ResearchOnline@JCU

This file is part of the following reference:

Susilawati, Tri Nugraha (2016) *Evaluation of next-generation sequencing technology in determining infectious causes of fever*. PhD thesis, James Cook University.

Access to this file is available from:

http://researchonline.jcu.edu.au/46187/

The author has certified to JCU that they have made a reasonable effort to gain permission and acknowledge the owner of any third party copyright material included in this document. If you believe that this is not the case, please contact ResearchOnline@jcu.edu.au and quote
http://researchonline.jcu.edu.au/46187/

JAMES COOK UNIVERSITY
AUSTRALIA

Evaluation of Next-Generation Sequencing Technology in Determining Infectious Causes of Fever

Thesis submitted by

Tri Nugraha Susilawati

Bachelor of Medicine, Medical Doctor, Master of Medicine

(Infection and Immunity)

For the Degree of Doctor of Philosophy

in the College of Medicine & Dentistry

James Cook University

January, 2016

Statement of Access

I, the undersigned, the author of this thesis, understand that James Cook University Library, by microfilm or by other means, allows access to users in other approved libraries. All users consulting this thesis will have to sign the following statement:

'In consulting this thesis, I agree not to copy or closely paraphrase it in whole or part without the consent of the author; and to make proper written acknowledgement for any assistance, which I have obtained from it.'

Beyond this I do not wish to place any restriction on access to this thesis.

Signed: Date: 7 January 2016

Statement of Sources Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published, unpublished or oral work of others has been acknowledged in the text and a list of references is given. Every reasonable effort has been made to gain permission and acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

Signed: Date: 7 January 2016

Statement of Contributions of Others

Nature of assistance	Contribution	Co-contributors
Intellectual support	The supervisory team provided assistance within the remit of their roles as described in the policy *The Role of the Advisory Panel in Providing Regular Guidance and Support to Research Student.* Editorial assistance was provided to the candidate by both the supervisory team and professional editors. Proofreading was restricted to correcting the presentation of the text to conform with standard usage and conventions and the provision of advice in the matters of structure, conventions of grammar and syntax, use of clear language, logical connections between phrases, clauses, sentences, paragraphs and sections, voice and tone and repetition.	Professor John McBride, MBBS, DTM&H, FRACP, FRCPA, PhD College of Medicine & Dentistry, James Cook University, Cairns, Australia Professor Alex Loukas, PhD Centre for Biodiscovery and Molecular Development of Therapeutics, Australian Institute of Tropical Health and Medicine, James Cook University, Cairns, Australia Aaron R. Jex, BSc, PhD Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Australia Cinzia Cantacessi, PhD Department of Veterinary Medicine, Cambridge Veterinary School, University of Cambridge, UK Elite Editing 213 Greenhill Road, Eastwood, South Australia 5063
Financial support	Project costs Stipend Supplementary academic support (thesis editing and conference presentation)	James Cook University Far North Queensland Hospital Foundation Australia Awards Scholarships

Acknowledgements

Study Participants

I would like to express my sincere appreciation to the participants of this study. The majority of participating patients were cooperative and happy to contribute. Contact with the study participants, especially with the patients, provided rewarding and unforgettable personal experiences.

PhD Supervisors

It was Professor John McBride who inspired me to keep up my spirit and persistence throughout this project. As my principal supervisor, he has guided me very well and was always available when I needed his help. He encouraged me to learn many things and try every opportunity. I thank him for his patience and careful supervision throughout the years of my candidature. My appreciation is also extended to Professor Alex Loukas, whose supervision has been always energetic, enthusiastic and thought provoking. I thank him for providing in-kind support and access to his laboratory. The biggest help in the molecular side of my project came from Dr Aaron Jex and Dr Cinzia Cantacessi, whose expertise has encouraged me to learn and learn more about the new fields in my life: molecular technique and genomic data analysis. Lastly, I would like to thank my former supervisor, Dr Jason Mulvenna, who had already made a major contribution in preparation of my proposal and literature review.

Funding Bodies

The study was made possible by an annual grant from the Far North Queensland Hospital Foundation and from the College of Medicine and Dentistry, James Cook University. I also thank AusAID for the Australia Award Scholarship that covered my tuition fees, living expenses and other necessities. A top-up scholarship from College of Medicine and Dentistry, James Cook University provided an additional stipend during my PhD.

Cairns Hospital Staff

I should highlight the outstanding contributions of the staff from the Clinical Research Unit: Sue Richmond, Sue Dixon, Donna Kreuter and Debra Horold. They provided great help in enrolment of participants and blood collection. The study was conducted with the permission and assistance of Dr Richard Stone (Director of the Emergency Department), Dr Peter Boyd (Director of the Medicine Department), Dr Drew Wenck (Director of the Intensive Care Unit), Mark Porton (Director of the Pathology Department), Peter Hiatt and Deborah Moffat (staff of the Pathology Department). I thank them for their cooperation in helping me obtain the samples I needed. I also thank the staff of the Medical Record Department of Cairns Hospital with their assistance in providing more than a thousand medical records that I requested.

Colleagues and Mentors

Over the last four and half years, I have received wonderful advice and encouragement from a number of people. They include Intansari Nurjannah, Mercy Rampengan, Lalu Adi Gunawan, Mangalasiri Jayathunge, Sri Warsini, Yvonne Hodder, Lyn Kerr, Dr Cindy Woods, Dr Gregory Maes, A/Prof Malcolm McDonald and Prof Jane Mills. I learnt a great deal from the wonderful mentors I had at the Queensland Tropical Health Alliance laboratory, including Ivana Ferreira, Darren Pickering, Dr Mark Pearson, Dr Severine Navarro, Dr Annette Dougall, Dr Javier Sotillo-Gallego and Dr Paul Giacomin. Thanks so much to all of you who helped me to survive PhD life.

Family

I sincerely thank my husband, Dr Atik Susianto, for his understanding, continuous support and encouragement during this seemingly never-ending journey. And, to our children, Fatyanaura Arleya, Elena Zedya Renata and Kinza Altair Rashad, thank you so much for giving me a rich life experience as a PhD student with three young children and for keeping my spirits high by simply looking at me with your sparkling eyes and wonderful smiles.

Abstract

Undifferentiated fever (UDF) is a common complaint in clinical practice, but its aetiology is not always determined due to non-specific symptoms and laboratory findings. While fever of unknown origin (FUO) is a common medical term for fever without obvious cause, this condition is distinguished from acute undifferentiated fever (AUF) in terms of duration, progression of illness and underlying causes. In FUO, fever must exist for more than 3 weeks and can persist for a very long period unless the underlying cause is found and eliminated. In contrast, AUF is more limited in duration and many episodes spontaneously resolve, presumably due to self-limiting infectious diseases.

The problem of determining the infectious causes of fever has received considerable attention, particularly in tropical countries. Previous studies in South and Southeast Asia reported high prevalence of infection as the main cause of AUF. This prompted the hypotheses that infection-related AUFs are common in the tropical region of Far North Queensland, Australia, and that a significant proportion of AUFs in this region are undiagnosed.

Diagnosing infectious causes of fever is a challenge for clinicians. With hundreds of possible agents and a limited number of specific tests that can be performed, it is very likely that doctors will miss the true cause of fever. Moreover, current diagnostic approaches rely on prior knowledge of the pathogens being sought, thus precluding the detection of unsought or novel pathogens. Thus, the prevalence of undiagnosed undifferentiated fever (UUDF) indicates either that clinicians are failing to order appropriate tests, that current diagnostic methods are inadequate, or that there are causes of fever that are yet to be discovered.

This diagnostic challenge necessitates good clinical skills, knowledge of the pattern of signs and symptoms associated with particular infections and a broad diagnostic tool for determining infectious causes of fever. Since there is a wide range of pathogens, the diagnostic tool should be able to distinguish one pathogen from another as well as identify multiple pathogens simultaneously. This can be achieved if the diagnostic tool can ‘read’ unique characteristics of pathogen(s) present in the sample as

distinguishable nucleic acid sequences so it can facilitate the identification of organism(s).

Next-generation sequencing (NGS) technology has the capability to produce large amounts of nucleic acid sequences in a relatively short time. This technology has been applied to the study of biological diversity in environmental samples. The extent of diagnostic problems and the capability of NGS technology to detect the presence of nucleic acids in any environment have brought the theory into clinical application. The potential of NGS as a broad-scale diagnostic tool prompted the hypothesis that NGS is a practical method for investigating pathogens causing fever.

There are several NGS platforms available, but little is known about their effectiveness and efficiency with regards to pathogen detection in clinical samples. The primary aim of this study was to assess the capacity, sensitivity, and more importantly, the specificity of the Illumina HiSeq platform for broad-scale characterisation of pathogens associated with fever. Secondary aims included: to describe the epidemiology of AUF and UUDF in Far North Queensland, Australia, and to optimise the preparation procedure for samples that will be subjected to NGS assay.

The study was conducted in three stages, comprising two preliminary studies and the main study. The first preliminary study was a retrospective study of fever patients presenting to Cairns Hospital over the three-year period between 1 July 2008 and 30 June 2011. The findings suggest that AUF is common in the population of Far North Queensland, Australia. A robust definition of UUDF is proposed based on the clinical and laboratory characteristics of patients reviewed in this study, including the following criteria: 1) a fever of ≥ 38.0 °C or symptoms suggestive of fever; 2) a duration of fever of ≤ 21 days; 3) a failure to reach a diagnosis after performing clinical evaluation and laboratory investigations, including complete blood count, serum biochemistry, urinalysis, blood culture, chest X-ray; 4) a request by the clinician of specific test for at least one infectious agent and; 5) a failure to make a specific diagnosis. The proportion of UUDF was 56.8% (193/340), indicating the need for a broad diagnostic tool to determine infectious causes of AUF. In general, the findings provide valuable information regarding the feasibility of conducting a fever study using NGS technology at Cairns Hospital.

The second preliminary study was conducted to determine the most suitable type of blood specimen for NGS analysis. It was anticipated that there would be small quantities of pathogen nucleic acids present among abundant human nucleic acids

background. Therefore, it was important to minimise the quantity of human nucleic acids in order to increase the sensitivity of detection of the pathogen. Six healthy volunteers participated in this second preliminary study, which aimed to measure levels of double-stranded DNA in plasma and serum. Specimens were taken using different methods of blood collection: with a syringe and with a vacuum system, and with and without applying a tourniquet. DNA concentration in the samples was measured using microplate fluorescence assays using SYBR Green I as the fluorescent dye. This study found that DNA concentration in plasma was significantly lower than that in serum ($p < 0.05$). However, the method of blood collection did not significantly affect DNA concentration.

A main component of this thesis was a prospective study involving the use of a NGS platform to determine infectious causes of AUFs. Isolation of DNA and RNA from plasma/serum samples was performed using QIAamp® DNA Mini Kit (Qiagen) and TRIzol® LS reagent (Life Technologies) respectively. Following nucleic acid extraction, amplification of DNA was conducted according to the SeqPlex Enhanced DNA Amplification Kit (SEQXE) protocol (Sigma-Aldrich). DNase treatment, cDNA synthesis and amplification were performed on RNA samples according to SeqPlex RNA Amplification Kit (SEQR) protocol (Sigma-Aldrich). Sequencing was conducted on 22 DNA/cDNA samples that met the standard input determined by the sequencing company. These samples originated from 17 patients, comprising seven positive control samples from patients who had specific diagnoses and 10 samples from patients for whom diagnoses had not been achieved. Data analysis was conducted using the Kraken program and the traditional assembly-alignment pipeline.

The study findings demonstrate the limitation and utility of NGS technology in determining the aetiology of AUF. Various viruses and bacteria were found in every sample, so considered selections were made on pathogens for which there were supporting reads consistent with clinical data and pathology findings. Aetiological diagnosis was verified in 85.7% (6/7) of controls. Among the undiagnosed participants, deep sequencing identified some plausible causes of fever in 60% (6/10) of subjects, including *Escherichia coli* bacteraemia and scrub typus that eluded conventional tests.

Further, the NGS technology generated valuable information for studying microbial diversity in human blood. Further work could be directed at optimising sample preparation and improving sequencing efficiency, as well as developing

efficient bioinformatics tools for analysing sequence data. It is hoped that NGS technology can be adopted in clinical practice at more affordable costs and with timely delivery of results.

Table of Contents

List of Tables

List of Figures

Glossary

Amplicon	any PCR amplification product
Cloud computing	remote computational resources available via internet
Coverage	the amount by which a genome is over-sampled; the ratio between the cumulative size of the set of reads and the size of the genome
De novo	from the beginning (i.e. without prior information)
Errors	percentage of confidently aligned reads that are mapped to the wrong location
E-value	Expect value, a parameter that describes the number of hits one can 'expect' to occur by chance when searching a sequence database of a particular size. An E-value of 1 means that it would be expected to find a match with a similar score simply by chance. The lower the E-value, or the closer it is to zero, the more significant the match.
Metagenomics	a novel field that deals with the sequencing and study of entire microbial communities isolated directly from a particular environment
Paired-end reads	DNA sequences from each end of DNA templates
Phage	a virus that infects bacteria
Reads	DNA fragments whose sequence is known
SEQR	SeqPlex RNA Amplification Kit
SEQXE	SeqPlex Enhanced DNA Amplification Kit
SYBR Green I	a fluorescent dye that bind to dsDNA

Abbreviations

AFI	Average fluorescence intensity
AGRF	Australian Genome Research Facility
ALT	Alanine aminotransferase
ASCII	American Standard Code for Information Interchange
AST	Aspartate aminotransferase
AUF	Acute undifferentiated fever
BGI	Beijing Genomics Institute
BLAST	Basic Local Alignment Search Tool
CaSS	Clinical and Statewide Service
CMV	Cytomegalovirus
CNA	Circulating nucleic acids
CNS	Central nervous system
CRP	C-reactive protein
CSF	Cerebrospinal fluid
CT	Computed tomography
DNA	Deoxyribonucleic acid
EBV	Epstein-Barr virus
ELISA	Enzyme-linked immunosorbent assay
FI	Fluorescence intensity
FID	Fever of intermediate duration
FUO	Fever of unknown origin
GAS	Group A Streptococcus
GP	General practitioner
HAV	Hepatitis A virus
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HIV	Human immunodefficiency virus
HPV	Human Papilloma virus
HREC	Human Research Ethics Committee
ICD	International Classification of Diseases

ICU	Intensive care unit
IV	Intravenous
JCU	James Cook University
MFA	Microplate fluorescence assay
NANBH	Non-A, non-B viral hepatitis
NCBI	National Centre for Biotechnology Information
NEAF	National Ethics Application Form
NGS	Next-generation sequencing
NHMRC	National Health and Medical Research Council
PCR	Polymerase chain reaction
PE	Paired-end
PGM	Personal Genome Machine
PUO	Pyrexia of unknown origin
QTHA	Queensland Tropical Health Alliance
RAM	Random-access memory
RGO	Research Governance Officer
RIN	RNA integrity number
RNA	Ribonucleic acid
SD	Standard deviation
SFG	Spotted fever group
SIV	Simian immunodeficiency virus
SPSS	Statistical Package for the Social Sciences
SSA	Site-Specific Assessment
TMV	Tobacco mosaic virus
TTMDV	Torque teno midi virus
TTV	Torque teno virus
UTI	Urinary tract infection
UDF	Undifferentiated fever
UUDF	Undiagnosed undifferentiated fever
WBC	White blood cell
WHO	World Health Organisation

Chapter 1: Introduction

1.1 Terminology

Fever is a common medical problem in clinical practice with various causes and diverse outcomes. It often poses challenges for clinicians because of its numerous associated diagnostic alternatives. Sometimes the cause of fever is unclear due to non-specific clinical manifestations and limited information available from the initial laboratory findings. In these cases, the condition is referred to as undifferentiated fever (UDF), and despite the advancement of medical technologies, a considerable proportion of UDF cases go undiagnosed. If the condition lasts for more than three weeks, it is then generally accepted that it meets the criteria for fever or pyrexia of unknown origin (FUO/PUO).[1]

Reid[2] defined PUO as an elevated body temperature of ≥ 38 °C on one occasion or ≥ 37.4 °C on three occasions in a patient over 14 years of age without adequate evidence of local symptoms and signs to be confidently diagnosed after initial examination, chest X-ray, and routine laboratory investigation. This definition did not specify fever duration as a criterion for diagnosing PUO; instead, all fever cases surpassing the temperature threshold were considered as PUO. According to Petersdof and Beeson,[3] FUO is defined as a temperature higher than 101 °F (38.3 °C) for more than three weeks without an identified cause after one week of hospital investigation. This conventional definition was modified in 1991 by Durack and Street,[4] who differentiated FUO into classical type and three other types: nosocomial, neutropenic and human immunodefficiency virus (HIV)-associated FUO. The classical type of FUO is not associated with prolonged fever acquired during hospital admission (nosocomial). It is also not associated with fever that is often experienced by patients who have an abnormally low level of neutrophils (a type of white blood cell) or those with HIV infection. They also suggested a shorter duration for investigation: three outpatient visits or three days of in-hospital evaluation.

The World Health Organisation (WHO) issues the International Classification of Diseases (ICD) as the standard diagnostic tool for epidemiology, health management and clinical purposes (http://www.who.int/classifications/icd/en/). The most current

version of the ICD, ICD-10 version 2015,[5] records 'fever of other and unknown origin', in which the aetiology of fever cannot be ascertained.

In contrast to FUO, which is clearly defined and widely studied, there is no internationally accepted consensus with regards to the diagnosis of short-term febrile illness with unclear aetiology. This condition is described as 'acute undifferentiated fever' (AUF) in this thesis. This term encompasses a variety of causes producing a range of clinical manifestations, with acute fever as a unifying symptom.[6] Most clinicians and researchers define acute fever as evidence of raised body temperature to 38 °C or higher for up to three weeks, without detection of systemic disease or the focus of infection or inflammation after initial clinical evaluation and basic laboratory investigations such as complete blood count and urinalysis. In malaria-endemic countries, case definition of AUF usually mandates that malaria is excluded, usually by microscopic examination of a thick blood smear.

In addition, a number of researchers refer to fever of intermediate duration (FID) to define fever higher than 38 °C that lasts between one and four weeks without a definite diagnosis after an initial approach.[7] The duration of fever defined for FID cases overlaps with that for the case definitions of AUF and FUO, which use a 3-week duration as a cut-off level for distinguishing the two conditions. Therefore, this thesis excludes FID and focuses only on AUF cases. Figure 1.1 depicts the outcomes of UDF and the terminology used in this thesis.[8]

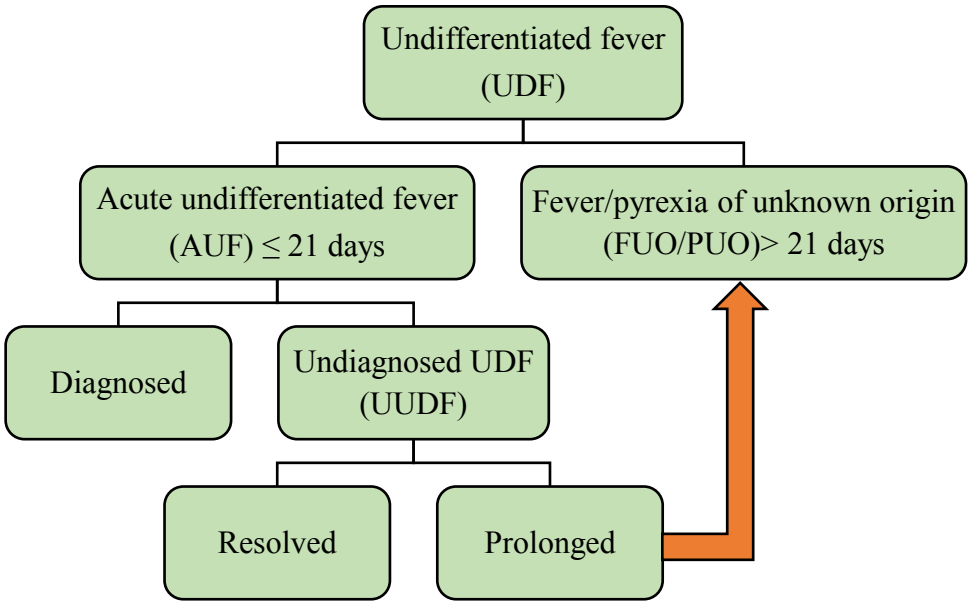


Figure 1.1: The outcomes of undifferentiated fever

- Undifferentiated fever (UDF) is any case of fever with unclear aetiology where the results of initial investigations (which include clinical examination, rapidly available pathology and/or radiological investigations) are not conclusive in achieving a diagnosis. Thus, the condition is characterised by a requirement for further investigation to explain the cause of fever and to consider differential diagnoses.
- AUF is a short-term febrile illness lasting up to 21 days without obvious source.
- Undiagnosed undifferentiated fever (UUDF) is a condition of AUF that remains undiagnosed after conducting laboratory and radiology investigation.
- Fever/pyrexia of unknown origin (FUO/PUO) is fever that exists for more than 3 weeks without definite diagnosis.

1.2 Study background

A presumptive diagnosis is frequently made based on focal signs and symptoms accompanying fever. However, sometimes a clear etiology is not found, making treatment problematic. In short-lived episodes, both patient and clinician are usually satisfied that further investigation is not warranted, and an explanation that a patient has a 'viral infection' is generally acceptable, although there is a paucity of data to support this common explanation. In many cases, intensive investigative efforts are performed for more prolonged or severe episodes only. This approach is efficient from the health-

economics point of view, but it is inappropriate in terms of the holistic management of diseases.

As the most common feature of infection, AUF requires careful attention because, although it may reflect minor and self-limited infection, it may also be a sign of potentially severe or lethal infection. Some fevers are not easily diagnosed, leading doctors to order several laboratory tests to ascertain an aetiological diagnosis, and the failure to diagnose may represent a failure to test for a pathogen that is already described. Therefore, a broad diagnostic tool that can detect a wide range of pathogens is needed to ascertain the aetiological cause of AUF.

Such a broad diagnostic tool is also important for surveillance purposes. Although many recently discovered ('new') pathogens may in fact have been present as causes of fever for many years, truly emerging diseases may also present as febrile illnesses. Some factors contributing to the emergence of new infectious diseases include global travel and environmental change. The increased exposure of humans to a habitat that was previously exclusive to animals may introduce a transfer of diseases from animals to humans. Likewise, the risk for the spread of new, emerging and re-emerging infections is ever increasing as a result of global travel. It is therefore vital to have methods and a systematic protocol in place that can be applied to new outbreaks of a presumed infectious disease. A method that can provide unbiased characterisation of pathogens may be the key for the rapid screening of infectious diseases, and will certainly reduce the prevalence of UUDF, which eventually will also reduce FUO cases.

1.3 Overview of next-generation sequencing

Recent years have seen an enormous reduction in the cost of high-throughput sequencing technologies, also known as next-generation sequencing (NGS). These technologies provide enormous capacity to produce large genomic sequence datasets over a relatively short period of time at a reasonable cost. The advancement of NGS technology has enabled a 1,000-fold reduction in sequencing costs, from $500 per megabase using Sanger sequencing to < $0.5 per megabase using an Illumina platform.[9, 10] It is now possible to sequence a complete human genome within days for < $1,000; this represents a massive advancement, considering that the first human

genome sequence was completed just over a decade ago after 13 years of work at a total cost of ~$3 billion.[11, 12]

NGS costs have now reduced to the point where they are feasible technologies for use in the clinical diagnosis of infections. Previous studies have shown that NGS can reliably identify microorganisms, including viruses, at levels beneath the detection sensitivity of conventional microscopy and/or serological tools.[13] The technology also facilitates the discovery of genes that serve as biomarkers in various pathological conditions. In fact, the application of NGS in the area of genetic diseases and cancer has facilitated the identification of genes that serve as biomarkers for diagnosing genetic abnormalities and certain types of cancers, as well as for monitoring disease progression.[14] In the field of microbiology, NGS provides huge amounts of sequence data for genetic profiling, the study of biodiversity and pathogen discovery. It has been reported that NGS technology has helped to elucidate a fatal infectious disease and revealed an occult Hepatitis B infection in an apparently healthy individual.[15, 16] As the technology is still evolving, there is a need to assess the reliability and practicality of NGS as a new tool for pathogen identification in patients with AUF.

