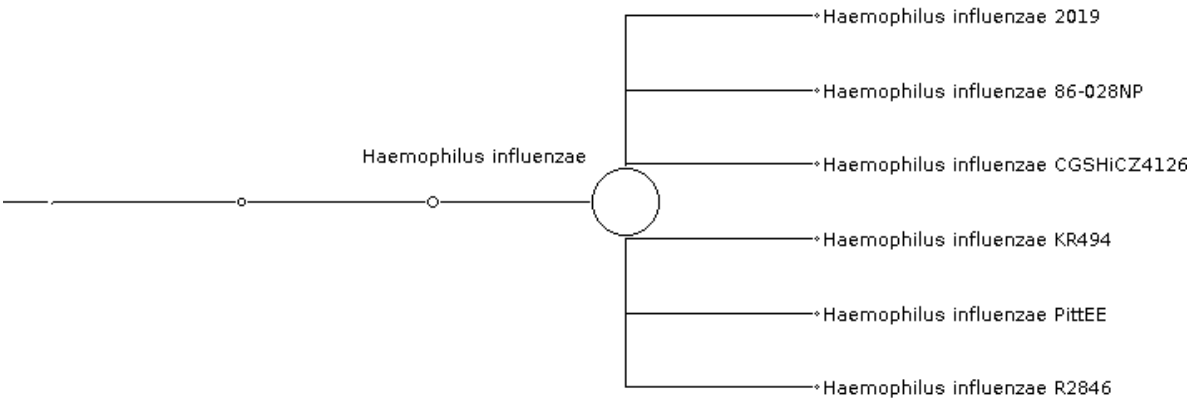


Figure 5-11 LCA analysis of *Haemophilus influenzae* sub-types after isolation from horse blood and sequencing

5.2.3.2 Hepatitis Model

The ϕ29 MDA produced a DNA concentration of 63ng/µl and produced 2365815 sequencing reads. After completion of the trimming pipeline 2,048,841 reads remained. The average read lengths for read one and read two were 94.4 and 86.8 bp. When examining the Megan LCA output 1,389,982 reads mapped to *H. influenzae*, suggesting a pre amplification contamination. Additionally, 354,389 reads mapped to *Macaca mulatta* (rhesus macaque), suggesting less efficient removal of host. There were 112,987 reads which mapped to HAV, which when *de novo* assembled covered 98.1% of the genome in 8 contigs

5.2.3.3 Mixed Brain Abscess Model

After ϕ29 MDA amplification the DNA concentration was 63 ng/µl, and after completion of sequencing there were 1,493,664 raw reads. After completion of trimming and host removal 1,326,822 reads remained.

587248 reads were identified using LCA as *Bacteroides* sp or higher, with 385,825 of these identified as *B. vulgatus,* with 298,498 reads further identified as ATCC 8482. When the reads were extracted from the fastq and mapped against the ATCC 8482 genome, 92.45% of the genome was covered in 589 contigs. When the *de novo* assembly was compared to the reference 81.0% of the genome was covered in 1128 contigs with 32 misassemblies. No resistance markers were found using ResFinder.

In total 757450 reads were identified as *Streptococcus* sp or higher **Figure 5-12**. The number of reads identified as *S. anginosa* was 228,770, with 98,710 reads identified as the subspecies 'C1050', when the extracted reads were mapped against this reference 87.65% of the genome was covered. When the *de novo* assembly was compared to the reference 79.65% of the reference was covered. No resistance markers were identified using ResFinder. The number of reads assigned to *S. mitis* was 19,633, and 5487 reads were further identified as strain 'B6', when this was used to reference assembly 85.89% of the reference was covered, and 74.87% was covered in the *de novo* assembly. ResFinder identified resistance markers conferring resistance to aminoglycosides and tetracycline **Table 5-10.** The number of reads identified as *S. oralis* was 153,093, with 3009 reads further identified as 'Uo5' strain. When Uo5 strain was used to reference assemble the extracted reads 93.3% of the genome was covered, and 81.4% was covered using *de novo* assembled reads. ResFinder identified resistance markers for macrolide and tetracycline resistance.

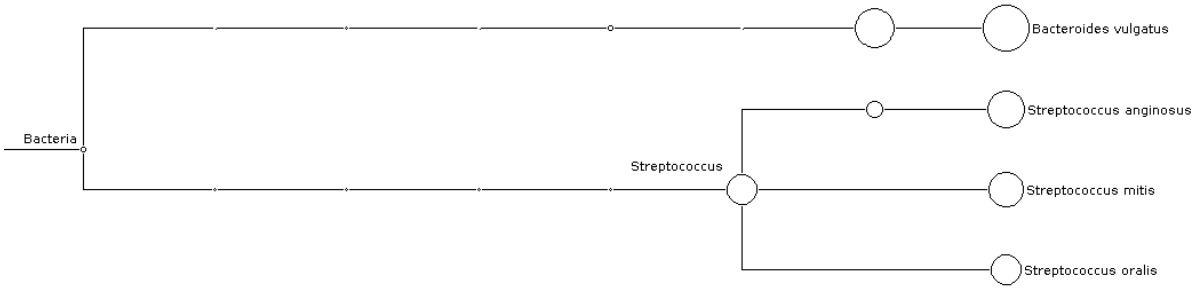


Figure 5-12 Bacterial read identification using LCA analysis on mixed brain abscess model

Resistance gene	%Identity	Query/HSP length	Predicted phenotype	Accession number
***B. vulgatus*-**No resistance markers found				
***S. anginosa*-**No resistance markers found				
S. mitis				
Aminoglycoside				
aac(6')-aph(2'')	100.00	1440 / 1440	Aminoglycoside resistance	M13771
aph(3')-III	99.75	795 / 795	Aminoglycoside resistance	M26832
Tetracycline				
tet(M)	99.48	1920 / 1920	Tetracycline resistance	FN433596
S. oralis				
MLS - Macrolide, Lincosamide and Streptogramin B				
erm(B)	100.00	738 / 738	Macrolide resistance	U86375
Tetracycline				
tet(M)	98.75	1920 / 1920	Tetracycline resistance	FN433596

Table 5-10 Resistance markers identified in bacteria used in mixed brain abscess model

5.2.3.4 Viral Pyelonephritis Model

After ϕ29 MDA amplification no DNA was detectable when using the Qubit, and so the sample was concentrated using an isopropanol clean up and the DNA reconstituted in 10 µl and submitted for sequencing. Sequencing produced 858,248 raw reads. 67.6% of these reads mapped to the human genome and so were removed during the trimming pipeline. After completion of the trimming pipeline 262,038 reads remained. Most of these reads (213,819) had no hits. The remaining reads were identified as *Macaca mulatta.* No reads were assigned to CMV.

5.2.4 Pathogen Detection in Urine

Culture negative urines from symptomatic patients were sent as part of another molecular study, 50 µl aliquots of these were blinded and tested using methods developed in this study. After amplification using ϕ29 MDA the amount of DNA present was 768, 156, 791 and 293 ng/µl for the four samples.

Two of the samples (1 and 3) had a large presence of vaginal flora including predominantly Lactobacillus sp (including *Lactobacillus acidophilus, Lactobacillus amylolyticus, Lactobacillus crispatus, Lactobacillus gallinarium, Lactobacillus gasseri, Lactobacillus iners, and Lactobacillus johnsoni*) suggesting female patients with poor quality sampling. In one of these samples, (3) it was possible to identify JC virus amongst the vaginal flora.

A large proportion of the reads in samples 2 and 4 were unidentified however it was possible to identify BK polyomavirus within these samples. No further identification of the BK virus was possible.

5.2.5 Pathogen Detection in CSF

After amplification the concentration of DNA for the four CSFs was 2, 142, 491 and 484 ng/µl. the number of raw reads produced after sequencing was 3085267, 1773960, 544302 and 2,113,620. The percentage of reads that mapped to the human genome was 94.1%, 93.9%, 61.2% and 99.2%. In CSF samples 1, 2 and 4 there was no pathogen identified with remaining reads being unassigned. In CSF there was 1.1% of reads which were identified as *Onchocerca volvulus,* when mapped against the genome 0.95% of the genome was covered. The remaining reads from CSF three were unassigned.

5.2.6 Sexually Transmitted Disease Detection

Four blinded samples were supplied by the sexually transmitted bacteria reference unit (STBRU) at PHE Colindale.

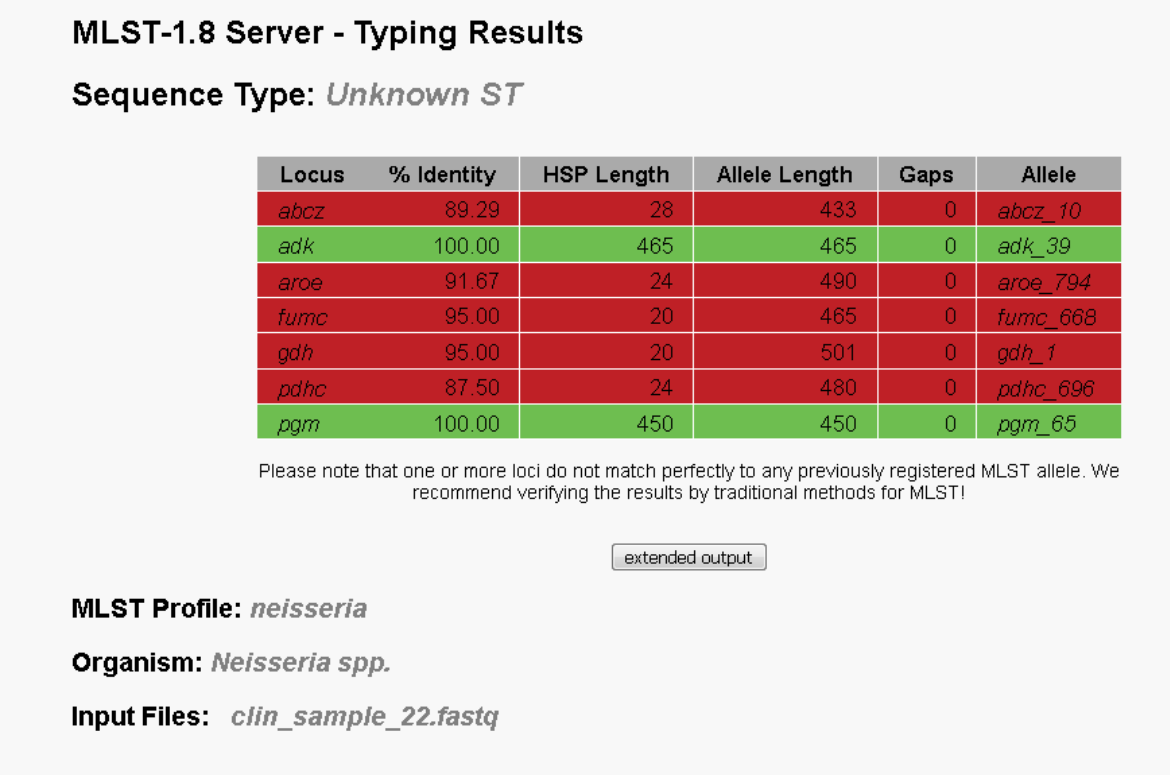
After amplification the concentration of the two samples in BD viper buffer were 1ng / µl for each sample, the two samples in COBAS buffer produced 361 and 251 ng/µl. After completion of sequencing for the first STI swab the percentage of reads mapping to host was 1.4%. When examining the LCA analysis, the majority of reads (1,254,897, 69.1%) mapped within the *Gammaproteobacteria* class, or the class Clostridia (324,946 reads, 17.9%). Other reads were either unidentified or represented other classes expected in the gut. It was not possible to detect *C. trachomatis* within this sample.

After completion of sequencing of STI sample two 2.6% of read mapped to the host. Within STI sample two 7.83% of reads mapped to the Bacteroides genus, with the main species within this was *Bacteroides uniformis* with 5.35% of all reads. Also identified in this sample was *Runinococcus torques* which represented 8.88% of all the reads. The largest species represented in this sample was *Sneathia sanguinegens* which represented 61.03% of all reads. *Chlamydia*

trachomatis was identified in this sample with 3.29% of reads. No further identification of the *C. trachomatis* was possible.

The number of raw reads produced after sequencing STI sample 3 was 2,104,335, with 408,962 mapping the human genome. After completion of the trimming pipeline 1,282,026 reads remained. When examining the LCA, it was possible to see a large variety of bacterial species represented. *Mobiluncus curtisii* represented 12.10% of all reads; Corynebacterium sp acounted 4.10% of reads, there was also a large presence of *Staphylococcus epidermidis* representing 4.44% of the reads. *Streptococcus* species were also heavily represented with a total read percentage of 10.85%; the genus Peptoniphilus sp was a highly represented with 31.85% of the reads. Within this family *Finegoldia magna* was present at the highest proportion with 12.16% of all reads. It was possible to identify *Neisseria gonorrhoeae* in the sample, with 3.54% of reads identified as this species. It was not possible to further subtype the *N. gonorrhoeae.* Also detected within this sample was *Mycoplasma genitalium*, which represented 3.83% of the overall reads, along with 1.38% of reads which were identified as *Trichomonas vaginalis.* A list of all genus and species detected in this sample at a level of greater than 0.2% of all reads can be seen in supplementary **Table 7-10.**

In the fourth STI swab sample the number of raw reads produced was 1778454 with 437691 of these mapping to human. After completion of the trimming pipeline, 889337 reads remained. When inspecting the LCA analysis, 89.32% of the reads were identified as *Gardnerlla vaginalis,* 0.67% of reads mapped to Alpha papillomavirus 8. Additionally, 9.08% of the reads were identified as *Neisseria gonorrhoeae*. A subtype was identified using LCA as MU_NG4, when this was used as a reference 38.3% of the genome was covered after reference assembly and 25.9% was covered using the *de novo* assembly. The MLST tool was unable to identify the sequence type as only two of the required genes had sufficient coverage **Figure 5-13,** ResFinder was unable to identify any resistance genes.



MLST-1.8 Server - Typing Results

Sequence Type: *Unknown ST*

Locus	% Identity	HSP Length	Allele Length	Gaps	Allele
abcz	89.29	28	433	0	*abcz_10*
adk	100.00	465	465	0	*adk_39*
aroe	91.67	24	490	0	*aroe_794*
fumc	95.00	20	465	0	*fumc_668*
gdh	95.00	20	501	0	*gdh_1*
pdhc	87.50	24	480	0	*pdhc_696*
pgm	100.00	450	450	0	*pgm_65*

Please note that one or more loci do not match perfectly to any previously registered MLST allele. We recommend verifying the results by traditional methods for MLST!

extended output

MLST Profile: *neisseria*

Organism: *Neisseria spp.*

Input Files: *clin_sample_22.fastq*

Figure 5-13 MLST output of *N. gonorrhoeae* reads identifed by LCA and *de novo* assembled

Five additional STI samples delivered from St Mary's hospital, three blood samples and two ulcer samples. After amplification using ϕ29 MDA the amount of DNA produced was 107, 53 and 184 ng/µl for the blood samples and 193 and 1 ng/µl for the ulcer samples. After sequencing the blood samples produced 2,470,807, 2,181,750 and 2,694,688 reads. Of these reads 85.1%, 98.9% and 97.6% mapped to the human genome. When examining the LCA analysis samples two and three all remaining reads were either mammalian or not assigned. In sample one, Torque teno virus accounted for 5% of the remaining reads, with other reads identified as mammalian or unassigned.

The ulcer samples produced raw 2,087,616 and 869 reads. From the first ulcer sample 2.5% reads mapped to the human genome. There was a large number of reads which mapped to the genus *Staphylococcus,* (54.9%) *with S. epidermidis* (43.0%) the majority species of this genus. Other genera heavily represented were *Corynebacterium*, *Acinetobacter* and *Streptococcus*. *Treponema Pallidum* was detected at a level of 0.05% of all the reads.

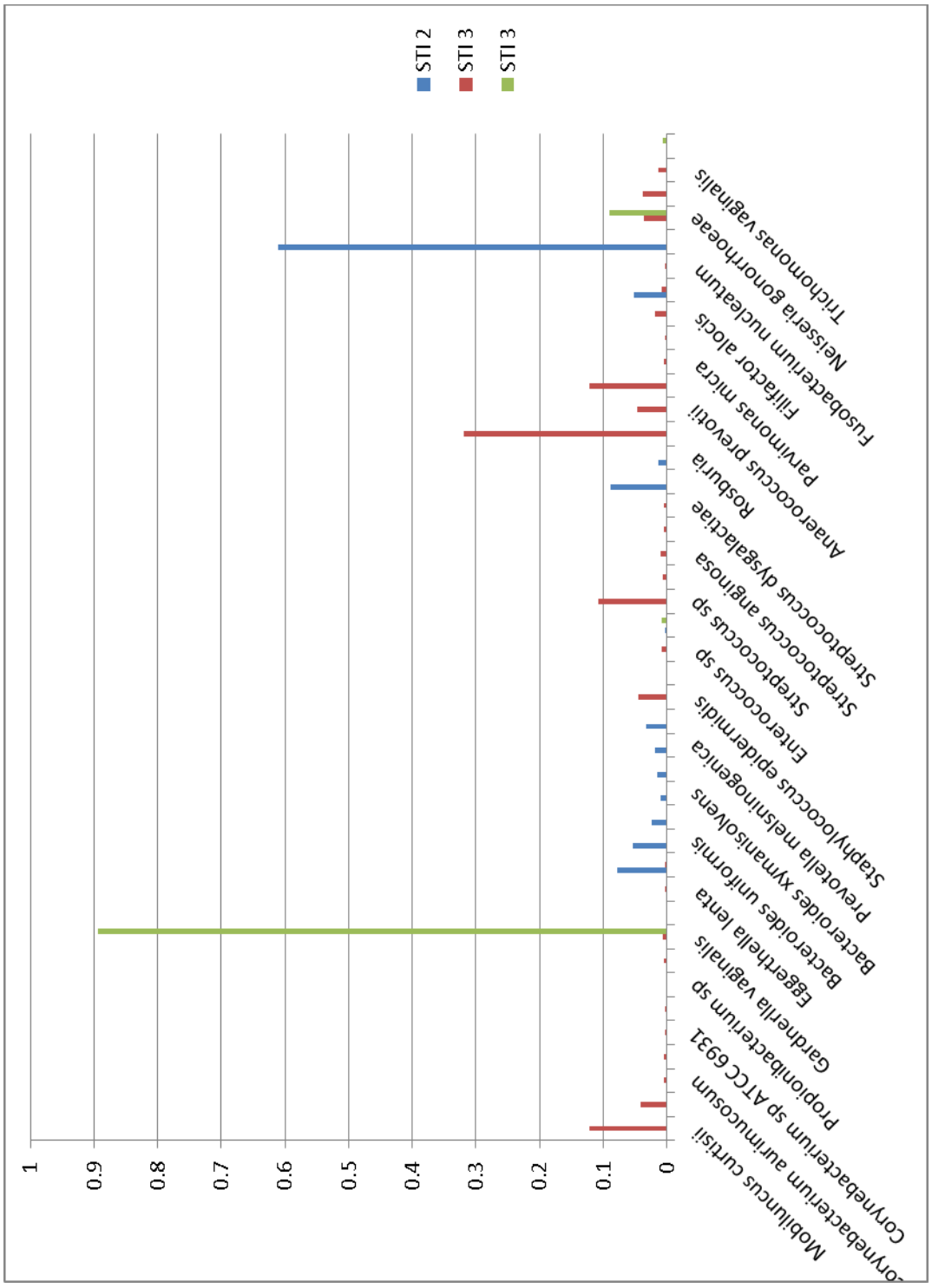


Figure 5-14 bacterial identification using LCA analysis: percentage of reads identified in three vagina swabs (reads >0.05% of all reads)

	Sample type	DNA Quant post amplification (ng/µl)	No: Raw Reads	No: reads Mapped to host	No: reads Mapped to contamination	No: reads after abundance trimming	No: Reads after pipeline	read length (bp)
1	Negative extraction	126	2779602	59100 53486	127 157	2280519 2280519	2254737 2254737	93.39 86.5
2	Negative amplification	68	2861044	57307 62505	108 144	2269259 2269259	2251228 2251228	93.43 87.14
3	Spiked blood	139	1524710	22633 21286	3191 2706	1444806 1444806	1436158 1436158	95.13 88.03
4	Spiked blood	143	1444641	4867 4335	467 412	255457 255457	1242940 1242940	92.72 80.1
5	Spiked blood	239	17732898	33574 31562	10027 9403	2132469 2132469	2117443 2117443	92.27 85.4
6	Spiked blood	394	1648391	36134 32345	61 100	1435310 1435310	1410913 1410913	94.25 86.37
7	Spiked heart	846	1709690	11699 10573	15016 13427	888193 888193	868754 868754	93.02 85.64
8	Spiked liver	973	2365815	34980 31376	12562 11898	2073516 2073516	2048841 2048841	94.4 86.8
9	Spiked brain	63	1493664	12259 11509	1949 1805	1332533 1332533	1326822 1326822	92.21 86.34
10	Spiked kidney	<	858248	579875 559886	682 855	264676 264676	262038 262038	98.68 99.7
11	Urine	768	2428095	43015 39354	111 138	2169474 2169474	2149803 2149803	93.71 86.91
12	Urine	156	2161570	39214 35532	158 169	1936518 1936518	1920385 1920385	93.82 87.1
13	Urine	791	1913768	32772 29133	59 101	1772817 1772817	1744528 1744528	94.12 86.03
14	Urine	293	2100457	36441 33440	247 251	1909599 1909599	1893709 1893709	94 87.2
15	CSF	2	3085267	2902191 2863550	19 297	79991 79991	79257 79257	89.55 84.4
16	CSF	142	1773960	1665264 1618340	2 4060	32290 32290	30909 30909	88.38 80.29
17	CSF	491	544302	333062 324408	3 3	126419 126419	114083 114083	92.06 76.23
18	CSF	484	2113620	2097739 2065525	2 5	3472 3472	3286 3286	80.32 76.69
19	STI swab	1	1938864	27826 26338	4239 3861	1827346 1827346	1816884 1816884	95.07 87.97