1.4 Overview of study design

This thesis describes efforts to identify the infectious causes of AUF when the symptoms do not exceed a three-week period. It presents three stages of research in which the ultimate objective is pathogen identification using a NGS platform. The first stage of the research was to study the epidemiology of AUF in Far North Queensland, Australia, and to identify the characteristics of patients who would subsequently be included in the main study employing the NGS method. The next phase of the study involved laboratory experiments to optimise the sample preparation technique for the NGS study. The proportion of nucleic acids within a clinical sample that belongs to a pathogen are miniscule, so it was essential to measure the levels of human deoxyribonucleic acid (DNA) in various blood specimens to determine which specimen contained the smallest amount of human DNA contaminant. The last and the main component of the research was a prospective study recruiting patients with AUF. Blood samples were collected from two groups of patients: those who were diagnosed with specific infectious disease and those who were undiagnosed. The samples were subjected to a NGS platform to obtain sequencing data for each individual. Analysis

was directed to identify the pathogen causing febrile illness by examining non-human nucleic acids sequences.

1.5 Organisation of thesis

The thesis is presented in seven chapters. It begins with an introduction (Chapter 1) followed by a review of the relevant literature (Chapter 2). Pathogenesis of fever, case definitions and the scope of undiagnosed AUF are described according to previous reports in the literature. In addition, the development of microbiology and diagnostic tools are discussed to formulate research questions and hypotheses. This is followed by a description of the study design and the three stages of the research (Chapter 3). Chapter 4 presents the first study, which describes the prevalence and characteristics of AUF in Far North Queensland, Australia. Chapter 5 covers the second study, which aimed to optimise the sample collection procedure for subsequent study using the NGS. Chapter 6 presents the main study, which investigates pathogen identification using a NGS platform. The chapter provides details on the study subjects, procedure for sample preparation, sequencing details and bioinformatic analysis. This is followed by interpretation of the results and a discussion of the findings of the main study. Chapter 7 offers general discussion unifying the themes and results from the previous chapters. This final chapter concludes the thesis and makes recommendations for future research.

Chapter 2: Literature Review

2.1 Overview

This chapter provides a comprehensive review of previous studies and available literature to highlight essential issues around AUF and investigations into the cause of this condition. Few articles have explored the topic of AUF and the use of NGS for its investigation. This lack of research on AUF is surprising given that it is a common presentation, though the use of NGS in this investigation is quite recent. This chapter outlines current understanding on the problem of AUF, identifies gaps in knowledge and provides context for this current study.

Initially, this chapter discusses the pathogenesis of fever and how infection can induce it. This is followed by a review of previous studies on AUF and a discussion around several issues related to AUF, including its epidemiology, investigations and diagnostic challenges. A review article about AUF in Asia was published as part of this section. The section is followed by a discussion of methods for pathogen detection and discovery and how causal links are established between microbes and diseases. An in-depth discussion of NGS technology is then presented, including the development of sequencing platforms, the principle of NGS and the implementation of NGS in various metagenomics studies (i.e. studies that apply DNA sequencing directly on a sample, bypassing the culture and clonal selection steps that are required in earlier sequencing techniques). Finally, the research questions proposed for this PhD are articulated.

2.2 Pathogenesis of fever

Fever is defined as a rise in body temperature above what is considered 'normal' (37 °C ± 1 °C), and it is a common symptom experienced by human beings as an adaptive response to various immune challenges of infectious or non-infectious origin. This response is regulated by the central nervous system (CNS) and involves endocrine, neurological, immunological and behavioural mechanisms.[1] In general, body temperature can increase as a result of physiological and pathological states. In physiological situations, an elevated body temperature is a reaction to an increase in internal or external temperature, for example, during exercise, pregnancy, hot weather and dehydration. On the other hand, infections and inflammatory diseases are the most

common pathological causes of fever, followed by malignancies and miscellaneous conditions such as medications (e.g., antibiotics and narcotics; drug-induced fevers can be due to adverse reactions or withdrawal), trauma or injury (e.g., heart attack, stroke, burns), autoimmune diseases (e.g., Guillian–Barre syndrome, lupus), hormone disorders (e.g., hyperthyroidism, adrenal insufficiency), embolisms, and various syndromes and diseases (e.g., Caroli's disease, Castleman's disease, Kawasaki's syndrome, Kikuchi's syndrome).[17–20]

Fever has been recognised as a major manifestation of inflammation since the sixth century BC.[21, 22] A causal relationship between infection and fever became clearer in the late eighteenth century through the works of Louis Pasteur and Robert Koch in the field of microbiology. William H. Welch established that the CNS was involved in regulating body temperature through his experiments, based on which others constructed the modern theory of the pathogenesis of fever. Welch identified the location of the thermoregulatory centre in the CNS and suggested the beneficial effect of fever, either directly by destroying microbes, or indirectly by increasing the host's resistance to infection.[21]

At present, it is known that fever is triggered by substances collectively called pyrogens, which may come from sources internal (endogenous) or external (exogenous) to the body. Endogenous pyrogens include cytokines (produced by phagocytic cells) such as interleukin 1 (IL-1), interleukin 6 (IL-6), and tumour necrosis factor-alpha (TNF-α). Lipopolysaccharide (LPS), a cell wall component of Gram-negative bacteria, is an example of an exogenous pyrogen. During infection, exogenous pyrogens cause the release of endogenous pyrogens, which, in turn, activate the arachidonic acid pathway to synthesise prostaglandin E2 (PGE2). The common pathway for fever pathogenesis is through the activation of thermoregulatory cells in the hypothalamus by PGE2, resulting in increased body temperature through various mechanisms, as illustrated in Figure 2.1.

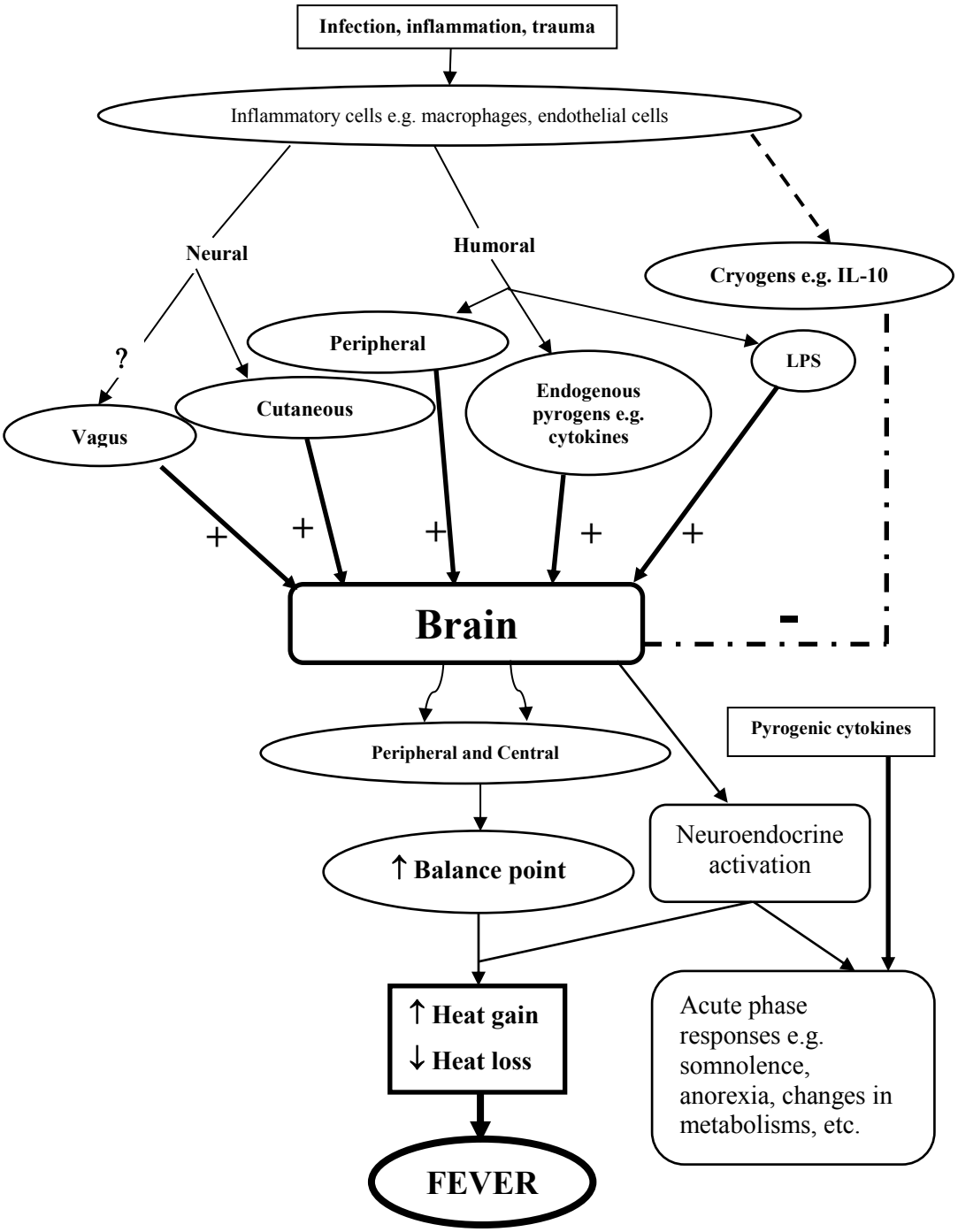


Figure 2.1: Pathway of fever pathogenesis

(Reproduced from[1] with permission)

2.3 Previous studies on acute undifferentiated fever

In the first stage of my research, a literature review was conducted in two stages to understand the problem of AUF and to build a study design that would further the current state of knowledge and contribute to the understanding of previously unidentified causes of AUFs. The first stage of the review was performed in January 2012, seeking insight into the global problem of acute fever with unknown infectious cause. The second stage was carried out in September 2012, aiming to identify specific causes of AUF and quantify the proportions of cases that remain undiagnosed.

In the first review, the following online databases were used: PubMed, Medline, Web of Science, CINAHL and Cochrane Database of Systematic Reviews. Keywords included (undiagnosed OR undifferentiated OR unknown OR unexplained OR unconfirmed) AND (fever OR febrile) AND acute AND infection. Additional articles were identified from the reference lists of retrieved articles and from articles that had cited the retrieved articles.

Studies were included if they were conducted in primary health care, hospital and laboratory settings, and reported a series of laboratory investigations for diagnosing infection as an underlying cause of AUF. The review was not limited to any country or region, but was restricted to articles published in English during the last decade (from 1 January 2001 to 31 December 2011).

Although AUF is frequently observed in clinical practice, there is a paucity of published reports on the condition. A total of 25 articles were reviewed. Most of the reviewed AUF studies were conducted in tropical and subtropical areas of the world. Of these, most originated in South and Southeast Asian countries, such as Thailand,[23–26] Vietnam,[27–29] India,[30–33] Nepal,[34] Sri Lanka,[35] Laos,[36] Indonesia[37] and Singapore.[38] Some AUF studies were conducted in African[39–41] and tropical South American countries.[42–45] Only two studies were conducted in developed countries in Europe.[46, 47] The most commonly reported infections in these studies included malaria,[25, 30] dengue,[32, 42, 44] scrub typhus,[23, 31] spotted fever rickettsial infections[35, 41] and leptospirosis.[24, 43, 44]

The search strategy did not discover any articles pertaining to AUF in Australia. Fever is not included in the national health priority areas that have been chosen for focused attention by Australian governments.[48] Only a small proportion of the Australian population live in areas where tropical infections are commonly found.

Despite this, tropical infections that manifest as acute fever are quite prevalent in some areas in Australia, particularly in areas situated in tropical and subtropical zones. Some Australian cities are major tourist destinations and may serve as points of entry for the spread of tropical infections as a result of increased global travel. Mosquito-borne pathogens such as the malaria parasite (*Plasmodium* spp.) and the dengue, Ross River and Barmah Forest viruses can be found in Australia, particularly in tropical North Queensland.[49–54] It has been reported that malaria and dengue were introduced to this region by travellers,[55] while the Ross River and Barmah Forest viruses are endemic in Queensland.[50] Other diseases that can present as an AUF, such as melioidosis, leptospirosis and rickettsiosis, are also found in Australia.[56, 57]

Although previous studies have evaluated the causes of AUF, limited data exists regarding the epidemiology of undiagnosed cases where pathogens other than that already described may be the cause. Those studies have focused on particular pathogens rather than the various causes of AUF. In addition, the proportion of AUF cases that remain undiagnosed was not specifically reported in most studies.

In September 2012, the literature review was updated by conducting the review systematically according to PRISMA guidelines (http://www.prisma-statement.org). The second review aimed to determine the case definition, investigations and aetiologies of AUF, and to determine the proportion of AUF cases that remain undiagnosed. It was shown in the first review that AUF studies have predominantly been conducted in Asia, so the second review focused on those studies that were conducted in Asian countries. This second review was limited to articles published in English during the period 1990–2012. This time period was chosen because nucleic acid testing began to be employed as a routine diagnostic tool after 1990. The following terms were used when searching for articles on PubMed database: Fever/etiology (Majr) OR Fever/microbiology (Majr) AND Asia (Mesh) AND Adult (Mesh) AND 1990/01/01 (PDAT): 2012/12/31 (PDAT) AND Journal Article (ptyp) AND English (lang). Literature search was also performed in other databases including Medline, Scopus and Web of Science. From 201 studies retrieved from the online databases, 9 were included in the review. This review was published in the *Southeast Asian Journal of Tropical Medicine and Public Health* in May 2014.[6]

Article: Acute undifferentiated fever in Asia: a review of the literature

Declaration of Authorship

Publication details	Nature and extent of the intellectual input of each author	Signature
Susilawati, TN & McBride, WJH. Acute undifferentiated fever in Asia: a review of the literature. *The Southeast Asian Journal of Tropical Medicine and Public Health.* 2014:45(3);719–726. *Accepted for publication 12 May 2014* **Published May 2014**	Developed the initial idea and design the review process, conducted literature search, prepared the manuscript.	Susilawati, TN
	Assisted with the article writing and editing.	McBride, WJH

Different case definitions were used for the different studies for AUF; including fever duration and temperature threshold. For example, four studies[23, 24, 30, 59] evaluated patients with a duration of fever of less than 14 days whereas another study[31] evaluated patients with fever of up to 21 days duration. In terms of the definition of fever, the AUF studies specified a fever cut-off level of ≥ 38 °C, 38.3 °C or 37.5 °C.[6]

To identify the etiologies of AUF, the studies employed non-specific and specific investigations. Non-specific investigations refer to blood analysis and other laboratory testing to describe the underlying cause of the disease without determining a specific microorganism, such as a complete blood count, serum biochemistry, urinalysis and chest X-ray. Specific investigations refer to laboratory testing investigating specific pathogens, such as malaria films, serological tests, polymerase chain reaction (PCR) assays and bacterial cultures. These methods identified specific infections, such as malaria, dengue fever, leptospirosis and rickettsioses as possible causes of AUF.

This review shows that prior studies have often used serological testing (measurement of immunoglobulin M and G [IgM and IgG] levels) as the main diagnostic method for determining infectious causes of AUFs. A drawback of this approach is that collection of convalescent serum is not possible from all patients, and analysis of acute samples only can lead to difficulties with interpretation. If only acute serum samples are collected, results may be falsely negative if collected too early, when antibody titres against a given pathogen are below detectable levels during the first few

day of illness. Additionally, there are problems associated with cross-reactivity of antibodies of closely related species or organisms, which can result in false positives.[58]

Direct methods based on pathogen detection by culture, polymerase chain reaction (PCR) or antigen detection provide more reliable results than antibody detection in the acute phase of illness.[59] A challenge with antigen detection is that patients may present at a stage of the illness after the antigen or agent causing the illness has been cleared by the immune system. This can result in false negatives and prevent a diagnosis from being established. Given these methodological difficulties, it is not surprising that a large number of AUFs remain undiagnosed. Despite the introduction of PCR as a routine diagnostic test, the identified aetiologies of fever and the proportion of undiagnosed fever cases are similar to those observed in prior serological studies.[60, 61]

Some conclusions can be drawn from the published article, which forms part of the literature review. First, during the 23-year period from 1 January 1990 to 31 December 2012, only a limited number of reports were published on AUF in Asia. Second, there has been no agreement on the case definition of short-term febrile illnesses with unclear aetiology. Third, the focus of the reviewed articles has largely been on detecting the common aetiologies of AUF, rather than exploring the aetiology of undiagnosed infections. Several gaps in knowledge were identified that underpin the need for further study.

The work presented in Chapter 4 goes towards filling the gaps in existing knowledge in several ways. First, the study describes the epidemiology of AUF by examining its prevalence and the proportion of fevers that go undiagnosed in the area of Far North Queensland, Australia. Second, the study compiles clinical and laboratory characteristics of patients to formulate a robust definition of UUDF. Third, the study puts emphasis on the elaboration of information that exists with regards to UUDF and proposes criteria for further investigation with broad diagnostic tools.

2.4 Identification of new pathogens in historical perspective

2.4.1 Methods for detecting microbes

Many tests have been developed for detecting bacterial and viral pathogens. Visualisation was the first method developed, representing an attempt to see the microbial world. The advances in light microscopy achieved by a Dutch botanist, van

Leeuwenhoek (1632–1723), introduced the existence of microbes to humans. The spectrum of known bacterial, fungal and protozoan pathogens has since been expanded with improved staining and culture techniques, in conjunction with the development of more advanced visualisation methods. For example, light microscopy has been refined with the use of immunohistochemical or immunofluorescent stains to detect specific molecules in the host or pathogen. Laser scanning confocal microscopy represents a further quantum improvement in resolution and signal quantification. Most recently, transmission and scanning electron microscopy has revealed microorganisms at very high resolution.[62]

Traditional methods for diagnosing bacterial infections include staining, culture and biochemical tests. The process of staining is performed by preparing samples on a glass slide. Clinical samples are usually obtained from fluids such as blood, cerebrospinal fluid (CSF), sputum, pus, ear or eye or nasal discharge, and urine. Samples can also be obtained from microbes grown in an artificial medium. Based on the dye used and the purpose of staining, bacterial staining can be categorised into simple staining and differential staining. Simple staining uses only one dye, such as carbol fuchsin, gentian violet or methylene blue. Simple staining is traditionally used to examine the morphology of the bacteria: on microscopic examination, the bacteria are stained more intensely than the background. Differential staining uses more than one dye and is useful for identifying bacteria based on their reaction to the stains. Examples of differential staining include Gram staining and acid-fast staining to differentiate Gram-positive/negative bacteria and to identify acid-resistant bacteria such as *Mycobacterium* sp. and *Nocardia* sp.

Bacterial and viral cultivation is an important method for detecting microbes. While many bacteria from clinical samples can be grown in artificial media, viruses and some bacteria (such as *Chlamydia* spp. and *Rickettsia* spp.) need to be cultivated in living cells. This often hampers the confirmation of a disease caused by those organisms. In clinical laboratories, biochemical testing and sensitivity testing are often performed following bacterial culture for determining the specific diagnosis and for choosing the appropriate antibiotics for treatment.

Serological assays are an indirect approach to diagnosis based on the specificity of the immunological response, and enable a clinician to diagnose infection in an individual patient and to study the epidemiology of microbes in host populations. These assays involve the detection of specific antibodies (IgM and IgG) or antigen.

Serological testing is based on the finding that specific antibodies are generated as a reaction to an infectious agent.[63, 64] This serological reaction can be measured qualitatively or quantitatively. The advantage of performing serological testing is that it can be performed quickly and relatively inexpensively. However, such assays have the potential to yield false positive or negative results, as discussed in the previous section.

Arguably, the most revolutionary advance to date in the biomedical sciences was the discovery of nucleic acids as the source of genetic information on the precise characterisation of an organism. The ability to detect pathogens' nucleic acid and to determine their nucleotide sequences created a powerful diagnostic means in the field of microbiology and infectious diseases.[11] Molecular tests, such as PCR, multiplex PCR, quantitative PCR, DNA sequencing and hybridisation techniques, were later developed for the detection of bacterial and viral genetic material.

2.4.2 Guidelines for determining disease causation

As a consequence of the technological advances that have facilitated pathogen detection and discovery, the number of putative pathogens reported has increased. At the same time, it is recognised that numerous microbes are normally associated with healthy humans, and indeed are required for our wellbeing.[65, 66] This has given rise to the question of how to distinguish between pathogenic microbes and commensal organisms. This issue is discussed further below.

In the 1880s, Koch[67] postulated the core principles that define the aetiologic role for a potential pathogen, which are known as 'Koch's postulates'. According to the postulates, to identify a pathogen as the causative agent of a particular disease, the following conditions must be met:

i. The pathogen must be present in all cases of the disease.
ii. The pathogen can be isolated from the diseased host and grown in pure culture.
iii. The pathogen from pure culture must cause the disease when inoculated into a healthy, susceptible host.
iv. The pathogen must be isolated from the new host and shown to be the same as the originally inoculated pathogen.

Koch's postulates have standardised the discovery of human pathogens by establishing a causal relationship between a microbe and a disease. In the nineteenth century, microbiologists such as Louis Pasteur, George Miller Sternberg, Robert Koch, Edwin Klebs and Richard Pfeiffer reported newly discovered microbes (e.g.,

Streptococcus pneumoniae, *Mycobacterium tuberculosis*, *Corynebacterium diphtheriae*, *Vibrio cholerae* and *Haemophilus influenzae*) after their success in isolating and culturing pathogens. During this era, the most common methods for bacterial identification included Gram staining, culture and biochemical tests.

The limitations of Koch's postulates were immediately identified for some diseases of which the pathogenesis could not be reconciled with the postulates. For instance, the first postulate, that 'the pathogen must be present in all cases of the disease', cannot be fulfilled in the case of disease caused by endotoxin. In this instance, the disease occurs without the presence of the pathogen because endotoxin is released by Gram-negative bacteria following the lysis of the bacterial cell wall.

The second Koch's postulate, that 'the pathogen can be isolated from the diseased host and grown in pure culture', is violated where an individual can be a healthy carrier of a pathogenic organism. For example, *V. cholerae, M. tuberculosis* and *Neisseria meningitidis* can be isolated from patients with diseases and can also be present in healthy subjects.[68] This postulate can also be questioned when the disease occurs due to the capability of the microbe to modulate disease-associated genes. For example, *S. pneumoniae* represents normal flora in the human nose, but can rearrange its genes to cause serious cases of pneumonia;[69] thus, this organism can be found in both healthy and diseased people. The most obvious violation of the second Koch's postulate is when the microbes cannot be grown in pure (lifeless or cell-free) culture; this applies to viruses, chlamydiae and rickettsiae, which require other cells for propagation. In addition, the inability to isolate *Plasmodium falciparum* and *M. leprae* from pure cultures also prevents the fulfilment of Koch's second postulate.[68]

The third and fourth Koch's postulates, that 'the pathogen from pure culture must cause the disease when inoculated into a healthy, susceptible host' and that 'the pathogen must be isolated from the new host and shown to be the same as the originally inoculated pathogen', are violated in cases of limited host availability for a particular microbe. For example, HIV hosts are restricted to humans, and it would be highly unethical to infect and reisolate HIV from an experimental human host. Although simian immunodeficiency virus (SIV), which infects monkeys and chimpanzees, bears a very close resemblance to HIV, these animals are not an ideal model for studying HIV because of the lack of pathogenic consequences despite infection.[70, 71]

As a consequence of the limitations outlined above, Koch's postulates clearly need reconsideration. Indeed, as Evans noted:

> failure to fulfil the Henle-Koch postulates does not eliminate the putative microbe from playing a causative role in a disease. It did not at the time of Koch's presentation in 1890 and it certainly does not today. Postulates of causation must change with the technology available to prove them and with our knowledge of the disease.[72]

The application of new technologies and advances in knowledge in the field of microbiology and infectious disease have indeed led to the revision of Koch's postulates for defining the causal relationship between a microbe and a disease. In 1937, Rivers[73] contended that Koch's postulates are not satisfied for viral diseases. However, he believed that certain conditions still must be met before the specific relation of a virus to a disease is established. Rivers' conditions included:

i. a specific virus must be associated with a disease with a degree of regularity;
ii. the virus must be shown to occur in the sick individual not as an incidental or accidental finding but as the cause of the disease under investigation.

Rivers' postulates differ from Koch's in that (i) the pathogenic virus does not need to be present in every case of the disease produced by it; (ii) the existence of the 'carrier state' with respect to viral disease is recognised; and (iii) the requirement for viral cultivation is abandoned.

Another guideline for establishing a causal link between a virus and a disease was proposed by Huebner in 1957.[74] Huebner recognised that some viruses can cause chronic or latent infections in humans and that simultaneous multiple viral infections are extremely common. Therefore, he introduced epidemiological and immunological aspects into the criteria used to judge disease causation. According to Huebner, in order that a virus be regarded as the cause of specific illness, the following factors are essential:

i. The virus should be real; that is, it should be well established on animal or tissue-culture passage.
ii. The virus must originate in the human specimens, not in the experimental animals, cells or media employed to grow it.
iii. The agent should evoke antibody response, as shown by an increase in neutralising antibodies or other serological tests.
iv. The virus should be characterised and compared with known agents, particularly immunological characterisations and comparisons.
v. There is constant association between the agent and the specific illness.

vi. In double-blind studies, the agent should reproduce clinical manifestation(s) consistent with the naturally occurring illness.

vii. Epidemiological studies are necessary to establish the etiological role of highly prevalent viruses in human disease.

viii. Prevention of the disease by specific vaccination is one of the best ways to establish the causal link between an agent and a disease.