20	STI swab	1	1270294	33891 32655	1978 1770	1224811 1224811	1220567 1220567	95.13 87.96
21	STI swab	361	2104335	408962 402582	5616 5161	1292173 1292173	1282026 1282026	94.81 87.99
22	STI swab	251	1778454	437691 428113	3966 3361	902435 902435	889337 889337	98.58 86.88
23	STI Blood	107	2470807	2103620 2071314	27 33	355004 355004	350672 350672	95.01 87.06
24	STI Blood	53	2181750	2157065 2115394	1 12	14265 14265	13930 13930	92.59 85.41
25	STI Blood	184	2694688	2630702 2544864	3 14	52915 52915	51426 51426	93.9 84.69
26	SnosStrip	193	2087616	40199 37114	52 99	1788958 1788958	1769905 1769905	94.17 87.04
27	SnosStrip	1	869	15 23	2 1	831 831	715 715	93.72 79.8

Table 5-11 DNA concentrations, sequencing outputs and pipeline outputs for modelled and clinical samples

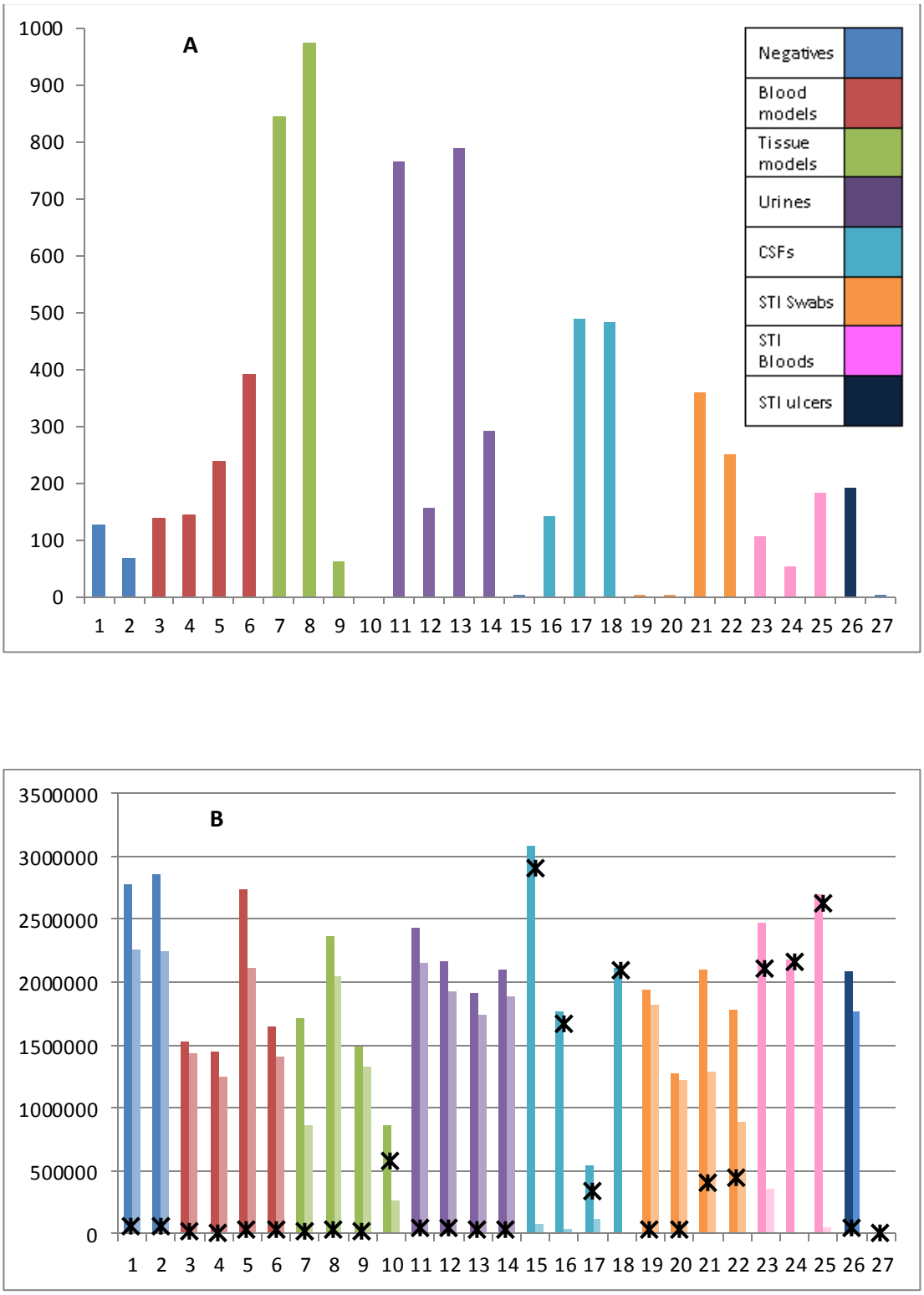


Figure 5-15 Results of amplification and sequencing of real and modelled clinical samples along with negative controls, grouped by sample type (A):DNA concentrations produced by ϕ29 MDA (B): number of raw reads (dark bar) and read number after completion of trimming (light bar), X illustrates the number of reads identified as human in each sample.

5.3 Chapter Discussion

5.3.1 Blood Culture Model-Development

Fresh horse blood was used to model bacteraemia, as whole blood was readily available. The initial experiments focused on bacterial survival in horse blood and the use of saponin to selectively lyse host cells. The aim of the isolation method was to keep bacterial cells whole so that the alkali extraction method could be used in order to prevent DNA shearing. Saponin has previously been used to lyse human cells [170]. Lysing human cells will release any intracellular pathogens and release host genetic material allowing removal using nucleases. Incubation of most bacterial cells in saponin showed no impact on the number of CFU recovered across the different bacterial species studied. *Paenibacillus anaericanus* showed the lowest survival rates (87.2%), however if the saponin was toxic to the bacteria a lower survival would be expected. The slight drop in survival could be accounted for by the bacteria's need to adapt to the aerobic conditions or some of the slow growing colonies not being visible and so not counted. Several bacteria did show poor survival in horse blood, particularly *S. pneumoniae, L. monocytogenes, S. sonnei and Paenobacterium.* This suggests that the horse blood is toxic to these bacteria, either due to specific antibodies, innate immune response or general toxicity of the blood. Other bacteria developed altered colony morphology after incubation with horse blood, suggesting some inhibition of growth, although not cell death. Knowledge that most of these bacteria survived well in horse blood allowed the isolation model to be developed, knowing to what degree the bacteria survived allowed informed decisions about which bacteria were spiked into the model and at what number. Two bacteria were selected to develop the isolation method, *E. coli* and *S. aureus*, representing both a Gram negative and Gram positive cell wall type, and different cell shapes and densities.

Red blood cells (RBCs) represent the largest proportion of the cellular makeup of blood (up to 96%[171]) and so the first stage was to remove them. By keeping the RBCs whole, it was hoped to not release cell contents which might be toxic to pathogens, or which may inhibit extraction or amplification. Originally isolation of bacteria in a density gradient using Percoll was investigated. The results showed poor separation of bacterial cells and RBCs, with *E. coli* cells being found at almost all points in the density gradient (**Figure 5-2).** The density of *E. coli* most likely varies greatly as the cell elongates before division, with short rod shaped bacteria being produced after cell division. Additionally, *S. aureus* was only found co-located with RBCs, suggesting that whilst they have a more constant cell density, it is in the same range as the RBCs. Density was shown to

be a poor method for isolating bacteria from RBCs, additionally, density of bacterial cells will vary among species, and even between bacteria of the same species at different life cycle stages.

HetaSep was then investigated as a method of isolating and removing RBCS from whole blood, this product causes the RBCs to form stacks, which increases the weight, allowing easy pelleting of the RBCs. The first stage of investigation was whether the bacteria and RBCs still co-pelleted, and also ensure that the use of HetaSep wasn't toxic to the bacterial cells. The manufacturer documented two methods for separating the stacked-RBCs from other blood components, either with centrifugation or by incubating at 37°C, both were investigated. When the sample was centrifuged there was co-sedimentation of the *E. coli* and *S. aureus* with the RBCs, probably through the action of centrifugation. Gravity sedimentation at 37°C, prevented co-sedimentation of bacterial cells and RBCs (**Table 5-3**).

Low numbers of bacterial cells (ten) were spiked into horse blood to investigate the process of bacterial isolation from whole blood. The process trialled included removal of RBCs using selective aggregation and lysis of other host cells. Lysis was achieved through the use of saponin coupled with an osmotic shock treatment, which was added to aid in host cell destruction, with the aim of improving intracellular pathogen recovery and increasing host cell breakdown. Once the host genetic material was released from the cells the aim was to remove it through the use of centrifugation and washing. At several points along the process bacterial cells were cultured to check bacterial viability and isolation. The water shock treatment proved detrimental to *E. coli* survival due to too much osmotic stress on the bacterial cell wall, and so the time the cells were exposed to the stress was reduced. The host cells would be selectively damaged by the saponin and so a short treatment should cause more damage to these cells than bacterial cells. The initial centrifugation speed of 3000 x g failed to pellet all cells of *S. aureus*, and so was increased to 4000 xg, a low speed was required so as to leave as much host DNA and cell debris in the supernatant as possible. After completion of the process the pellet was suspended in PBS and the free DNA quantified, showing that there was still host DNA present (the *E. coli* sample had a higher concentration possibly due to release of bacterial DNA due to osmotic cell lysis). The large genome size and complex structure of the host genome along with accompanying proteins and cellular debris may have caused the DNA to pellet alongside the bacteria. A DNAse stage was added to the process, aiming to avoid use of heat to deactivate the enzyme to prevent sample coagulation and release of potentially toxic cellular components. The combination of Mg^{+} chelation with EDTA and several centrifuge and washing stages successfully removed DNAse activity without damage to the bacteria. Any small traces of DNAse would be inactivated during bacterial DNA extraction.

When the improvements were trialled they showed good results for both *E. coli* and *S. aureus*. Once the isolation method was optimised, the method was tested by spiking single bacterial cells into 1ml horse blood, to represent a low level bacteraemia. The bacterial DNA was then extracted and sequenced, with the reads being input into the pipeline previously developed.

After sequencing of the *S. aureus* isolated from blood, 6.76% of reads were identified as Equus, and 62.1% of the reads were identified as *Staphylococcus*. The remaining Equus reads were most likely due to remaining free DNA as whole host cells would produce more reads than the *S. aureus* due to a larger starting genome. Additionally, LCA analysis identified the presence of a horse tapeworm, suggesting the horse had a current or previous infection with *Parascaris equorum*. The LCA analysis provided subspecies identification of the *S. aureus,* 5.5% of the *S. aureus* reads were identified as a single subspecies, which most likely represented a *S. aureus* lineage that was very closely related to the clinical isolate. The *de novo* assembly was then used to predict the antibiogram using a tool specifically designed for *S. aureus,* Mykrobe. The antibiogram matched for 11/12 of the antibiotics available in the genotypic and phenotypic methods. The genotypic results for Ciprofloxacin were inconclusive, suggesting either poor coverage of the targets, or insufficient information in the database to call the isolate resistant. With 92% of the resistance markers being called correctly, data quality is unlikely to be the cause of the Ciprofloxacin resistance mismatch. The *de novo* assembly only covered 83% of the genome, suggesting that the gene might be missing, however 13% of the *S. aureus* genome has been shown to be repetitive[172], and so most of the absent genome could be accounted for if these repeated elements were misplaced during assembly. Three mechanisms of fluoroquinolone resistance have been proposed in *S. aureus,* Topoisomerase IV gene mutations, DNA gyrase gene mutations and an active efflux pump (NorA)[173]. The complexity of predicting ciprofloxacin resistance suggests that the database may be lacking in its ability to predict ciprofloxacin resistance, and so this is the most likely cause of the inconclusive result for ciprofloxacin resistance. Additionally the creators of Mykrobe (Bradley et al[167]) found a false negativity rate of 4.6% for ciprofloxacin resistance. Overall there was good concordance of phenotypic and genotypic results show the potential for rapid genotypic prediction of antibiotic resistance from a single cell in 1 ml of host blood, however the inconclusive ciprofloxacin results demonstrate the need for improved understanding of the mechanisms of resistance. The fact equine tapeworm DNA was identified in the blood of this animal highlights the ability of this method to discover pathogens in a non-targeted manner.

There were fewer antibiotic non-susceptibilities predicted using the genomic data compared with phenotypic testing, but this is most likely a reflection that the BD phoenix is a generic tool, with the same panel of antibiotics being tested for all Gram positive bacteria, and Mykrobe being a specific tool for *S. aureus*. In addition to identifying the isolates resistance to beta-lactams the database was able to identify the blaZ gene and MecA gene. Genotypic testing will never entirely replace phenotypic susceptibility testing, due to its inability to identify novel resistance determinants and the comprehensive nature of phenotypic testing. However, in this scenario, of invasive sepsis, the gain in speed provided by not having to culture the organism to determine susceptibility could be life-saving. When Prokka was used to annotate the resulting genomes more CDS were predicted than in the reference used, which may reflect the fact that the clinical isolate was not an exact match to the reference sequence used, or the fact that only 92% of the genome was covered when the reference mapping was used.

When the LCA analysis from *E. coli* spiked blood was analysed it was not possible to assign a single genotype, with a small proportion of reads being assigned to two different subtypes. This could be that both of these named genotypes are actually the same strain but with different names, or that the clinical isolate is not an exact match for any genome in the database. Additionally, there is the chance that there was low genome coverage at the points where genotypes differ. However, when using one of the genotypes identified good genome coverage was achieved (93.5%) when the data was reference assembled, and the *de novo* assembly also performed well covering 89% of the genome. Multiple antibiotic resistance factors were identified, which gave good concordance with the phenotypic data. Using the genotypic data, it was possible to rapidly identify the beta-lactamase present as BlaCMY-2. The rapid identification of the specific resistance genes in bacteria could help identify outbreaks by providing more information than a simple antibiogram. Additionally, it could help monitor novel resistance genes, or genes that are becoming increasingly common. Large amounts of horizontal genome transfer amongst Gram negative bacteria has the potential to cause outbreaks of resistance bacteria through gene or plasmid outbreaks[174], which would be more complex to track. Rapid identification of genes causing the resistance in isolates could help inform epidemiological and outbreak studies which could involve several species of bacteria. The output for the resistance prediction was quite difficult to interpret with several genes identified which didn't always have specific drug resistance associated. This highlights a down side of generic databases, as they are often time-consuming to interpret. A study of NGS sequencing data from bacteraemia isolates of *E. coli* have shown resistance prediction specificity of 97%[6], if this was coupled with direct from sample sequencing genotypic prediction could inform treatment more rapidly than phenotypic test.

When using whole genome sequencing on NGS platforms the biggest limitation is the data output of the platform use, which for the 454 junior is 0.04 GB/run[175]. This means that when pure bacteria are investigated it is possible to get good depth of coverage, allowing robust prediction and good genome coverage. However, when host and environmental DNA is also sequenced there is an impact on the number of reads assigned to the pathogen, potentially lowering coverage and decreasing depth. At several points during library preparation only a small portion of the DNA is carried forward to the next stage. For example, after completion of ϕ29 MDA there is often >100 ng/µl DNA in a 50 µl reaction. Only 500 ng of this is needed for the library preparation. There are also additional points during library preparation when small amounts of DNA are needed. These multiple stages will mean that at every point random pieces of DNA will be lost, which may lead to technical variation in the method. This phenomenon will be exacerbated when looking a pathogen signals that originated in clinical material if host DNA is present. However, despite these limitations, we have shown that this method of isolation, amplification and sequence interpretation produces good data is produced allowing pathogens to be identified, even on a sequencing platform which is showing its age, the 454 Junior. Good quality antibiotic resistance predication has also been possible using unbiased amplification and *de novo* data analysis methods.

There are several limitations to the model created for bacteraemia, firstly by using a model several things are controlled for which may not be possible in real samples. This includes the time from inoculation to isolation of the bacteria, the temperature of the blood, and the inoculum itself. Additionally, bacteria spiked into horse blood may differ in their behaviour compared to bacteria causing disease in humans. Horse blood may also differ from human blood in several ways including cellular make up, and human blood which is reacting to infection may contain additionally components which are toxic to bacteria and make isolating whole bacterial cells more challenging. However, the main limiting factor for this model is the estimation of a single cell based on dilution studies, which may be incorrect and mean that more than a single cell was added due to human error or aggregation of bacteria leading to more cells being added. To overcome all these limitations would require a large study and ethical approval for using an unbiased amplification method on real human samples, which would be challenging to achieve. However, some limitations could be more easily over come, such as the use of human blood in place of horse blood, which would require volunteers donating blood. Additionally, to ensure accurate single cell inoculum microfluidics could be used to isolate single cells to be spiked into the samples.

The process developed and tested for the bacteraemia model was used to inform a single sample process which could be applied to multiple sample types. This also included to viral sample preparation, the two processes were kept separate due to the small size of viruses, and not wanting to exposure enveloped viruses to saponin. The two halves were joined at the ϕ29 MDA stage which allowed one amplification and sequencing per sample. The aim of the work flow was to be able to identify pathogens without need for any prior knowledge other than the sample type. The success of this approach was highlighted by the discovery of a parasite in horse blood, a pathogen type for which no explicit optimisation was carried out. The intention was to remove as much host as possible using laboratory methods without compromising the pathogen integrity.

5.3.2 Application to Multiple Sample Types

Several different samples including both real and modelled clinical samples were processed in parallel to test the applicability of the method. The sequencing was performed using the Illumina HiSeq, which allowed all samples to be tested on a single run as the data output is higher than that of the 454 Junior. By switching to an alternative platform the applicability of the method to multiple platforms was demonstrated. The first samples to be analysed were two negative samples, an extraction and amplification control. Using two controls allowed monitoring of the processing, and if contamination was identified as a problem it would be possible to identify potential sources. DNA was produced by both negative samples, with the extraction negative producing more, which would be expected as there were more opportunities to pick up contaminants. However, both had very few reads that were assigned an identity (0.45% and 0.39%) showing very low level contamination, which were most likely introduced after amplification. With such low numbers it could be that the reads that were assigned weren't the true identity, and were actually random DNA that happened to be similar to organism signals, however as the reads were similar in both would seem to suggest that these were true identities. The most likely source would be sequencing preparation kits the low numbers suggest the source was degraded DNA. A very low number of reads from the new negative samples mapped to the original negative library previously prepared, several factors could account for this, a new ϕ29 MDA kit was used to perform the final set of samples, a different laboratory was used and additional reagents were used. Furthermore, the change to a new sequencing platform and library preparation methods alter the process which may have selected for different contamination points.

Monitoring of kit, environmental and instrument contamination is vital to prevent false association of microorganisms with disease. Strong et al compared RNAseq data from two different studies on the same, presumed bacterial free, cell lines[176], with bacterial reads being

identified in all samples at a level of up to 10,0000 reads per million human mapped reads (RPMH). Interestingly the two studies identified different taxa, suggesting bacterial reads didn't originate in the samples. They presented further evidence by sequencing the same EBV-positive lymphoma at six different quality controlled study laboratories. Again bacterial reads were identified in all samples, but the number of reads identified varied by up to 30-fold in number, with distinct taxa identified at each site. Additionally, when sample sequencing was repeated at the same laboratory data showed taxa clustering with the original sequencing data. This study put doubt on another study which reported an association of *Fusobacterium nucleatum* with human colorectal cancer[177], which found a level of 861 RPMH. This evidence from multiple study sites highlights a need for care when drawing conclusions of association, especially when entire study sets are batched into single sequencing runs. Negative controls at multiple sample preparation points will increase the confidence with which conclusions can be drawn by providing evidence that pathogen reads originated in the sample and are not derived during sample processing.