In addition to these factors, Huebner also highlighted that viral research is very expensive, and thus that financial support is absolutely necessary for proving causal links between virus and disease. In fact, Huebner suggested that this financial factor deserves to be called a postulate.

Integrative guidelines considering environmental factors in the development of a disease were proposed by Hill, 1965.[75] Hill proposed nine epidemiological criteria for evaluating a possible causal relationship, including:

i. strength of the association between the cause and the disease;

ii. consistency of the association observed by different persons, in different circumstances and times;

iii. specificity of the association: that is, that the association is unique to particular types of disease;

iv. temporality of the association: that is, that exposure precedes the outcome;

v. biological gradient: that is, evidence of a dose-response relationship;

vi. plausibility: that is, that the association between the cause and the disease is biologically plausible;

vii. coherence: that is, that the causal association is compatible and does not conflict with existing knowledge of the disease;

viii. that experimental or semi-experimental evidence is available to support the causation hypothesis;

ix. analogy: that is, that the causal relationship conforms to a previously described relationship.

When serological assays became widely used for investigating diseases, a causal link between microbe and disease could be drawn based on the purification of viral antigen and detection of specific antibody. Following this development, Evans proposed his ‘Elements of Immunological Proof of Causation’, which were derived from experience with Epstein-Barr virus (EBV), where the causal relationship between the virus and Burkitt’s tumour was assigned on the basis of the population-based

features of sero-reactivity to EBV antigen.[76, 77] Evans' criteria for establishing causal relationships are as follows:

i. Antibody to the agent is regularly absent prior to the disease and exposure to the agent (i.e., before the incubation period).
ii. Antibody to the agent regularly appears during illness and includes both IgG and IgM classes.
iii. The presence of antibody to the agent indicates immunity to the clinical disease associated with primary infection by the agent.
iv. The absence of antibody to the agent indicates susceptibility to both infection and the disease by the agent.
v. Antibody to no other agent should be similarly associated with the disease unless it is a cofactor in its production.

The discovery of 'slow virus' infections of the nervous system, such as the agents of Kuru and Creutzfeldt–Jakob disease, posed a challenge for establishing disease aetiology using the available guidelines. The agents could not be seen, grown in the laboratory or monitored by serological responses, and thus failed to fulfil Koch's postulates or to meet many of the criteria for disease causation by viruses suggested by Rivers and Huebner. It is now known that Kuru and Creutzfeldt–Jakob disease are caused by prions, which are infectious, proteinaceous particles. Unlike viruses, prions lack nucleic acids. In order to establish a causal link between prion and disease, one may consider the criteria proposed by Walker in 2006:[78]

i. The protein must be invariably present in a disease-specific form and arrangement in the diseased tissue. It is necessary to confirm the presence of protein accumulation (such as in the form of inclusion bodies or extracellular protein deposits) in certain tissues.
ii. The physicochemical characteristics that confer infectivity on a specific protein must be established. The characterisation of the infectious protein includes the primary amino acid sequence, the secondary, tertiary, and quaternary structure, and post-translational modifications or processing.
iii. The genetic, biochemical and cellular characteristics that render the host susceptible to infection by a specific proteinaceous agent must be established. The most crucial characteristic to be determined is the amino acid sequence of the host protein.

iv. The disease process must be induced in a susceptible organism by the pure agent in its infectious form. However, producing a purified and infectious proteinaceous agent remains a challenge.

v. The protein must be recovered in its infectious form from the animal that was experimentally infected with the purified agent.

The emergence of nucleic acid amplification assays and sequencing technologies has enabled the identification of novel and previously uncharacterised microorganisms, including those that cannot be propagated *in vitro*. In this situation, Koch's postulates cannot be fulfilled because disease cannot be replicated. Further, when used under conditions of maximum sensitivity, nucleic acid amplification methods might detect even a small number of pathogenic microorganisms that may occur in the absence of pathology, thereby calling into question the specificity of the parasite–disease association that is demanded by Koch's postulates. Fredricks and Relman[68] offered guidelines for the establishment of causal relationships between microbes and diseases based on the detection of nucleic acid sequences. These criteria are as follows:

i. A nucleic acid sequence identifiable as belonging to a putative pathogen should be present in most cases of an infectious disease. Microbial nucleic acids should be preferentially detectable in sites known to be diseased and not in organs that lack pathology.

ii. Fewer or no copy numbers of pathogen-associated nucleic acid sequences should be detected in hosts or tissues without disease.

iii. With resolution of disease, the copy number of pathogen-associated nucleic acid sequences should decrease or become undetectable. In cases of clinical relapse, the opposite should occur.

iv. When sequence detection predates disease, or sequence copy number correlates with severity of disease or pathology, the sequence-disease association is more likely to be a causal relationship.

v. The nature of the microorganism inferred from the available sequence should be consistent with the known biological characteristics of that group of organisms.

vi. Tissue-sequence correlates should be sought at the cellular level: efforts should be made to demonstrate specific in situ hybridisation of microbial sequence to areas of tissue pathology and to visible microorganisms or to areas where microorganisms are presumed to be located.

vii. The above mentioned evidence for microbial causation should be reproducible.

Koch's postulates have been superseded in the twenty-first century as a result of the wide use of advanced molecular technologies for pathogen identification. In his review, Lipkin[79] classified the proof of causation into three different categories, that is, possible, probable and definite causal relationship. Lipkin also discussed the challenges of pathogen discovery in relation to the limitations of the investigation methods and the complexities of pathogenic mechanisms that make it difficult to establish causal links, as previously discussed.

A staged strategy for discovering viruses in particular was also proposed by Lipkin.[79] This strategy involves culture and molecular methods, which are performed in parallel to detect a novel agent. *In vitro* culture and multiplex PCR are conducted as the first attempts to discover new microbes. If there is no growth *in vitro*, culture can be performed *in vivo.* If there is no yield in multiplex PCR, the investigation continues using DNA microarray and unbiased high-throughput sequencing, consecutively. Positive culture and positive molecular testing can be followed by experiments using animal models and pathology testing, such as *in situ* hybridisation or immunohistochemistry. Detailed characterisation of the agent can be performed using electron microscopy (viruses) and sequencing. Epidemiological studies can be further performed using PCR and serology. Ultimately, clinical trials for developing vaccines and therapeutics can be performed after the causal link between pathogen and disease has been proven.

Finally, a modification to Koch's postulates in the era of metagenomics has been well presented by Mokili et al,[80] and includes the following criteria:

i. The metagenomic traits (such as sequence reads, assembled contigs, genes or full genomes) in diseased subjects must be significantly different from those of healthy subjects.

ii. Inoculation of samples from a diseased animal into the healthy control must lead to the induction of the disease state in the latter. Comparison of the metagenomes before and after inoculation should suggest the acquisition or increase of new metagenomics traits. New traits can be purified by methods such as serial dilution or time-point sampling of specimens from a diseased animal.

iii. Inoculation of the suspected purified traits into a healthy animal will induce disease if the traits correspond to the aetiology of the disease.

From the description above, it can be concluded that detection of a putative pathogen is only the first step in understanding the cause of a disease. There is no set of criteria that can provide absolute proof of causation, but the various sets of guidelines discussed above can and should be used to carefully weigh evidence. Some detailed examples of the application of various methods for pathogen discovery are presented in the following section. The discussion is limited to the discovery of viruses to illustrate the complexity of investigations aiming to establish causal links between viruses and diseases.

2.4.3 Viral discovery: challenges and success stories

Early evidence of human diseases caused by viruses can be found in ancient records. Polio, smallpox and influenza are some examples of 'ancient' diseases caused by viruses. Clues to the existence of poliovirus infection exist in Egyptian hieroglyphics dating back to 1300 BC, depicting a young man with typical clinical signs of paralytic poliomyelitis.[79] Smallpox skin lesions are observed in the well-preserved mummy of Pharaoh Ramses V, King of Egypt, who died in 1157 BC.[81] Although human diseases such as poliomyelitis and smallpox—later shown to be caused by viruses—have been recognised since ancient times, the discovery of viruses as a special class of microorganisms occurred at the end of the nineteenth century.

In the 1880s, a German scientist, Adolf Mayer, discovered that extracts taken from the mottled leaves that develop on tobacco plants were able to transmit the same mottled appearance to new plants when rubbed onto the surface of a leaf or injected into the phloem.[82] This condition was later shown to be caused by the tobacco mosaic virus (TMV), and the virus particles were isolated in 1935.[83] The first human virus to be identified was the agent of yellow fever disease. This virus was identified through transmission experiments in 1901.[84] Between 1901 and 2005, 188 virus species have been reported to infect humans.[85]

The majority of viruses are smaller than the highest resolution of a light microscope and cannot be retained by filters with pore sizes of 1–5 microns in diameter. In addition, viruses are unable to grow in cell-free culture media because they rely on living cells for their propagation. A significant leap forward in viral research occurred with the development of virus cultivation in eggs in 1931.[86] Another milestone in microbiology research in the twentieth century was the invention of the electron microscope in 1938. The development of electron microscopes and membrane filters

led to further studies of the size and shape of viruses. In addition, the introduction of cell culture technologies using flasks in the 1920s contributed to viral isolation and the development of vaccines. However, many viruses (e.g., all hepatitis viruses and human rhinovirus C) are not easily cultured, thus making their identification difficult;[87, 88] more importantly, uncultivatable viruses pose a challenge for determining disease causation according to Koch's and later postulates.

In order to meet these postulates, there are a number of additional challenges in proving that a virus is the aetiologic cause of specific syndromes. Some of these challenges include prolonged viral shedding after acute illness (e.g., enteroviruses); latent infection and asymptomatic shedding (e.g., herpesviruses); clinical disease in a minority of infected individuals (e.g., poliovirus); and recurrent asymptomatic infections of immune adults (e.g., respiratory syncytial virus).[89] In such situations, these postulates are clearly not applicable to the establishment of a causal link between viruses and diseases. Moreover, the application of molecular methods for viral discovery begs a revision of existing guidelines for proving disease causation, as previously discussed.

Prior to the use of molecular techniques, viral discovery often involved the use of several methods, and it was often a long journey to discovery. Classical approaches to the characterisation of novel viruses have involved multiple methods: (i) *in vitro* viral amplification, followed by observation of the virus by electron microscopy; (ii) use of reference serum from previously infected or vaccinated hosts; and (iii) cell culture combined with visual observation for cytopathic effects, followed by testing for immunological cross-reactivity using large panels of sera. Such approaches are labour intensive and time consuming, and are often inadequate for the characterisation of novel viruses due to the intrinsic difficulties in amplifying them in cell culture and limited antigenic/serological cross-reactivity.[90]

Since PCR was developed by Kary Mullis in the late 1980s,[91] it has become a key technique in molecular biology, and is now widely used in research and clinical microbiology laboratories for microbe detection. Isolation of a virus can be performed by inoculating cell culture and detecting viral ribonucleic acid (RNA) or deoxyribonucleic acid (DNA) by PCR assay, if the sequence is known. Subsequently, electron microscopy and nucleic acid sequencing can be used to further characterise the pathogen.

The following paragraphs present some examples of the utilisation of viral discovery methods. These examples include the discovery of the yellow fever virus, Japanese encephalitis virus, West Nile virus, Ross River virus, Hepatitis C virus and coronavirus causing severe acute respiratory syndrome (SARS-CoV virus).

The yellow fever virus was the first human virus discovered. The first epidemic was reported in 1648, but the causal agent was elucidated nearly three centuries later in 1928.[92] Methods that led to discovery of the yellow fever virus included transmission experiments and filtration (1881–1901), animal models (1927) and animal culture (1928).[92, 93] Through transmission experiments involving the introduction of non-immune volunteers to various exposures, in 1901 Reed proved that yellow fever virus was transmitted by mosquitoes.[84] In 1928, 27 years later, Max Theiler developed a method to propagate the virus by inoculating mice intracerebrally.[93]

The first outbreak of the Japanese encephalitis virus occurred in Japan in 1871, but the Nakayama strain of the virus was not isolated until 63 years later (in 1934) via the cultivation of clinical specimens in the brains of suckling mice. In 1950, mosquitoes were recognised as vectors of this virus.[94]

The discovery of the West Nile virus was reported by Smithburn et al[95] in 1940, three years after the collection of a blood sample from an African woman in 1937. The steps in discovery included filtration, animal inoculation, immunological and histopathological studies. It was recognised that the agent could readily pass through filters with pore diameters of 79 millimicrons or greater. The filtrates and the original serum were then inoculated into animals via various routes of injection, including intracerebral, intraperitoneal, intranasal, intracorneal, subcutaneous and intravenous. The virus caused fatality in mice and rhesus monkeys (*Macaca mulatta*), but did not induce encephalitis in the African monkeys (*Cercopithecus ethiops centralis*), rabbits, guinea pigs and hedgehogs. Neutralisation tests using acute and convalescent serum of the same African woman indicated that the virus was derived from her blood and was not obtained accidentally. Further immunological studies showed that antibodies against this virus were different from those against viruses causing St Louis encephalitis, Japanese B encephalitis and louping ill. Histopathological studies on visceral organs and brain sections of mice and monkeys showed that the West Nile virus was neurotropic.

The first reported epidemic of polyarthritis due to Ross River virus occurred in 1928 in Australia. After a second epidemic of polyarthritis occurred in the Murray

Valley in 1956, investigators noted the similarities and differences between the clinical manifestations of Ross River virus and Chikungunya infection. Following an extensive immunological study conducted by Anderson[96], it was concluded that the cause of this epidemic polyarthritis was an unknown group A arbovirus; subsequently, the Ross River virus was isolated in 1972, following several passages of intracerebral inoculation of mice. There was a 44-year gap (1928–1972) between the first reported syndrome and the identification of Ross River virus.

Non-A, non-B viral hepatitis (NANBH) was first identified in 1975 from cases of transfusion-associated hepatitis characterised by the lack of serological markers of either Hepatitis A or B viruses.[97, 98] However, epidemiological studies demonstrated the sporadic occurrence of the disease in an urban US population and in the absence of transfusions.[99] Studies in chimpanzees allowed researchers to understand the nature of NANBH infection,[100] while the invention of tissue culture methods in conjunction with electron microscopy, as well as the development of molecular techniques, accelerated the discovery of the Hepatitis C virus. Six years (1982–1988) of intensive investigations using the chimpanzee model, tissue culture, electron microscopy, hybridisation techniques and immunoscreening methods enabled the isolation of a single cDNA clone of a novel Flavivirus, designated Hepatitis C virus.[101]

Coronavirus causing severe acute respiratory syndromes (SARS-CoV) is a virulent and highly transmissible virus that first emerged in southern China in the autumn of 2002. Within a few months, it had spread to more than 30 countries. Tissue culture, immunohistochemistry, electron microscopy and reverse transcriptase PCR (RT-PCR) assays were employed to discover the aetiologic agent of SARS.[102] The application of sequencing technologies and phylogenetic analyses led to the identification of the agent causing SARS.[103, 104]

From the illustrations above, it is clear that molecular techniques have played an important role in facilitating the discovery of pathogens such as the Hepatitis C and SARS-CoV viruses. These techniques have reduced the temporal gap between the recognition of the initial symptoms and the identification of the causal agent. Between the discoveries of the yellow fever and SARS-CoV viruses, this gap has been reduced from 300 years to less than a month. Current molecular technologies are able to simultaneously detect multiple pathogens present in a single sample using techniques such as multiplex PCR.

A method for the comprehensive analysis of viral presence in biological specimens is now available using long oligonucleotide (70-mer) DNA microarray, which has the potential to simultaneously detect hundreds of viruses. The 'pan-viral microarray' not only allows detection of known viruses, but also the discovery of new viruses, if their nucleic acid sequences are sufficiently similar to allow cross-hybridisation with probes for known viruses and/or conserved higher taxa (e.g., viral families). However, this does not allow for the detection of as-yet undescribed viruses whose nucleotide sequences do not share homology to previously known agents.[90, 105] Moreover, the array-hybridisation approach requires constant refinement to cope with rapid mutation of viruses, which can result in failure of hybridisation.[104]

Since the identity of a pathogen is often not known *a priori*, a random, unbiased, and sequence-independent method for 'universal' amplification has been a significant advance for pathogen discovery.[106] Various sequencing platforms have been developed to meet these demands. The field of viral discovery has greatly benefited from the advancement of nucleic acid sequencing technologies, resulting in a significant increase in papers reporting discoveries of new viruses over the last decade.[80]

A recent study incorporating NGS and reverse-transcriptase quantitative PCR (RT-qPCR) has successfully identified a previously unknown bornavirus infecting three patients in Germany. Those patients were breeders of variegated squirrels (*Sciurus variegatoides*) and had progressive encephalitis or meningoencephalitis that led to fatality within 2–4 months of the onset of clinical symptoms. With the use of metagenomic approach, the virus causing disease was detected in brain samples of the patients and in a contact squirrel. The virus was tentatively named variegated squirrel 1 bornavirus (VSBV-1).[107]

2.5 Next-generation sequencing (NGS)

2.5.1 Development of sequencing technology

Ever since DNA was discovered as the genetic code of all biological life on earth, scientists have sought methods for reading it to better understand the origins of life. The structure of DNA consists of sugar, phosphate and nucleotides. There are four nucleotide bases: adenine (A), guanine (G), cytosine (C) and thymine (T). The positions of these nucleotides in the DNA follow a certain rule: G always matches to C, and T

always matches to A. Therefore, the complement strand of a known single DNA strand can be predicted. Each nucleotide sequence is unique to a particular organism, such that an organism can be distinguished from others based on the sequence of its DNA.

The first nucleotide sequencing technology was introduced in 1977 by Fred Sanger and colleagues.[108] In Sanger sequencing, as it is known, a DNA sequence is determined by selective incorporation between a single-stranded DNA template and dideoxynucleotides triphospate (ddATP, ddGTP, ddCTP or ddTTP) for 'chain termination' in the presence of primer, DNA polymerase and standard deoxynucleotides triphosphate (dATP, dGTP, dCTP or dTTP). Dideoxynucleotides lack the molecules required for the formation of a junction between two nucleotides, causing discontinuation of DNA extension and generating fragments of different lengths ending in ddATP, ddGTP, ddCTP or ddTTP. In Sanger sequencing, the DNA polymerase facilitates the extension of the template DNA from the bound primer; this process is conducted in four separate reactions, with each reaction containing one of the four dideoxynucleotides and three other standard dideoxynucleotides. The resulting DNA fragments are separated by size in a gel electrophoresis with each of the four reactions run in one of four individual lanes (lane A, T, G and C). The DNA bands are then visualised by autoradiography or ultraviolet light, and the DNA sequence can be directly read from the X-ray film or gel image as the complement of the bands containing labelled strands.

Automated Sanger sequencing was developed by Leroy Hood and co-workers.[109] The dideoxynucleotydes are labelled by fluorescent dyes that permit sequencing in a single reaction. The resulting DNA fragments pass through capillary electrophoresis, and the fluorescence is detected in four-colour plots representing four different nucleotides (A, C, T and G). The Sanger sequencer is considered the 'first-generation' sequencing technology, and was the only DNA sequencing method for nearly three decades (1977–2004).[110] This technology has the capacity to read through long DNA fragments (up to 1200 bp), which aids pathogen discovery;[111] however, its low throughput makes it unsuitable for assisting the diagnosis of severe infectious diseases, particularly during an outbreak.

The advent of the shotgun sequencing technique accelerated the generation of sequencing data by combining new sequencing methods with computational analysis.[112] In the shotgun sequencing technique, very long DNA or RNA fragments (millions to billions of base pairs in length), or potentially the entire genome of an organism, are

broken up randomly into several small overlapping segments. These smaller segments are then sequenced and reassembled to reconstruct the whole sequence of the original genome using computer programs called genome assemblers. Shotgun sequencing was the most advanced technique for genome sequencing in the period 1995–2005. Although the shotgun sequencing technique accelerated the completion of the human genome sequencing project, it took over ten years to complete a human genome sequence (using the Sanger sequencer). Sanger sequencing was used to obtain the first consensus of the human genome in 2001, and the first individual human diploid sequence (Craig Venter) in 2007.[111]

A revolution in sequencing technology began in 2004 with the launch of the 454 GS20 (followed by the GS-FLX in 2005) platform developed by Roche. The second complete genome of an individual (James D. Watson) was sequenced using this platform.[111] Following the invention of the 454 platform, other platforms were introduced, including Solexa/Illumina in 2006 and the SOLiD technology in 2007.[112–114] These sequencing technologies are referred to as 'next-generation' sequencing (NGS) platforms, and are now also available as benchtop instruments, including the 454 GS Junior (Roche), MiSeq (Illumina) and Ion Proton and Ion Torrent Personal Genome Machines (Life Technologies). With increasing speed and rapidly decreasing costs, it is now possible to conduct sequencing projects with relative ease. Previous reviews[9–11, 111, 114–116] have discussed each platform in detail; Table 2.1 summarises the features of the various NGS platforms currently available in the marketplace.[111, 114, 115, 117–119]

Table 2.1: Next-generation sequencing platforms

Platform (manufacturer)	Chemistry	Read length	Output volume	Run time	Advantages	Limitations
454 Genome Sequencer (GS) FLX (Roche)	Pyro-sequencing	mode 700 bp up to 1 kb	1 Gb	23 hours	Long reads	High homopolymer error rate
454 GS Junior System (Roche)	Pyro-sequencing	500 bp	35 Mb	10 hours	Dekstop machine; longest read length of desktop sequencers	High homopolymer error rate; lower depth compared to GS-FLX
HiSeq (Illumina)	Reversible terminator	36–150 bp	95–600 Gb	40 hrs-11 days	High-throughput/ low cost; low error rate; paired-end reads	Short reads; long run time; decreasing read quality towards ends
MiSeq (Illumina)	Reversible terminator	25–300 bp	13–15 Gb	5 hrs	Dekstop machine; lowest error rate of dekstop sequencers; low cost; paired-end reads; short run time	Short reads; decreasing read quality towards ends
SOLiD (Life Technologies)	Ligation	50–75 bp	120–160 Gb	8 days	Low error rate	Short reads; long run time
Ion Proton (Life Technologies)	Proton detection	< 200 bp	10 Gb	2–4 hr	Dekstop machine; short run time	Short reads; chimeras; homopolymer errors
Ion Torrent with Personal Genome Machine (PGM) (Life Technologies)	Proton detection	200–400 bp	30 Mb–1 Gb	2–7 hr	Dekstop machine; short run time	Short reads; homopolymer errors

2.5.2 Principle of NGS

The NGS technology involves various strategies that rely on the combination of three steps, as illustrated in Figure 2.2. These steps are library preparation, sequencing and imaging and data analysis. Library preparation is a key step enabling NGS to produce unprecedented amounts of data. In this stage, the entire genome is broken up into small pieces from which templates are created; those templates are subsequently attached to a solid surface. Each NGS platform uses a different strategy for template immobilisation, but the purpose is the same. The immobilisation of spatially separated template sites allows thousands to billions of sequencing reactions to be performed simultaneously. The second step, sequencing, involves reading the DNA templates randomly along the entire genome. The sequence of the target genetic material is determined using sequence by synthesis (using labelled nucleotides [Illumina platforms] or pyro-sequencing [Roche platforms] or proton [Ion Proton/Ion Torrent platforms] for detection) or sequence by ligation (SOLiD platform). Sequencing is conducted in a massively parallel fashion, and sequence information is captured by a computer. The final step is imaging and data analysis, in which the fragmented DNA is reconstructed to form a complete or nearly complete genome. The data analysis step usually includes genome alignment and assembly using bioinformatics tools.