Further blood spiking experiments were undertaken to explore the potential of the method for more challenging situations. Most clinical diagnostics separate the viral and bacterial elements of infectious disease diagnosis, however there situations where the two can be linked, for example bacterial pneumonia following influenza infection[178]. The first model was a mix of bacteria and virus in the blood. An enveloped RNA virus was used as it is the most challenging pathogen type as the enveloped nature of the virus means it is fragile during sample processing and the RNA needs additional conversion before amplification. The method allowed identification of both the influenza and *S. pyogenes* in the sample showing it is possible to merge investigation of viruses and bacteria. For future applications this could be important to look at disease associations, when looking at the 2009 influenza pandemic S. *pyogenes* re-emerged as a bacterial superinfection[179]. Using methods developed in this thesis, both infections could be diagnosed simultaneously, which could aid in the appropriate treatment of this co-infection which has a very high associated morbidity[180].

When spiking the horse blood, a higher inoculum of *Shigella sonnei* was used than with other bacteria, this was to compensate for the poor survival of the bacteria in horse blood. Diagnostic identification of *E. coli* and *Shigella* can be challenging with some methods not being able to differentiate them (e.g. MALDI-TOF) as the two are very closely related. The difference is historically based on disease profile, toxin production and serological differences. As genomic methods are being increasingly adopted separation of *Shigella* and *E. coli* is becoming less fundamental, as genomic data shows how related they are. In this sample most reads were

identified as Escherichia, with reads being speciated as both *E. coli* and *Shigella sonnei*, which is to be expected with how closely related they are. Very few genes differentiate the two species, the IpaH gene is the main method of separating enteroinvasive *E. coli* (EIEC) and *Shigella* sp from other *E. coli*[181]. This method was adopted bioinformatically, with good mapping results. The challenge of this sample was to see if it was possible to differentiate *E. coli* and *Shigella* in a clinical sample. Because of their mostly shared genome it would always be expected to never separate the two completely, but if the LCA analysis from the *Shigella* (Figure 5-10) and the *E. coli* (Figure 5-6) are compared they show some difference, with *Shigella* only identified in the *Shigella* sample. By double checking for the presence of the IpaH gene the *Shigella* in this case can be confidently call *Shigella. sonnei* above the *E. coli* however if the two were mixed (e.g. in faeces) it would be a more complex picture with the *Shigella* representing a much lower proportion and the identification may not be so clear cut. It was possible to sequence type using an *E. coli* database, and the trimethoprim resistance was predicted from the genome. The other resistances reported on the phenotypic report were either intrinsic resistances or cases where antibiotics were not clinically applicable. These were not reported using database resistance reporting, as the database was written to be pan-pathogen applicable and so species specific interpretation is not reported. If bioinformatic tools are to be used to predict antibiograms in bacteria it is important to build in a database of species specific intrinsic resistances and report this alongside any acquired resistance genes. Distinguishing between different pathogenic subtypes of *E. coli* and *Shigella* can be important due to the potentially serious nature of infection. These challenges have previously been described[66], and the ability to on some level differentiate this in a sterile site shows the potential of this method for use in future outbreaks.

The genotypic method applied to the *Salmonella* gave better results than the phenotypic test undertaken at the hospital laboratory which was unable to speciate the isolate using biochemical methods. The genotypic method was able to identify, sequence type and provide accurate resistance information directly from the sample, which would have been more rapid in real-time than the current method which required the reference laboratory to assign a species to this isolate. The application of real-time typing using sequencing data has been shown during a Salmonella hospital outbreak [182], using methods developed in this thesis genomic data could be obtained without the need for culture, speeding up time to results.

When investigating the clinical isolate of *P. aeruginosa* several resistance markers were initially identified using the *de novo* assembly on the reads identified as *P. aeruginosa* using Blastn and LCA analysis. However, the beta-lactamase identified using PCR at the reference laboratory was not found. When the unassigned reads were investigated the blaVIM-7 gene was identified,

which demonstrates the need for not discarding these, as they may contain reads that are plasmid, not obviously belonging to a specific species or be of poor quality. This may be more of a problem with very short read technologies as the reads won't span whole genes and therefore are less likely to span genus/species specific elements. It also demonstrates this method is poor at plasmid classification and might suggest that further pipeline development is needed in order to reliably identify plasmid presence. However, this *de novo* method did allow an additional beta-lactamase to be identified, Oxa-50, which is not one of the targets amplified in the reference laboratory PCR. This demonstrates a weakness in targeted processes for resistance identification. Interestingly although Oxa-50 is a fairly narrow spectrum oxacillinase, it has been reported to have low level capacity to hydrolyse imipenem[183]. Evolution of resistance in bacteria can be rapid, especially in Gram negative species, and so using targeted methods may miss newly evolved genes or SNPs conferring novel or adapted resistance mechanisms. One emerging area that this technique could be applied is for plasmid outbreaks, which have recently been reported in *Klebsiella spp*[174]. This is particularly important as Klebsiella spp are the 5th most common cause of bacteraemia in the UK[93]. Using the methods developed in this thesis the exact isolate causing a bacteraemia could be sequenced, rapidly and without need for culture, and be used to inform further epidemiological study where an outbreak is suspected. The method has been shown to be able to sequence plasmids, although further bioinformatic work may need to be done to increase detection of plasmids.

The value of the developed method for host DNA depletion from samples can be seen when the results of this study are compared to those found by Grumaz et al [184]. When they used NGS to investigate bacteraemia an average of 96% of their reads mapped to the human genome. This is in stark contrast to the average 1.5% of Illumina reads mapping to host from the spiked blood samples in this thesis. Grumaz et al relied only on centrifugation without a selective lysis step before extraction and amplification of the DNA. In order to identify bacteria species in the sample they relied on very high depth sequencing (average of 26 million reads per sample), which was an order of magnitude higher than the number of reads achieved in this thesis (average1.8 million reads). This level of sequencing output wouldn't be practical in a clinical setting due to the prohibitive cost associated, additionally time to bioinformatic analysis completion would also increase. The results achieved using the selective lysis presented in this thesis also compare favourably to those achieved using ϕ29 amplification to investigate HPV from cervical swabs[185], which had 64.5% reads identified as human.

Four tissue models were investigated to test the suitability of the sample processing procedure to solid cellular samples, which is potentially the biggest challenge to this method as there are high levels of host in a difficult to extract form. The first sample was *H. influenzae* spiked into heart tissue. We were able to recover *H. influenzae* reads, showing that the sample processing worked for this tissue type. The LCA analysis was unable to sub-type the bacterium which probably reflects the lack of similar sequence available in the database. It was possible to use the MLST database using the *de novo* assembly, showing this is a more suitable method for sub-typing pathogens.

The hepatitis model showed high level contamination with *H. influenzae* indicating pre-amplification contamination, probably during sample preparation. This problem highlights the downside of very sensitive methods, in that they are very open to contamination events, as even a single cell can be amplified. However, it was still possible to detect HAV in the sample with good genome coverage. The brain model showed good isolation of all four bacteria and good separation of the three *Streptococcus* species, which is often time consuming where using biochemical test. An identical resistance marker was identified in both the *S. oralis* and the *S. mitis* isolate, which could either mean that the gene is conserved, rapidly spread through the Streptococcus genus, or that the LCA analysis mistakenly placed the gene in both bacterial species. Separation of *S. mitis* and *S. oralis* is important due to their differing antibiotic susceptibilities[186]. Differentiation can be difficult, for example the sequence homology of the 16S gene is greater than 99%[187] and routine MALDI-TOF techniques also often struggle[188]. There have been developments, with Friedrichs et al [186]showing that using a bioinformatic approach known as support vector machine (SVM) analysis of MALDI-TOF spectrum they can confidently differentiate the two species, however this analysis is beyond the scope of clinical laboratories.

The kidney model sample had undetectable levels of DNA after ϕ29 MDA suggesting that no amplification occurred. This could be because of a technical error (e.g. no enzyme was added), the ϕ29 was inhibited or the DNA was lost during treatment with S1 nuclease. All identifiable reads mapped the Macaca, the host species and the pathogen of interest was not detectable.

These tissue models were only designed to test the suitability of the isolation method and DNA amplification from tissues, and do not truly reflect a deep tissue infection. When applied to real samples it might be necessary to increase the depth of sequencing and isolating the pathogen may be more challenging.

Four clinical urine samples were tested to investigate the potential for whole genome pathogen analysis from urine. Two of these urines were of poor quality, with high levels of vaginal

flora present, suggesting female patients and poor sample collecting. Female urines are often problematic to take due to the high floral content, when using traditional methods selective agar and purity plates can be used to help isolate the pathogens above regional flora. In the two samples with less bacterial contamination there was a large number of unassigned reads, suggesting that this was random background contamination. Additionally, the urines were old and had been freeze thawed several times. There is also evidence that urine is an inhibitory sample for molecular methods [189], and so this combination of factors suggests that this method is not suitable for diagnostics from urine.

When examining the CSF data there was very little pathogen recovery. There are several possible explanations for this, firstly the samples came from a collection of patients which had neurological symptoms but no proven infection, and so there is the possibility of a non-infectious cause of the symptoms. Additionally these were old samples that may have been freeze thawed causing pathogen degradation, which may have particular impact on pathogens present at low levels in the sample [190]. There is also the factor that there were high levels of human signals, which may have masked smaller pathogen signals present at low levels. The decision to not treat these samples with nucleases was taken as the integrity of the sample was unknown and there was potential that lysed pathogen signals would have been destroyed. This high level of host even in a fairly low cellular sample shows how important laboratory host depletion is. One of the CSFs had 1.1% of reads assigned to *Onchocerca volvulus* covering less than 1% of the genome, a nematode that is a causative agent of river blindness. The nematode has been associated with a neurological condition known as 'nodding syndrome' through an epidemiological and IgG study [191] in Uganda. There was no clinical data available alongside these samples and so it is difficult to comment on the possibility of this being a true cause of the neurological syndrome. There was very low genome coverage of the pathogen, so no strong evidence that it was a true positive. However other studies have inferred clinical significance from the presence of as little as two viral reads, from a total read number of 15 million reads [192]. In this case for further conclusions to be drawn patient information on syndrome and geographical exposure would be needed. This pathogen would not have been tested for when investigating neurological symptoms in the UK.

Four STI eluents were provided from the Sexually Transmitted Bacteria Reference Unit at PHE Colindale after they had completed routine testing. The samples in BD viper buffer produced very little DNA in the ϕ29 MDA amplification. DNA quantified in these first two STI swabs could have been the original DNA extracted from the samples and not amplified the ϕ29 MDA reaction. The first sample was a rectal swab which had previously been shown to be positive for *C. trachomatis*. Using ϕ29 MDA amplification is was not possible to detect the *C. trachomatis* above rectal flora,

however this may have been possible if the sample was deep sequenced. The second STI swab was a vaginal swab and it was possible to find the *C. trachomatis* in the ϕ29 MDA sample; however, it was not possible to further identify the pathogen. There was a high level of anaerobic bacteria present in line with those found in culture based investigations[193] . Additionally *Sneathia sanguinegens* was identified in this sample, which has previously been suggested as a pathogen[194] . *Sneathia s*pp have only recently been associated with a variety of female reproductive orders, due to its fastidious nature[194]. The ability of the methods developed to identify a putative pathogen, shows the potential value in pathogen discovery and characterisation. Additionally, the GC content of Sneathia spp being as low was 28% provides further evidence that this method is robust to extremes of GC content. In the third STI swab sample *N. gonorrhoeae* was identified concurring with the reference laboratory. Additionally, the pathogen *T. vaginalis* was identified; there was no clinical information available so it was not possible to tell if this was a known *T. vaginalis* infection. *T. vaginalis* is a poorly diagnosed pathogen with most diagnostics dependant on wet mount microscopy or culture, and visual identification of the pathogen, both of which need viable parasites and have sensitivity ranging from 35% to 60% [195]. *Mycoplasma genitalium* was also identified, which has increasingly been considered a pathogen of the genital tract[196], *M. genitalium* can be difficult to identify with traditional means as it is an intracellular pathogen and no commercial PCR systems are available. The fourth STI swab had a single species (*Gardnerlla vaginalis*) which represented 89.32% of identified reads, which shows a difference in floral composition from the mixed one seen in the other samples, suggesting a disease state that has altered the flora state. As well as identifying the expected *N. gonorrhoeae* a human papilloma virus (HPV8) was also identified, which although not know to be pathogenic itself, showed that the method would be suitable for detection of viral STIs.

The female genital tract is a complex interaction of colonising bacteria[197] which are in dynamic response to hormonal changes, sexual exposure and disease state of the host. In the three vaginal samples sequenced a large number of different species were found **Figure 5-14.** Increasingly, the line between 'pathogen' and 'commensal' is becoming blurred, especially in this site, as what constitutes a pathogen depends not only on the virulence of the bacteria itself but also with its interaction with the complex flora present, as well as the host state[193]. The samples sequenced here represented samples that are referred to the reference laboratory because they show a more complex infection, perhaps due to treatment failure or unusual or severe symptoms, which might explain why multiple pathogens were present in each vaginal sample. Sequencing analysis of complex presentations of infections will allow insight into causes of symptoms moving on from the single-pathogen diagnostic system currently used to a more dynamic, ecological understanding of the disease state[198]. This will bring many challenges in interpretation, and a high

level of training will need to be provided to take full advantage of the potential benefits of this technology.

The three STI blood samples appeared to be lysed when they were received, most likely due to freeze thawing action. This process would also affect the integrity of pathogens present. The method was developed to be applied to whole fresh blood, and so it wasn't possible to remove the RBCs from the blood samples as intended. *T. pallidum* was not identified in any of these samples, however even if the sample integrity wasn't compromised, *T. pallidum* only shows transient blood infections. In one of the samples the Torque teno virus was identified, which has previously been described as a 'ubiquitous virus' [199], with no evidence of pathogenicity. The ulcer samples also showed poor amplification, possibly because the transport media had high salt levels. It was possible to identify skin flora bacteria. In one of the samples 0.05% of reads were identified as *T. pallidum.*

Overall the methods developed in this thesis have been shown to be suitable for fresh whole blood, providing sensitive genome amplification directly from sample. When looking at isolates sequenced directly from blood there was consistently comparable results to the phenotypic results. In some cases, (*Salmonella and P. aeruginosa*) additional identification was achieved through genomic analysis compared to phenotypic results. The methods developed also performed well in clinical STI samples, a clinical CSF and spiked tissue samples. Previous work done with similar techniques have focussed on a targeted pathogen such as HPV[200] or single sample type such as viruses from skin swabs[201]. Another approach to overcome human background is to use a capture array technique such as SureSelect (Agilent Technologies), this has been used successfully to sequence TB in smear positive samples[202]. However, an approach such as this would not have been able to identify *Onchocerca volvulus* in the CSF.

The urines in this study produced poor results, because of contamination and probably sample freeze-thawing. Use of nuclease treatment to remove host signals is an important stage in the laboratory process and this process is most suited to samples with low host cell number. Sample quality is important for good quality pathogen isolation and sequencing directly from samples. It is possible to use this method to look at changes in the microbiome of samples and to simultaneously identify bacterial and viral pathogens present in blood

The failure of the method to work in some sample types highlights the challenges of applying modern molecular biology techniques to samples collected for traditional processes. What we can learn from this failure is that this technique may be most effectively deployed close to the patient, by teams who appreciate the need for specialised sample handling.

Overall there was a noticeable poorer performance of real clinical samples when compared to prepared models in this thesis. There are several reasons for this, firstly the basis for the sample processing was entirely based on the creation of freshly spiked horse blood. Real samples may differ from the horse blood in several important ways such as cellular make up, volume and presence of inhibitory molecules. Sample and pathogen volumes were tightly controlled in the models, which would not be possible in real samples. Additionally, the calculated inoculums of bacteria and viruses were estimated based on dilution studies, and so there may have been more bacteria present than estimated. Problems with the tested clinical samples mainly stem from the acquisition of samples which had been submitted to the reference laboratory, after all reference work had ben completed. This means that all samples had been processed, stored and transported lowering sample integrity. Additionally, low volumes were often available and storage of samples would normally involve at least one freeze thaw. To overcome the issue of sample integrity a clinical trial would need to be undertaken to allow additionally samples to be taken from patients in order to compare the traditional diagnostics and direct from sample next generation sequencing to be compared. Ethics would need to be sort, and a plan in place for the potential public health outcomes of new and novel pathogen discovery. This would involve large planning over a long period of time and would be an expensive study, possibly involving several staff including patient interactions. This kind of study is beyond the bounds of this PhD thesis, and would take a long time to plan, be commissioned and run.

Conclusions

6. General Conclusions

Currently, laboratory techniques for infectious disease diagnosis are based on culture and PCR amplicon detection. These techniques have inherent biases towards known aetiologies of infection, and would be slow in responding to novel outbreaks. In depth characterisation and epidemiological investigation is often time and labour intensive. The rapid improvements in technology and data outputs of next generation sequencers, offers the potential for improved diagnosis and scientific understanding of infectious diseases. Currently the application of NGS to microbiology is often limited to whole genome sequencing of cultured bacteria or DNA amplicons. The use of an unbiased method for direct from sample amplification alongside a robust data analysis pipeline could provide improved identification, characterisation and understanding of infections.

Primer free amplification was investigated by attempting to initiate DNA replication at DNA nicking points. Although this had previously been successful using other enzymes capable of displacement amplification, it was not successful when ϕ29 was trialled. This failure was most likely down to the low working temperatures of ϕ29 meaning little DNA dissociation. Additionally, as discovered in this thesis, the enzymes requires of at least 5 bases of single stranded DNA in order to initiate replication. Combination with other displacement enzymes was also unsuccessful most likely due to different enzymatic working conditions.

Attachment of a DNA tag to DNA as initiation points for DNA replication was successful when more than five bases was available, but lost efficiency when 15 single stranded bases were present. However, it was not possible to initiate reverse transcription using this method. Amplification from tagging lowered the efficiency with which the DNA was produced, due to less DNA replication initiation points. For these reasons and the need for fragmented DNA this was not deemed suitable for rapid sensitive amplification of pathogen nucleic acid. It did however show potential, and could be used to produce 'bar-coded' DNA so exact initial abundance of samples could be established after sequencing in situations where rapid amplification is less important.

The use of ϕ29 MDA with random primers allows rapid production of DNA for sequencing without the need of prior target knowledge. It was proven to be successful at amplifying single cells, with results being comparable to whole colony input (estimated $1x10^6$ cells) and even non-amplification culture based sequencing. It was also shown that reduction of amplification time to just two hours still allowed sufficient amplification for whole genome sequencing without

impacting on genome coverage. The method also proved to be applicable to a range of bacterial genome GC contents and to low number mixed bacterial cells. The extraction method was successful at extracting nucleic acid from a variety of cell wall types and encapsulated viruses. Use of the alkali extraction method rapidly produced single stranded DNA with limited shearing, which improved the quality of the assemblies compared to column based extraction methods.

Bioinformatic removal of host signals through reference mapping is a rapid and effective way of lowering data complexity and should be performed on all sequencing samples as the first stage of analysis. Likewise, it is important to monitor presence of contamination of environment and kits through the use of negative amplification. This allows a contamination library to be created which allows mapping based methods to be used to rapidly remove known contaminants, therefore simplifying the data set and removing the chance of false incorporation in *de novo* assemblies. Use of very strict error trimming is detrimental to the dataset as too much information is lost; the best combination for this dataset is a contamination of average reads and end trimming at a cut off of Q20. Use of abundance trimming lowers the size and complexity of the dataset without losing information and allows better quantification of input, most likely due to removing repetitive genome elements. There is a need to test and optimise the *de novo* assembler for datasets based on input, DNA preparation method and sequencing platform. For this dataset the best performing assembler was SPAdes which combines a *de Bruijn* algorithm with read distance relationships, allowing advantage to be taken of long reads. It also has the advantage of using several kmer lengths, allowing the optimal kmer length to be found by the programme for each dataset and doesn't need to be set by the user. Characterisation using *de novo* assembled data in multiple contigs is comparable to results produced from whole polished reference genomes. Virulence and antibiotic resistance prediction is mostly database dependant and so attention should be used when interpreting the outputs.

It is possible to have a processing pathway that allows a single amplification reaction and sequencing run to be performed on both viral and bacterial pathogens. There is however a need to separate the isolation processes due to differences in size, particle stability and nucleic acid makeup. The sample processing pipeline developed in this thesis allows both sets to be performed simultaneously. When the process was performed on fresh whole blood it was possible to produce accurate information on pathogen identification, sequence type and antibiotic resistances. This is included an example of both a bacteria and enveloped RNA virus.

As we approach the dawn of the era of genomic diagnostics it is important to re-evaluate how pathogens are classified, such as our current system of defining *Shigella* sp and *E. coli* as different

species despite their close genetic relationship. Use of untargeted resistance detection allowed an additional resistance gene to be identified in the *P. aeruginosa*. Resistance in Gram negative pathogens evolves quickly, and targeted methods will limit those detected to those looked for. This also highlights the presence of multiple resistance genes in single isolates, which may not always be investigated. This sample also highlighted that LCA analysis is not suitable for plasmid detection, and so additional plasmid-specific analysis may be required.

Contamination can have serious impact on the detection of low level pathogens, and so extra care and vigilance should be taken during sample processing and data analysis. Removing host signals is important so that pathogens signals can be detected above them. Sample and pathogen integrity are hugely important when using amplification methods, especially when they are used in combination with nuclease methods for removal of host signals.

Looking at the microbiome of samples, especially when symptom presentation is complex, will allow more insight into the disease state of the patient. By increasing data production, it is hoped that more understanding of the microbial environment can be achieved. potentially leading to a re-evaluation of how a pathogen is defined, accounting for not just the virulence of a single bacterium but also its impact and interaction with the microbiome. Overall the sample processing, amplification and data interpretation techniques are most suited to sterile site infections, and can allow identification of pathogens from the presence of a single copy of the genome.

6.1 Future Developments and Applications

Microfluidics allows highly precise manipulation of small particles in small volumes, and they has been increasingly used for concentration and isolation of bacteria from fluid samples[203],[204]. This has been shown to be successful in environmental studies[205], with magnetic beads used to selectively isolate *E. coli* to a sensitivity of 100 cfu/ml in milk and food samples using a 3D printed microfluidic device. Microfluidic isolation has also been coupled with nucleic acid techniques[206] and used to isolate bacteria from blood[170]. With the increased availability of 3D printing, and the development of more sophisticated yet affordable microfluidic systems, coupling this isolation system with ϕ29 MDA could produce rapid, accurate sequencing, and more importantly lead to a higher throughput system with less hands on time for isolation of low numbers of bacterial cells from clinical samples.

Other capture methods for pathogens could involve using magnetic beads in combination with a pan-pathogen binding molecular. One candidate is mucin, which is produced by epithelial cells, which is the site of most pathogen binding. It has previously proven affective as a capture method for Norovirus[207,208], in house (unpublished) work has also shown that this methods is affective for capture of a number of Enteroviruses including Ev-D68 which is a respiratory pathogen. Additionally it has been shown that mucins are the first molecule that enteric bacteria interact with[209] suggesting potential for pan-pathogen capture. Magnetic capture allows simultaneous selection and concentration of pathogens, but would have to be as broad as possible to take advantage of the un-targeted nature of the approach developed here.

An alternative approach to capturing pathogen would be the use ϕ29 MDA with semi-specific primers which would allow whole genome capture without knowing the whole genome sequence. For example, these primers could be developed for species specific targets, but because of the long amplicons produced and the displacement activity of the enzyme whole genomes could be amplified including novel genome elements such as insertions or mutations.

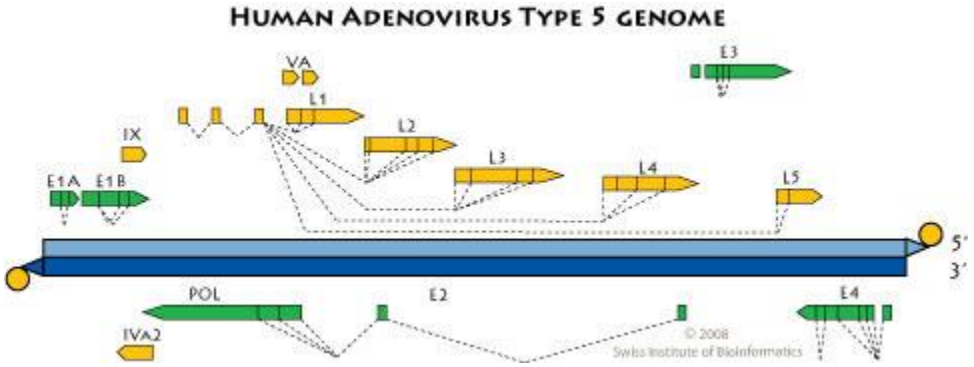
Additional work would need to be performed on the bioinformatic processes, to allow the results to be more reliable and easier to interpret. Current pipeline outputs are complex to interpret due to them using pan-pathogen databases (such as Prokka, ResFinder), species specific databases which are created and curated by experts in the field and would allow genomic predictions to be more accurate and reliable.

Although short read technologies are the most used techniques for producing bacterial sequence data, difficulty in placing repetitive elements and fully mapping chromosomes means that it is not always possible to fully identify novel resistance and virulence islands in pathogens.

Long read technologies such as the MinION nanopore have been used to fully characterise these gene structures [61]. One novel finding of this work was that DNA produced by ϕ29 MDA gave better results in bioinformatic analysis than traditional culture extraction. As there is no apparent, fundamental limit on the read lengths that can be produced with the MinION, one exciting possibility is coupling the method here with MinION sequencing. The long reads would be ideal for the MinION library prep, where DNA shearing is a problem. DNA production using ϕ29 MDA would lead to large numbers of long reads being sequenced, which could improve the efficiency of MinION sequencing. It would also be interesting to apply this method to low biomass microbiomes, as the DNA produced by ϕ29 MDA shows impressive abundance correlation with the input DNA.

Additionally, short read technologies often have a long run time with no access to data before the completion of the run, they also require samples to be processed in bulk to provide economic data acquisition. The MinION allows real-time data collection and analysis, when this technology was applied to a hospital outbreak of *Salmonella*[174], the species was identified within 20 minutes of sample loading. Furthermore, it was possible to assign an individual isolate to outbreak cluster or non-outbreak within 100 minutes, allowing informed infection control measures to be implemented. MinION data was also used during the Ebola outbreak in Guinea in 2015[175], allowing real-time phylogenetic analysis to be performed. When optimum processing was achieved results were available with 24 hours of sample receipt. One advantage was the ability to performed remote bioinformatic analysis, minimising the amount of equipment and staff needed at the site of the outbreak. However, data uploading did have the disadvantage of requiring internet connection, which is not always predicable in remote locations. However, this work relied on amplification of pathogen signals which was reliant on prior knowledge of the aetiology. The long DNA strands produced by the untargeted methods developed in this thesis would allow direct from sample real-time identification, characterisation and possibly phylogenetic analysis of infections for real-time field epidemiology of novel outbreaks.

7. Supplementary Material



	Elements only in adeno_mda :	Elements only in adeno_ref :
E1A	"Early E1A 27 kDa protein"	"Early E1A protein"
E1B	"E1B protein, large T-antigen"	"Adenovirus EB1 55K protein / large t-antigen"
IX	"Hexon-interlacing protein"	"Adenovirus hexon-associated protein (IX)"
E2A	"DNA-binding protein"	"Viral DNA-binding protein, zinc binding domain"
	"Early E2A DNA-binding protein"	
E2B	"Preterminal protein"	
E3	"Early 3 14.7 kDa protein"	"Adenovirus 15.3kD protein in E3 region"
	"Pre-early 3 receptor internalization and	"Adenovirus E3B protein"
U	"U exon protein"	
E4	"Splicing factor"	"Mastadenovirus E4 ORF3 protein"
	"putative early E4 11 kDa protein"	"Mastadenovirus early E4 13 kDa protein"
	"Early 4 ORF4 protein"	
	"Early 4 ORF6 protein"	
L1	"Pre-capsid vertex protein"	"Adenoviral protein L1 52/55-kDa"
L2	"Core-capsid bridging protein"	"Adenovirus minor core protein PV"
	"Late L2 mu core protein precursor"	"Adenovirus late L2 mu core protein (Protein X)"
	"Pre-histone-like nucleoprotein"	"Adenoviral core protein VII"
	"Penton protein"	"Adenovirus penton base protein"
		"Minor capsid protein VI"
L3	"Hexon protein"	"Adenoviral DNA terminal protein"
	"Protease"	"Adenovirus endoprotease"
L4	"Shutoff protein"	"Late 100kD protein"
	"Pre-hexon-linking protein"	"Hexon-associated protein (IIIa)"
		"Adenovirus hexon associated protein, protein
L5	"Fiber protein 1"	"Adenoviral fibre protein (repeat/shaft region)"
	"Fiber protein 2"	
	"Packaging protein 3"	"Adenovirus IVa2 protein"
IVa2	"Packaging protein 1"	

Table 7-1 all the genes identified by Prokka in Adenovirus reference and MDA amplified adenovirus, arranged by transcript

	Phage associated genes	Plasmid associated genes	Putative genes
1	"Phage lysozyme"	"Plasmid SOS inhibition protein (PsiB)"	"putative acyl transferase"
2	"Phage holin family (Lysis protein S)"	"Plasmid-derived single-stranded DNA-binding	"putative ATP-dependent helicase Lhr"
3	"Phage integrase family protein"	"Conjugative relaxosome accessory transposon	"putative signal peptide"
4	"Lambda phage tail tape-measure protein	"Type-F conjugative transfer system pilin	"putative deoxyribonuclease RhsA"
5	"Bacteriophage CII protein"	"Conjugal transfer protein TrbE"	"putative deoxyribonuclease RhsB"
6	Bacteriophage Lambda NinG protein"	"conjugal transfer mating pair stabilization	"putative zinc-type alcohol dehydrogenase-like
7	"Phage portal protein, lambda family"	"Type-F conjugative transfer system protein	"Putative dimethyl sulfoxide reductase chain YnfE
8	"Phage terminase large subunit (GpA)"	"F pilus assembly Type-IV secretion system for	"putative dimethyl sulfoxide reductase chain YnfF
9	"Phage minor tail protein L	"Bacterial conjugation TrbI-like protein"	"putative nucleotide-binding protein containing
10	"Phage minor tail protein"		"putative oxidoreductase YciK"
11	"Prophage tail length tape measure protein"		"Putative metal chaperone YciC"
12	"Bacteriophage lambda minor tail protein (GpG)"		"putative fimbrial-like protein YcbV precursor"
13	"Phage minor tail protein U"		"putative dehydratase"
14	"Prophage minor tail protein Z (GPZ)"		"putative Nudix hydrolase NudL"
15	"Phage Head-Tail Attachment"		
16	"Phage major capsid protein E"		
17	"Bacteriophage lambda head decoration protein D"		
18	"Bacteriophage lambda Kil protein"		
19	"Lambda Phage CIII"		
20	"Bacteriophage replication protein O"		
21	"Phage NinH protein"		
22	"Phage portal protein, lambda family"		
23	"Bacteriophage lambda tail assembly protein I"		
24	"Phage minor tail protein L"		

Table 7-2 All genes found in MDA and Control K12 but not in the reference

E. coli	*C. difficile*	*A. naeslundii*
"6-N-hydroxylaminopurine resistance protein"	**"Aluminium resistance protein"**	**"6-N-hydroxylaminopurine resistance protein"**
"Arsenical resistance operon repressor"	"Bicyclomycin resistance protein"	**"Antiseptic resistance protein"**
"Bicyclomycin resistance protein"	"Daunorubicin/doxorubicin resistance ABC	**"Copper resistance protein A precursor"**
"Bifunctional polymyxin resistance protein ArnA"	"Fusaric acid resistance protein family protein"	"Daunorubicin/doxorubicin resistance ABC
"Copper resistance protein C precursor"	**"Mercuric resistance operon regulatory protein"**	"Daunorubicin/doxorubicin resistance ATP-binding
"Daunorubicin/doxorubicin resistance ATP-binding	"Methicillin resistance mecR1 protein"	**"Mercuric resistance operon regulatory protein"**
"Fosmidomycin resistance protein"	"Methicillin resistance regulatory protein MecI"	"Multidrug resistance protein 3"
"Multidrug resistance-like ATP-binding protein	"Multidrug resistance operon repressor"	"Multidrug resistance protein EbrB"
"Multidrug resistance outer membrane protein MdtP	"Multidrug resistance protein 3"	"Multidrug resistance protein MdtA"
"Multidrug resistance protein D"	"Multidrug resistance protein MdtA"	"Multidrug resistance protein MdtK"
"Multidrug resistance protein MdtA precursor"	"Multidrug resistance protein MdtG"	"Multidrug resistance protein stp"
"Multidrug resistance protein MdtC"	"Multidrug resistance protein MdtK"	**"tellurite resistance protein TehB"**
"Multidrug resistance protein MdtE precursor"	"Multidrug resistance protein NorM"	"Vancomycin B-type resistance protein VanW"
"Multidrug resistance protein MdtF"	"Multidrug resistance protein stp"	
"Multidrug resistance protein MdtH"	"Multiple antibiotic resistance protein MarA"	
"Multidrug resistance protein MdtK"	"Non-motile and phage-resistance protein"	
"Multidrug resistance protein MdtL"	**"Organic hydroperoxide resistance transcriptional**	
"Multidrug resistance protein MdtM"	"putative multidrug resistance ABC transporter	
"Multidrug resistance protein MdtN"	"putative multidrug resistance protein EmrY"	
"Multidrug resistance protein MdtO"	"**Quaternary ammonium compound-resistance protein**	
"Multidrug resistance protein stp"	**"tellurite resistance protein TehB"**	
"Multiple antibiotic resistance protein MarA"	**"Tellurium resistance protein"**	
"Multiple antibiotic resistance protein MarR"	"Tetracycline resistance protein TetM"	
"Multiple stress resistance protein BhsA	"Tetracycline resistance protein TetO"	
"putative multidrug resistance ABC transporter	"Vancomycin B-type resistance protein VanB"	
"putative multidrug resistance protein EmrK"	"Vancomycin B-type resistance protein VanW"	
"putative multidrug resistance protein EmrK"		
"putative multidrug resistance protein EmrY"		
"Putative multidrug resistance protein MdtD"		
"Tellurite resistance protein TehA"		
"tellurite resistance protein TehB"		
"Tetracycline resistance protein, class B"		
"Zinc resistance-associated protein precursor"		

Table 7-3 output of initial resistance prediction using Prokka output word search

E. coli	*C. difficile*	*A. naeslundii*
"bicyclomycin/multidrug efflux system"	**"Ferrous-iron efflux pump FieF"**	**"Ferrous-iron efflux pump FieF"**
"Cation efflux system protein CusA"	**"Glutathione-regulated potassium-efflux system**	**"Glutathione-regulated potassium-efflux system**
"Cation efflux system protein CusC precursor"	**"Magnesium and cobalt efflux protein CorC"**	**"Magnesium and cobalt efflux protein CorC"**
"Cation efflux system protein CusF precursor"	**"manganese efflux pump MntP"**	**"Vacuole effluxer Atg22 like protein"**
"Cysteine/O-acetylserine efflux protein"	"multidrug efflux protein"	
"Ferrous-iron efflux pump FieF"	"Multidrug-efflux transporter 1 regulator"	
"Glutathione-regulated potassium-efflux system	"Outer membrane efflux protein"	
"Homoserine/homoserine lactone efflux protein"	**"Purine efflux pump PbuE"**	
"Leucine efflux protein"	**"putative cation efflux system proteinc/MT2084"**	
"Magnesium and cobalt efflux protein CorC"	"Putative efflux system component YknX"	
"Multidrug efflux pump accessory protein AcrZ"	**"Vacuole effluxer Atg22 like protein"**	
"Multidrug efflux pump subunit AcrA precursor"		
"Multidrug efflux pump subunit AcrB"	**"Glutathione-regulated potassium-efflux system**	
"Multidrug efflux pump subunit AcrB"	**"Magnesium and cobalt efflux protein CorC"**	
"Nickel/cobalt efflux system RcnA"	**"manganese efflux pump MntP"**	
"p-hydroxybenzoic acid efflux pump subunit AaeA"	"multidrug efflux protein"	
"p-hydroxybenzoic acid efflux pump subunit AaeB"	"Multidrug-efflux transporter 1 regulator"	
"Purine ribonucleoside efflux pump NepI"	"Outer membrane efflux protein"	
"putative amino-acid metabolite efflux pump"	**"Purine efflux pump PbuE"**	
"Sugar efflux transporter"	**"putative cation efflux system proteinc/MT2084"**	
"Sugar efflux transporter A"	"Putative efflux system component YknX"	
"Sugar efflux transporter B"	**"Vacuole effluxer Atg22 like protein"**	
"Sugar efflux transporter C"	**"Ferrous-iron efflux pump FieF"**	
"Threonine efflux protein"	**"Glutathione-regulated potassium-efflux system**	

Table 7-4 output of all efflux pumps in using Prokka annotations

Common elements in E_coli_ref E_coli_con E_coli_MDA :	
"Beta-lactamase precursor"	
"bicyclomycin/multidrug efflux system"	**Bicyclomycin** antibiotic with unique structure, transmembrane transporter activity
"Bicyclomycin resistance protein"	
"Colicin I receptor precursor"	**Colicin** is a type of bacteriocin produced by and toxic to some strains of Escherichia coli.
"Colicin V production protein"	
"Daunorubicin/doxorubicin resistance ATP-binding	**Daunorubicin:** a synthetic antibiotic that interferes with DNA synthesis
"Fosmidomycin resistance protein"	**Fosmidomycin** is an antibiotic that was originally isolated from culture broths of bacteria of the genus Streptomyces
"Modulator of drug activity B"	
"Multidrug efflux pump accessory protein AcrZ"	AcrB and its homologues are the principal multidrug transporters in Gram-negative bacteria and are important in antibiotic drug tolerance[210]
"Multidrug efflux pump subunit AcrA precursor"	
"Multidrug efflux pump subunit AcrB"	
"Multidrug export protein AcrE precursor"	
"Multidrug export protein AcrF"	
"Multidrug export protein EmrA"	efflux system EmrAB-TolC, which confers resistance to antibiotics such as CCCP, FCCP, 2,4-dinitrophenol and nalidixic acid[211]
"Multidrug export protein EmrB"	
"Multidrug transporter EmrE"	Multidrug transporter that expels positively charged hydrophobic drugs across the inner membrane of *E. coli*. thereby conferring resistance to a wide range of toxic compounds. The drug efflux is coupled to an influx of protons. Is involved in the resistance of *E. coli* cells to methyl viologen, ethidium bromide and acriflavine. Is also able to transport tetraphenylphosphonium (TPP^+) and benzalkonium[212]
"putative multidrug resistance protein EmrK"	Part of the tripartite efflux system EmrYK-TolC, which confers resistance to various drugs
"putative multidrug resistance protein EmrY"	
"Multidrug resistance protein MdtA precursor"	The MdtABC tripartite complex confers resistance against novobiocin and deoxycholate[213]
"Multidrug transporter MdtB"	
"Multidrug resistance protein MdtC"	
"Putative multidrug resistance protein MdtD"	Putative transmembrane protein
"Multidrug resistance protein MdtE precursor"	Part of the tripartite efflux system MdtEF-TolC, which confers resistance to compounds such as rhodamine 6G, erythromycin, doxorubicin, ethidium bromide, TPP, SDS, deoxycholate, crystal violet and benzalkonium[214]
"Multidrug resistance protein MdtF"	
"Multidrug resistance protein MdtH"	Confers resistance to norfloxacin and enoxacin[214]
"Multidrug resistance protein MdtK"	Multidrug efflux pump that functions probably as a Na^+/drug antiporter. Confers resistance to many drugs such as fluoroquinolones (norfloxacin, ciprofloxacin, enoxacin), tetraphenylphosphonium ion (TPP), deoxycholate, doxorubicin, trimethoprim, chloramphenicol, fosfomycin, acriflavine, ethidium bromide and benzalkonium[214]
"Multidrug resistance protein MdtL"	Confers resistance to norfloxacin and enoxacin[214]
"Multidrug resistance protein MdtM"	Confers resistance to acriflavine, chloramphenicol, norfloxacin, ethidium bromide and TPP[214]
"Multidrug resistance protein MdtN"	Operon: mdtNOP: Could be involved in resistance to puromycin, acriflavine and tetraphenylarsonium chloride
"Multidrug resistance protein MdtO"	
"Multidrug resistance outer membrane protein MdtP	
"Multidrug resistance protein stp"	Contributes to spectinomycin and tetracycline resistance[215]

	(common in mycobacterium)
"Multiple antibiotic resistance protein MarA"	May be a transcriptional activator of genes involved in the multiple antibiotic resistance (Mar) phenotype[216]
"Multiple antibiotic resistance protein MarR"	Repressor of the marRAB operon which is involved in the activation of both antibiotic resistance and oxidative stress genes[216]
"Penicillin-binding protein 1A"	AKA mrcA and mrcB Cell wall formation. Synthesis of cross-linked peptidoglycan from the lipid intermediates. The enzyme has a penicillin-insensitive transglycosylase N-terminal domain (formation of linear glycan strands) and a penicillin-sensitive transpeptidase C-terminal domain (cross-linking of the peptide subunits)[217]
"Penicillin-binding protein 1B"	
"Penicillin-binding protein 1C"	AKA pbpC Cell wall formation.[217]
"Penicillin-binding protein 2D"	Involved in the polymerization and cross-linking of spore peptidoglycan. May be required for synthesis of the spore germ cell wall, the first layer of peptidoglycan synthesized on the surface of the inner forespore membrane
"Penicillin-binding protein activator LpoA	Regulator of peptidoglycan synthesis that is essential for the function of penicillin-binding protein 1A/1B Stimulates transpeptidase activity of PBP1a/PBP1b in vitro[218]
"Penicillin-binding protein activator LpoB	
"Penicillin-insensitive murein endopeptidase	Aka mepA Murein is a another name for peptidoglycan Involved in the removal of murein from the sacculus.[219]
"Peptide antibiotic transporter SbmA"	Uptake of antimicrobial peptides. Required for the transport of microcin B17 (MccB17), microcin 25 (Mcc25) and proline-rich antimicrobial peptides into the cell[220]
"putative antibiotic transporter"	
"Putative beta-lactamase HcpD precursor"	May hydrolyze 6-aminopenicillinic acid and 7-aminocephalosporanic acid (ACA) derivatives (By similarity). Binds to penicillin. (UniProt)
"putative multidrug resistance ABC transporter	
"Putative phosphinothricin acetyltransferase	
"Ribosome-recycling factor"	
"RNA-splicing ligase RtcB"	
"vancomycin high temperature exclusion protein"	AKA sanA Participates in the barrier function of the cell envelope.
Common elements in E_coli_ref E_coli_MDA :	
Multidrug resistance protein D"	AKA emrD Multidrug resistance pump that participates in a low energy shock adaptive response
Tetracycline resistance protein, class B"	AKA tetA: Resistance to tetracycline by an active tetracycline efflux. This is an energy-dependent process that decreases the accumulation of the antibiotic in whole cells. This protein functions as a metal-tetracycline/H^+ antiporter