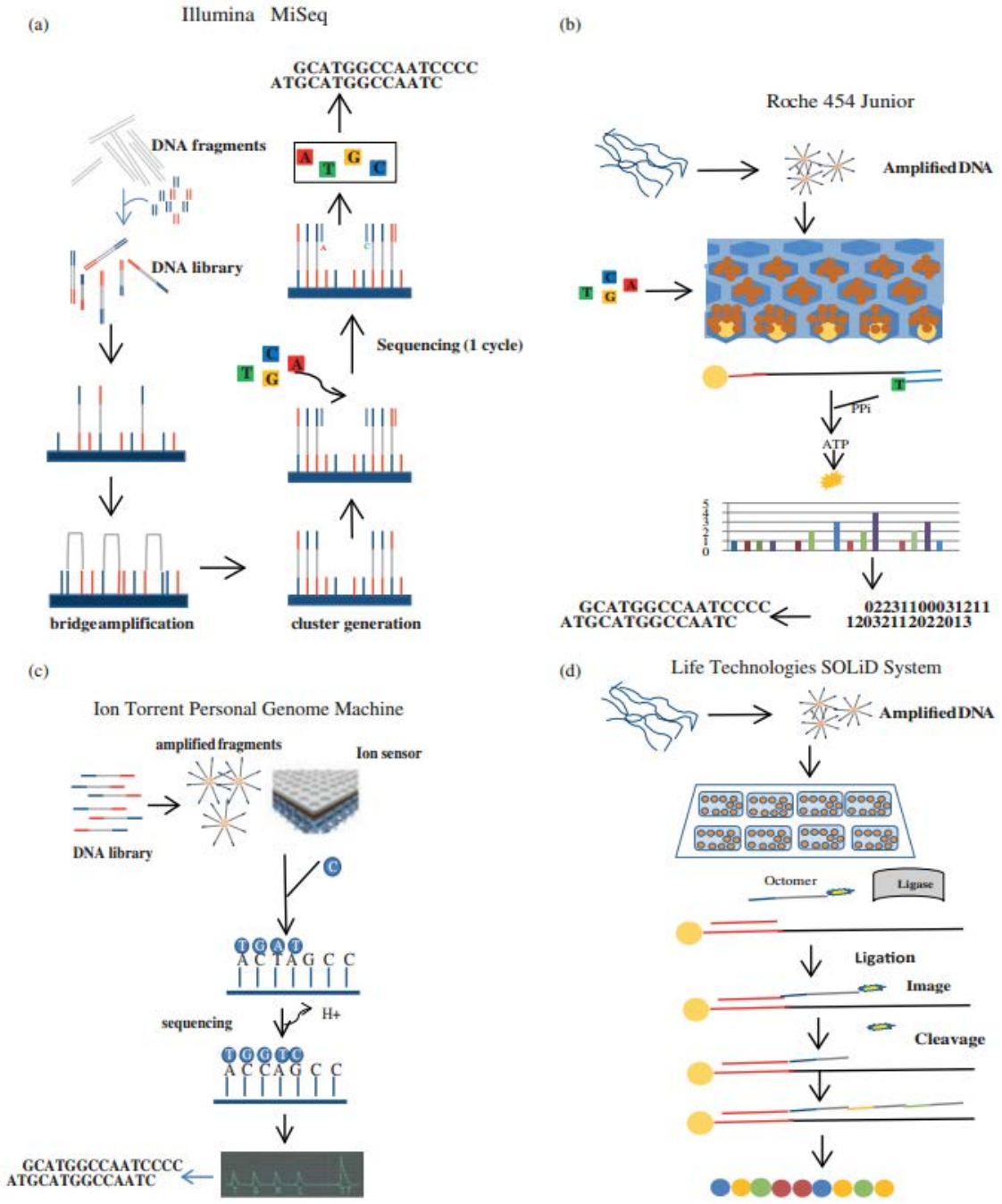


Figure 2.2: Schematic workflow in different NGS platforms

(Reproduced from[118] with permission)

(a) Illumina MiSeq: DNA to be sequenced is fragmented and adapters are ligated to both ends. Fragments randomly attach to the surface of the flowcell lanes coated with primers complementary to that present on the fragment. Unlabeled nucleotides and enzyme are added to initiate solid-phase bridge amplification. Resulting double-stranded DNA fragments are denatured to leave single-stranded fragments tethered to the surface. Several million clusters are generated in each lane of the flow cell. Sequencing is performed in cycles wherein four labeled reversible terminators (indicated by coloured boxes), primers, and DNA polymerase are added. After laser excitation, the emitted fluorescence from each cluster is captured and the nucleotides are identified. (b) Roche 454 GS Junior: DNA to be

sequenced is fragmented, ligated to adapters and denatured to yield single strands. The fragments are bound to beads and amplified using emulsion polymerase chain reaction (PCR). The beads carrying single-stranded DNA fragments are deposited into wells of a Pico titer plate. A mixture of smaller beads that carry bound ATP sulfurylase and luciferase required to generate light from free pyrophosphate are also loaded into the wells. Nucleotides are added in a fixed sequence. Addition of a nucleotide is recorded by the release of photons. The sequence information is generated as a numeric string which needs to be decoded to generate sequence of nucleotides. (c) Ion Torrent's Personal Genome Machine: Adapters are ligated to the ends of DNA fragments. The fragments are then bound to Ion Sphere particles and clonally amplified. These particles are then loaded into the Ion Chip and sequencing performed. Nucleotides are added in a predefined order and addition of a nucleotide is detected as change in voltage due to release of hydrogen ion. The change in voltage is proportional to the number of nucleotides incorporated during each cycle. (d) Life Technologies' SOLiD System: DNA is first fragmented, denatured into single strands and adapters added to the fragments. The fragments are captured on beads and amplified by emulsion PCR. The beads are deposited on glass slides to generate a disordered array of amplified fragments. Sequencing is performed by using DNA ligase and a set of fluorescently labeled octamers. Each octamer queries two adjacent nucleotides simultaneously and each dinucleotide is represented by a colour corresponding to the dinucleotide identity. Multiple cycles are conducted to generate read of fixed length (35–75 bp). The raw data is generated in colour space form that needs to be converted to nucleotide space.

The massively parallel nature of NGS technologies inexpensively produces large volumes of sequence data within a short period. This is the primary advantage of NGS platforms compared to the conventional sequencing method. In fact, these technologies outperform the older Sanger-sequencing technology by a factor of 100–1,000 in daily throughput, and at the same time reduce the cost of sequencing one million nucleotides (1Mb) to 0.1–4% of that associated with Sanger sequencing.[9] It is believed that in the next few years, rapid advances in sequencing technology will provide sequencing tools with faster sequencing speed, improved accuracy and more affordable prices.

Although NGS technology (later called 'second-generation' sequencing) has helped scientists to answer many biological questions, the technology still involves amplification steps, which may introduce bias and lead to false interpretation of sequencing data. To avoid this problem, the 'third-generation' sequencing platforms have been developed. These platforms are capable of directly sequencing a single molecule without the need for an amplification step. Some companies that have developed the 'third-generation' sequencing technology include Pacific Biosciences (PacBio; Menlo Park, CA), Life Technologies (Carlsbad, CA), Oxford Nanopore (Oxford, UK) and Ion Torrent (Gilford, CT).[120] Discussions of 'third-generation' single-molecule sequencing are beyond the scope of this literature review.

2.5.3 Application of NGS in metagenomics studies

A general approach to the study of the total genetic information of microbes and microbial diversity in an environment is called 'metagenomics', a term first used by Handelsman et al in 1998.[121] This approach aims to identify any genetic material present in a given sample by employing non-specific amplification and sequencing of nucleic acid. Metagenomics has been used extensively to identify already-known and novel viruses in seawater, near shore sediments, faeces, serum, plasma and respiratory secretions.[90] This approach has helped scientists to study extraordinarily diverse and previously unexplored microorganisms, which has led to the realisation that there are many organisms yet to be discovered.[122]

The application of NGS in metagenomics has allowed for the discovery of organisms that cannot be grown in culture, and has uncovered an unexpected scale of biodiversity. It has been reported that 200 litres of surface seawater contain more than 5,000 different viruses.[13, 123] Studies conducted by Breitbart et al[124, 125] found more than

1,000 different viruses in human faeces, of which the majority were new species, including plant viruses. NGS technology has also been applied in hospital environments to reveal substantial microbial diversity on inanimate surfaces.[126]

Recently, NGS technology has been used to detect infectious agents in clinical specimens (see Table 2.2). This approach is commonly defined as ‘deep sequencing’, which refers to a ‘needle-in-a-haystack’ approach involving the analysis of millions of sequences derived from nucleic acids present in clinical specimens, in order to detect rare sequences corresponding to candidate pathogens. Given the low amounts of input nucleic acids in clinical samples, a universal amplification is typically performed during NGS library generation. The high-throughput nature of NGS makes it a potential tool for identifying both known but unexpected agents and highly divergent novel agents. This technology is thus particularly attractive for the identification of novel emerging viruses, which can exhibit high inherent sequence diversity and rapid rates of mutation, recombination or reassortment.[106] Unprecedented amounts of sequence data can be analysed to show the relation of newly discovered viruses to other known viruses, resulting in the characterisation of novel pathogens.

Table 2.2: Recent applications of next-generation sequencing technology for detecting infectious agents causing fever

Sample	Detection platform	Bioinformatics approach[a]	Agents detected	Disease association (reference)
Serum	Illumina	Subtraction and BLAST search	SFTS (severe fever with thrombocytopenia syndrome) virus, a bunyavirus	Severe fever with thrombocytopenia syndrome[127]
Cell culture media infected by patients' blood	454	BLAST search and *de novo* gene assembly	Heartland virus, a phlebovirus	Severe febrile illness[128]
Serum	Illumina	Subtraction, BLAST search, and *de novo* gene assembly	BASV (Bas-Congo virus), a rhabdovirus	Acute haemorrhagic fever[129]
Serum, cerebrospinal fluid, brain, liver, kidney	454	Subtraction and BLAST search	Arenavirus associated with lymphocytic choriomeningitis virus (Dandenong arenavirus)	Fatal febrile illness in transplant patients[15]
Serum and tissues	454	Subtraction and BLAST search	Lujo arenavirus	Acute haemorrhagic fever[130]
Cell culture media infected by patients' serum	454	BLAST search and *de novo* gene assembly	Zungarococha virus	Acute febrile illness[131]
Serum	Illumina	Subtraction and BLAST search	Human Herpesvirus 6 and other viruses	Dengue-like illness[132]

[a] Subtraction: computational 'digital' subtraction of host background sequences from NGS data; BLAST: basic local alignment search tool; *de novo:* reference-free

The studies summarised in Table 2.2 show that methods involving the sequencing of entire microbial communities, or metagenomics, in human blood samples seem promising for use in the investigation of fever. This approach offers the opportunity to provide broad-spectrum diagnostic tools for rapid characterisation of previously known or even novel pathogens, which is highly valuable for accurately treating the root cause of fever. In other words, metagenomics studies using NGS technology may yield new insights into the causes of undiagnosed fevers, as these methods are capable of generating unbiased characterisations of pathogens existing in human blood. The evidence that NGS is a practical and powerful tool for diagnosing infectious diseases could lead to a significant contribution to the assessment and management of undifferentiated fever in the future.

2.6 Research questions

Surprisingly, the review of the literature[6] found that a significant proportion (8–80%) of fever cases have gone undiagnosed. Further, there has been no published research on AUF in Australia within a 22 year period (1990-2012), especially in the tropical region of Far North Queensland. This raises the following questions for investigation:

1. How common is AUF in the population of Far North Queensland?
2. How is this condition investigated?
3. What are the frequent diagnoses?
4. What is the proportion of undiagnosed cases and what information exists with regards to this condition?

To answer these questions, a study involving a medical chart audit was conducted; the results are presented in Chapter 4.

When the purpose of a metagenomics study is to produce a description of complete or nearly complete pathogen genomes, a high frequency of pathogen nucleic acids relative to human nucleic acids is required in order to generate large quantities of overlapping sequences. Therefore, specimens with the smallest proportions of human nucleic acids are desirable. Chapter 5 presents a study that was conducted to answer the following methodological questions:

1. What type of blood specimen contains the least human DNA?
2. Does blood sampling technique affect the concentration of circulating DNA?

As there are several NGS platforms available, the decision to use the most suitable platform to investigate undiagnosed fever should consider the balance of the priorities of cost, read length, data volume and rate of data generation. Ideally, a NGS platform that provides long reads will greatly facilitate a reliable bioinformatics assembly of new pathogens where a reference sequence is not available. The ideal platform should also be high-throughput but economical for use in routine diagnostics. Thus, the best combination of low operating costs, long reads and high sequencing throughput creates the most desirable method. The main study, presented in Chapter 6, answers the following research questions:

1. Is a deep sequencing approach using NGS technology a reliable method for diagnosing infectious diseases?
2. Is deep sequencing a practical approach for identifying unknown pathogens in human blood?

2.7 Chapter summary

This chapter has discussed various aspects of the essential elements that constitute the background of the study. First, this chapter discussed the pathogenesis of fever and the issues related to AUF identified in previous studies. A review on the epidemiology of AUF presented the common aetiologies of AUF, methods of investigation and diagnostic challenges. This was followed by a discussion of the various methods of pathogen detection and discovery, including the latest developments in sequencing techniques. A comprehensive understanding of the scope of the problem and the potential use of NGS as a broad diagnostic tool in infectious diseases led to the formulation of the research questions. Chapter 3 discusses the methodology and ethical considerations involved in the present studies.

Chapter 3: General Methodology

The main project of this thesis is a prospective study investigating the infectious causes of AUF using NGS technology. This study aimed to assess the capacity and sensitivity of NGS technology for deep exploration and broad-scale characterisation of pathogens associated with fever. To achieve this purpose, the careful selection of study participants and optimisation of specimens were crucial. Therefore, two pilot studies were conducted. This chapter describes the general methodology used in the pilot and main studies, including study design, aims, hypotheses and ethical clearance. Further details on the materials and methods used in each study are included separately in subsequent chapters (Chapter 4, 5 and 6).

3.1 Study design and aims

The first study consisted of an audit of patient notes, aiming to understand the epidemiology of AUF in Far North Queensland, Australia. This study defined the prevalence, aetiologies and investigations of AUF cases presenting to Cairns Hospital (formerly Cairns Base Hospital) over a three-year period, from 1 July 2008 to 30 June 2011. Cairns Hospital is a tertiary hospital in Cairns, Queensland, a regional city on the east coast of Australia with 170,586 inhabitants.[133] As the main referral hospital for Far North Queensland, Cairns Hospital serves a broader population of about 400,000 residents in the surrounding districts and the broader catchment areas of Cape York and Torres Strait.[133, 134] This hospital has an extensive diagnostic capacity, including an advanced laboratory, infectious disease specialists and microbiologists, providing valuable resources for investigations such as that of the cause of undifferentiated fever.

Data for the first study was collected from medical notes. Additional information was sought from various databases to complete the data on individual patients and to verify the data obtained from their medical notes. These databases included AUSLAB® and Auscare®, which record the details of pathology findings; Merlin Web®, which records radiology findings; and Viewer®, which summarises patient details, clinical information, important investigations, significant findings, diagnosis, treatment and follow-up plans. Descriptive statistics and cross-tabulations were performed for presenting data. The proportion of undiagnosed cases was

determined and the information with regards to these conditions was explored, such as clinical and laboratory characteristics, working diagnosis, investigations, findings and follow-up records. The information obtained from this study was used to estimate the sample size and to define the characteristics of patients that could be included in the main study.

The second study included quantification of circulating DNA in blood samples collected from six healthy volunteers using different techniques. In an infectious disease, the proportion of pathogen nucleic acid compared to human DNA is very small; thus it is considered crucial for NGS studies to collect specimens that contain the smallest proportions of background human DNA in order to increase the sensitivity of detection of the pathogen. Because the majority of human nucleic acid is located intracellularly, the NGS study had to use either plasma or serum samples, which are cell-free. This pilot study aimed to compare the DNA levels in plasma and serum and to examine whether blood stasis and the speed of blood aspiration altered the levels of circulating DNA.

Blood samples were collected at the James Cook University (JCU) Cairns Clinical School, located at Cairns Hospital. Blood samples were taken under four different conditions: with or without the use of a tourniquet, and using standard syringes and needles or using a vacuum container system (Vacuette® tubes, Greiner Bio-One). After collection, the samples were transferred in a cooler box to the Queensland Tropical Health Alliance (QTHA) laboratory located at JCU Smithfield, 16 km from Cairns Hospital. Plasma and serum samples were prepared in QTHA laboratory, followed by nucleic acids extraction and DNA quantification. Plasma and serum were prepared by centrifugation of whole bloods for 15 minutes at 2,000 g within 3 hours after collection. Total nucleic acids were extracted from whole blood, plasma and serum using a commercial kit, the High Pure Viral Nucleic Acid kit (Roche Applied Science, catalogue number 11858874001). The concentration of DNA in the samples was measured using a spectrofluorometer (POLARstar® Omega, BMG LABTECH) fitted with a 485/520 nm excitation/emission filter set. The values of fluorescence intensity obtained from the spectrofluorometer were recorded and transferred into a Microsoft Excel spreadsheet. From these, the DNA concentrations were calculated and compared using descriptive statistics and non-parametric tests from the Statistical Package for the Social Sciences (SPSS) version 20. Although this study focused on comparing DNA in cell-free samples (plasma and serum), the DNA level in

whole blood was also measured to demonstrate the level of human DNA in samples that contain cells (red blood cells, white blood cells and platelets). Some pathogens are intracellular or cell-associated, and some baseline data were sought on levels of DNA in whole blood, should blood specimens be considered for use with NGS in the future. The findings of this study justified the most appropriate type of sample and the proper method of blood collection for the NGS study.

In the third and main study, blood samples were collected from patients presenting to Cairns Hospital with AUF. Some patients had already been given specific diagnoses (controls) and some were undiagnosed (study subjects). Plasma and serum samples were prepared at the hospital, but nucleic acid extraction and amplification were performed at the QTHA laboratory. Total nucleic acids from plasma or serum samples were subjected to deep sequencing using a NGS platform available at the Australian Genome Research Facility (AGRF), the Illumina HiSeq 2000. This platform was chosen because it delivers a relatively high volume of data within a relatively short period, thus providing a more economical cost per megabase compared to other platforms (see Table 2.1). Sequence data obtained was analysed using bioinformatics tools to identify the pathogen(s) associated with fever. The results of the bioinformatic analysis were interpreted in conjunction with the available clinical, laboratory and radiology findings to produce the most likely diagnosis.

The reliability and feasibility of deep sequencing as a broad general approach for identifying infectious causes of AUF were evaluated. The success of the project was measured by the ability of this approach to validate the diagnosis obtained using conventional methods in control subjects and its ability to identify unknown (novel) or unpredicted causes of fever in study subjects. The conceptual framework and connections between the three stages of the study are presented in Figure 3.1.

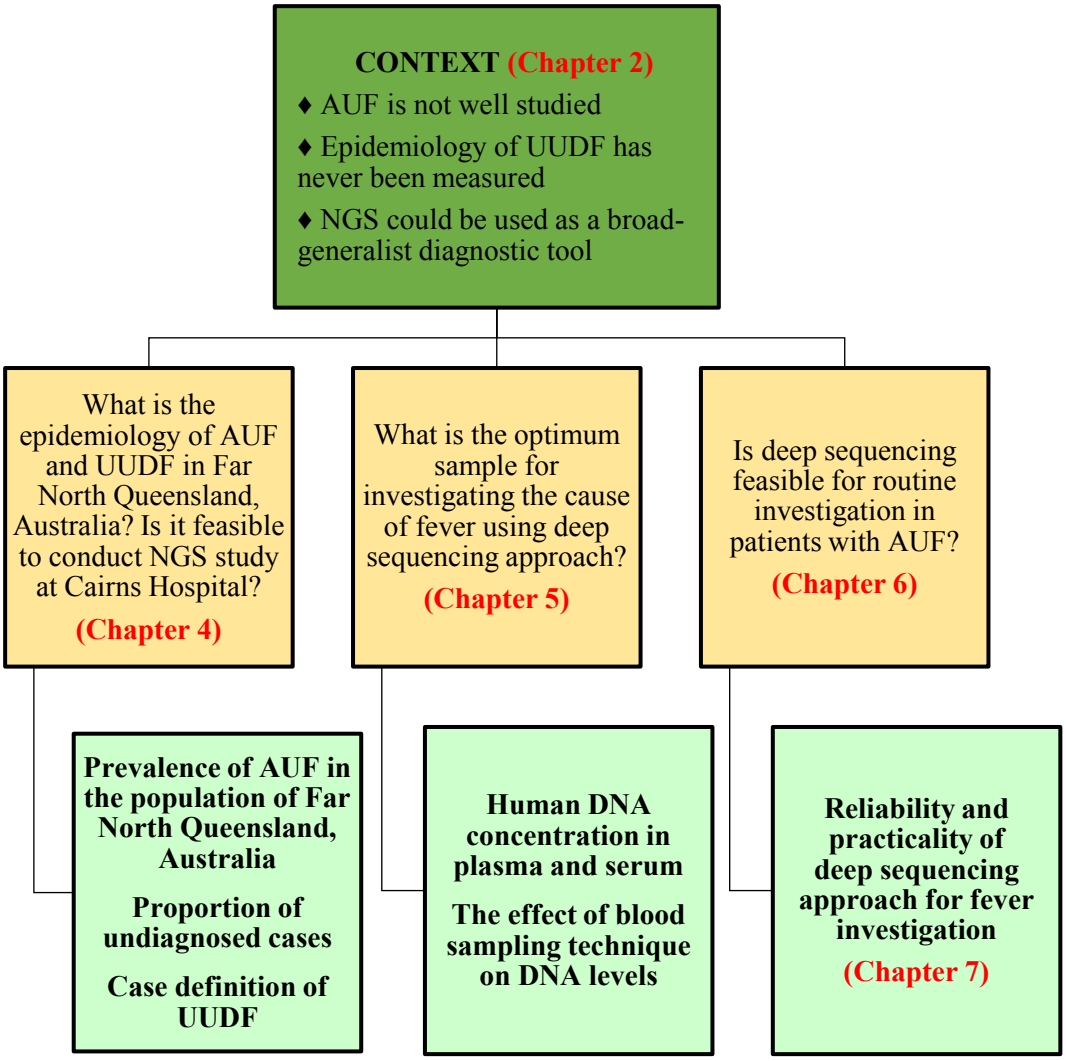


Figure 3.1: Conceptual framework of the study

3.2 Hypotheses

In the first pilot study, the following hypotheses were proposed:

1. AUFs are common in the population of Far North Queensland and a significant proportion of these do not result in a specific diagnosis.
2. UUDFs occur frequently enough at Cairns Hospital to justify a subsequent NGS study looking for unpredicted and unknown (novel) infectious agents at this site.
3. A robust definition of UUDF can be developed, based on the findings of this study, in order to compare the incidence of this entity across different geographical sites.

The second pilot study tested the following hypotheses:

1. Levels of human DNA in plasma and serum differ.
2. The DNA concentrations in plasma and serum samples are affected by the method of blood collection.

The main study tested the following hypotheses:

1. For those patients for whom a diagnosis can be achieved using contemporary diagnostic methods (controls), the use of NGS technology will identify pathogens that match their diagnosis.
2. The use of NGS technology in patients with UUDF can inform diagnosis by detecting genome sequence(s) of previously known (unpredicted) or unknown (novel) pathogens.

3.3 Ethical clearance

3.3.1 First study: Undifferentiated fever in Cairns (Base) Hospital: a retrospective study

The ethics approval for this study was sought from the Cairns and Hinterland Health Service District Human Research Ethics Committee (HREC). This study was assessed under negligible and low-risk ethical review processes, under which applications are exempt from the full HREC review. According to the National Health and Medical Research Council's (NHMRC) *National Statement on Ethical Conduct in Human Research 2007*,[135] a research project is described as having 'negligible risk' where there is no foreseeable risk of harm or discomfort, and any foreseeable risk is of not more than inconvenience to the participants. NHMRC categorises a research project as 'low risk' if the only foreseeable risk is one of discomfort, such as research that involves the use of existing collections of patients' data or records.

There were two separate processes involved in the ethical clearance of this study (see Figure 3.2). The first was the ethics process, in which the low risk application went to the ethics administrator, who forwarded the documents to the Chairperson of the ethics committee, who then judged whether or not the research could be considered 'low risk'. As the Chairperson deemed it a low risk application, it did not go to a full HREC review. This project was approved under the low risk

approval process by the Chairperson of the HREC and was assigned Cairns and Hinterland Health Service District reference number HREC/11/QCH/102.

A second process, the Site-Specific Assessment (SSA), was a component of research governance and the mechanisms for financial accountability and transparency in Queensland Health. This process included assessment of the suitability of the study site and the resources required for the conduct and completion of the project.[136] Although SSA was a separate process and not associated with ethics process, the HREC approval was a pre-requisite for the SSA submission. On receipt of the ethics approval letter, the completed SSA component was submitted to a Research Governance Officer (RGO) to be processed and signed off by the District Chief Executive Officer (DCEO). This study was given SSA reference number SSA/11/QCH/105 (see Appendix A).

Ethics approval for this study was also sought from JCU by forwarding the external HREC approval and SSA approval to JCU HREC. The Chairperson of the JCU HREC reviewed the documentation, and this study was given JCU approval number H4550 (see Appendix B).