Table 7-5 all resistance factors identified using Prokka in *E. coli*

Gene identified	**Function**
Common elements in C_diff_ref C_dif_con C_dif_MDA :	
Beta-lactamase	
Beta-lactamase HcpA precursor	Slowly hydrolyzes 6-aminopenicillinic acid and 7-aminocephalosporanic acid (ACA) derivatives. May be involved in the synthesis of the cell wall peptidoglycan (By similarity).
Beta-lactamase precursor	
Bicyclomycin resistance protein	
Daunorubicin/doxorubicin resistance ABC	Part of the ABC transporter complex DrrAB involved in daunorubicin and doxorubicin resistance. Probably responsible for the translocation of the substrate across the membrane.
Daunorubicin/doxorubicin resistance ATP-binding	Part of the ABC transporter complex DrrAB involved in daunorubicin and doxorubicin resistance. Responsible for energy coupling to the transport system. Binds ATP or GTP
Metallo-beta-lactamase superfamily protein	Functionally predicted beta-lactamase
Methicillin resistance regulatory protein MecI	Transcriptional repressor that constitutively blocks the transcription of the gene for the penicillin-binding protein MecA. Binds palindromic DNA with the sequence 5'-TACA-[AT]-N-TGTA-3'. Regulates genes involved in antibiotic resistance. Binds DNA as a dimer. (In S.aureus)
Multidrug-efflux transporter 1 regulator	MerR family transcriptional regulator
Multidrug export protein EmrB	Drug resistance transporter, EmrB/QacA subfamily
Multidrug export protein MepA	Multidrug resistance efflux protein involved in transporting several clinically relevant monovalent and divalent biocides and the fluoroquinolone antimicrobial agents norfloxacin and ciprofloxacin
Multidrug resistance operon repressor	
Multidrug resistance protein 3	
Multidrug resistance protein MdtA	The MdtABC tripartite complex confers resistance against novobiocin and deoxycholate
Multidrug resistance protein MdtG	Confers resistance to fosfomycin and deoxycholate
Multidrug resistance protein MdtK	
Multidrug resistance protein NorM	Multidrug efflux pump that functions as a Na^+/drug antiporter. Confers resistance to several drugs, such as norfloxacin, ciprofloxacin, ethidium, kanamycin and streptomycin
Multidrug resistance protein stp	Contributes to spectinomycin and tetracycline resistance[215] (common in mycobacterium)
Multiple antibiotic resistance protein MarA	May be a transcriptional activator of genes involved in the multiple antibiotic resistance (Mar) phenotype.
Penicillinase repressor	Transcriptional repressor that constitutively blocks expression of beta-lactamase. Regulates genes involved in antibiotic resistance. Binds DNA as a dimer
Penicillin-binding protein 1A	AKA mrcA Cell wall formation. Synthesis of cross-linked peptidoglycan from the lipid intermediates. The enzyme has a penicillin-insensitive transglycosylase N-terminal domain (formation of linear glycan strands) and a penicillin-sensitive transpeptidase C-terminal domain (cross-linking of the

	peptide subunits)[217]
Putative multidrug export ATP-binding/permease	May be involved in multidrug export. Transmembrane domains (TMD) form a pore in the cell membrane and the ATP-binding domain (NBD) is responsible for energy generation
putative multidrug resistance protein EmrY	Part of the tripartite efflux system EmrYK-TolC, which confers resistance to various drugs
Putative small multi-drug export protein	
RNA-splicing ligase RtcB	
Tetracycline resistance protein TetM	Abolishes the inhibitory effect of tetracyclin on protein synthesis by a non-covalent modification of the ribosomes.
Vancomycin B-type resistance protein VanB	D-alanine-- ligase, This protein is involved in the pathway peptidoglycan biosynthesis, which is part of Cell wall biogenesis
Vancomycin B-type resistance protein VanW	Homology based
Common elements in C_diff_ref C_dif_con	
Demethylrebeccamycin-D-glucose	Glycosyl O-methyltransferase that catalyses the final step in the biosynthesis of rebeccamycin, an indolocarbazole alkaloid that inhibits topoisomerase 1. Has broad substrate specificity and functions as glycosyl O-methyltransferase on a number of rebeccamycin analogs
Penicillin-binding protein 1A/1B	Cell wall formation. Synthesis of cross-linked peptidoglycan from the lipid intermediates. The enzyme has a penicillin-insensitive transglycosylase N-terminal domain (formation of linear glycan strands) and a penicillin-sensitive transpeptidase C-terminal domain (cross-linking of the peptide subunits)[217]
Ribosome-recycling factor	Responsible for the release of ribosomes from messenger RNA at the termination of protein biosynthesis
Streptothricin hydrolase	
Elements only in C_dif_con :	
vancomycin high temperature exclusion protein	Participates in the barrier function of the cell envelope.
Common elements in C_diff_ref C_dif_MDA :	
Methicillin resistance mecR1 protein	Penicillin-interactive protein and potential antirepressor
multidrug efflux protein	
Oleandomycin glycosyltransferase	Oleandomycin macrolide antibiotic Specifically inactivates oleandomycin via 2'-O-glycosylation using UDP-glucose.
putative multidrug resistance ABC transporter	
Tetracycline resistance protein TetM	Abolishes the inhibitory effect of tetracycline on protein synthesis by a non-covalent modification of the ribosomes
Tetracycline resistance protein TetO	
Elements only in C_dif_MDA :	
Penicillin binding protein transpeptidase domain	AKA penicillin binding protein 1B

Table 7-6 all resistance factors identified using Prokka in *C. difficile*

Common elements in A_naes_ref A_naes_con A_naes_MDA :	
Daunorubicin/doxorubicin resistance ABC	Part of the ABC transporter complex DrrAB involved in daunorubicin and doxorubicin resistance. Probably responsible for the translocation of the substrate across the membrane.
Daunorubicin/doxorubicin resistance ATP-binding	Part of the ABC transporter complex DrrAB involved in daunorubicin and doxorubicin resistance. Responsible for energy coupling to the transport system. Binds ATP or GTP
Multidrug export protein EmrB	art of the tripartite efflux system EmrAB-TolC, which confers resistance to antibiotics such as CCCP, FCCP, 2,4-dinitrophenol and nalidixic acid
Multidrug resistance protein 3	
Multidrug resistance protein MdtA	The MdtABC tripartite complex confers resistance against novobiocin and deoxycholate
Multidrug resistance protein stp	Contributes to spectinomycin and tetracycline resistance
Penicillin-binding protein 1A	Cell wall formation. Synthesis of cross-linked peptidoglycan from the lipid intermediates. The enzyme has a penicillin-insensitive transglycosylase N-terminal domain (formation of linear glycan strands) and a penicillin-sensitive transpeptidase C-terminal domain (cross-linking of the peptide subunits) (By similarity).
Putative multidrug export ATP-binding/permease	May be involved in multidrug export. Transmembrane domains (TMD) form a pore in the cell membrane and the ATP-binding domain (NBD) is responsible for energy generation
Ribosome-recycling factor	Responsible for the release of ribosomes from messenger RNA at the termination of protein biosynthesis. May increase the efficiency of translation by recycling ribosomes from one round of translation to another.
vancomycin high temperature exclusion protein	Participates in the barrier function of the cell envelope.
Elements only in A_naes_ref :	
Maf-like protein YceF	Unknown: maf Involved in septum formation.
Penicillin-binding protein 2D	Involved in the polymerization and cross-linking of spore peptidoglycan. May be required for synthesis of the spore germ cell wall, the first layer of peptidoglycan synthesized on the surface of the inner forespore membrane
Penicillin-binding protein PbpB	Penicillin-binding proteins (PBPs) function in the late steps of murein biosynthesis. PBP-2B is required for vegetative cell division and sporulation septation. Beta-lactamase inactivates the PBPs by acylating an essential serine residue in the active site of these

	proteins, thereby interrupting normal cell wall synthesis.
Common elements in A_naes_ref A_naes_MDA :	
Multidrug resistance protein EbrB	Part of a multidrug efflux pump. Confers resistance to cationic lipophilic dyes such as ethidium bromide, acriflavine, pyronine Y and safranin O. The efflux is probably coupled to an influx of protons
Multidrug resistance protein MdtK	Multidrug efflux pump that functions probably as a Na^+/drug antiporter. Confers resistance to many drugs such as fluoroquinolones (norfloxacin, ciprofloxacin, enoxacin), tetraphenylphosphonium ion (TPP), deoxycholate, doxorubicin, trimethoprim, chloramphenicol, fosfomycin, acriflavine, ethidium bromide and benzalkonium
Penicillin-binding protein A	Cell wall formation. Synthesis of cross-linked peptidoglycan from the lipid intermediates. The enzyme has a penicillin-insensitive transglycosylase N-terminal domain (formation of linear glycan strands) and a penicillin-sensitive transpeptidase C-terminal domain (cross-linking of the peptide subunits) (By similarity).
Vancomycin B-type resistance protein VanW	By similarity
Common elements in A_naes_con A_naes_MDA :	
Beta-lactamase	
Beta-lactamase precursor	
Bicyclomycin resistance protein	Involved in sulfonamide (sulfathiazole) and bicyclomycin resistance. Probable membrane translocase
Multidrug resistance operon repressor	

Table 7-7 all resistance factors identified using Prokka in *A. naeslundii*

gene	Identification	origin
entF	Enterobactin_synthetase_component_F	Escherichia_coli_CFT073
fimD	Outer_membrane_usher_protein_fimD_precursor	Escherichia_coli_CFT073
ecpC	putative_enzyme	Escherichia_coli_O157:H7
sap	Sap_kinase	*Shigella* _flexneri_(serotype_2a)
fepA	Ferrienterobactin_receptor_precursor	Escherichia_coli_CFT073
ibeC	membrane_protein_YijP	Escherichia_coli_CFT073
ecpD	putative_receptor	Escherichia_coli_O157:H7
entE	Enterobactin_synthetase_component_E	Escherichia_coli_CFT073
orf	conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
fecA	FecA	*Shigella* _flexneri_(serotype_2a)
orf49	hypothetical_protein	Escherichia_coli_536
orf48	hypothetical_protein	Escherichia_coli_536
ibeB	IbeB	Escherichia_coli
aslA	putative_arylsulfatase	Escherichia_coli
entC	Isochorismate_synthase_entC	Escherichia_coli_CFT073
orf	conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
intB	bacteriophage_P4_integrase	Escherichia_coli_536
orf30	unknown	*Shigella* _flexneri_(serotype_2a)
ompA	outer_membrane_protein_A	Escherichia_coli
fecC	FecC	*Shigella* _flexneri_(serotype_2a)
fecR	FecR	*Shigella* _flexneri_(serotype_2a)
fepE	Ferric_enterobactin_transport_protein_fepE	Escherichia_coli_CFT073
fecD	FecD	*Shigella* _flexneri_(serotype_2a)
yi22_	orf,_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
yi22	orf,_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
fepD	Ferric_enterobactin_transport_system_permease_protein_fepD	Escherichia_coli_CFT073
fecB	FecB	*Shigella* _flexneri_(serotype_2a)

fepB	Ferrienterobactin-binding_periplasmic_protein_precursor	Escherichia_coli_CFT073
yjgX	hypothetical_protein	Escherichia_coli_536
fimH	FimH_protein_precursor	Escherichia_coli_CFT073
orf50	hypothetical_protein	Escherichia_coli_536
fepG	Ferric_enterobactin_transport_system_permease_protein_fepG	Escherichia_coli_CFT073
entB	Isochorismatase	Escherichia_coli_CFT073
gspL	putative_general_secretion_pathway_for_protein_export	*Shigella* _dysenteriae_Sd197
rpoS	sigma_S_(sigma_38)_factor_of_RNA_polymerase,_major_sigmafactor_during_stationary_phase	*Salmonella*_enterica_(serovar_typhimurium)_LT2
yi22g5	orf,_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
fecE	ATP-binding_protein_FecE_	*Shigella* _flexneri_(serotype_2a)
fepC	Ferric_enterobactin_transport_ATP-binding_protein_fepC	Escherichia_coli_CFT073
c3559	Hypothetical_protein	Escherichia_coli_CFT073
gspC	putative_general_secretion_protein_GspC	*Shigella* _dysenteriae_Sd197
fimC	Chaperone_protein_fimC_precursor	Escherichia_coli_CFT073
fecE	ATP-binding_protein_FecE	*Shigella* _flexneri_(serotype_2a)
fepC	Ferric_enterobactin_transport_ATP-binding_protein_fepC	Escherichia_coli_CFT073
c3559	Hypothetical_protein	Escherichia_coli_CFT073
gspC	putative_general_secretion_protein_GspC	*Shigella* _dysenteriae_Sd197
fimC	Chaperone_protein_fimC_precursor	Escherichia_coli_CFT073
ecpE	hypothetical_protein	Escherichia_coli_O157:H7
entD	4'-phosphopantetheinyl_transferase_entD	Escherichia_coli_CFT073
entA	2,3-dihydro-2,3-dihydroxybenzoate_dehydrogenase	Escherichia_coli_CFT073
ecpB	hypothetical_protein	Escherichia_coli_O157:H7
ecpA	hypothetical_protein	Escherichia_coli_O157:H7
fimE	Type_1_fimbriae_Regulatory_protein_fimE	Escherichia_coli_CFT073

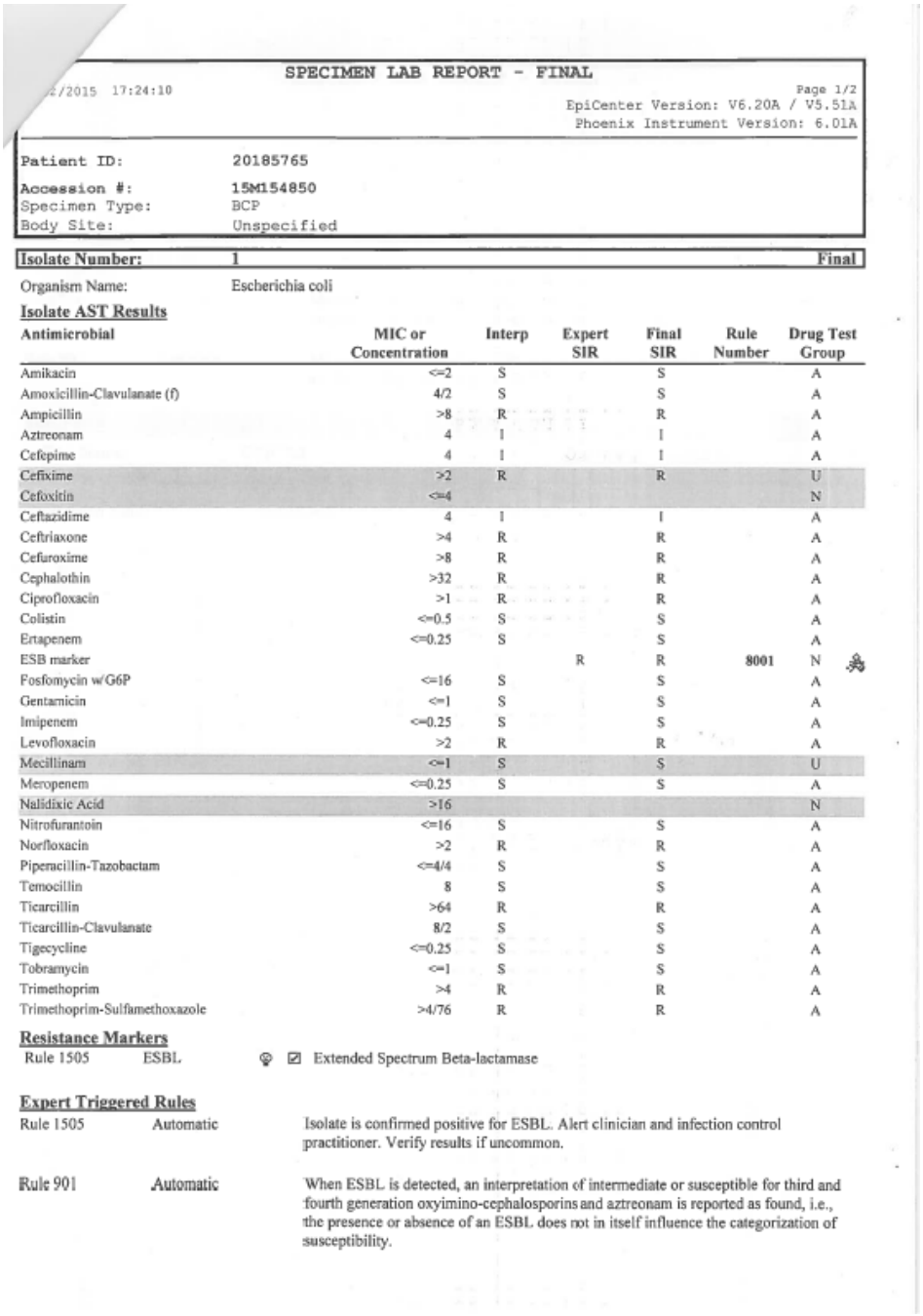
fimB	Type_1_fimbriae_Regulatory_protein_fimB	Escherichia_coli_CFT073
ecpR	putative_regulator	Escherichia_coli_O157:H7
fecI	FecI	*Shigella* _flexneri_(serotype_2a)
fimI	Fimbrin-like_protein_fimI_precursor	Escherichia_coli_CFT073
orf46	unknown	*Shigella* _flexneri_(serotype_2a)
fimF	FimF_protein_precursor	Escherichia_coli_CFT073
fimG	FimG_protein_precursor	Escherichia_coli_CFT073
tnpH	orf,_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
fimA	Type-1_fimbrial_protein,_A_chain_precursor	Escherichia_coli_CFT073
z1217	Z1217_protein	Escherichia_coli_536
SF2976	orf,_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
orf83	hypothetical_protein	Escherichia_coli_536
orf50	unknown	*Shigella* _flexneri_(serotype_2a)
yi21	g2_-_orf,_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
gspM	putative_secretion_pathway_protein	*Shigella* _dysenteriae_Sd197
yi21	g2_-_orf,_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
insG	transposase_InsG_of_insertion_element_IS4	Escherichia_coli_536
orf25	unknown	*Shigella* _flexneri_(serotype_2a)
orf58	hypothetical_protein	Escherichia_coli_536
yeeU	hypothetical_protein	Escherichia_coli_536
z5091	Z5091_protein	Escherichia_coli_536
Z5091	unknown_protein_encoded_within_prophage_CP-933L	Escherichia_coli_O157:H7
c3576	Unknown_in_IS	Escherichia_coli_CFT07
l7045	L7045	Escherichia_coli_536
SF2977	orf,_partial_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
yeeT	YeeT_protein	Escherichia_coli_536
SF3713	orf,_partial_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)

c3614	Hypothetical_protein	Escherichia_coli_CFT073
z5092	Z5092_protein	Escherichia_coli_536
orf88	hypothetical_protein	Escherichia_coli_536
SF3720	orf,_partial_conserved_hypothetical_protein	*Shigella* _flexneri_(serotype_2a)
Only found in MDA and control		
c3611	Transposase_insD_for_insertion_element_IS2A/D/F/H/I/K	Escherichia_coli_CFT073
c3612	Transposase_insC_for_insertion_element_IS2A/D/F/H/I/K	Escherichia_coli_CFT073
pB171ORF51	ORF51_protein_of_pB171	Escherichia_coli_536