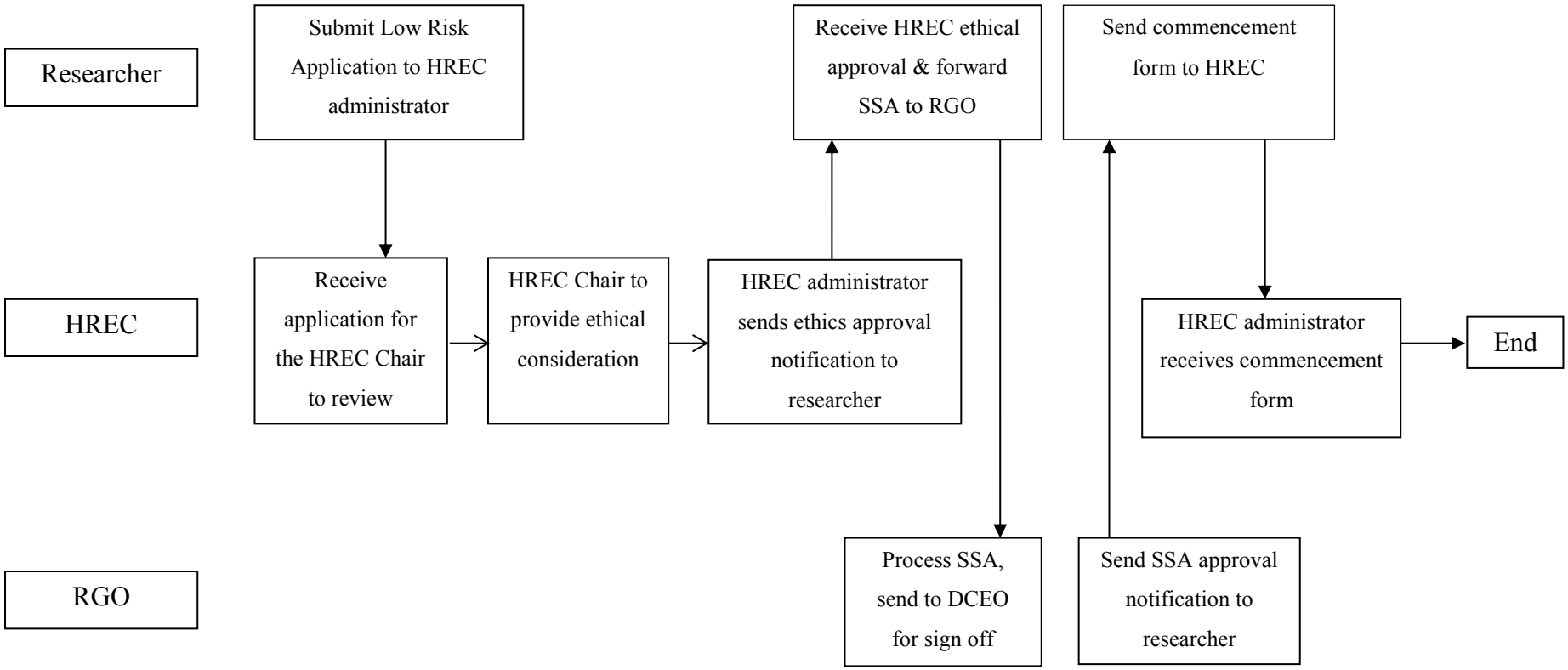


Figure 3.2: Process of submission and authorisation of low risk research

HREC: Human Research Ethics Committee; RGO: Research Governance Officer; SSA: Site-Specific Assessment; DCEO: District Chief Executive Officer

3.3.2 Second study: Quantitative analysis of nucleic acid in serum and plasma

The ethics approval for conducting this experiment was sought from the JCU HREC. This project was assessed under Category 3, which means that the project has the potential to cause mild psychological distress or physical stress. In anticipation of this risk, the participants were informed that there would be some discomfort associated with the use of a needle and vacuum system to collect blood, with bruising as a potential complication. The participants were also advised that free counselling and medical treatment would be provided if they felt distressed or experienced any ongoing discomfort due to their involvement in the project.

The second and third studies included laboratory work for sample processing. Prior to any laboratory work, it was mandatory for JCU students and staff to enrol in biosafety training and to pass an assessment task. The researcher's biosafety training was completed on 16 February 2012 (see Appendix C) and the reference number of biosafety approval for the project was Med 39. The ethics approval number assigned for the second study was H4456 (see Appendix D).

3.3.3 Third (main) study: Evaluation of NGS technology in determining infectious causes of human febrile illness

All ethics applications for research involving patients of a Health Service District, whether they reside in hospital or in the community, must be reviewed by the appropriate Health Service District HREC. The ethics application for the main study went through a formal National Ethics Application Form (NEAF) (https://au.ethicsform.org/SignIn.aspx), which was significantly more detailed than the previous assessments, and went to the full HREC for review. At the committee's recommendation, amendments were made, particularly to the information sheet and consent form. The amendments included simplification of the language, defining next-of-kin who could consent if the patient was incapacitated, and description of the time delay for returning results to patients. Other issues for clarification included who would collect the blood, host DNA storage for mature minors (16–18 years) and incapacitated patients, and sample size. In addition, the HREC was uncomfortable with next-of-kin giving consent for collection of host genetic information, since this has far-reaching implications well beyond the usual power of attorney. The documents were amended; the following Sections 3.4 and 3.5 discuss the strategies employed to address these issues.

It was anticipated that some samples would not be obtained directly from the patient and would need to be retrospectively collected from the Pathology Department. In order to obtain samples from Pathology, an approval from Clinical and Statewide Service (CaSS) was needed. Figure 3.3 provides a flow chart of the ethical clearance process for this third phase of study. The study protocol was approved by Cairns and Hinterland Health Service District HREC (reference number: HREC/12/QCH/7—765; date of approval: 24 February 2012). The SSA was approved on 18 October 2012 (reference number: SSA/12/QCH/102—Lead 85; see Appendix E).

Once the project has received approval from Cairns and Hinterland Health Service District HREC, the ethics documents were then forwarded to the JCU Human Ethics and Grants Administrator. In accordance with the *National Statement on Ethical Conduct in Human Research 2007*,[135] as this project had already been reviewed by an external HREC, the JCU HREC accepted the decisions of the external committee and released a JCU ethics approval after an internal review (reference number: H4882; date of approval: 7 November 2012; see Appendix F).

The process of obtaining ethical clearance and SSA approval for this study was quite challenging and time consuming, it took about one year to complete the process illustrated in Figure 3.3. To make up for the delays, it was anticipated that additional participants would need to be recruited from Cairns Private Hospital. Ethics approval for conducting the study at Cairns Private Hospital was obtained from Greenslopes HREC (reference number: Protocol 13/02; date of approval: 19 March 2013; see Appendix G).

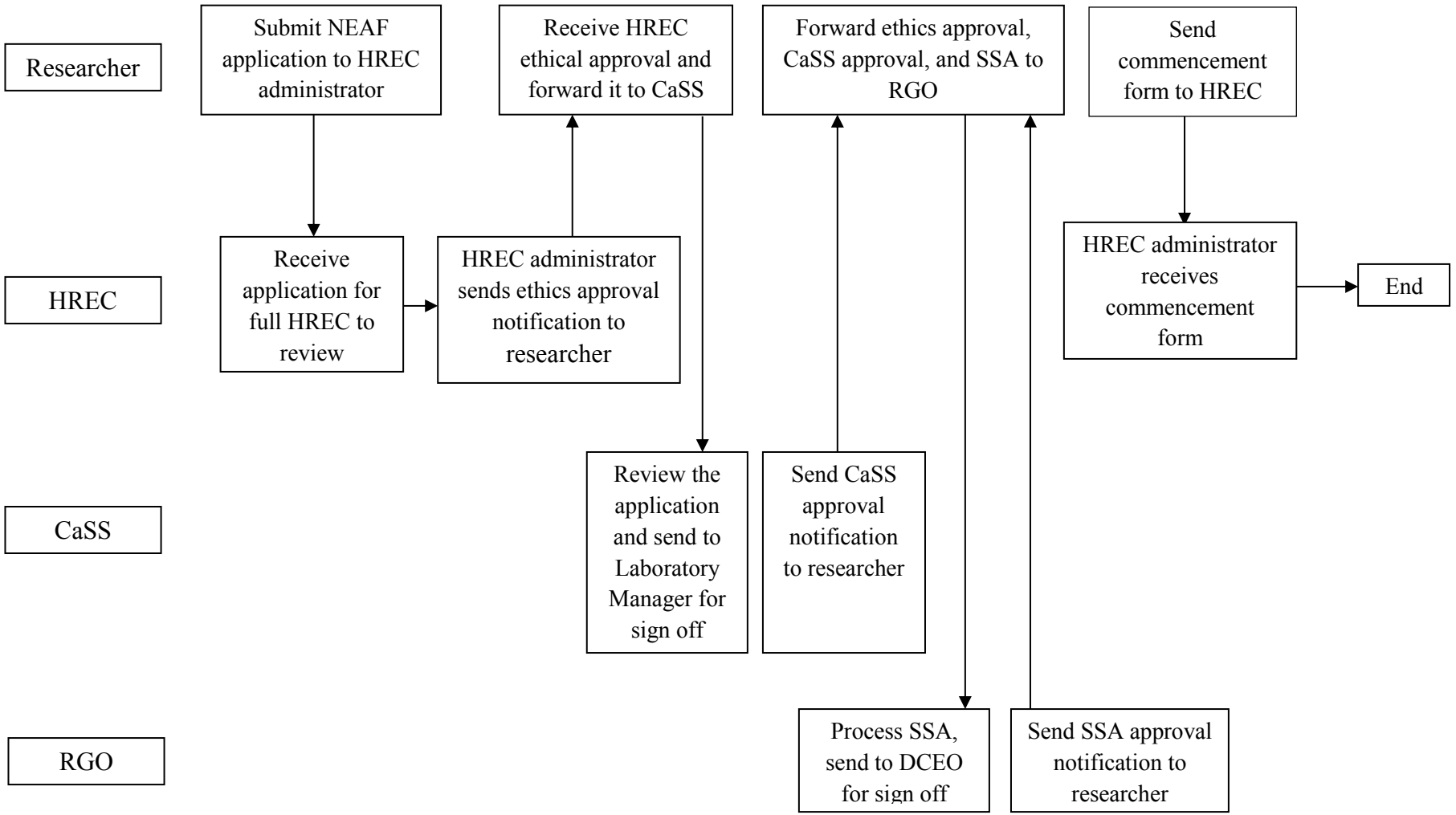


Figure 3.3: Process of submission and authorisation of main study

HREC: Human Research Ethics Committee; RGO: Research Governance Officer; SSA: Site-Specific Application; DCEO: District CEO

3.4 Consent and enrolment

The enrolment process commenced by screening potential participants through the hospital database. The information collected during the screening included clinical data that focused on the evidence of fever, laboratory investigations, radiology findings and working diagnosis. The next step involved approaching the doctor or nurse in attendance to obtain permission to peruse the patient's medical record and take the patient's consent.

Seeking consent for genetic and genomic research can be challenging. Challenges include explaining the details of the research and the potential uses of both samples and data. In order for participants to give truly informed consent to take part in research, researchers must disclose all critical information about a study, and prospective participants should understand the information in terms of potential risks, potential benefits and what participation will entail.[137] Great care needs to be taken when explaining complex information to participants to facilitate understanding.

Generally, genomic and genetic research involves concepts that are challenging for a lay audience. Patients that are unwell may not be able to absorb complex information, so there is a need to explain the study in a simplified way. However, it is important for researchers to ensure that the informed consent captures the essence of issues that could be of concern to the participants. It is important to recognise that some aspects of complex genomic research may be of less interest to research participants and less well understood by them.[138]

In this study, consent was sought from patients aged 16–65 years old. If the patient was incapable of giving informed consent (e.g., patients in intensive care), consent to participate in the study was obtained from the patient's next-of-kin, such as a guardian, attorney, spouse, carer, relative or close friend of the patient. In cases of next-of kin giving consent for incapacitated participants, the participants would be asked to re-consent once they recovered from their illness.

Even where they understand the aim of the research, patients' concerns may cause them to decline to participate. The patients' main concern regarding participation in this study was that NGS technology would generate all genetic information present in the sample. Some patients were concerned about incidental findings and other potential health problems that might be identified from genetic data. In addition, patients were aware that genetic data is highly identifying and can reveal information

relevant to family members. To address these concerns, the principal investigator explained that although NGS yields data on all nucleic acid sequences present in a sample, including a participant's own DNA/RNA sequences, the data pertaining to the host were not analysed in this study. Thus, most ethical issues of privacy and confidentiality, potential stigmatisation or potential abuse of data generated from the research would not apply, because the findings would pertain only to non-human nucleic acid and not to individuals per se. It was essential to assure the participants that there was no risk that the investigators would be able to detect genetic abnormalities or traits in the participant or their family.

Another perception that affected participation was the high expectations concerning the diagnostic capacity of a tertiary hospital, which often led to the judgment that research is not required. The strategy for overcoming this issue was to explain the findings of the pilot study, which showed a high proportion of undiagnosed fever at Cairns Hospital.[8] After understanding the significance of undiagnosed fever, patients were more likely to acknowledge the importance of evaluating the new approach (NGS study) for diagnosing undifferentiated fever.

After obtaining the patient's consent, the next step was sample collection. Although the principal investigator is a medical doctor, the investigator's overseas qualification is not recognised in Australia and the patients were informed of this. While the principal investigator could give information related to the research and obtain patient consent, the investigator was not allowed to draw blood samples from the patients. Therefore, a registered nurse from the clinical research unit was delegated to collect patients' samples.

It was expected that the enrolment process would be completed in 12 months to fit into the PhD timeline. As this was an exploratory study with a limited timeframe for sample collection, 40 was considered a sufficient number of participants to test the hypothesis. These 40 participants included controls, who already had specific diagnoses, and study subjects, who were as yet undiagnosed at the time of enrolment. As the study was prospective, it was not known at the time of enrolment to which group each participant belonged.

3.5 Handling and storage of data

In general, a hard copy of participants' data is kept securely in a designated locked filing cabinet located at the Cairns Clinical School of JCU. Electronic data is stored in JCU Tropical Data Hub, which is a catalogue for research data produced by and/or held at JCU. Login authentication is required to access the data stored at JCU Tropical data Hub. After 15 years of storage, paper records are shredded and electronic databases are deleted. Participants' data can be re-identified using their medical record number or assigned identification (ID) number, but their identities are removed before any publication or presentation of results. At this stage, data are non-identifiable and identifiers have been permanently removed, such that no specific individual can be identified from the dataset. The details regarding confidentiality and data security for each stage of the project are described in the following paragraphs.

In the first pilot study, the following steps were taken to guarantee confidentiality and anonymity of participant data. First, during the data collection stage, individually identifiable data were recorded, including individual names, dates of birth and addresses. Afterwards, data were stored in re-identifiable format using hospital Unit Record (UR) numbers, from which identifiers had been removed and replaced by a code, but it remained possible to re-identify a specific individual.

In the second pilot study, each participant was assigned an ID number and tubes containing blood specimens were labelled with the ID number and date of collection. Six months after the date of blood collection, all samples were discarded. Data concerning the concentration of human nucleic acids were stored in non-identifiable format.

In comparison to other data types in medical tests, the security and privacy of genomic data can become a particular issue for several reasons: the data could have implications for others besides the patient/participant; the data could have an impact throughout the patient/participant's lifetime; and the data will not change throughout a participant's lifetime.[139] The sensitive nature of genomic data with regards to identifiability and privacy, and the issues that could result from data breaches, were a major concern in this study, while the physical harms of participating were minimal (e.g., the pain experienced due to blood draw). In addition, participants had little conception of the potential downstream uses of the genomic data generated from their samples.

Since NGS technology generates a significant amount of data from an individual study participant, it is important to carefully consider procedures for data handling and storage. Leaving security and privacy issues aside, the storage of data provides opportunities for those data to continue their contribution to scientific advancement through future studies. While the main interest of this study was the genetic information of pathogens, there was a choice as to whether to delete or store the host genetic information that was partially generated.

In the main study, during the informed consent process, each participant had the option to choose whether to give consent for the investigator to store their human genetic information for future study, or to request that the data be deleted. For participants who were under 18 years old or unable to re-consent after consent was given by their next-of-kin, their genetic information was deleted. Therefore, this study only stored the genetic information of people over 18 who had given their personal consent.

All data generated during this PhD project are kept under restricted access. The principal investigator and the data manager at JCU Tropical Data Hub should be contacted to negotiate access to the data. In the future, with patient consent and ethics approval, these data might be used to study host–pathogen interaction, and the principal supervisor will determine who can have access to the data for this purpose.

3.6 Returning the results

Most participants expected that medical testing should immediately answer questions about the cause of their fever. In fact, most thought that results would be available in 1–2 days. The projected time delay in obtaining the results of NGS study may have hindered participation because of the perception that the study findings would not change the outcome for the patient. While it is true that the benefits of genetic research may not occur at an individual level and may not be immediate, study outcomes have the potential to yield knowledge leading to general improvement in human health.[137] The results of this study have the potential to change the management of AUF and may benefit the wider community in the future. It was important to stress that results would be experimental and that the study participants could opt to be informed of the individual research results.

Participants in this study received the information related to infectious diseases only, provided that they had opted to receive such information (as indicated on the consent form). For those participants who had opted to receive results, the results of the study were communicated via email or telephone. Appointments were also offered if the patients wished to discuss the study findings further.

3.7 Chapter summary

This chapter discussed the research methodology, which involved three phases of study. The process of ethical clearance was also described, including the practical issues involved in conducting an NGS study and providing informed consent to participants. Although the study was carefully planned, some unexpected delays occurred, primarily due to the lengthy process of obtaining SSA approval and the data analysis during the main study. It is hoped that this experience will be valuable for other researchers setting up similar projects in the region.

Chapter 4: Undiagnosed Undifferentiated Fever (UUDF) in Far North Queensland, Australia

4.1 Introduction

Fever is a common complaint in healthcare settings and has various possible aetiologies, including infection, connective tissue disorders, malignancies and a number of miscellaneous conditions. The cause of fever may not be immediately obvious; in these cases, the condition is referred to as undifferentiated fever (UDF). There is a broad differential diagnosis for UDF, usually influenced by the geographical location, which may necessitate further laboratory investigations to determine the cause of fever. Sometimes, despite investigation, UDFs remain undiagnosed, and while some undiagnosed cases resolve spontaneously, others may be associated with considerable morbidity and even mortality. Outcomes may include AUF, FUO/PUO and UUDFs (for an illustration of the outcomes of UDF and case definition, see Figure 1.1).

This chapter presents the first phase of the research. A published paper[8] is incorporated into this chapter; adjustments have been made to make it more reader-friendly, including the removal of any redundancy alongside the previous chapters of the thesis, insertion of links to the other chapters, cross-referencing and re-labelling of figures and tables to match the chapter structure of the thesis. In this chapter, methods and results are also presented in more detail than they were in the publication.

Article: Undiagnosed undifferentiated fever in Far North Queensland, Australia: a retrospective study

Declaration of Authorship

Publication details	Nature and extent of the intellectual input of each author	Signature
Susilawati, TN & McBride, WJH. Undiagnosed undifferentiated fever in Far North Queensland, Australia: a retrospective study. *International Journal of Infectious Diseases.* 2014:27;59–64. *Accepted for publication 20 May 2014* **Published October 2014**	Developed the study design, collected and analysed the data, prepared the manuscript.	Susilawati, TN
	Assisted with the development of the study design, assisted with data analysis, assisted with article writing.	McBride, WJH

It has long been known that infections are the most common cause of acute fever, and that other conditions become more frequent causes as fever duration increases.[140] Nevertheless, diagnosing infectious causes of fevers is a challenge, as many infections present with a similar clinical picture. Current diagnostic approaches often fail to detect the aetiology of fever, with physicians attempting to minimise laboratory investigations by only requesting tests for the most likely aetiologies. Broad-spectrum diagnostic tools could improve the diagnostic yield.

Situated in a tropical zone, and a major tourist destination in Australia, the Cairns region is endemic for a range of tropical infections and is susceptible to the introduction of infections from other countries. Some of the known prevalent diseases in this area are leptospirosis, scrub typhus, spotted fever, melioidosis and infections caused by mosquito-borne viruses.[56] *Aedes aegypti* is present in North Queensland's urban areas, and dengue outbreaks are frequently reported.[52, 56]

The objectives of this study were: (1) to provide information about the epidemiology of AUF in the population of Far North Queensland, Australia; and (2) to understand the scope of the problem of UUDF in this region by elaborating the information related to this syndrome (i.e., proportion, characteristics, eventual outcome and adequacy of investigation). The ultimate purpose of this study was to gather information for a prospective study investigating the causes of AUF using NGS technology.

4.2 Methods

This study included a retrospective review of the medical charts of patients, aged 15–65 years, who presented to Cairns Hospital between 1 July 2008 and 30 June 2011. AUF is defined as a raised body temperature of ≥ 38 °C, or a history of fever (with chills or shivering) of duration up to 21 days, without an immediately obvious cause on the basis of clinical findings, rapidly available (in 6 hours after admission) pathology or radiological investigations, and not associated with focal infection, nosocomial infection, neutropenia or immunosuppressing conditions.

Potential AUF cases were identified by searching AUSLAB® (a laboratory management software system in Queensland) for test requests to diagnose one or more specific pathogens. Examples of tests often performed in the evaluation of AUF are malaria screening as well as serology and PCR for dengue, *Leptospira*, Q fever and rickettsial infection. Following the identification of potential subjects, the medical charts were reviewed to determine patients who met the criteria for AUF.

The following information was retrieved from medical records:

- Demographic data: age, gender, date of birth, residential address;
- Clinical data: details of any referral, symptoms, fever duration prior to hospital presentation, highest recorded body temperature, duration of hospitalisation, admission to intensive care;
- Laboratory findings (white blood cell [WBC] count, neutrophil count, lymphocyte count, platelet count, C-reactive protein [CRP] level, urea, creatinine, alanine aminotransferase [ALT], aspartate aminotransferase [AST], blood culture results, CSF analysis, serology and any other specific investigations);
- Radiology findings;
- Diagnoses made and follow-up records.

Diagnoses were categorised into two groups: (1) provisional clinical diagnosis, which was recorded from the discharge record or from the working diagnosis in the emergency room if the discharge diagnosis was not available; and (2) the final diagnosis that was made after the results of investigations and follow-up visits were available.

A laboratory-confirmed case was defined as one that met one or more of the following criteria:

- The isolation of pathogen from a clinical specimen;
- The detection of pathogen nucleic acids in a clinical specimen during the acute phase of the illness;
- The detection of a four-fold rise in serum IgG antibodies by indirect immunofluorescence assay, or neutralisation, and/or seroconversion on enzyme-linked immunosorbent assay (ELISA) on testing of paired sera. If a paired serum analysis was not performed, a single raised IgM test, together with consistent clinical, laboratory and radiology investigations formed the basis of a final diagnosis.

Data were incorporated into a Microsoft Excel spreadsheet. Statistical analyses were performed using IBM SPSS version 20 software (IBM Corp., Armonk, NY, USA). Descriptive statistics and cross-tabulations were produced for the presenting data. The normality of the data distribution was assessed using Kolmogorov–Smirnov and Shapiro–Wilk tests. Inter-group comparisons were made using the Pearson chi-square test for categorical variables and the Mann–Whitney test for continuous variables. A value of $p < 0.05$ was considered statistically significant.

4.3 Results

The study flow chart is shown in Figure 4.1. During the period between 1 July 2008 and 30 June 2011, 970 requests were made to investigate one or more infectious agent(s) recorded by AUSLAB®. Of these, 340 cases met the definition of AUF. The most common clinical diagnoses on admission were dengue (80/340, 23.5%), viral infection (73/340, 21.5%) and PUO (35/340, 10.3%).

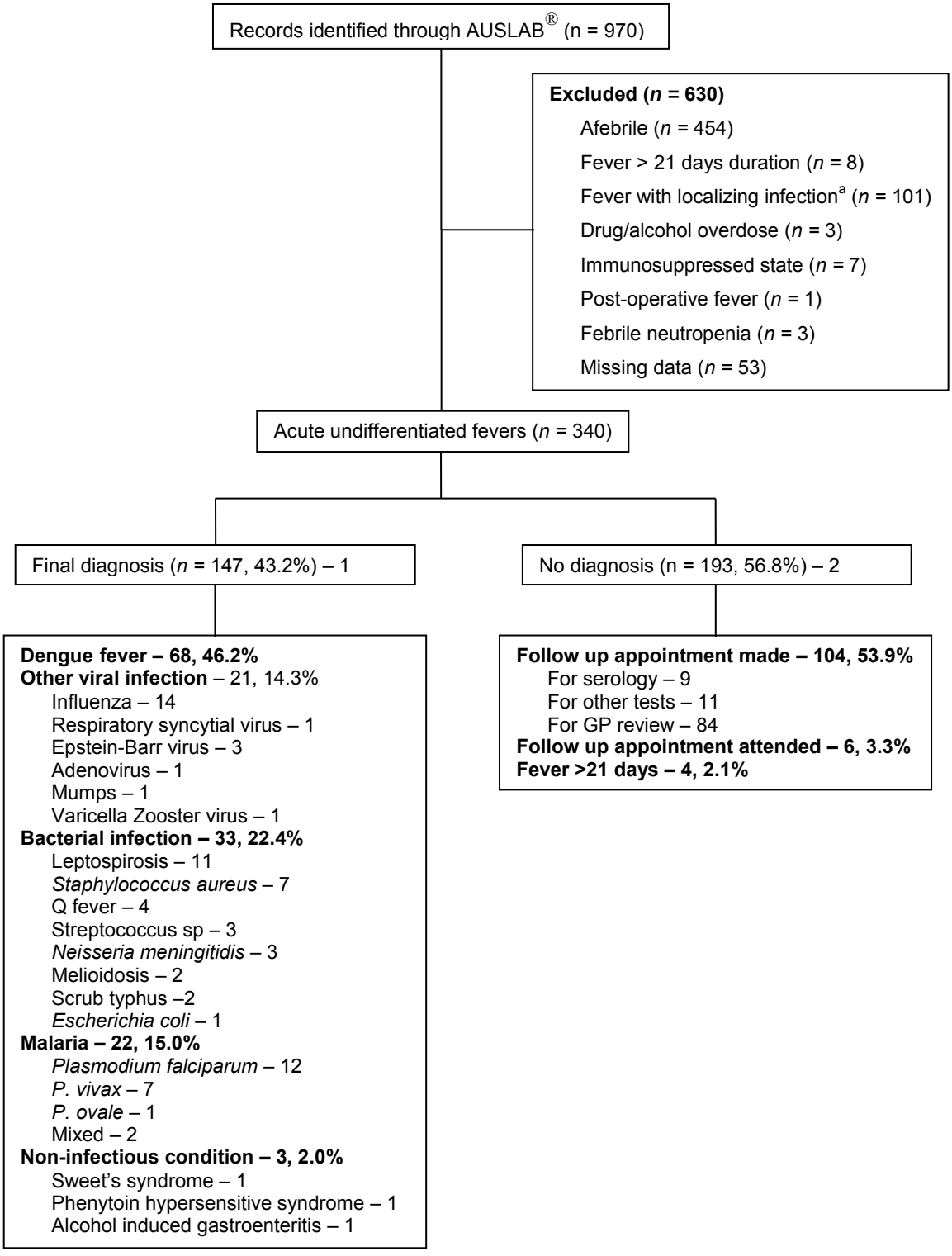


Figure 4.1: Study flow chart

[a] Including community acquired pneumonia, skin and soft tissue infection, bone and dental infection, intra-abdominal infection, pelvic inflammatory disease and urinary tract infection

Seasonal variation influenced the incidences of AUF and mosquito-borne diseases. Rainfall data from the Bureau of Meteorology (BOM) database[141] was retrieved to show that the occurrence of AUFs and dengue predominated during the wet season (see Figure 4.2). Almost all (66/68, 97%) of the dengue cases occurred during an outbreak in late 2008 and early 2009. During this period, there was a high incidence of AUFs; 83 cases had specific diagnoses and 66 cases were undiagnosed.

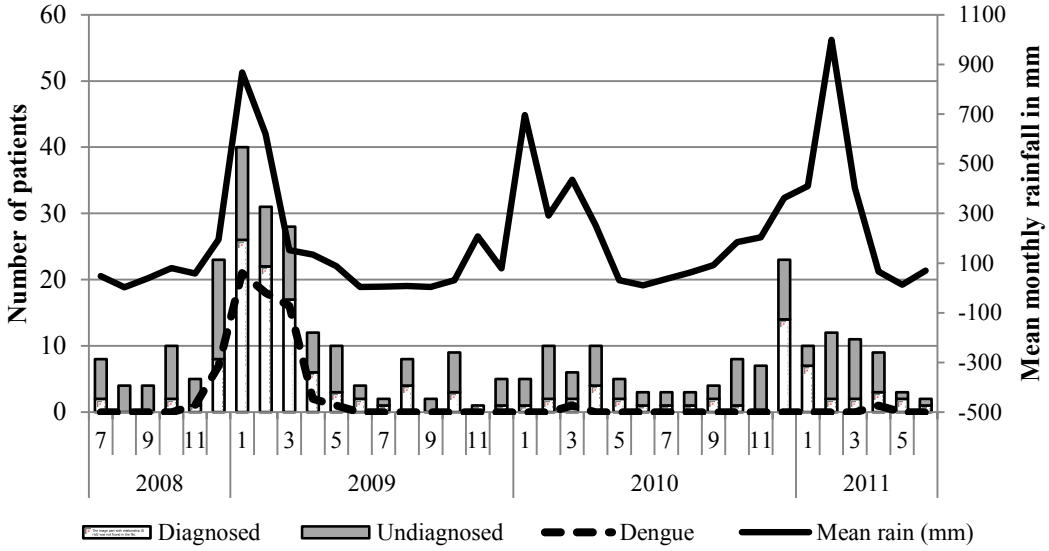


Figure 4.2: Seasonal variations of dengue and acute undifferentiated fever (diagnosed and undiagnosed cases) in Cairns Hospital, Far North Queensland, Australia, from 1 July 2008 to 30 June 2011

Specific investigatons included malaria screening (microscopy, *P. falciparum* antigen, pan malarial antigen), *Mycobacterium tuberculosis* (microscopy and culture) as well as serology and/or PCR for Flavivirus, hepatitis viruses, HIV, EBV, cytomegalovirus (CMV), respiratory viruses, Varicella Zooster virus, Herpes Simplex virus, *Leptospira, Rickettsia,* Q fever, *Streptococcus sp*, *Legionella, Burkholderia pseudomallei, Mycoplasma pneumoniae, N. meningitidis, Salmonella typhi* and *Cryptococcus.* Around half of AUFs ($n = 166$, 48.8%) were tested for one to three agents; over a quarter of patients ($n = 94$, 27.8%) were tested for four to six agents; and the remainder ($n = 80$, 23.5%) were tested for six to 20 agents. Most patients with AUFs were investigated for dengue ($n = 267$, 78.5%), and many for leptospirosis ($n = 137$, 40.3%) and malaria ($n = 84$, 24.7%). A final diagnosis was possible in 147 (43.2%) patients. Eighteen patients were admitted to intensive care due to leptospirosis ($n = 3$), meningitis ($n = 2$), Staphylococcal sepsis ($n = 2$), melioidosis ($n = 2$), scrub

typhus (n = 1), H1N1 infection (n = 1), varicella pneumonitis (n = 1) and unspecific diagnosis (n = 6). There were three deaths attributed to sepsis. One death was due to *Staphylococcus aureus* septicaemia, while the causes of the other two deaths were not identified.

Low platelet count was detected in 37% (126/340) of AUFs. Sixty percent (204/340) of patients were tested with blood culture, and the results were positive in only 5.3% (18/340) cases, suggesting that bacteraemia has not been under-diagnosed. Likewise, 48.8% (166/340) of AUFs were screened for respiratory infections using chest X-ray, and abnormalities were detected in only 13.2% (45/340) of cases.

Aetiological diagnoses were made in 147 patients (43.2%), with dengue (n = 68), malaria (n = 22) and influenza (n = 14) being the most frequent entities. Among AUFs with specific diagnoses, viral infection was more common than bacterial or parasite infection, accounting for 60.5% (89/147) of diagnosed cases. The details of the investigations and specific diagnoses are presented in the following subsections.

4.3.1 Dengue infection

There were 267 requests for dengue tests, though only 68 (25.5%) of those tested had confirmed dengue through serology and/or PCR results. Of those confirmed having dengue, 15 patients had an incorrect or non-specific clinical diagnosis, including viral infection (n = 10), PUO (n = 3), hypotension postural (n = 1) and viral gastroenteritis (n = 1). Five patients did not have a clinical diagnosis. One patient had acute renal failure as a complication of dengue infection.

4.3.2 Other viral infection

PCR assay was the main method for detecting respiratory viruses. Of the 28 samples from nasal swab, influenza virus was detected in 14 samples, adenovirus in one sample and respiratory syncytial virus in one sample. Three patients were highly suspected of having infectious mononucleosis based on clinical presentation (sore throat, odinophagia) and positive IgM ELISA for EBV. One patient had bilateral parotitis and was diagnosed with mumps. One patient had chest radiographic results suggesting pneumonitis and positive varicella zoster IgM.

4.3.3 Leptospirosis

Leptospirosis was confirmed by serology, PCR or culture in nine patients and highly suspected in two patients. Clinically, patients showed fever and few physical signs; two patients had dark urine and oliguria, nine patients had thrombocytopenia and six patients had abnormal chest X-ray, suggesting leptospiral pulmonary involvement.

4.3.4 Central nervous system (CNS) infection

A total of 46 patients had fever and CNS symptoms and signs (e.g., altered mental status, seizure, neck stiffness) consistent with an encephalopathy or meningitis. Of these, 25 patients underwent lumbar puncture. Diagnosis of bacterial meningitis was achieved on two CSF samples that had positive cultures (one for *N. meningitidis* and one for *Streptococcus constellatus*), and one sample that had intracellular Gram-negative diplococci, suggesting *N. meningitidis* infection. Another patient had a positive blood culture for *N. meningitidis*, which supports a diagnosis of bacterial meningitis even though this patient did not undergo CSF analysis. In addition, 4 patients had lymphocytic predominant CSF, which was probably consistent with viral meningitis or encephalitis.

4.3.5 Other bacterial infection

Blood cultures were obtained from 60% (204/340) of patients with UDF. The results of blood cultures confirmed *S. aureus* infection in seven patients (two of whom were methicillin resistant) and *Escherichia coli* infection in one patient. Two patients were diagnosed with melioidosis based on the growth of *B. pseudomallei* in blood and sputum culture. Both patients had a cough and abnormal chest X-ray (consolidation, pleural effusion). One patient had positive streptococcal serology and abnormal chest X-ray as evidence of recent infection with Group A Streptococcus (GAS). Another patient had acute rheumatic fever with primary AV block on electrocardiogram, and culture of throat swab was positive for GAS. Q fever was serologically confirmed in 3 patients and highly suspected in one patient, based on clinical presentation and very high IgM titre in acute serum samples. Two patients had jaundice and three patients were thrombocytopenic. All patients with Q fever had normal leukocyte count, but increased levels of ALT and AST. Two patients were diagnosed with scrub typhus. In both patients, clinical presentations were unremarkable and no eschars were observed. However, a lesion that was thought to be a tick bite was found on one patient, and this

patient developed sepsis, acute respiratory distress syndrome and acute renal failure with serologically confirmed scrub typhus. The other patient was diagnosed based on single raised antibody to *Orientia tsutsugamushi* and positive response to doxycycline. Both patients had very high CRP levels (389 and 356 mg/L).

4.3.6 Malaria

Malaria screening was performed by microscopic test and detection of malarial antigen in serum. Of the 84 patients screened, 22 were positive for malaria: 12 had *Plasmodium falciparum*, 7 had *P. vivax*, 1 had *P. ovale*, 1 had mixed *P. falciparum/P. vivax* and 1 had *P. vivax/P. malariae* co-infection.

4.3.7 Non-infectious condition

A woman who was 24 weeks pregnant presented with pustular rash and had a skin specimen taken that was consistent with Sweet's syndrome, likely secondary to pregnancy or streptococcal infection. Other non-infectious cases included phenytoin hypersensitivity syndrome and alcohol-induced gastroenteritis.

4.3.8 Undiagnosed cases

The aetiology of fever remained unknown in 193 (56.8%) patients; these cases were classified as UUDF. Table 4.1 shows the demographic and laboratory characteristics of the patients with diagnosed and undiagnosed undifferentiated fever. Patients with UUDF were admitted for a shorter period, while patients with lower platelet and WBC counts but with higher liver transaminases were more likely to have specific diagnoses made.

The symptoms of UUDF were non-specific, with a high prevalence of constitutional and gastrointestinal symptoms. The most common symptoms of UUDF were headache (135/193, 69.9%), muscle pain (105/193, 54.4%), joint pain (95/193, 49.2%), nausea (81/193, 41.9%) and vomiting (76/193, 39.4%).

The majority of patients with UUDF had normal results of full blood count (130/193, 67.4%) and renal function tests (173/193, 89.6%). Notable abnormalities on laboratory testing were elevated levels of hepatic aminotransferases and CRP. Among the 189 patients tested for aminotransferases, 102 (53.9%) had increased levels of ALT and/or AST. Nearly all patients tested (88/89, 98.8%) had elevated CRP. Among those tested for platelet count and/or WBC, around one quarter (51/187, 27.3%) of patients

with UUDF were thrombocytopenic and nearly one third (61/191, 31.9%) had a high leukocyte count. Less than 10% of patients tested had increased urea or creatinine levels. All patients with UUDF had normal results of urinalysis. Nineteen patients had subtle abnormalities on chest X-ray, such as evidence of hyperinflated lungs, increased lung markings, peribronchial thickening and pleural effusion. Four patients with UUDF went on to fulfil FUO criteria after their fever duration exceeded 21 days without a definite diagnosis made during hospitalisation.

Table 4.1: Demographic and significant laboratory characteristics of patients with diagnosed and undiagnosed undifferentiated fever[a]

	Diagnosed patients (*n* = 147)	Undiagnosed patients (*n* = 193)	*p* value
Age, years	36 (25–47)	32 (24–44.5)	0.122
Male sex	60.5	60.6	0.988
Residential address			0.500
Suburban	76.9	74.1	
Rural	17.7	17.1	
Overseas	5.4	8.8	
Body temperature, °C	38.9 (38.3–39.4)	38.8 (38.1–39.4)	0.371
Duration of fever before hospital presentation, days	3 (2–5)	3 (2–5)	0.068
Type of referral			0.914
Self-referred	68.7	67.4	
Local doctors	22.4	24.4	
Other hospital	8.8	8.3	
Length of hospital stay, days	1 (0–5)	0 (0–3)	0.001
Number of agents tested for	3 (2–7)	4 (2–6)	0.716
Dengue test done	74.1	81.9	0.086
Platelet count, x 10^9/l	131 (67.5–180.5)	185 (134–242)	<0.001
WBC, x 10^9/l	5.35 (3.325–8.175)	8.3 (5.6–12.3)	<0.001
ALT level, U/l	80.5 (29.5–137.25)	37 (20–74.5)	<0.001
AST level, U/l	70 (29–147.5)	32 (22–70)	<0.001

WBC, white blood cell; ALT, alanine aminotransferase; AST, aspartate aminotransferase.

[a] Results are presented as the median (interquartile range) or as the percentage.

4.3.9 Fatal cases

A 60-year-old man died due to staphylococcal infection. This patient presented in hospital with 7 days of fever, with the highest temperature being 40.1 °C, back pain, dyspnoea and palpitations. A chest X-ray identified pleural effusion. Routine blood tests showed an elevated CRP as well as serum urea, creatinine, ALT and AST (162 mg/l, 19.2 mmol/l, 219 µmol/l, 87 U/l, and 217 U/l, respectively). Leukocyte count was high (40.9 x 10^9/l) with neutrophil predominance (96.2%). Specific investigations included tests for *Rickettsia* sp., *Leptospira* sp., dengue, *Legionella* sp. and *M. pneumoniae*; all results were negative. The patient was admitted to intensive care and *S. aureus* was isolated from blood culture. A computed tomography (CT) scan showed multi-organ abnormalities, including brain ischemia, pleural effusions, splenic infarcts, cirrhosis and cholelithiasis. He died on day 8 in hospital.

Two deaths of unknown cause were recorded, involving a boy with intellectual impairment and a woman from Western Province, Papua New Guinea. Autopsy reports of those patients were not available in the hospital database or medical notes, thus it was assumed that autopsy was not performed in either case.

The first fatal case was a teenage boy with Lennox–Gastaut syndrome, a medical problem characterised by frequent seizures and mental deficiency. This patient went to a local doctor with 1 day of fever reaching 41°C. The initial diagnosis was an upper respiratory tract infection and conjunctivitis. At home, the patient developed a seizure and was sent to a local hospital before being transferred to Cairns Hospital. He had an elevated CRP as well as elevated serum creatinine, ALT and AST (44 mg/l, 219 µmol/l, 378 U/l and 1220 U/l respectively). Platelet count was very low (35 x 10^9/l) and WBC increased (25 x 10^9/l) with lymphocyte predominance (54.2%). No abnormality was detected on chest X-ray, and PCR tests were negative for *N. meningitidis* and *S. pneumoniae*. The patient died after several hours in intensive care.

The second fatality was a 21-year-old woman who was transferred from Thursday Island Hospital with 14 days of fever reaching 38.6 °C, night sweats, anorexia, cough, dyspnoea, pleuritic chest pain and abdominal pain. Laboratory tests showed haemolytic anaemia, increased CRP and liver aminotransaferases (CRP 85 mg/l, ALT 108 U/l, AST 75 U/l), leukocytosis (25 x 10^9/l) with neutrophil predominance (90.2%) and progressive thrombocytopenia (platelets decreased from 76 x 10^9/l to 9 x 10^9/l over 5 days). Blood culture was sterile and specific tests were negative for numerous pathogens including malaria, *Streptococcus* sp., *Legionella* sp.,

Mycoplasma sp., *B. pseudomallei*, HIV and Hepatitis viruses (HAV, HBV, HCV). *Mycobacterium* sp. was not found in sputum, pleural fluid, ascites or a bone marrow specimen. Bone marrow biopsy ruled out myelodysplastic syndromes, and there was no evidence of lymphoma or haemophagocytic disorder. Chest X-ray and chest CT showed bilateral effusions, while abdomen CT showed massive splenomegaly and a large amount of ascites. Despite having localised signs, this patient was included in this study because the cause of her fever was obscure. However, by the time she died (day 9 in hospital), she had fulfilled the criteria for FUO and was thus excluded from the UUDF series.

4.4 Discussion

4.4.1 Aetiologies of AUF in Far North Queensland, Australia

Despite being common and more prevalent than FUO, AUF is not well described. This study investigated this common clinical presentation and described the causes of AUF. A number of smaller hospitals serve the region of Far North Queensland, so not all patients meeting the definition of AUF in the region were seen at Cairns Hospital. Among those patients presenting with AUF to Cairns Hospital, the diagnosed diseases were consistent with what has previously been described in this region.[53–57]

The results of the study underline the importance of dengue, malaria and leptospirosis in patients with AUF. The findings are consistent with similar studies conducted in tropical countries in Asia,[23–25, 30, 31, 34, 59, 142, 143] West Africa[40] and South America.[43] Other studies, conducted in European countries, have found Q fever to be a significant cause of AUF.[46, 144] The finding that less than half of AUF patients had a confirmed clinical diagnosis is also consistent with some of the studies previously mentioned.[23, 25, 34, 59]

This study identified a high incidence of both diagnosed and undiagnosed undifferentiated fever in late 2008, which coincided with a dengue outbreak in Cairns during that period. It is possible that some of the UUDF patients did indeed have dengue, which would indicate that there were missed diagnoses. Alternatively, there may have been an increased presentation of febrile patients who would not have otherwise presented to hospital, or there may have been concurrent circulation of an unrecognised cause of fever.

4.4.2 Definition of UUDF

In any consideration of undiagnosed fever, it is necessary to ascertain whether there is an accepted meaning of the term. There is certainly no official terminology regarding the undiagnosed short-term febrile illness so that the incidence of such cases is not easy to determine. This is because even in hospital cases the admission diagnosis may be vague and varied, such as PUO, febrile illness, or viral infection. Although this syndrome has clinical similarities with PUO, the shorter duration of fever and the differing aetiologies necessitates a different term. The UUDFs that are associated with severe illness and intensive investigation are clearly important illnesses and the causes may have important public health implications. The occurrence of this syndrome, particularly if its incidence rises above a baseline threshold, could be an early indication of the emergence of a condition that requires recognition.

This study has quantified and described a syndrome of undiagnosed short-term fever in Far North Queensland. This condition is referred to as UUDF and defined as: 1) a fever of ≥ 38.0 °C or symptoms suggestive of fever; 2) a duration of fever of ≤ 21 days; 3) a failure to reach a diagnosis after performing clinical evaluation and laboratory investigations, including complete blood count, serum biochemistry, urinalysis, blood culture, chest X-ray; 4) a request by the clinician of specific test for at least one infectious agent and; 5) a failure to make a specific diagnosis.

4.4.3 Diagnostic challenges

This study illustrates the challenges faced by physicians in diagnosing infectious causes of fever. Because of its non-specific clinical picture, AUF is difficult to diagnose on clinical grounds only. Viruses are frequently suspected as the aetiology of AUF, but this is often difficult to prove, because clinical laboratories have limited capacity to detect a wide variety of viruses causing fever. This study demonstrates that the available resources in a tertiary hospital in a developed country are inadequate for diagnosing infectious diseases.

In this study, conventional methods contributed to identifying pathogens in less than half of the patients with fever. The non-specific clinical features led to frequent requests for testing for well-recognised pathogens. As an example, dengue testing was recorded at a high level throughout the study regardless of the epidemic and indicates a high level of awareness among clinicians at the hospital of this particular disease.

Despite the frequent requests for dengue testing in this study, only 68/267 (25.5%) of patients had a positive result. Overemphasis of dengue may have caused other infectious agents to be neglected and underinvestigated.

Current investigation methods often fail to detect the aetiology of AUF, where physicians, attempting to minimise laboratory investigations, only request tests for the most likely aetiologies. Therefore, the knowledge and experience of the treating physician is the starting point for requesting specific tests in order to confirm clinical diagnosis or to eliminate differential diagnosis. Conventional testing methods, such as culture, serology or targeted nucleic acid-based testing, such as specific PCR, rely on prior knowledge of the pathogen under investigation, and thus do not permit the detection of unpredicted or novel pathogens. All these factors lead to a limited scope of investigation resulting in failure to detect infectious causes of fever in a significant fraction of cases. Thus, the high rate of undiagnosed cases in this study implies that AUFs were not investigated thoroughly, possibly because of short stays in the hospital or limited tests on patients with minor symptoms. Other possible explanations are that clinicians failed to order appropriate tests, that current diagnostic methods are not adequate, or that there are causes of fever that are yet to be discovered.

4.4.4 Study strengths and weaknesses

One weakness of this study was that a significant amount of data was missing, including lost or incomplete medical histories, absent discharge summaries, and the failure to match some patients identified on AUSLAB® to their medical records. Further, this study did not include febrile cases that were not investigated for infectious agents. The retrospective design of this study meant that there was no establishment of a standardised testing regimen for subjects, and this was likely to have resulted in fewer patients included in the sample with a specific diagnosis. On the other hand, the study does provide insights on how this condition is currently managed.

Despite these weaknesses, useful information was obtained. The majority of patients presented early and required only short periods of hospitalisation, suggesting that acute infection is the main cause of AUF. Moreover, frequent requests for tests for arboviruses and leptospirosis suggest that the assessment of AUF is influenced by the local occurrence of infectious diseases. The diagnostic approach in Northern Queensland, and elsewhere, should be tailored to the local epidemiology of the known infectious aetiologies.

4.4.5 Suggestions for implementation

Using the definition of UUDF, a wide variation in disease severity can be identified further using a grading system to help identify patients for whom additional diagnostic measures are required. Patients could be scored on the basis of several criteria, as suggested in Table 4.2. Application of the scoring system in this study demonstrated that the average score of diagnosed cases was higher than that of undiagnosed fevers: 5.42 (standard deviation: 1.66) compared to 4.71 (standard deviation: 1.68) respectively.

Table 4.2: Scoring system to determine the significance of undifferentiated fever

Criteria	Score
Fever duration—from time of recording ≥ 38 °C or historically suggestive fever onset to time of fever lysis (recorded temperature not exceeding 37.5 °C for 48 hours)	
1 day	0
1–3 days	1
4–21 days	2
Hospital admission duration	
< 1 day	0
1–3 days	1
> 3 days	2
Thoroughness of investigation (number of investigations for specific agents that are appropriate for known regional diseases or for patients' clinical signs and symptoms)	
1 agent	0
2–3 agents	1
4 or more agents	2
Laboratory abnormalities (thrombocytopaenia; leucocytes, lymphocytes and/or neutrophils outside of normal range; elevated AST and/or ALT; elevated CRP; abnormal renal function)	
No abnormal values	0
1–2 abnormalities	1
3 or more abnormalities	2

It was noted that the occurrence of a dengue outbreak (November 2008–April 2009) influenced the diagnostic approach at Cairns Hospital, as evidenced by frequent requests for dengue tests. On average, during this period, doctors ordered 3.32 specific tests ($SD = 1.83$, median = 2, interquartile range = 2–4). There was no evidence of increased testing on the individual during the epidemic. In the authors' opinion, this will not significantly affect the third criterion of the scoring system, that is, thoroughness of investigation, because the number of agents tested during the dengue outbreak was similar ($p = 0.716$) to those applied to all cases (see 'Number of agents tested for' in Table 4.1). If anything, there were fewer tests done during the dengue epidemic, probably because clinicians were focused on this disease. This could serve to lower the score for a patient with UUDF in some circumstances.