Table 7-8 all virulence factors identified by VFDB in *E. coli* K12

Elements only in de_novo :	Elements only in reference :
Phosphoribosylformylglycinamidine synthase 1	putative amino acid permease YhdG
Type II secretion system protein E	Peptidoglycan DL-endopeptidase CwlO precursor
Glycine betaine/carnitine transport binding	Phosphoribosylformylglycinamidine synthase
Phosphoadenosine phosphosulfate reductase family	tRNA-Trp
RNA-binding protein	tRNA-His
Iron-uptake system permease protein FeuC	Tyrosine--tRNA ligase
Phosphoglycerate transporter protein	Putative type II secretion system protein E
Heparin-sulfate lyase precursor	Type II secretion system protein G precursor
PTS system N-acetylgalactosamine-specific	Deoxyguanosine kinase
Sorbose-specific phosphotransferase enzyme IIB	Collagen adhesin precursor
Bacterial SH3 domain protein	Cna protein B-type domain protein
glutathionine S-transferase	Bone sialoprotein-binding protein precursor
Autoinducer 2 import system permease protein	Glycine betaine/carnitine transport permease
Cardiolipin synthase	chromosome segregation protein
NADH oxidase	Oligopeptide transport ATP-binding protein OppF
SkfA peptide export ATP-binding protein SkfE	RNA-binding protein YhbY
Melibiose carrier protein	Regulatory protein MgsR
DNA polymerase III subunits gamma and tau	Lactate 2-monooxygenase
GDSL-like Lipase/Acylhydrolase	putative oxidoreductase YdgJ
Type IV leader peptidase family protein	Glycerol-3-phosphate transporter
Arabinose operon regulatory protein	Putative metallo-hydrolase YycJ
CRISPR-associated protein Cas4/endonuclease Cas1	Ribonuclease 3
HTH-type transcriptional regulator CynR	Xylose isomerase-like TIM barrel
3-oxoadipate CoA-transferase subunit A	Putative bacilysin exporter BacE
Sulfate/thiosulfate import ATP-binding protein	Heparinase II/III-like protein
Alpha/beta hydrolase family protein	Fructose-specific phosphotransferase enzyme IIB
M protein, serotype 2 precursor	DNA-invertase hin
Arginine/agmatine antiporter	Methyltransferase domain protein
Multidrug resistance protein MdtH	Modification methylase HindIII
	D12 class N6 adenine-specific DNA
	Phage terminase, small subunit
	L-serine dehydratase, beta chain

Table 7-9 list of genes present in either the de novo or reference assembly of S.Pyogenes isolated from horse blood



SPECIMEN LAB REPORT - FINAL

/2015 17:24:10 Page 1/2

EpiCenter Version: V6.20A / V5.51A
Phoenix Instrument Version: 6.01A

Patient ID: 20185765

Accession #: 15M154850
Specimen Type: BCP
Body Site: Unspecified

Isolate Number: 1 **Final**

Organism Name: Escherichia coli

Isolate AST Results

Antimicrobial	MIC or Concentration	Interp	Expert SIR	Final SIR	Rule Number	Drug Test Group
Amikacin	<=2	S		S		A
Amoxicillin-Clavulanate (f)	4/2	S		S		A
Ampicillin	>8	R		R		A
Aztreonam	4	I		I		A
Cefepime	4	I		I		A
Cefixime	>2	R		R		U
Cefoxitin	<=4					N
Ceftazidime	4	I		I		A
Ceftriaxone	>4	R		R		A
Cefuroxime	>8	R		R		A
Cephalothin	>32	R		R		A
Ciprofloxacin	>1	R		R		A
Colistin	<=0.5	S		S		A
Ertapenem	<=0.25	S		S		A
ESB marker			R	R	8001	N
Fosfomycin w/G6P	<=16	S		S		A
Gentamicin	<=1	S		S		A
Imipenem	<=0.25	S		S		A
Levofloxacin	>2	R		R		A
Mecillinam	<=1	S		S		U
Meropenem	<=0.25	S		S		A
Nalidixic Acid	>16					N
Nitrofurantoin	<=16	S		S		A
Norfloxacin	>2	R		R		A
Piperacillin-Tazobactam	<=4/4	S		S		A
Temocillin	8	S		S		A
Ticarcillin	>64	R		R		A
Ticarcillin-Clavulanate	8/2	S		S		A
Tigecycline	<=0.25	S		S		A
Tobramycin	<=1	S		S		A
Trimethoprim	>4	R		R		A
Trimethoprim-Sulfamethoxazole	>4/76	R		R		A

Resistance Markers

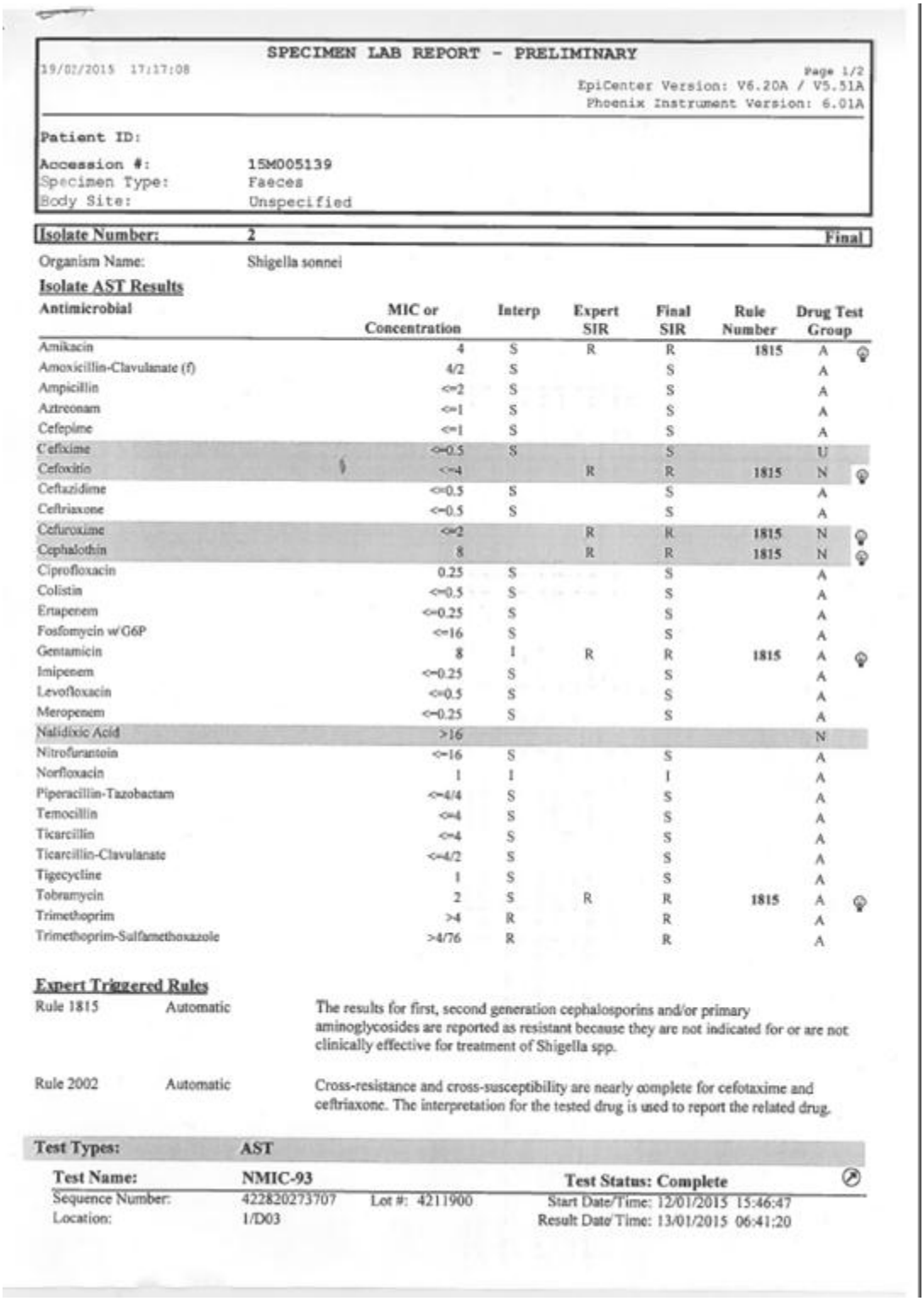
Rule 1505 ESBL ☑ Extended Spectrum Beta-lactamase

Expert Triggered Rules

Rule 1505 Automatic — Isolate is confirmed positive for ESBL. Alert clinician and infection control practitioner. Verify results if uncommon.

Rule 901 Automatic — When ESBL is detected, an interpretation of intermediate or susceptible for third and fourth generation oxyimino-cephalosporins and aztreonam is reported as found, i.e., the presence or absence of an ESBL does not in itself influence the categorization of susceptibility.

Figure 7-1 phenotypic results of clinical *E. coli*



SPECIMEN LAB REPORT - PRELIMINARY

19/02/2015 17:17:08 Page 1/2

EpiCenter Version: V6.20A / V5.51A
Phoenix Instrument Version: 6.01A

Patient ID:

Accession #: 15M005139
Specimen Type: Faeces
Body Site: Unspecified

Isolate Number: 2 Final

Organism Name: Shigella sonnei

Isolate AST Results

Antimicrobial	MIC or Concentration	Interp	Expert SIR	Final SIR	Rule Number	Drug Test Group
Amikacin	4	S	R	R	1815	A
Amoxicillin-Clavulanate (f)	4/2	S		S		A
Ampicillin	<=2	S		S		A
Aztreonam	<=1	S		S		A
Cefepime	<=1	S		S		A
Cefixime	<=0.5	S		S		U
Cefoxitin	<=4		R	R	1815	N
Ceftazidime	<=0.5	S		S		A
Ceftriaxone	<=0.5	S		S		A
Cefuroxime	<=2		R	R	1815	N
Cephalothin	8		R	R	1815	N
Ciprofloxacin	0.25	S		S		A
Colistin	<=0.5	S		S		A
Ertapenem	<=0.25	S		S		A
Fosfomycin w/G6P	<=16	S		S		A
Gentamicin	8	I	R	R	1815	A
Imipenem	<=0.25	S		S		A
Levofloxacin	<=0.5	S		S		A
Meropenem	<=0.25	S		S		A
Nalidixic Acid	>16					N
Nitrofurantoin	<=16	S		S		A
Norfloxacin	1	I		I		A
Piperacillin-Tazobactam	<=4/4	S		S		A
Temocillin	<=4	S		S		A
Ticarcillin	<=4	S		S		A
Ticarcillin-Clavulanate	<=4/2	S		S		A
Tigecycline	1	S		S		A
Tobramycin	2	S	R	R	1815	A
Trimethoprim	>4	R		R		A
Trimethoprim-Sulfamethoxazole	>4/76	R		R		A

Expert Triggered Rules

Rule 1815 Automatic — The results for first, second generation cephalosporins and/or primary aminoglycosides are reported as resistant because they are not indicated for or are not clinically effective for treatment of Shigella spp.

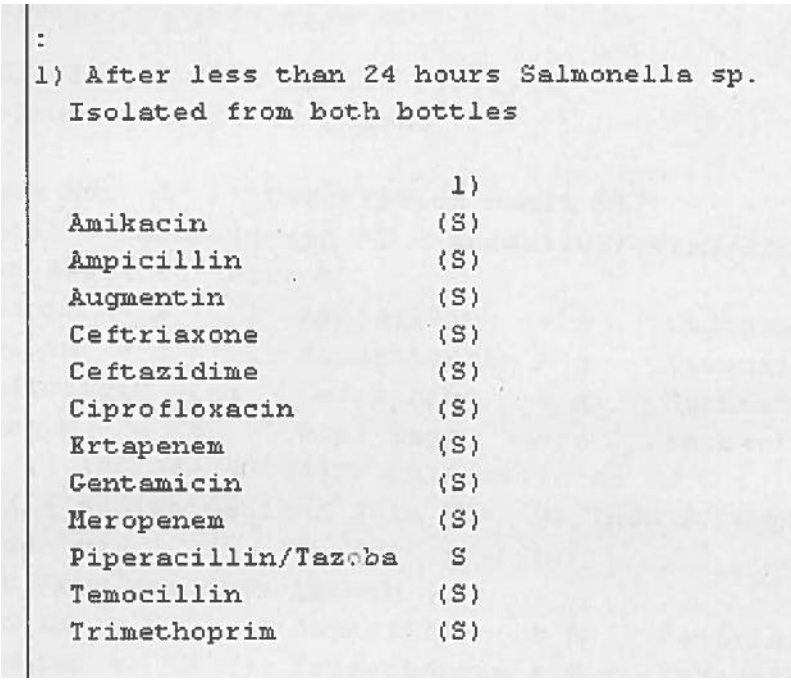
Rule 2002 Automatic — Cross-resistance and cross-susceptibility are nearly complete for cefotaxime and ceftriaxone. The interpretation for the tested drug is used to report the related drug.

Test Types: AST

Test Name: NMIC-93 Test Status: Complete

Sequence Number: 422820273707 Lot #: 4211900 Start Date/Time: 12/01/2015 15:46:47
Location: 1/D03 Result Date/Time: 13/01/2015 06:41:20

Figure 7-2 results of phenotypic resistance testing for clinical *Shigella sonnei*

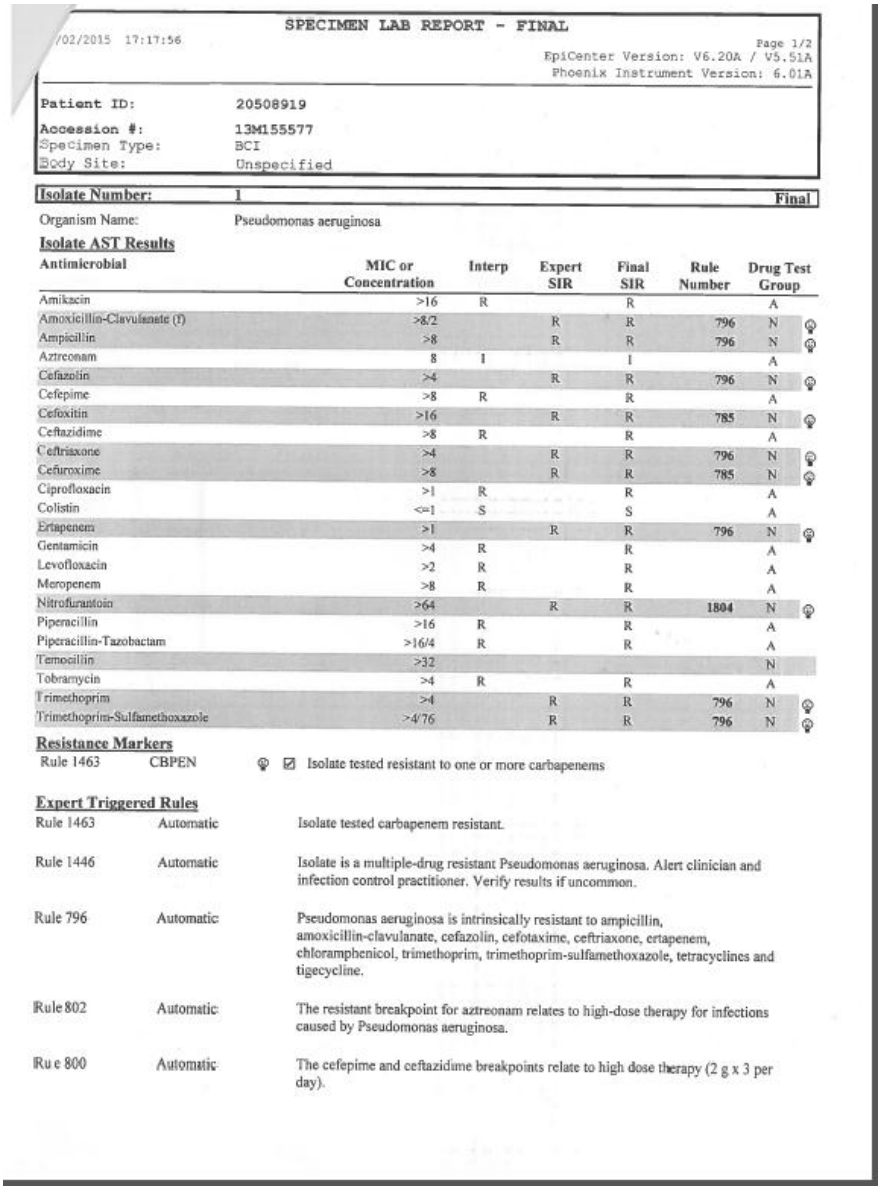


:

1) After less than 24 hours Salmonella sp.
Isolated from both bottles

	1)
Amikacin	(S)
Ampicillin	(S)
Augmentin	(S)
Ceftriaxone	(S)
Ceftazidime	(S)
Ciprofloxacin	(S)
Ertapenem	(S)
Gentamicin	(S)
Meropenem	(S)
Piperacillin/Tazoba	S
Temocillin	(S)
Trimethoprim	(S)

Figure 7-3 Phenotypic resistance testing for clinical *Salmonella sp*



SPECIMEN LAB REPORT - FINAL

/02/2015 17:17:56 — Page 1/2

EpiCenter Version: V6.20A / V5.51A
Phoenix Instrument Version: 6.01A

Patient ID: 20508919
Accession #: 13M155577
Specimen Type: BCI
Body Site: Unspecified

Isolate Number: 1 — Final

Organism Name: Pseudomonas aeruginosa

Isolate AST Results

Antimicrobial	MIC or Concentration	Interp	Expert SIR	Final SIR	Rule Number	Drug Test Group
Amikacin	>16	R		R		A
Amoxicillin-Clavulanate (I)	>8/2		R	R	796	N
Ampicillin	>8		R	R	796	N
Aztreonam	8	I		I		A
Cefazolin	>4		R	R	796	N
Cefepime	>8	R		R		A
Cefoxitin	>16		R	R	785	N
Ceftazidime	>8	R		R		A
Ceftriaxone	>4		R	R	796	N
Cefuroxime	>8		R	R	785	N
Ciprofloxacin	>1	R		R		A
Colistin	<=1	S		S		A
Ertapenem	>1		R	R	796	N
Gentamicin	>4	R		R		A
Levofloxacin	>2	R		R		A
Meropenem	>8	R		R		A
Nitrofurantoin	>64		R	R	1804	N
Piperacillin	>16	R		R		A
Piperacillin-Tazobactam	>16/4	R		R		A
Temocillin	>32					N
Tobramycin	>4	R		R		A
Trimethoprim	>4		R	R	796	N
Trimethoprim-Sulfamethoxazole	>4/76		R	R	796	N

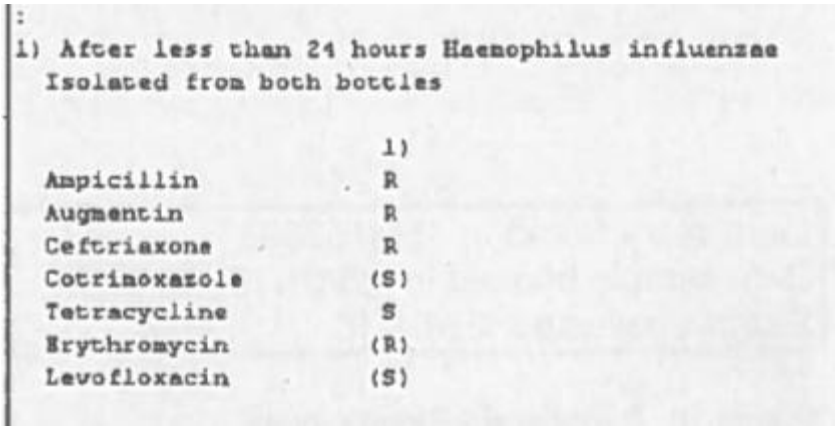
Resistance Markers

Rule 1463 CBPEN — ☑ Isolate tested resistant to one or more carbapenems

Expert Triggered Rules

Rule 1463	Automatic	Isolate tested carbapenem resistant.
Rule 1446	Automatic	Isolate is a multiple-drug resistant Pseudomonas aeruginosa. Alert clinician and infection control practitioner. Verify results if uncommon.
Rule 796	Automatic	Pseudomonas aeruginosa is intrinsically resistant to ampicillin, amoxicillin-clavulanate, cefazolin, cefotaxime, ceftriaxone, ertapenem, chloramphenicol, trimethoprim, trimethoprim-sulfamethoxazole, tetracyclines and tigecycline.
Rule 802	Automatic	The resistant breakpoint for aztreonam relates to high-dose therapy for infections caused by Pseudomonas aeruginosa.
Rule 800	Automatic	The cefepime and ceftazidime breakpoints relate to high dose therapy (2 g x 3 per day).