A possible diagnostic approach to UUDF would be to use metagenomics. In terms of cost-effectiveness, NGS reduces the cost of sequencing one million nucleotides (1Mb) to 0.1–4% of that associated with Sanger sequencing.[9] For example, in 2014, a single lane of the Illumina HiSeq platform could produce 35 Gb of sequencing data from up to 24 samples at a cost of approximately USD $3,000 in total. The cost of sequencing is rapidly decreasing over time, so it is inevitable that NGS will become more affordable in the near future. The scoring system proposed may assist in selecting the most clinically significant samples for more intensive investigations. Patients scoring 5 or more points should be considered for investigation using a NGS platform. Alternatively, the use of proteomics, multiplex PCR or microarray hybridisation might be useful to improve diagnostic yield.

Regardless of sensitivity and specificity, it is imperative to use a diagnostic tool that can deliver results immediately to improve patient outcomes. NGS technology is available on several platforms, and the run time varies from several hours to more than a week depending on the volume of output data (see Table 2.1). Clinicians who need rapid answers on individual patients may prefer to use Illumina MiSeq than the HiSeq platform. The MiSeq platform produces 13–15 Gb of sequencing data in 5 hours, as opposed to the HiSeq platform, which can produce 95–600 Gb of data in periods ranging from 40 hours to 11 days.

4.5 Chapter summary

This chapter has presented the first phase of the research and provided important information for planning the main study, which seeks to determine the infectious causes of AUF using NGS technology. Data on the prevalence of AUF and UUDF was used to pragmatically determine the number of participants that were recruited into the main study. It was demonstrated that AUFs are common in the population of Far North Queensland, Australia. Over the three-year study period (2008–2011), there were 340 cases of AUF presenting to Cairns Hospital, of which more than half were undiagnosed, despite the availability of extensive diagnostic facilities at a tertiary referral hospital. This means that UUDFs occur frequently enough at Cairns Hospital to justify subsequent research into unpredicted and unknown (novel) infectious agents at this site. The preliminary study findings provide insight into clinical and laboratory characteristics of AUF, as well as how this condition has been investigated. This understanding was important in the development of criteria for recruiting patients into the main study. In addition, the study findings also helped to develop a robust definition of UUDF that would be useful for further study comparing the incidence of this entity between different geographical sites and over time.

Chapter 5: Quantitative Analysis of Circulating DNA in Plasma and Serum

5.1 Introduction

Nucleic acids are essential molecules for living organisms, as they carry genetic information. These molecules are made from nucleotides. Each nucleotide consists of three components: a 5-carbon sugar, a phosphate group, and a nitrogenous base. There are two types of nucleic acids: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). The two differ in the structure of the sugar in their nucleotides: DNA contains 2'-deoxyribose, while RNA contains ribose. Also, the nitrogenous bases found in the two nucleic acid types are different: adenine, cytosine and guanine are found in both RNA and DNA, while thymine occurs in DNA and uracil occurs in RNA. All living cells contain both DNA and RNA, while viruses contain either DNA or RNA, but usually not both. In most cases, naturally occurring DNA molecules are double-stranded and RNA molecules are single-stranded. However, some viruses have genomes made of double-stranded RNA (e.g., Reoviridae, Birnaviridae), while other viruses have single-stranded DNA genomes (e.g., Parvoviridae, Circoviridae).

It is commonly thought that a human's DNA is found within the nucleus of a cell, and that RNA is confined within a cell. However, research shows that DNA and RNA fractions can be isolated from cell-free samples such as plasma, serum and urine.[14] Plasma is the supernatant fluid obtained when anti-coagulated blood has been centrifuged. Serum is blood plasma without fibrinogen or other clotting factors. Serum is clearer than plasma because it has a lower protein concentration than plasma. Both DNA and RNA isolated from serum and plasma are commonly referred to as circulating nucleic acids (CNA).[16]

Little is known about the origin of CNA. It probably derives from a combination of apoptosis, necrosis and release from tumour cells, active release of newly synthesised DNA into circulation, breakdown of blood cells, breakdown of pathogens such as bacteria or viruses, and leukocyte surface DNA.[16, 145, 146] The mean quantity of plasma-circulating DNA in normal subjects varies from less than 10 ng/ml to more than 1500 ng/ml.[146] It has been observed that levels of circulating DNA are higher among individuals with pregnancy, certain cancers (e.g., breast, lung and

prostate cancers) and inflammatory disorders (e.g., systemic lupus erythematosus, pancreatitis, inflammatory bowel disease) compared with healthy individuals.[14, 147–149] It is also known that specific nucleic acid fragments can serve as biomarkers for particular disorders such as diabetes, cancer, myocardial infarction and stroke, which facilitates their early diagnosis.[14]

The increase of CNA in the state of infection has not yet been routinely investigated. A study conducted by Ha et al in 2011[150] showed that plasma DNA levels were significantly higher in patients with dengue virus infection than those with other non-dengue febrile illnesses (i.e. patients who were suspected to have dengue fever, but returned negative diagnostic tests for dengue infection) and healthy controls. Remarkably, increasing plasma DNA levels correlated with the severity of dengue.

During an infection, a small proportion of the pathogen's nucleic acids can be found in peripheral human blood.[16] NGS technology offers the possibility of identifying microorganisms in circulating blood by producing a small amount of pathogen sequence among an abundant background of human genomic information. As the subsequent investigation using NGS would only analyse a small portion of the genomic data, a proportionally lower amount of human DNA background is desirable in order to increase the sensitivity of detection of pathogen nucleic acids. This chapter presents the second pilot study, which aimed to compare concentrations of circulating DNA in plasma and serum samples collected using different techniques. The output of this research provided important information for the planning of sample collection and preparation for the subsequent NGS study, which sought to identify pathogen nucleic acids in the circulation of patients with AUF.

5.2 Materials and methods

5.2.1 Study participants

The recruitment process was initiated by posting a flyer on several JCU noticeboards with the investigator's contact number and email address. Six healthy volunteers—three males and three females aged 25–45 years—were recruited into the study. Among these, five participants were JCU postgraduate students and one participant was a JCU administration staff member. Prior to blood collection, a written explanatory statement or information sheet (see Appendix H) was provided, and each

participant was invited to ask questions before signing a written consent form (see Appendix I).

5.2.2 Sample collection and processing

Blood samples were collected at the JCU Cairns Clinical School, Cairns Hospital. Two types of tube were prepared to collect the different types of sample: 4 ml K2EDTA Vacuette® tubes (purple-top tube, Greiner Bio-One) were used to contain plasma and whole blood, while 5 ml Vacuette® tubes containing clot activator (red-top tube, Greiner Bio-One) were used for serum samples. Each participant provided six specimens through venous puncture in both arms, using a butterfly needle and vacuum system in the first arm and using a standard syringe and needle for the second arm.

The procedure for blood collection is detailed as follows. Prior to blood collection, the investigator verified the participant's health status and checked for any allergies to materials used during the procedure (i.e., antiseptics, adhesives or latex). During blood collection, the participant was asked to sit in a chair and hyperextend his/her arm. Area of needle insertion was prepared with alcohol 70%. In the first arm, the use of the vacuum system enabled quick aspiration of the blood. Initially, a cuff or tourniquet was applied on the upper arm, and then a butterfly needle was inserted into the median cubital vein located in the anterior surface of the elbow. Blood specimens were collected into three different tubes designated for whole blood, plasma and serum. Following this, the tourniquet was released. Two to three minutes after the release of the tourniquet, blood was drawn into the other two tubes designated for whole blood and plasma. Finally, a gauze pad was placed over the puncture site upon removal of the butterfly needle. When bleeding stopped, a bandage was applied. In the second arm, with the application of the tourniquet, 5 ml of blood was gently aspirated from the median cubital vein with a needle and syringe. After this, the tourniquet was released, the needle was removed and a gauze pad was placed over the puncture site. When bleeding stopped, a fresh bandage was applied. Lastly, the blood was slowly poured into a red-top tube for further processing to obtain serum sample. Soon after blood collection, the tubes were inverted carefully 5–10 times to mix blood and anticoagulant (in the purple-top tubes) or to mix blood and clot activator (in the red-top tubes).

Serum and plasma were prepared within 3 hours after blood draw by centrifugation for 15 minutes at 2,000 *g* at the QTHA laboratory, JCU Smithfield. After centrifugation, the supernatants (plasma and serum) were aspirated carefully with a

clean sterile pipette tip and pooled into centrifuge tubes. Turbid samples were centrifuged and aspirated again as above, and then all specimens (whole blood, plasma and serum) were immediately used for DNA extraction. Following specimen processing, each participant had contributed six different specimens, as described in Table 5.1.

Table 5.1: Specimens obtained from each participant

Specimen	Blood collection method			Specimen code
	Tourniquet applied (T)	Vacuum system applied (V)	Syringe and needle applied ($)	
Whole blood (W)	√	√	-	W.T.V
Whole blood (W)	-	√	-	W.N.V
Plasma (P)	√	√	-	P.T.V
Plasma (P)	-	√	-	P.N.V
Serum (S)	√	√	-	S.T.V
Serum (S)	√	-	√	S.T.$

5.2.3 Extraction of total nucleic acids

There are many commercial kits available for nucleic acid isolation. Most kits have specific application in regards to the isolation of particular types of human nucleic acids, such as genomic DNA, circulating DNA, or RNA. The kits are also specifically designed for application on particular samples, that is, with or without cellular components in the samples. The use of appropriate kits is critical for the success of this study. The kits should be carefully selected with particular regard to the specimen type.

Although plasma and serum samples are free from any cells, whole blood contains white blood cells, which have a nucleus and therefore contain genomic DNA and expressed RNA. There are few kits that can isolate nucleic acids from both cell-free and cellular samples. In this experiment, the High Pure Viral Nucleic Acid kit (Roche Applied Science, catalogue number 11858874001) was used because the kit is suitable for the isolation of total nucleic acids (DNA and RNA) from liquid samples (plasma and serum) and from samples containing cells (whole blood).

Nucleic acid extraction was performed according to the manufacturer's instructions. The detailed procedure was as follows. First, 200 µl of working solution (binding buffer supplemented with poly-A carrier RNA) and 50 µl of proteinase K were

added to a 200 µl sample of serum, plasma or whole blood in a sterile Eppendorf tube, mixed and incubated for 10 minutes at 72 °C. After incubation, 100 µl of binding buffer was added. The filter and collection tube were combined and the sample was pipetted into the upper reservoir, followed by centrifugation for 1 minute at 8,000 *g*, after which the flowthrough and collection tube were discarded. Next, 500 µl of inhibitor removal buffer was added into the assembled filter and collection tube, followed by centrifugation for 1 minute at 8,000 *g*, after which the flowthrough and collection tube were discarded. The filter was washed twice with the wash buffer with centrifugation for 1 minute at 8,000 *g*. Finally, centrifugation was performed for 10 seconds at full speed (13,000 *g*) to remove the entire residual wash buffer. Each collection tube was discarded and a clean, nuclease-free 1.5 ml tube was used to collect the eluted nucleic acids in 50 µl of elution buffer. In this study, nucleic acid extraction was performed in duplicates. Nucleic acid preparations were stored at -20 °C until quantification.

5.2.4 Quantification of double-stranded DNA (dsDNA)

Working with CNA can be problematic due to its low concentration and short fragment size. Low concentrations of CNA are often undetectable using conventional methods such as ultraviolet (UV) spectrometry/spectrophotometry or real-time PCR. CNA is also highly fragmented, often less than 200 bp, and the circulatory RNA is severely degraded (the RNA Integrity Number [RIN] of circulatory RNA is often less than 2, whereas RNA of good quality should have RIN more than 7), so that sample processing demands careful attention to prevent further fragmentation and further degradation of CNA. This study focused on the measurement of dsDNA levels, which are more stable than single-stranded nucleic acids (ssDNA and RNA). In this experiment, concentrations of dsDNA in the samples were determined by conducting a microplate fluorescence assay (MFA) using SYBR Green I dye (Life Technologies). The principle of MFA is that samples mixed with SYBR Green I in the wells of a microtitre plate produced fluorescence in proportion with dsDNA concentration. The intensity of fluorescence was measured using a spectrofluorometer. The fluorescence intensity values were then used to determine the levels of dsDNA in the samples.

To demonstrate the linear correlation between fluorescence intensity and dsDNA concentration, a standard curve was generated using a commercially sold DNA ladder of known DNA concentration diluted serially in buffer (EB buffer, Qiagen). SYBR Green I (diluted to 1:1250 in EB buffer) was used to bind the dsDNA in the

sample. The detailed procedure is as follows. GeneRuler Express DNA ladder (Fermentas) containing 0.5 µg/µl of DNA was serially diluted with EB buffer to produce the following concentrations of DNA standards: 0.25, 0.0125, 0.063, 0.031, 0.016 and 0.008 ng/µl. EB buffer alone was used as a 'blank'. Then, 25 µl of each DNA standard was prepared in the 384-well microplates (Greiner) and mixed with an equal volume (25µl) of SYBR Green I to give 50 µl per well. For the assay of test samples, 5 µl of each DNA from each sample was diluted in 20 µl EB buffer in the microplate well and mixed with 25 µl SYBR Green I solution. The plates were incubated for 10 minutes in the dark at room temperature, and the fluorescence intensity in each well was measured at room temperature using a spectrofluorometer (POLARstar® Omega) fitted with a 485/520 nm excitation/emission filter set. The experiment was conducted in duplicate and the average reading of the two measures of fluorescence intensity was used to calculate the levels of dsDNA in test samples.

5.3 Results

Figure 5.1 shows a standard curve generated by plotting the value of fluorescence intensity against the DNA concentration for a series of DNA concentration standards. The standard curve demonstrates that MFA has good linearity and sensitivity for the quantification of low levels of DNA. The 'blank' (buffer alone) generates 20 units of fluorescence intensity.

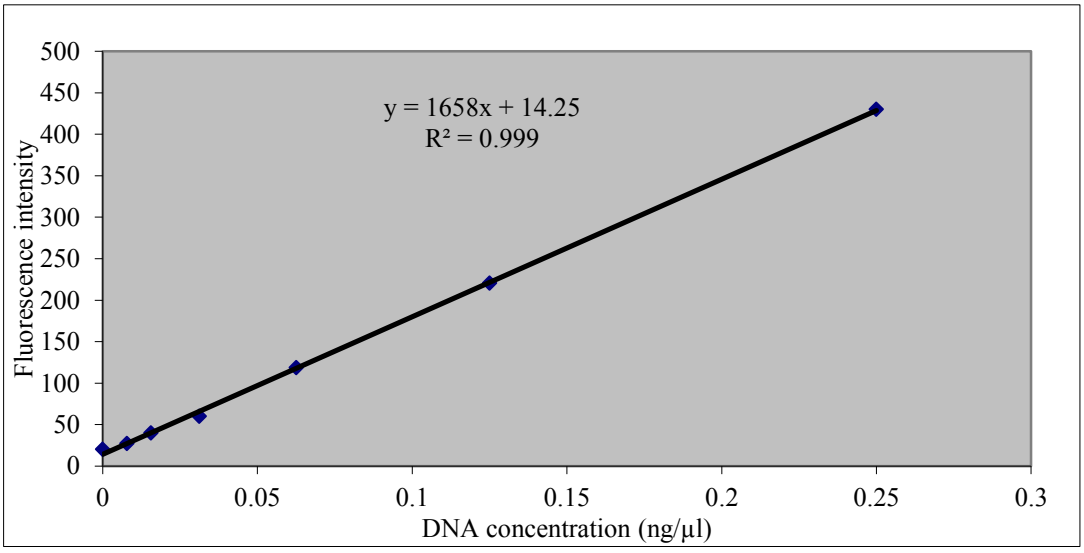


Figure 5.1: Standard curve

Although the experiment was performed successfully in DNA standards, Table 5.2 reveals some problems with this assay in the test samples (see shaded cells). Whole blood samples from Participant 3 (W.T.V) and from Participant 4 (W.T.V and W.N.V) had considerably lower FI values than those obtained from other participants. In fact, some of the FI values were lower than the 'blank' (less than 20), indicating that there was no detectable dsDNA in these samples. Another problem with this assay was the inconsistency of FI values between duplicates, as observed in the W.N.V sample from Participant 6 and the S.T.V sample from Participant 1.

Table 5.2: Fluorescence intensity (FI) of test samples in duplicates

Test samples*		Participant ID					
		1	2	3	4	5	6
W.T.V	FI [1]	304	268	73*	17*	308	298
	FI [2]	327	290	82*	18*	287	282
W.N.V	FI [1]	386	303	446	24*	293	196†
	FI [2]	355	296	447	16*	290	535†
P.T.V	FI [1]	126	155	142	143	161	164
	FI [2]	139	160	168	120	153	141
P.N.V	FI [1]	159	147	150	177	164	152
	FI [2]	160	156	149	157	154	129
S.T.V	FI [1]	162†	186	139	165	154	159
	FI [2]	21†	185	272	165	158	162
S.T.$	FI [1]	152	139	166	167	222	157
	FI [2]	197	170	152	166	202	165

* W.T.V: whole blood collected with the use of tourniquet and vacuum system; W.N.V: whole blood collected by vacuum system without the application of tourniquet; P.T.V: plasma collected with the use of tourniquet and vacuum system; P.N.V: plasma collected by vacuum system without the application of tourniquet; S.T.V: serum, collected with the use of tourniquet and vacuum system; S.T.$: serum collected by syringe and needle with tourniquet applied.

* The FI values are similar to the 'blank', indicating that the dsDNA was barely detectable in these samples.

† Inconsistency of the FI values between duplicates.

Due to the problems mentioned above, the FI values in the W.T.V sample from Participant 3 as well as those in the W.T.V and W.N.V samples from Participant 4 were not applicable for the calculation of DNA concentration in test samples. With regards to the inconsistency of FI values between duplicates, as observed in the W.N.V sample from Participant 6 and the S.T.V sample from Participant 1, the DNA levels in these samples were determined by analysing the first FI value only, because the second FI value is either too high or too low compared to the 'normal' FI values for all W.NV or S.T.V samples. Thus, the first FI value is considered as the average FI value (the second FI value was excluded) in the W.N.V sample from Participant 6 and the S.T.V sample from Participant 1.

Table 5.3 shows the DNA concentrations of each specimen, which were calculated using the formula generated from the standard curve ($y = 1658x + 14.25$). The concentration of dsDNA in the test sample (x) was calculated by inputting the average of FI (AFI) from the test sample after 'blank' values were subtracted (y) into the standard curve equation and solving for x. The final dsDNA concentration (ng/µl) was obtained by multiplying x by 10, the dilution factor, because 5 µl of test sample was added to a final volume of 50 µl in each microplate well.

Table 5.3: Average fluorescence intensity (AFI) after subtraction of 'blank' values and the concentration of DNA (ng/µl) in test samples

Test samples*		Participant's ID (Gender)					
		1 (F)	2 (M)	3 (M)	4 (M)	5 (F)	6 (F)
W.T.V	AFI	295.5	259	NA	NA	277.5	270
	[DNA]	1.7	1.48	NA	NA	1.59	1.54
W.N.V	AFI	350.5	279.5	426.5	NA	271.5	176
	[DNA]	2.03	1.6	2.49	NA	1.55	1.1
P.T.V	AFI	112.5	137.5	135	111.5	137	132.5
	[DNA]	0.59	0.74	0.73	0.59	0.74	0.71
P.N.V	AFI	139.5	131.5	129.5	147	139	120.5
	[DNA]	0.76	0.71	0.69	0.8	0.75	0.64
S.T.V	AFI	142	165.5	185.5	145	136	140.5
	[DNA]	0.77	0.91	1.03	0.79	0.73	0.76
S.T.$	AFI	154.5	134.5	139	146.5	192	141
	[DNA]	0.85	0.73	0.75	0.8	1.07	0.76

* [DNA]: concentration of dsDNA in test samples (ng/µl); F: female; M: male; NA: not applicable; W.T.V: whole blood, tourniquet applied, vacuum system; W.T.V: whole blood collected with the use of tourniquet and vacuum system; W.N.V: whole blood collected by vacuum system without the application of tourniquet; P.T.V: plasma collected with the use of tourniquet and vacuum system; P.N.V: plasma collected by vacuum system without the application of tourniquet; S.T.V: serum, collected with the use of tourniquet and vacuum system; S.T.$: serum collected by syringe and needle with tourniquet applied

Further analysis using SPSS 20 (see Figure 5.2) showed that the concentration of dsDNA in plasma and serum was below 1.5 ng/µl (or less than 1500 ng/ml). The level of dsDNA in whole blood samples was evidently higher than those in plasma and serum, demonstrating the high level of human DNA contamination in samples that contain white blood cells and platelets.

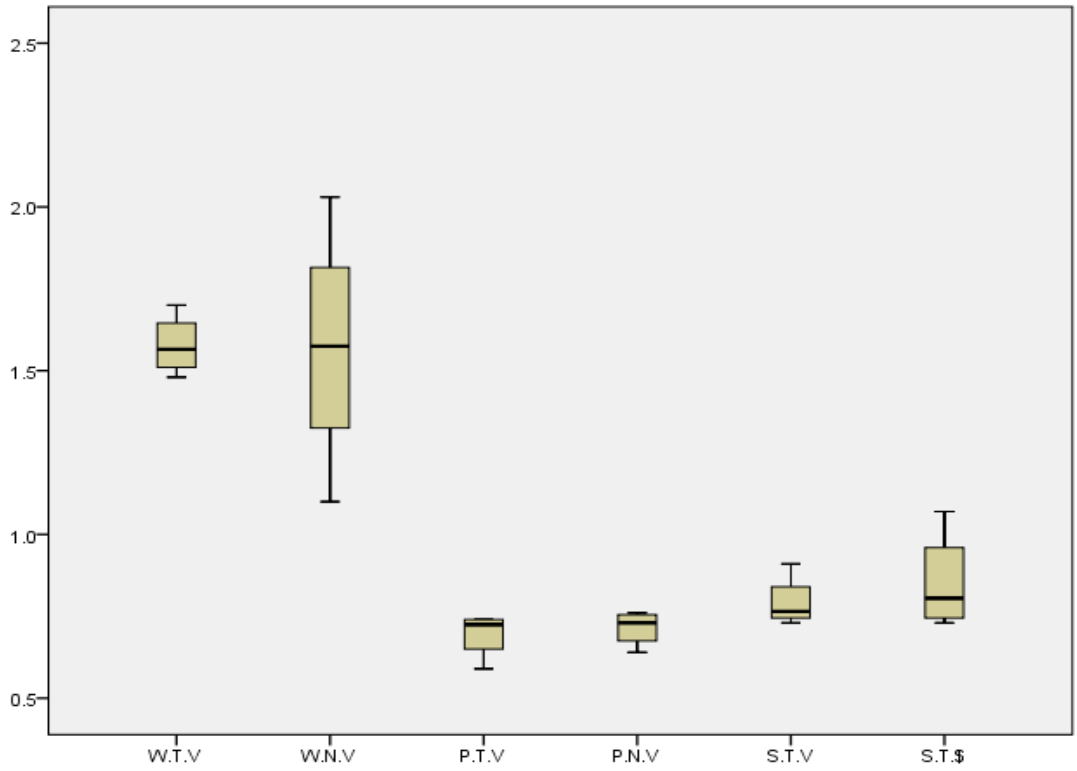


Figure 5.2: dsDNA concentrations (ng/µl) in various blood specimens

W.T.V: whole blood collected with the use of tourniquet and vacuum system; W.N.V: whole blood collected by vacuum system without the application of tourniquet; P.T.V: plasma collected with the use of tourniquet and vacuum system; P.N.V: plasma collected by vacuum system without the application of tourniquet; S.T.V: serum, collected with the use of tourniquet and vacuum system; S.T.$: serum collected by syringe and needle with tourniquet applied

In this study, the population data were not normally distributed, and the variances of the populations to be compared were not equal. Therefore, a non-parametric test (the Wilcoxon Signed-Rank test) was used to compare concentrations of cell-free DNA measured under two different conditions. The comparisons made were: dsDNA levels in plasma versus those in serum that were collected using the same method (P.T.V v. S.T.V); dsDNA levels in plasma collected with the application of tourniquet versus those collected without the application of tourniquet (P.T.V v. P.N.V); and dsDNA levels in serum samples that were collected using the vacuum system versus those collected using the standard syringe (S.T.V v. S.T.$) (see Table 5.4). There were higher levels of dsDNA in serum compared to those in plasma samples ($p < 0.05$). No significant difference was observed in DNA concentrations between specimens obtained with and without the application of tourniquet. Further, the

use of the vacuum system or the standard syringe and needle did not significantly alter the levels of DNA in plasma and serum samples.