Figure 7-4 phenotypic results of resistance testing of *P.aeruginosa*



1) After less than 24 hours Haemophilus influenzae
Isolated from both bottles

	1)
Ampicillin	R
Augmentin	R
Ceftriaxone	R
Cotrimoxazole	(S)
Tetracycline	S
Erythromycin	(R)
Levofloxacin	(S)

Figure 7-5 phenotypic resistance of *H.influenzae*

Genus/Species	% reads identified using LCA analysis		
	Sample 2	Sample 3	Sample 4
Mobiluncus curtisii		12.10%	
Corynebacterium sp		4.10%	
Corynebacterium aurimucosum		0.47%	
Corynebacterium riegelii		0.48%	
Corynebacterium sp ATCC 6931		0.34%	
Corynebacterium ureicelerivorans		0.23%	
Propionibacterium sp		0.20%	
Bifidobacteriaceae adolescentis		0.52%	
Gardnerlla vaginalis		0.61%	89.32%
Atopobium vaginae			0.11%
Eggerthella lenta		0.26%	
Bacteroides sp	7.83%	0.36%	
Bacteroides uniformis	5.35%		
Bacteroides dorei	2.32%		
Bacteroides xymanisolvens	1.08%		
Odoribacter splanchnicus	1.54%		
Prevotella melsninogenica	1.79%		
Chlamydia trachomatis	3.29%		
Staphylococcus epidermidis		4.44%	
Aerococcus sp		0.18%	
Enterococcus sp		0.80%	
Lactobacillus	0.28%		0.82%
Streptococcus sp		10.85%	
Streptococcus agalactiae		0.60%	
Streptococcus anginosa		1.04%	
Streptococcus constellatus		0.52%	
Streptococcus dysgalactiae		0.52%	
Runinococcus torques	8.88%	0.00%	
Rosburia	1.37%	0.00%	
Peptoniphilus sp		31.85%	
Anaerococcus prevotii		4.71%	
Finegoldia magna		12.16%	
Parvimonas micra		0.46%	
Peptoniphilus asaccharolyticus		0.35%	
Filifactor alocis		1.89%	
Clostridiales genomosp	5.23%	0.86%	
Fusobacterium nucleatum		0.22%	
Sneathia sanguinegens	61.03%	0.20%	
Neisseria gonorrhoeae		3.54%	9.08%
Mycoplasma genitalium		3.83%	
Trichomonas vaginalis		1.31%	
Alpha papillomavirus 8			0.67%

Table 7-10 all species representing > 0.2% of reads in the three vaginal STI samples

8. Figures and Tables

8.1 Figures

8.2 Tables

8.3 Commands

9. References

1. Glaser, C. A. *et al.* In search of encephalitis etiologies: diagnostic challenges in the California Encephalitis Project, 1998-2000. *Clin. Infect. Dis.* **36,** 731–742 (2003).

2. Regev-yochay, G. *et al.* Nasopharyngeal Carriage of Streptococcus pneumoniae by Adults and Children in Community and Family Settings. *Clin. Infect. Dis.* **38,** (2004).

3. Riedel, S. & Carroll, K. C. Blood cultures: key elements for best practices and future directions. *J. Infect. Chemother.* **16,** 301–16 (2010).

4. Drancourt, M. Detection of microorganisms in blood specimens using matrix-assisted laser desorption ionization time-of-flight mass spectrometry: A review. *Clin. Microbiol. Infect.* **16,** 1620–1625 (2010).

5. Das, M., Badley, A. D., Cockerill, F. R., Steckelberg, J. M. & Wilson, W. R. Infective endocarditis caused by HACEK microorganisms. *Annu. Rev. Med.* **48,** 25–33 (1997).

6. Stoesser, N. *et al.* Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. *J. Antimicrob. Chemother.* **68,** 2234–44 (2013).

7. Stevenson, L. G., Drake, S. K. & Murray, P. R. Rapid identification of bacteria in positive blood culture broths by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J. Clin. Microbiol.* **48,** 444–7 (2010).

8. Leland, D. S. & Ginocchio, C. C. Role of cell culture for virus detection in the age of technology. *Clin. Microbiol. Rev.* **20,** 49–78 (2007).

9. Lawn, S. D. *et al.* Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-care test. *Lancet Infect. Dis.* **13,** 349–61 (2013).

10. Lessells, R. J. *et al.* Impact of a novel molecular TB diagnostic system in patients at high risk of TB mortality in rural South Africa (Uchwepheshe): study protocol for a cluster randomised trial. *Trials* **14,** 170 (2013).

11. Kulkarni, S. *et al.* An in-house multiplex PCR test for the detection of Mycobacterium tuberculosis, its validation & comparison with a single target TB-PCR kit. *Indian J. Med. Res.* **135,** 788–94 (2012).

12. Flores, L. L., Pai, M., Colford, J. M. & Riley, L. W. In-house nucleic acid amplification tests for the detection of Mycobacterium tuberculosis in sputum specimens: meta-analysis and meta-regression. *BMC Microbiol.* **5,** 55 (2005).

13. Causse, M., Ruiz, P., Gutiérrez-Aroca, J. B. & Casal, M. Comparison of two molecular methods for rapid diagnosis of extrapulmonary tuberculosis. *J. Clin. Microbiol.* **49,** 3065–7 (2011).

14. Lawn, S. D. & Nicol, M. P. Xpert® MTB/RIF assay: development, evaluation and implementation of a new rapid molecular diagnostic for tuberculosis and rifampicin resistance. *Future Microbiol.* **6,** 1067–82 (2011).

15. Weyer, K. *et al.* Rapid molecular TB diagnosis: evidence, policy making and global implementation of Xpert MTB/RIF. *Eur. Respir. J.* **42,** 252–71 (2013).

16. Wu, H. M. *et al.* Accuracy of real-time PCR, Gram stain and culture for Streptococcus pneumoniae, Neisseria meningitidis and Haemophilus influenzae meningitis diagnosis. *BMC Infect. Dis.* **13,** 26 (2013).

17. Nolte, F. S. *et al.* Evaluation of a rapid and completely automated real-time reverse transcriptase PCR assay for diagnosis of enteroviral meningitis. *J. Clin. Microbiol.* **49,** 528–33 (2011).

18. Archimbaud, C. *et al.* Improvement of the management of infants, children and adults with a molecular diagnosis of Enterovirus meningitis during two observational study periods. *PLoS One* **8,** e68571 (2013).

19. Huizing, K. M. N., Swanink, C. M. a, Landstra, A. M., van Zwet, A. a & van Setten, P. a. Rapid enterovirus molecular testing in cerebrospinal fluid reduces length of hospitalization and duration of antibiotic therapy in children with aseptic meningitis. *Pediatr. Infect. Dis. J.* **30,** 1107–9 (2011).

20. Dupuis, M. *et al.* Molecular detection of viral causes of encephalitis and meningitis in New York State. *J. Med. Virol.* **83,** 2172–2181 (2011).

21. Chang, H. *et al.* Comparison of real-time polymerase chain reaction and serological tests for the confirmation of Mycoplasma pneumoniae infection in children with clinical diagnosis of atypical pneumonia. *J. Microbiol. Immunol. Infect.* **47,** 137–144 (2014).

22. Enoch, D. A. *et al.* Epidemiology of extended-spectrum beta-lactamase- producing Enterobacteriaceae in a UK district hospital; an observational study. *J. Hosp. Infect.* **81,** 270–277 (2012).

23. Dallenne, C., Da Costa, A., Decré, D., Favier, C. & Arlet, G. Development of a set of multiplex PCR assays for the detection of genes encoding important beta-lactamases in Enterobacteriaceae. *J. Antimicrob. Chemother.* **65,** 490–5 (2010).

24. Willemsen, I., Hille, L., Vrolijk, A., Bergmans, A. & Kluytmans, J. Evaluation of a commercial real-time PCR for the detection of extended spectrum β-lactamase genes. *J. Med. Microbiol.* **63,** 540–3 (2014).

25. Vandamme, A. *et al.* European recommendations for the clinical use of HIV drug resistance testing: 2011 update. *Aids Rev.* **13,** 77–108 (2004).

26. Mansky, L. M. Retrovirus mutation rates and their role in genetic variation. *J. Gen. Virol.* **79 (Pt 6),** 1337–45 (1998).

27. Shafer, R. W. Genotypic testing for human immunodeficiency virus type 1 drug resistance. *Clin. Microbiol. Rev*. **15,** 247–277 (2002).

28. Mohamed, S. *et al.* Comparison of ultra-deep versus Sanger sequencing detection of minority mutations on the HIV-1 drug resistance interpretations after virological failure. *AIDS* **28,** 1315–24 (2014).

29. Johnson, V. a *et al.* Update of the drug resistance mutations in HIV-1: March 2013. *Top. Antivir. Med.* **21,** 6–14 (2013).

30. Sabat, A. J. *et al.* Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill. Bull. Eur. sur les Mal. Transm. = Eur. Commun. Dis. Bull.* **18,** 20380 (2013).

31. Arbeit, R. D. *et al.* Resolution of recent evolutionary divergence among escherichia coli from related lineages: The application of pulsed field electrophoresis to molecular epidemiology. *J. Infect. Dis.* **161,** 230–235 (1990).

32. Gordillo, M. E., Singh, K. V., Baker, C. J. & Murray, B. E. Typing of group B streptococci: Comparison of pulsed-field gel electrophoresis and conventional electrophoresis. *J. Clin. Microbiol.* **31,** 1430–1434 (1993).

33. Tosh, P. K. *et al.* Outbreak of Pseudomonas aeruginosa surgical site infections after arthroscopic procedures: Texas, 2009. *Infect. Control Hosp. Epidemiol.* **32,** 1179–86 (2011).

34. Swaminathan, B., Barrett, T. J., Hunter, S. B. & Tauxe, R. V. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg. Infect. Dis.* **7,** 382–389 (2001).

35. Overdevest, I. T. M. a *et al.* Evaluation of the DiversiLab typing method in a multicenter study assessing horizontal spread of highly resistant gram-negative rods. *J. Clin. Microbiol.* **49,** 3551–4 (2011).

36. Grundmann, H. *et al.* Geographic distribution of Staphylococcus aureus causing invasive infections in Europe: a molecular-epidemiological analysis. *PLoS Med.* **7,** e1000215 (2010).

37. Maiden, M. C. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 3140–5 (1998).

38. Miller, M. B. & Tang, Y. Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology. *Clin. Microbiol. Rev*. **22,** 611–633 (2009).

39. McCarthy, A. J. & Lindsay, J. A. The distribution of plasmids that carry virulence and resistance genes in Staphylococcus aureus is lineage associated. *BMC Microbiol.* **12,** 104 (2012).

40. Jackson, S. a *et al.* Rapid genomic-scale analysis of Escherichia coli O104:H4 by using high-resolution alternative methods to next-generation sequencing. *Appl. Environ. Microbiol.* **78,** 1601–5 (2012).

41. Sontakke, S., Cadenas, M. B., Maggi, R. G., Diniz, P. P. V. P. & Breitschwerdt, E. B. Use of broad range16S rDNA PCR in clinical microbiology. *J. Microbiol. Methods* **76,** 217–25 (2009).

42. Rampini, S. K. *et al.* Broad-range 16S rRNA gene polymerase chain reaction for diagnosis of culture-negative bacterial infections. *Clin. Infect. Dis.* **53,** 1245–51 (2011).

43. Mahlen, S. D. & Clarridge, J. E. Evaluation of a selection strategy before use of 16S rRNA gene sequencing for the identification of clinically significant gram-negative rods and coccobacilli. *Am. J. Clin. Pathol.* **136,** 381–8 (2011).

44. Böttger, E. C. *et al.* Disseminated "Mycobacterium genavense" infection in patients with AIDS. *Lancet (London, England)* **340,** 76–80 (1992).

45. Griffen, A. L. *et al.* Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J.* **6,** 1176–85 (2012).

46. Huse, S. M., Ye, Y., Zhou, Y. & Fodor, A. a. A core human microbiome as viewed through 16S rRNA sequence clusters. *PLoS One* **7,** e34242 (2012).

47. Flanagan, J. L. *et al.* Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with Pseudomonas aeruginosa. *J. Clin. Microbiol.* **45,** 1954–62 (2007).

48. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320,** 1647–1651 (2008).

49. Mollet, C., Drancourt, M. & Raoult, D. rpoB sequence analysis as a novel basis for bacterial identification. *Mol. Microbiol.* **26,** 1005–1011 (1997).

50. Valiunas, D., Jomantiene, R. & Davis, R. E. Evaluation of the DNA-dependent RNA polymerase beta-subunit gene (rpoB) for phytoplasma classification and phylogeny. *Int. J. Syst. Evol. Microbiol.* **63,** 3904–3914 (2013).

51. Case, R. J. *et al.* Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies □. *J. Clin. Microbiol.* **73,** 278–288 (2007).

52. Kuczynski, J., Lauber, C. L., Walters, W. A. & Parfrey, L. W. Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* **13,** 47–58 (2011).

53. Lysholm, F. *et al.* Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS One* **7,** (2012).

54. Yolken, R. H. *et al.* Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. *Proc. Natl. Acad. Sci.* **111,** 16106–16111 (2014).

55. Sanger, F., Nicklen, S. & Coulson, a R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–7 (1977).

56. Fleischmann, R. D. *et al.* Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae. *Science (80-.).* **269,** 496–512 (1997).

57. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46,** 305–309 (2014).

58. Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. & Seelow, D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One* **6,** e28240 (2011).

59. Loman, N. J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.* **10,** 599–606 (2012).

60. Di Bella, J. M., Bao, Y., Gloor, G. B., Burton, J. P. & Reid, G. High throughput sequencing methods and analysis for microbiome research. *J. Microbiol. Methods* **95,** 401–14 (2013).

61. Ashton, P. M. *et al.* MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* **33,** (2015).

62. English, A. C. *et al.* Mind the Gap : Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* **7,** 1–12 (2012).

63. Hendriksen, R. S. *et al.* Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* **2,** (2011).

64. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477,** 462–5 (2011).

65. Rasko, D. A. *et al.* Origins of the E. coli Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany. *N. Engl. J. Med.* **365,** 709–717 (2011).

66. Loman, N. J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *Jama* **309,** 1502–1510 (2013).

67. Holger Rohde, M.D., Junjie Qin, Ph.D., Yujun Cui, Ph.D., Dongfang Li, M.E., Nicholas J. Loman, M.B., B.S., Moritz Hentschke, M.D., Wentong Chen, B.S., Fei Pu, B.S., Yangqing Peng, B.S., Junhua Li, B.E., Feng Xi, B.E., Shenghui Li, B.S., Yin Li, B.S., Zhao, and *the E. coli O. G. A. C.-S. C. Open-Source Genomic Analysis of Shiga-Toxin–Producing. *N. Engl. J. Med.* (2011).

68. Mellmann, A. *et al.* Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* **6,** e22751 (2011).

69. Kos, V. N. *et al.* Comparative genomics of vancomycin-resistant staphylococcus aureus strains and their positions within the clade most commonly associated with methicillin-resistant s. aureus hospital-acquired infection in the United States. *MBio* **3,** (2012).

70. Daum, L. T. *et al.* Next-generation ion torrent sequencing of drug resistance mutations in Mycobacterium tuberculosis strains. *J. Clin. Microbiol.* **50,** 3831–7 (2012).

71. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67,** 2640–4 (2012).

72. Zankari, E. *et al.* Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J. Antimicrob. Chemother.* **68,** 771–7 (2013).

73. Hornsey, M. *et al.* Whole-genome comparison of two Acinetobacter baumannii isolates from a single patient, where resistance developed during tigecycline therapy. *J. Antimicrob. Chemother.* **66,** 1499–503 (2011).

74. Feng, J. *et al.* Genome sequencing of linezolid-resistant Streptococcus pneumoniae mutants reveals novel mechanisms of resistance. *Genome Res.* **19,** 1214–1223 (2009).

75. Howard, C. R. & Fletcher, N. F. Emerging virus diseases: can we ever expect the unexpected? *Emerg. Microbes Infect.* **1,** e46 (2012).

76. Levesque-Sergerie, J.-P., Duquette, M., Thibault, C., Delbecchi, L. & Bissonnette, N. Detection limits of several commercial reverse transcriptase enzymes: impact on the low- and high-abundance transcript levels assessed by quantitative RT-PCR. *BMC Mol. Biol.* **8,** 93 (2007).

77. Schoenfeld, T. *et al.* Assembly of viral metagenomes from yellowstone hot springs. *Appl. Environ. Microbiol.* **74,** 4164–74 (2008).

78. Moser, M. J. *et al.* Thermostable DNA polymerase from a viral metagenome is a potent RT-PCR enzyme. *PLoS One* **7,** e38371 (2012).

79. Garcia-Diaz, M. & Bebenek, K. Multiple functions of DNA polymerases. *CRC. Crit. Rev. Plant Sci.* **26,** 105–122 (2007).

80. Weiner, J. H., Bertsch, L. L. & Kornberg, A. The deoxyribonucleic acid unwinding protein of Escherichia coli. Properties and functions in replication. *J. Biol. Chem.* **250,** 1972–1980 (1975).

81. Garg, P. & Burgers, P. M. J. DNA polymerases that propagate the eukaryotic DNA replication fork. *Crit. Rev. Biochem. Mol. Biol.* **40,** 115–28 (2005).

82. Ramadan, K., Shevelev, I. & Hübscher, U. The DNA-polymerase-X family: controllers of DNA quality? *Nat. Rev. Mol. Cell Biol.* **5,** 1038–1043 (2004).

83. Ling, H., Boudsocq, F., Woodgate, R. & Yang, W. Crystal structure of a Y-family DNA polymerase in action: A mechanism for error-prone and lesion-bypass replication. *Cell* **107,** 91–102 (2001).

84. Lasken, R. S. Single-cell genomic sequencing using Multiple Displacement Amplification. *Curr. Opin. Microbiol.* **10,** 510–6 (2007).

85. Garmendia, C., Bernad, a, Esteban, J. a, Blanco, L. & Salas, M. The bacteriophage phi 29 DNA polymerase, a proofreading enzyme. *J. Biol. Chem.* **267,** 2594–9 (1992).

86. Blanco, L. *et al.* Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **264,** 8935–8940 (1989).

87. Hutchison, C. A., Smith, H. O., Pfannkoch, C. & Venter, J. C. Cell-free cloning using phi29 DNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 17332–6 (2005).

88. Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11,** 1095–9 (2001).

89. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71,** 3342–3347 (2005).

90. Aviel-Ronen, S. *et al.* Large fragment Bst DNA polymerase for whole genome amplification of DNA from formalin-fixed paraffin-embedded tissues. *BMC Genomics* **7,** 312 (2006).

91. Tanner, N.Evans, T. C. A. Loop-Mediated Isothermal Amplification for Detection of Nucleic Acids. *Curr. Protoc. Mol. Biol.* **105,** (2014).

92. Notomi, T. *et al.* Loop-mediated isothermal amplification of DNA. *Nucleic acids Res.* **28,** (2000).

93. Wilson, J. *et al.* Trends among pathogens reported as causing bacteraemia in England, 2004-2008. *Clin. Microbiol. Infect.* **17,** 451–8 (2011).

94. Folks T M, D, P., M, L. & Koenig S, Fauci A S, Benn S, Rabson A, Daughert D , Gendelman H E, and H. M. D. Biological and biochemical characterization of a cloned Leu-3- cell surviving infection with the acquired immune deficiency syndrome retrovirus. *J. Exp. Med.* **164,** 280–290 (1986).

95. Louwagie, J. *et al.* Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of African origin. *J. Virol.* **69,** 263–271 (1995).

96. Tebbe, C. C. & Vahjen, W. Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Appl. Environ. Microbiol.* **59,** 2657–2665 (1993).

97. Walker, G. T., Little, M. C., Nadeau, J. G. & Shank, D. D. Isothermal in vitro amplification of DNA by a restriction enzyme/DNA polymerase system. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 392–396 (1992).

98. Ding, F. *et al.* Single-molecule mechanical identification and sequencing. *Nat. Methods* **9,** 367–372 (2012).

99. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

100. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945 (2000).

101. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072–5 (2013).

102. Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21,** 1552–1560 (2011).

103. Salter, S. *et al.* Reagent contamination can critically impact sequence-based microbiome analyses. *bioRxiv* 007187 (2014). doi:10.1101/007187

104. Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F. M. An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS One* **8,** 1–13 (2013).

105. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–4 (2011).

106. Martin, M. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011. Date of access 05/08/2015. *Bioinforma. action* **17,** 10–12 (2011).

107. Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *arXiv* **1203.4802,** 1–18 (2012).

108. Zhang, W. *et al.* A practical comparison of De Novo genome assembly software tools for next-generation sequencing technologies. *PLoS One* **6,** (2011).

109. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95,** 315–327 (2010).

110. Warren, R. L., Sutton, G. G., Jones, S. J. M. & Holt, R. a. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23,** 500–501 (2007).

111. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–1123 (2009).

112. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).

113. Boisvert, S., Laviolette, F. & Corbeil, J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol.* **17,** 1519–1533 (2010).

114. Chevreux, B., Wetter, T. & Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information Computer Science and Biology. *Proc. Ger. Conf. Bioinforma.* **99,** 45–56 (1999).

115. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–9 (2014).

116. Bardou, P., Mariette, J., Escudié, F., Djemiel, C. & Klopp, C. jvenn : an interactive Venn diagram viewer. *BMC Bioinformatics* 1–7 (2014).

117. Chen, L. *et al.* VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33,** D325–8 (2005).

118. Yang, J., Chen, L., Sun, L., Yu, J. & Jin, Q. VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res.* **36,** D539–42 (2008).

119. Chen, L., Xiong, Z., Sun, L., Yang, J. & Jin, Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* **40,** D641–5 (2012).

120. Liu, B. & Pop, M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37,** D443–7 (2009).

121. Kohno, T. *et al.* A new improved method for the concentration of HIV-1 infective particles. *J. Virol. Methods* **106,** 167–173 (2002).

122. Ramaswamy, M., Smith, M. & Geretti, A. M. Detection and typing of herpes simplex DNA in genital swabs by real-time polymerase chain reaction. *J. Virol. Methods* **126,** 203–206 (2005).

123. Ramaswamy, M. *et al.* Diagnosis of genital herpes by real time PCR in routine clinical practice. *Sex. Transm. Infect.* **80,** 406–410 (2004).

124. Enriquez C E, G. C. P. Concentration of enteric adenovirus 40 from tap, sea and waste water. *Water Res.* **29,** 2554–2560 (1995).

125. Boisvert, M., Fernandes, S. & Tijssen, P. Multiple pathways involved in porcine parvovirus cellular entry and trafficking toward the nucleus. *J. Virol.* **84,** 7782–7792 (2010).

126. Dandri, M., Burda, M. R., Will, H. & Petersen, J. Increased hepatocyte turnover and inhibition of woodchuck hepatitis B virus replication by adefovir in vitro do not lead to reduction of the closed circular DNA. *Hepatology* **32,** 139–146 (2000).

127. Lenhoff, R. J. & Summers, J. Coordinate regulation of replication and virus assembly by the large envelope protein of an avian hepadnavirus. *J. Virol.* **288,** 9i–9 (1994).

128. Garcia, D. F. *et al.* Human metapneumovirus and respiratory syncytial virus infections in older children with cystic fibrosis. *Pediatr. Pulmonol.* **42,** 66–74 (2007).

129. Deboosere, N. *et al.* Development and validation of a concentration method for the detection of influenza a viruses from large volumes of surface water. *Appl. Environ. Microbiol.* **77,** 3802–3808 (2011).

130. Guévremont, E., Brassard, J., Houde, A., Simard, C. & Trottier, Y. L. Development of an extraction and concentration procedure and comparison of RT-PCR primer systems for the detection of hepatitis A virus and norovirus GII in green onions. *J. Virol. Methods* **134,** 130–135 (2006).

131. Lewis, G. D. & Metcalf, T. G. Polyethylene glycol precipitation for recovery of pathogenic viruses, including hepatitis A virus and human rotavirus, from oyster, water, and sediment samples. *Appl. Environ. Microbiol.* **54,** 1983–1988 (1988).

132. Huang, P. W. *et al.* Concentration and detection of caliciviruses in water samples by reverse transcription-PCR. *Appl. Environ. Microbiol.* **66,** 4383–4388 (2000).

133. Blok, J., Henchal, E. a. & Gorman, B. M. Comparison of dengue viruses and some other flaviviruses by cDNA-RNA hybridization analysis and detection of a close relationship between dengue virus serotype 2 and Edge Hill virus. *J. Gen. Virol.* **65,** 2173–2181 (1984).

134. Joneja, A. & Huang, X. Linear nicking endonuclease-mediated strand-displacement DNA amplification. *Anal. Biochem.* **414,** 58–69 (2011).

135. Spargo, C. a *et al.* Detection of M. tuberculosis DNA using thermophilic strand displacement amplification. *Mol. Cell. Probes* **10,** 247–56 (1996).

136. Sebaihia, M. *et al.* The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nat. Genet.* **38,** 779–786 (2006).

137. Pineda-Peña, A.-C. *et al.* Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol.* **19,** 337–348 (2013).

138. Danial, J. & Child, J. A. Epidemiology and costs associated with norovirus outbreaks in NHS Lothian, Scotland 2007–2009. *J. Hosp. Indection* **79,** 354–358 (2010).

139. Berthet, N. *et al.* Phi29 polymerase based random amplification of viral RNA as an alternative to random RT-PCR. *BMC Mol. Biol.* **9,** 77 (2008).

140. Kumar, S., Gangoliya, S. R., Berri, M., Rodolakis, A. & Alam, S. I. Whole genome amplification of the obligate intracellular pathogen Coxiella burnetii using multiple displacement amplification. *J. Microbiol. Methods* **95,** 368–372 (2013).

141. Kong, H., Kucera, R. B. & Jack, W. E. Characterization of a DNA polymerase from the hyperthermophile archaea Thermococcus litoralis. *J. Biol. Chem.* **268,** 1965–1975 (1993).

142. He, Y. & Jiang, T. Nickase-dependent isothermal DNA amplification. *Adv. Biosci. Biotechnol.* **2013,** 539–542 (2013).

143. Dickson, K. S., Burns, C. M. & Richardson, J. P. Determination of the free-energy change for repair of a DNA phosphodiester bond. *J. Biol. Chem.* **275,** 15828–15831 (2000).

144. Lee, M. S. *et al.* One-step reverse-transcription loop-mediated isothermal amplification for detection of infectious bursal disease virus. *Can. J. Vet. Res.* **75,** 122–127 (2011).

145. Arakaki, A., Shibusawa, M., Hosokawa, M. & Matsunaga, T. Preparation of genomic DNA from a single species of uncultured magnetotactic bacterium by multiple-displacement amplification. *Appl. Environ. Microbiol.* **76,** 1480–5 (2010).

146. Erlandsson, L., Rosenstierne, M. W., McLoughlin, K., Jaing, C. & Fomsgaard, A. The microbial detection array combined with random Phi29-Amplification used as a diagnostic tool for virus detection in clinical samples. *PLoS One* **6,** 2–11 (2011).

147. Lasken, R. S. Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10,** 631–40 (2012).

148. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.* **7,** 19 (2007).

149. Stepanauskas, R. & Sieracki, M. E. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 9052–7 (2007).

150. Rodrigue, S. *et al.* Whole genome amplification and de novo assembly of single bacterial cells. *PLoS One* **4,** e6864 (2009).

151. Kashtan, N. *et al.* Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Sci. (New York, NY)* **344,** 416–420 (2014).

152. Willner, D. *et al.* Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One* **4,** (2009).

153. Marine, R. *et al.* Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome* **2,** 3 (2014).

154. H.C.Bimboim, J. D. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* **7,** (1979).

155. Laurence, M., Hatzis, C. & Brash, D. E. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One* **9,** 1–8 (2014).

156. Lo, S. C. *et al.* Isolation of Novel Afipia septicemium and Identification of Previously Unknown Bacteria Bradyrhizobium sp. OHSU_III from Blood of Patients with Poorly Defined Illnesses. *PLoS One* **8,** (2013).

157. Tariq, M. a. *et al.* A metagenomic approach to characterize temperate bacteriophage populations from Cystic Fibrosis and non-Cystic Fibrosis bronchiectasis patients. *Front. Microbiol.* **6,** 1–12 (2015).

158. Khudyakov, J. I., Preeyanon, L., Champagne, C. D., Ortiz, R. M. & Crocker, D. E. Transcriptome analysis of northern elephant seal (Mirounga angustirostris) muscle tissue provides a novel molecular resource and physiological insights. *BMC Genomics* **16,** 64 (2015).

159. Reznicek, O., Luesken, F., Facey, S. J. & Hauer, B. Draft Genome Sequence of Phenylobacterium immobile Strain E (DSM 1986), Isolated from Uncontaminated Soil in Ecuador. *Genome Announc.* **3,** 10–11 (2015).

160. Blattner, F. *et al.* The Complete Genome Sequence of Escherichia coli K-12. *Science (80-.).* **2771613,** 1453–1462 (1997).

161. Dige, I., Raarup, M. K., Nyengaard, J. R., Kilian, M. & Nyvad, B. Actinomyces naeslundii in initial dental biofilm formation. *Microbiology* **155,** 2116–2126 (2009).

162. Richardson, E. J. & Watson, M. The automatic annotation of bacterial genomes. *Brief. Bioinform.* **14,** 1–12 (2013).

163. Zuo, G., Xu, Z. & Hao, B. Shigella Strains Are Not Clones of Escherichia coli but Sister Species in the Genus Escherichia. *Genomics, Proteomics Bioinforma.* **11,** 61–65 (2013).

164. Dobson, S. R. M. & Edwards, M. S. Extensive Actinomyces naeslundii infection in a child. *J. Clin. Microbiol.* **25,** 1327–1329 (1987).

165. Heider, L. C. *et al.* Genetic and phenotypic characterization of the bla(CMY) gene from Escherichia coli and Salmonella enterica isolated from food-producing animals, humans, the environment, and retail meat. *Foodborne Pathog. Dis.* **6,** 1235–1240 (2009).

166. Hansen, L. H., Johannesen, E., Burm??lle, M., S??rensen, A. H. & S??rensen, S. J. Plasmid-encoded multidrug efflux pump conferring resistance to olaquindox in Escherichia coli. *Antimicrob. Agents Chemother.* **48,** 3332–3337 (2004).

167. Bradley, P. *et al.* Rapid antibiotic resistance predictions from genome sequence data for S. aureus and M. tuberculosis. *bioRxiv* 018564 (2015). doi:10.1101/018564

168. Consortium, T. U. UniProt : a hub for protein information. *Nucleic Acids Res.* **43,** 204–212 (2015).

169. Larsen, M. V. *et al.* Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria. *J. Clin. Microbiol.* **50,** 1355–1361 (2012).

170. Zelenin, S. *et al.* Microfluidic-based isolation of bacteria from whole blood for sepsis diagnostics. *Biotechnol. Lett.* **37,** 825–830 (2015).

171. Alberts B, Johnson A, L. J. *Molecular Biology of the Cell. 4th edition.* Table 22–1 (2002).

172. Monod, I. J. & Jussieu, P. Origin and fate of repeats in bacteria. *Nucleic Acids Res.* **30,** 2987–2994 (2002).

173. Tanaka, M., Wang, T., Onodera, Y., Uchida, Y. & Sato, K. Mechanism of quinolone resistance in Staphylococcus aureus. *J. Infect. Chemother.* **6,** 131–139 (2000).

174. Tofteland, S., Naseer, U., Lislevand, J. H., Sundsfjord, A. & Samuelsen, ??rjan. A Long-Term Low-Frequency Hospital Outbreak of KPC-Producing Klebsiella pneumoniae Involving Intergenus Plasmid Diffusion and a Persisting Environmental Reservoir. *PLoS One* **8,** 1–8 (2013).

175. Loman, N. J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.* **10,** 599–606 (2012).

176. Strong, M. J. *et al.* Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathog.* **10,** 1–6 (2014).

177. Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.* **22,** 299–306 (2012).

178. Peltola, V. T. & Mccullers, J. A. Respiratory viruses predisposing to bacterial infections : role of neuraminidase. *pediactric Infect. Dis. J.* **23,** 87–97 (2004).

179. Wong, S. S. & Yuen, K.-Y. Streptococcus pyogenes and re-emergence of scarlet fever as a public health problem. *Emerg. Microbes Infect.* **1,** e2 (2012).

180. Scaber, J. *et al.* Group A streptococcal infections during the seasonal influenza outbreak 2010/11 in South East England. *Euro Surveill* **16,** 1–4 (2011).

181. Van Den Beld, M. J. C. & Reubsaet, F. a G. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli. *Eur. J. Clin. Microbiol. Infect. Dis.* **31,** 899–904 (2012).

182. Quick, J. *et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.* **16,** 114 (2015).

183. Girlich, D., Naas, T. & Nordmann, P. Biochemical Characterization of the Naturally Occurring Oxacillinase OXA-50 of Pseudomonas aeruginosa These include : Biochemical Characterization of the Naturally Occurring Oxacillinase OXA-50 of Pseudomonas aeruginosa. *Antimicrob. Agents Chemother.* **48,** 2043–2048 (2004).

184. Grumaz, S. *et al.* Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med.* **8,** 73 (2016).

185. Meiring, T. L. *et al.* Next-generation sequencing of cervical DNA detects human papillomavirus types not detected by commercial kits. *Virol. J.* **9,** 1 (2012).

186. Friedrichs, C., Rodloff, a. C., Chhatwal, G. S., Schellenberger, W. & Eschrich, K. Rapid identification of viridans streptococci by mass spectrometric discrimination. *J. Clin. Microbiol.* **45,** 2392–2397 (2007).

187. Kawamura, Y., Hou, X. G., Sultana, F., Miura, H. & Ezaki, T. Determination of 16S rRNA sequences of Streptococcus mitis and Streptococcus gordonii and phylogenetic relationships among members of the genus Streptococcus. *Int J Syst Bacteriol* **45,** 406–408 (1995).

188. Carbonnelle, E. *et al.* MALDI-TOF mass spectrometry tools for bacterial identification in clinical microbiology laboratory. *Clin. Biochem.* **44,** 104–109 (2011).

189. G. Khan , H. O. Kangro, P. J. Coates, R. B. H. Inhibitory effects of urine on the polymerase chain reaction for cytomegalovirus DNA. . *J. Clin. Pathol* **. 44,** 360–365 (1991).

190. Cuthbertson, L. *et al.* Implications of multiple freeze-thawing on respiratory samples for culture-independent analyses. *J. Cyst. Fibros.* **14,** 464–7 (2015).

191. Foltz, J. L. *et al.* An Epidemiologic Investigation of Potential Risk Factors for Nodding Syndrome in Kitgum District, Uganda. *PLoS One* **8,** (2013).

192. Smelov, V. *et al.* Detection of DNA viruses in prostate cancer. *Sci Rep* **6,** 25235 (2016).

193. Larsen, B. & Monif, G. R. Understanding the bacterial flora of the female genital tract. *Clin. Infect. Dis.* **32,** e69–e77 (2001).

194. Harwich, M. D. *et al.* Genomic sequence analysis and characterization of Sneathia amnii sp. nov. *BMC Genomics* **13 Suppl 8,** S4 (2012).

195. Van Der Pol, B. Trichomonas vaginalis infection: the most prevalent nonviral sexually transmitted infection receives the least public health attention. *Clin. Infect. Dis.* **44,** 23–25 (2007).

196. Ross, J. D. C. & Jensen, J. S. Mycoplasma genitalium as a sexually transmitted infection: implications for screening, testing, and treatment. *Sex. Transm. Infect.* **82,** 269–271 (2006).

197. Ma, B., Forney, L. J. & Ravel, J. Vaginal microbiome: rethinking health and disease. *Annu. Rev. Microbiol.* **66,** 371–89 (2012).

198. Rogers, G. B., Hoffman, L. R., Carroll, M. P. & Bruce, K. D. Interpreting infective microbiota: The importance of an ecological perspective. *Trends Microbiol.* **21,** 271–276 (2013).

199. Brajão de Oliveira, K. Torque teno virus: A ubiquitous virus. *Rev. Bras. Hematol. Hemoter.* **37,** 357–358 (2015).

200. Lowe, B. *et al.* HPV Genotype Detection Using Hybrid Capture Sample Preparation Combined with Whole Genome Amplification and Multiplex Detection with Luminex XMAP. *J. Mol. Diagnostics* **12,** 847–853 (2010).

201. Bzhalava, D. *et al.* Unbiased Approach for Virus Detection in Skin Lesions. *PLoS One* **8,** 33–36 (2013).

202. Brown, A. C. *et al.* Rapid whole-genome sequencing of mycobacterium tuberculosis isolates directly from clinical samples. *J. Clin. Microbiol.* **53,** 2230–2237 (2015).

203. Cattoni, D. I., Fiche, J. B., Valeri, A., Mignot, T. & Nöllmann, M. Super-Resolution Imaging of Bacteria in a Microfluidics Device. *PLoS One* **8,** (2013).

204. Boedicker, J. Q. *et al.* Microfluidic Confinement of Single Cells of Bacteria in Small Volumes Initiates High-Density Behavior of Quorum Sensing and Growth and Reveals Its Variability. *Angew Chem Int Ed Engl* **48,** 5908–5911 (2010).

205. Lee, W., Kwon, D., Choi, W., Jung, G. Y. & Jeon, S. 3D-printed microfluidic device for the detection of pathogenic bacteria using size-based separation in helical channel with trapezoid cross-section. *Sci. Rep.* **5,** 7717 (2015).

206. Lui, C., Cady, N. C. & Batt, C. a. Nucleic Acid-based Detection of Bacterial Pathogens Using Integrated Microfluidic Platform Systems. *Sensors (Basel).* **9,** 3713–44 (2009).

207. Tian, P., Engelbrektson, A. & Mandrell, R. Two-log increase in sensitivity for detection of norovirus in complex samples by concentration with porcine gastric mucin conjugated to magnetic beads. *Appl. Environ. Microbiol.* **74,** 4271–4276 (2008).

208. Gandhi, K. M., Mandrell, R. E. & Tian, P. Binding of virus-like particles of Norwalk virus to romaine lettuce veins. *Appl. Environ. Microbiol.* **76,** 7997–8003 (2010).

209. Mcguckin, M. a, Lindén, S. K., Sutton, P. & Florin, T. H. Mucin dynamics and enteric pathogens. *Nat. Rev. Microbiol.* **9,** 265–278 (2011).

210. Nakashima, R., Sakurai, K., Yamasaki, S., Nishino, K. & Yamaguchi, A. Structures of the multidrug exporter AcrB reveal a proximal multisite drug-binding pocket. *Nature* **480,** 565–569 (2011).

211. Lomovskaya, O. & Lewis, K. I. M. emr, an Escherichia coli locus for multidrug resistance. *Cell Biol.* **89,** 8938–8942 (1992).

212. Rotem, D. & Schuldiner, S. EmrE, a multidrug transporter from Escherichia coli, transports monovalent and divalent substrates with the same stoichiometry. *J. Biol. Chem.* **279,** 48787–48793 (2004).

213. Nagakubo, S., Nishino, K. & Hirata, T. The Putative Response Regulator BaeR Stimulates Multidrug Resistance of Escherichia coli via a Novel Multidrug Exporter System , MdtABC The Putative Response Regulator BaeR Stimulates Multidrug Resistance of Escherichia coli via a Novel Multidrug Exporter. *J. Bacteriol.* **184-no.15,** 4161–4167 (2002).

214. Nishino, K. & Yamaguchi, A. Analysis of a complete library of putative drug transporter genes in Escherichia coli. *J. Bacteriol.* **183,** 5803–5812 (2001).

215. Ramón-García, S., Martín, C., De Rossi, E. & Aínsa, J. a. Contribution of the Rv2333c efflux pump (the Stp protein) from Mycobacterium tuberculosis to intrinsic antibiotic resistance in Mycobacterium bovis BCG. *J. Antimicrob. Chemother.* **59,** 544–547 (2007).

216. Barbosa, T. M. & Levy, S. B. Activation of the Escherichia coli nfnB gene by MarA through a highly divergent marbox in a class II promoter. *Mol. Microbiol.* **45,** 191–202 (2002).

217. Charpentier, X., Chalut, C., Rémy, M. H. & Masson, J. M. Penicillin-binding proteins 1a and 1b form independent dimers in Escherichia coli. *J. Bacteriol.* **184,** 3749–3752 (2002).

218. Typas, A. *et al.* Regulation of peptidoglycan synthesis by outer-membrane proteins. *Cell* **143,** 1097–1109 (2010).

219. Marcyjaniak, M., Odintsov, S. G., Sabala, I. & Bochtler, M. Peptidoglycan amidase MepA is a LAS metallopeptidase. *J. Biol. Chem.* **279,** 43982–43989 (2004).

220. Mattiuzzo, M. *et al.* Role of the Escherichia coli SbmA in the antimicrobial activity of proline-rich peptides. *Mol. Microbiol.* **66,** 151–163 (2007).

Online references

1. Endmemo (copy number calculator) http://www.endmemo.com/bio/dnacopynum.php

2. OligoAnalyzer 3.1 http://eu.idtdna.com/calc/analyzer

10. Abbreviation list

ABC transporter	ATP-Binding Cassette Transporters
AIDS	Acquired Immune Deficiency Syndrome
CDC	Centre For Disease Control And Prevention
CPE	Cytopathic Effects
CFU	Colony-Forming Unit
CMV	Cytomegalovirus
CNS	Central Nervous System
CSF	Cerebrospinal Fluid
ELISA	Enzyme-Linked Immunosorbent Assay
ESBL	Extended Spectrum Beta Lactamases
HACEK	Haemophilus, Aggregatibacter (Previously Actinobacillus), Cardiobacterium, Eikenella Corrodens, Kingella
HBV	Hepatitis B Virus
HCV	Hepatitis C Virus
HIV	Human Immunodeficiency Virus
HUS	Haemolytic Uremic Syndrome
LAMP	Loop Mediated Isothermal Amplification
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionization Time Of Flight

MDA	Multiple Displacement Amplification
MDR TB	Multi-Drug Resistant Tuberculosis
MIC	Minimum Inhibitory Concentration
MLST	Multi-Locus Sequence Type
MRSA	Meticillin Resistant *Staphylococcus aureus*
NDM-1	New Delhi Metallo-Beta-Lactamase-1
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PGM	Personal Genome Machine
PSGE	Pulsed-Field Gel Electrophoresis
SFF	Standard flowgram format
SNP	Single Nucleotide Polymorphisms
TB	Tuberculosis
TDR TB	Totally Drug-Resistant Tuberculosis
VNTR	Variable Number Tandem Repeat
WGS	Whole Genome Sequencing
XDR TB	Extensively Drug-Resistant Tuberculosis