Table 5.4: Statistical analysis[a] comparing DNA concentration in various specimens

	P.T.V–S.T.V	P.T.V–P.N.V	S.T.V–S.T.$
Z	-1.992[b]	-.314[b]	-.135[b]
Asymp. Sig. (2-tailed)	.046	.753	.893

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

P.T.V: plasma collected with the use of tourniquet and vacuum system; P.N.V: plasma collected by vacuum system without the application of tourniquet; S.T.V: serum, collected with the use of tourniquet and vacuum system; S.T.$: serum collected by syringe and needle with tourniquet applied

5.4 Discussion

Metagenomics is a new approach to studying biodiversity in a particular environment. The development of NGS technology has enabled the identification of previously unknown species by providing an enormous amount of genetic information from massively parallel sequencing data. This study was conducted to determine the most appropriate type of blood specimen and the blood collection technique that would produce the minimum amount of human DNA contaminant. These parameters were required for the subsequent metagenomics study, which used samples containing scarce genetic material of the pathogens associated with fever.

It has been shown in previous studies that the plasma and serum of healthy individuals contain very small amounts of nucleic acids, ranging from undetectable to hundreds of nanograms per millilitre of sample.[145, 147, 148, 151, 152] The study findings confirm those of previous studies reporting that serum contains a higher concentration of DNA than plasma, possibly because of the release of DNA from the blood cells during the clotting process.[153–155] Thus, it can perhaps be argued that it might have been more reflective of the in vivo situation in the circulation if plasma, rather than serum DNA, is studied using massively parallel sequencing. In addition, this excessive 'contamination' of human DNA in serum samples may interfere with the sensitivity of NGS. Hence, extracting CNA from plasma rather than serum may provide a superior method for detecting small amounts of pathogen nucleic acids in human blood.

Ong et al[156] suggested that factors causing increased sample haemolysis may include pressure differences and needle size, prolonged time between sample collection and analysis, size of collection tubes, difficulty of blood drawing and the use of a vacutainer system. Among those factors, it was contended that the use of vacutainer was associated with the highest rates of haemolysis. However, the findings of the present study contradict this theory. This study demonstrates that blood collection technique (i.e. using a vacuum system compared to a standard syringe and needle) is not associated with levels of circulating DNA. Further, levels of circulating DNA were similar across samples taken with and without the application of a tourniquet.

There is no standard protocol for extracting CNA. Previous research implementing various methods of extraction has shown that the older chemical methods yield larger amounts of CNA than the matrix-binding methods.[157] Traditional methods of extracting DNA and RNA have centred on the separation of protein and other non-DNA/RNA components with phenol/chloroform/alcohol preparations. While still considered to provide the best quantity and quality DNA and RNA for sequencing and cloning studies, this method has drawbacks, especially for clinical identification work. It is slow and laborious, and phenol and chloroform are noxious and require fume hood facilities. The development of kit-based techniques for DNA and RNA extraction has simplified and reduced the time involved in extraction, while providing good quality DNA and RNA for downstream analysis.

The High Pure Viral Nucleic Acid kit (Roche Applied Science) was used in this study because the kit is designed specifically for isolating total nucleic acids (DNA and RNA) from whole blood, plasma and serum samples. Whole blood samples from two participants yielded very low values of FI, indicating scarce/undetectable DNA. The present study hypothesised that the DNA was undetectable because of a failure during DNA extraction or unsuccessful DNA–dye binding that reduced the FI. DNA isolation failed in two whole blood samples, possibly because the blood clot in the whole blood samples was not dissolved completely during the initial stage of DNA isolation (when adding the lysis/binding buffer). Incomplete dissolution of the blood clot inhibited the work of proteinase K and subsequently precluded the yield of DNA. The clot itself might be caused by delay in the DNA extraction.

After blood was drawn, the samples were stored at room temperature in JCU Cairns Clinical School, Cairns Hospital. Plasma/serum separation and DNA extraction were processed at different time intervals (up to 3 hours) after venepuncture. Sample

processing and quantification were performed at QTHA laboratory, JCU Smithfiled campus. Improper mixing of blood and anticoagulant, together with delay in specimen processing, may trigger clotting. This clotting is invisible if it occurs in the centre of the tube, and 10 minutes of incubation with proteinase K might not be sufficient in this case. Other possibilities explaining the low levels of DNA and inconsistency of FI values between duplicates include human error, such as inaccurate pipetting or missing steps, incomplete mixing of reagents and samples, or the presence of inhibitory substances in the samples. Last but not least, the researcher contacted Roche Application Support Centre and was advised that the kit is designed for isolating nucleic acids for PCR or reverse-transcription PCR (RT-PCR) analyses, so its performance might not be optimal for isolating DNA for MFA. This may explain the presence of undetectable DNA, as the kit is not equipped with reagent for eliminating the substances in whole blood samples that inhibit DNA–dye binding. If this is the case, the DNA concentration in other whole blood samples should be much higher than that in plasma and serum samples.

A number of different methods have been used to quantify DNA, including spectrophotometry, fluorometry and quantitative real-time PCR (qRT-PCR). Spectrophotometry based on ultraviolet absorption is an easy and quick method for quantifying DNA. However, this method has several weaknesses: it requires large sample volumes, has a narrow dynamic range and poor sensitivity.[158] It cannot detect DNA below nanogram levels.[159] Moreover, the calculation of DNA concentration is significantly affected by the presence of contaminants such as free nucleotides, single-stranded DNA, RNA, proteins and carbohydrates.[160] These contaminants exhibit significant absorbance at 260 nm, similar to that for DNA, which causes an overestimation of measured DNA. The presence of salts and pH also interferes with DNA concentration.[161] Thus, the spectrophotometry method is accurate and reproducible if the samples are highly purified and available in adequate amount.

Fluorometric assays have been developed for the quantitation of nucleic acids with the use of dyes (fluorophores) that bind to particular types of nucleic acids. Fluorometric measurement of DNA concentration is a simple and quick method, and is more sensitive than absorbance measurement using a spectrophotometer, allowing for the detection of picogram to nanogram levels of dsDNA.[162] A study conducted by Szpechcinski[159] suggested that the total DNA content determined by PicoGreen in a MFA was around 10 times higher than the amplifiable DNA amount measured by qRT-

PCR. The study implied that fluorometry methods were able to enhance the sensitivity of detection of small amounts of CNA, especially in conjunction with a highly sensitive dye such as PicoGreen. Further, Leggate et al[163] demonstrated that the MFA can accurately detect DNA at picogram levels (0.25–2500 pg/µl).

In comparison with qRT-PCR, the sensitivity of fluorometry depends on the dyes used. While the use of fluorescent dyes for sensitive DNA quantitation is costly, the benefits often outweigh the cost disadvantage. Also, there are several dyes that are specific to dsDNA, ssDNA, or RNA. Dyes that bind to dsDNA include ethidium bromide, Hoechst 33258, SYBR Green I and PicoGreen. Below, the features of each dye are discussed, in summary of Sections 8.1 and 8.3 of the Molecular Probes® Handbook[164] and other references.

Ethidium bromide (EtBr) dye is inexpensive and has good sensitivity for detecting DNA and RNA. It can detect as little as 1 ng of nucleic acid on agarose gel electrophoresis (https://www.thermofisher.com/order/catalog/product/A25645). This dye should be used with extreme precautions, as it is a toxic mutagen with possible carcinogenic properties. In addition, EtBr only fluoresces under UV light[165] and causes DNA damage,[166] potentially leading to poor quality DNA. For this reason, EtBr is clearly not suitable for DNA quantification using a microplate reader. This may pose an issue where there are limited samples for downstream analysis that require high-quality DNA, such as PCR or sequencing.

Hoechst 33258 is also inexpensive, but is the least sensitive of the dyes discussed, as detectability is limited by the absence of adenine-thymine (AT) base pairs and the length of the DNA sequence. Hoechst dye contains bisbenzimidazole derivatives, which are supravital minor groove-binding DNA stains with AT selectivity. The dyes bind to all nucleic acids, but AT-rich dsDNA strands enhance fluorescence approximately two times more than guanine-cytosine (GC)-rich strands. Hoechst 33258 can be used to quantitate DNA down to 3 ng in the presence of RNA in agarose gel electrophoresis (https://tools.thermofisher.com/content/sfs/manuals/mp21486.pdf). This dye does not show a significant increase in fluorescence in the presence of proteins. Sodium dodecyl sulfate (SDS) causes significant increase in fluorescence. Another factor in fluorescence yield is pH: a pH of 5 will give a much higher fluorescent yield than will a pH of 8.

PicoGreen is the most sensitive dye for DNA detection. This dye results in a very strong increase in fluorescence (> 1000 times) in the presence of dsDNA.

PicoGreen has been shown to have a sensitivity for detecting DNA down to 250 pg/ml.[167] The assay is not affected by the base content of the DNA sample or by common contaminants such as salts and alcohol. In addition, it does not show a significant increase in fluorescence in the presence of proteins, carbohydrates, ssDNA, RNA or free nucleotides. This enables the quantitation of DNA without purification after PCR amplification. However, PicoGreen is expensive, and high concentrations of the dye are required for analysis.

SYBR Green I exhibits detectabilities almost identical to those of PicoGreen, but is approximately 20–30 times less expensive when using SYBR Green I at a concentration of 1:5,000 or 1:10,000.[158, 163] The sensitivities of SYBR Green I and PicoGreen are more than 1,000 times higher than those of ethidium bromide and Hoechst 33258.[158] SYBR Green I is sufficiently sensitive to measure low concentrations of dsDNA, and the use of SYBR Green I in MFA provides a broad dynamic range of detection (from 2ng/ml to 2 μg/ml) without being affected by the presence of common contaminants such as salts, proteins, and alcohol.[168]

In this pilot study, SYBR Green I was considered reasonably sensitive and economical for detecting low levels of circulating DNA. The use of the MFA method and SYBR Green I dye was applied for ease and cost-effectiveness, with adequate sensitivity within the expected dsDNA concentration range. The study found that the concentrations of circulating DNA were very low, just under 1.5 ng/μl (1,500 ng/ml). These findings were consistent with those of previous cancer studies reporting concentrations of circulating DNA in healthy controls ranging from < 4 ng/ml to > 500 ng/ml.[147–150, 152, 159, 169, 170] This variation was likely due to subject variability, the level of enzyme that degrades DNA (DNase) in the blood, condition of specimen (haemolysis or not) and method used during DNA extraction and quantification.

5.5 Chapter summary

This study measured DNA concentrations in various blood specimens (whole blood, plasma, serum) and sought to determine the effects of various blood collection techniques on the levels of circulating DNA. The study findings provided a rationale for the use of plasma samples for the main study, which utilized NGS technology to investigate pathogens associated with fever. Additional information was obtained with regards to the use of tourniquets and vacuum systems or the standard syringe and

needle for blood aspiration. It was shown that the application of a tourniquet and the speed of blood aspiration did not significantly affect the levels of circulating DNA in healthy volunteers. Since different methods of blood collection did not affect the concentration of human DNA, this study did not yield any recommendation for which blood collection technique should be applied in the subsequent NGS study, with plasma as the preferred specimen.

Chapter 6: Fever Investigation Using Deep Sequencing Approach

6.1 Introduction

Undifferentiated fevers pose a diagnostic challenge for clinicians due to their non-specific clinical features and the indistinctive profiles of routine blood tests. Without any specific diagnosis made, the treatment of fever is often based on an educated guess or syndromic approach.[171, 172] This approach relies on knowledge of the local prevalence of infections, and often leads to inappropriate treatment, especially when the cause of fever is unpredicted or unknown. Some patients may be undertreated, which can increase morbidity and mortality, while others may be overtreated with unnecessary antibiotics, contributing to the emergence of microbial resistance.

It has long been known that infection is the main cause of fever, particularly in the acute stage. Unfortunately, there are hundreds of possible aetiologies of fever, such that conventional diagnostic tools are often either unavailable or restricted to a subset of the 'most likely' infectious causes due to the costs associated with laboratory testing. The wide availability of nucleic acid (i.e. PCR-based) assays in clinical laboratories provides sensitive and specific detection of pathogens. However, while techniques such as multiplex PCR can provide simultaneous detection of multiple pathogens, this approach is impractical for more than a handful of pathogens in any one assay,[173, 174] and is not capable of detecting novel pathogens.[175] Diagnostic microarrays can expand detection capacity considerably, allowing for the simultaneous detection of tens of pathogens or more,[176, 177] but these too are extremely limited for the detection of novel or emerging pathogens.

Broad-spectrum diagnostic tools with enhanced detection capacity are needed to inform diagnosis and to facilitate more effective treatment for AUF. The advent of NGS[115] provides a basis for unbiased AUF diagnosis with the potential to detect any known cause of AUF, as well as the capacity to identify novel and emerging pathogens.[132, 178, 179]

This chapter presents the main study, which characterises the aetiologies of AUFs using a NGS platform. This method is referred to as a deep sequencing approach, and has been used in previous studies to determine the diagnosis of dengue-like

illnesses[132] and acute haemorrhagic fever.[129] The hypothesis that a deep sequencing approach is feasible for routine investigation in patients with AUF, if proven, could have profound implications for the diagnosis and management of AUF worldwide.

6.2 Materials and methods

6.2.1 Participants and samples

The study was approved by the HREC of the Cairns and Hinterland Health Service District, as well as the Greenslopes HREC and JCU HREC (see Chapter 3, Section 3.3.3) to recruit participants of both genders who met these inclusion criteria:

1. 16–65 years of age;
2. presence of fever for 21 days or less;
3. documented temperature of at least 38 °C or history of fever associated with symptoms of feeling cold or shivering;
4. no obvious cause of fever after initial investigations (available within 6 hours);
5. have undergone diagnostic tests for at least one specific infectious agent (can include rapid serological tests and malaria parasite screening);
6. willing to provide, under informed consent, acute and convalescent blood samples for NGS study and validation of NGS results, which may include conventional tests such as blood culture, PCR and serology.

Recruitment was conducted between 30 November 2012 and 5 December 2013. Hospital databases were used to identify potential cases of AUF. These databases included AUSLAB® and Auscare®, which record the details of pathology findings, and the Merlin Web®, which records radiology findings. Initially, potential participants were identified from AUSLAB® and Auscare® through requests for tests for infectious agents. Subsequently, the Merlin Web® database was used to exclude patients with obvious focal infections, such as community-acquired pneumonia or tuberculosis. Following the identification of a potential participant, the patient's notes were reviewed and the patient was interviewed to assess their eligibility for the study. The interview results were collated on a data collection form (see Appendix J).

Patients were excluded if their cause of fever was immediately identified by the attending physician. Further, patients were excluded if they had immunosuppressive conditions, suspected nosocomial infections (infection acquired during hospitalisation) or febrile neutropenia. Information about the study (see Appendix K) was provided to

patients who met the inclusion criteria, and then informed consent (see Appendix L – O) was requested from those patients.

There was no competing interest in this study. The funding and support providers did not have a financial interest in the outcome of the research. The investigatory team was not involved in patient care and did not have a relationship with the patients. Participants were free to withdraw from the research at any time. Non-study tests were ordered by the treating doctor and no additional tests were ordered by the investigatory team. Thus, a specific diagnosis at the time of patient recruitment was unavailable. Subsequent investigation(s) determined by the attending doctors ascertained an aetiological diagnosis in a subset of study participants (control subjects), while other participants (test subjects) remained undiagnosed. Figure 6.1 provides the flow chart of the patient selection process.

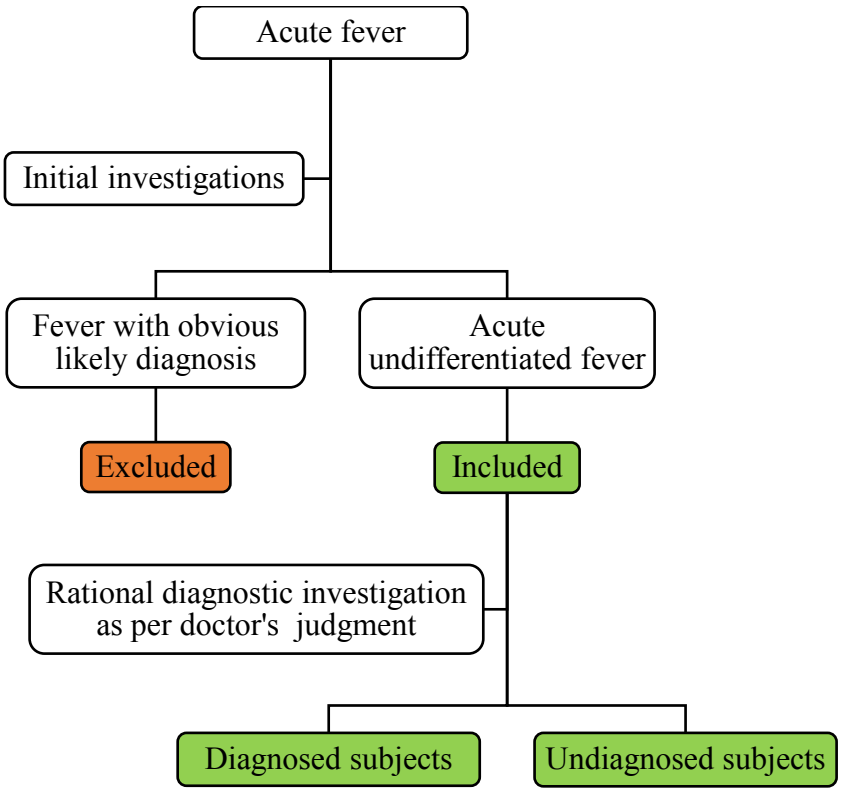


Figure 6.1: Flow chart of patient selection

Definitions:

Acute fever is an increase in body temperature of 38 °C or more for a period of 21 days or less. **Initial investigations** refers to comprehensive clinical assessment and basic laboratory and radiology tests; this included tests that are normally reported within 6 hours. **Comprehensive clinical assessment** includes complete history taking and thorough physical examination. **Basic laboratory tests** usually include complete blood count and urinalysis. **Basic radiology tests** could include chest X-ray, abdomen and pelvic X-ray, and ultrasonography. **Fever with obvious likely diagnosis** is any case of fever with definite diagnosis immediately after initial investigations. This includes fever cases with an obvious focus of infection or local inflammation, such as community acquired pneumonia, urinary tract infection, skin and soft tissue infection, bone and dental infection, pelvic inflammatory disease and intra-abdominal infection. **Acute undifferentiated fever** is any case of acute fever with unclear aetiology and the results of initial investigations are not conclusive in achieving a diagnosis. Thus, the condition is characterised by a requirement for further investigation to explain the cause of fever and to consider differential diagnoses. **Rational diagnostic investigations** are further tests as judged necessary by an attending doctor to determine the cause of fever, such as further serology, CSF analysis and/or advanced radiology tests (e.g., CT scan and magnetic resonance imaging [MRI]). Samples were collected from both groups of participants (diagnosed subjects and undiagnosed subjects) for fever investigation using the deep sequencing approach.

Ten millilitres of peripheral blood samples were collected on presentation (acute sample) and after 2–3 weeks (convalescent sample). Samples were collected either directly from patients or from the Pathology Department of Cairns Hospital. For those obtained directly from the patients, the blood samples were collected by venous puncture into EDTA-containing (purple-top) Vacuette® tubes (Greiner Bio-One). Plasma separation was conducted at room temperature by centrifugation of blood samples at 2,000 *g* for 15 minutes within 1 hour of collection. Plasma aliquots were immediately stored at -70 °C until used for DNA and RNA isolation.

Samples were collected from the Pathology Department if (a) the patient was eager to participate in the study but refused to undergo phlebotomy, or (b) the patient was willing to undergo phlebotomy but the nurse could not draw a sufficient volume of blood, or (c) the patient signed the consent form several days after the onset of their illness. In such cases, requesting retrospective samples from Pathology Department was deemed necessary to obtain an optimal sample with the presumed presence of a pathogen.

The procedures for plasma and serum separation by the Pathology Department were as follows. Plasma separation was conducted with centrifugation of blood samples in EDTA-containing (purple-top) Vacuette® tubes (Greiner Bio-One) for 15 minutes at 2,000 *g* at room temperature. Serum samples were obtained by drawing blood into red-top Vacuette® tubes (Greiner Bio-One) containing clot activator and left to coagulate before being centrifuged for 10 minutes at 2,000 *g*. The isolated plasma and serum were stored at 4 °C at the Pathology Department of Cairns Hospital up to two weeks, and were given to the investigator after all requested tests were completed. Upon receipt from Pathology Department, those samples were centrifuged for 5 minutes at 2,000 *g* at room temperature, aliquoted and immediately stored at -70 °C until used for DNA and RNA isolation.

6.2.2 Sample preparation and sequencing

Plasma and serum samples were transferred in a cooler box to the QTHA laboratory for further processing. In this university laboratory, DNA and RNA preparations for sequencing were performed in duplicate according to the workflow illustrated in Figure 6.2.

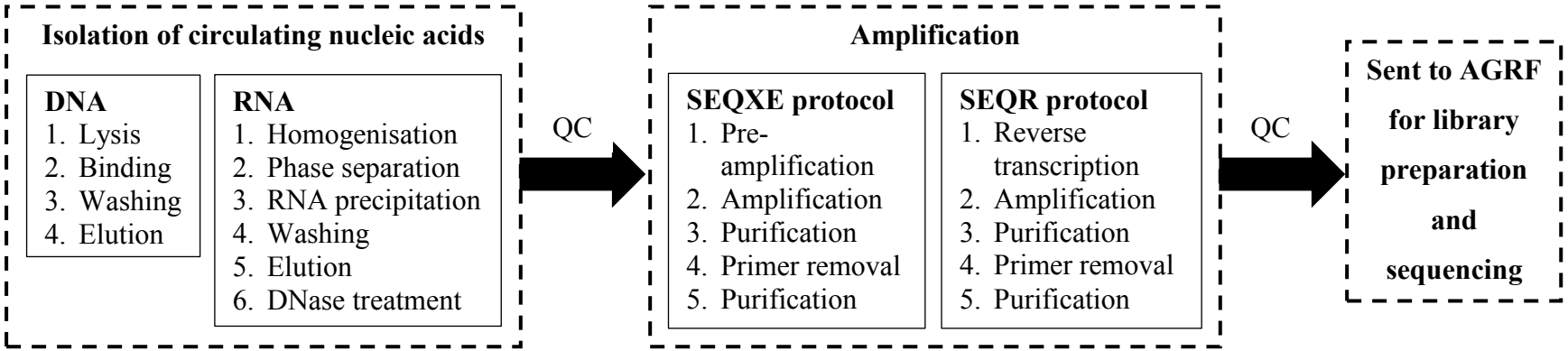


Figure 6.2: Workflow of sample preparation for sequencing

DNA was isolated from 200 µl of sample using the QIAamp® DNA Mini Kit (Qiagen, catalogue number 51304) according to the manufacturer's spin protocol for DNA purification from blood or body fluids. DNA was eluted in 20 µl of water and stored at -30 °C. RNA was isolated from 250 µl of sample mixed with 750 µl of TRIzol® LS reagent (Life Technologies, catalogue number 10296-010) with the addition of 3 M sodium acetate and RNase-free glycogen (Thermo Scientific, catalogue number R0551) during RNA precipitation. Following RNA isolation, genomic DNA was removed from RNA samples using DNase I, Amplification Grade (Sigma-Aldrich, catalogue number: AMPD 1) as per manufacturer's instructions. The isolated DNA-free RNA was suspended in 20 µl of RNase-free water and stored at -70 °C.

The concentration of isolated nucleic acids and the RNA integrity were determined using the Agilent 2100 Bioanalyzer in conjunction with a suitable kit from Agilent Technologies (Waldbronn, Germany). The recommended protocol for bioanalyser analysis is as follows. Briefly, 1 µl of DNA/RNA sample was added into the sample well loaded with gel-dye mix and buffer. The chip was vortexed for 1 minute and run on the bioanalyser. Total DNA/RNA quantity and RNA integrity were determined against internal standards using the 2100 expert software tool (Agilent Technologies, Waldbronn, Germany).

Amplification of DNA and RNA was conducted according to the SeqPlex Enhanced DNA Amplification Kit protocol and SeqPlex RNA Amplification Kit protocol (Sigma-Aldrich, catalogue numbers SEQXE and SEQR, respectively) (see Figures 6.3 and 6.4).

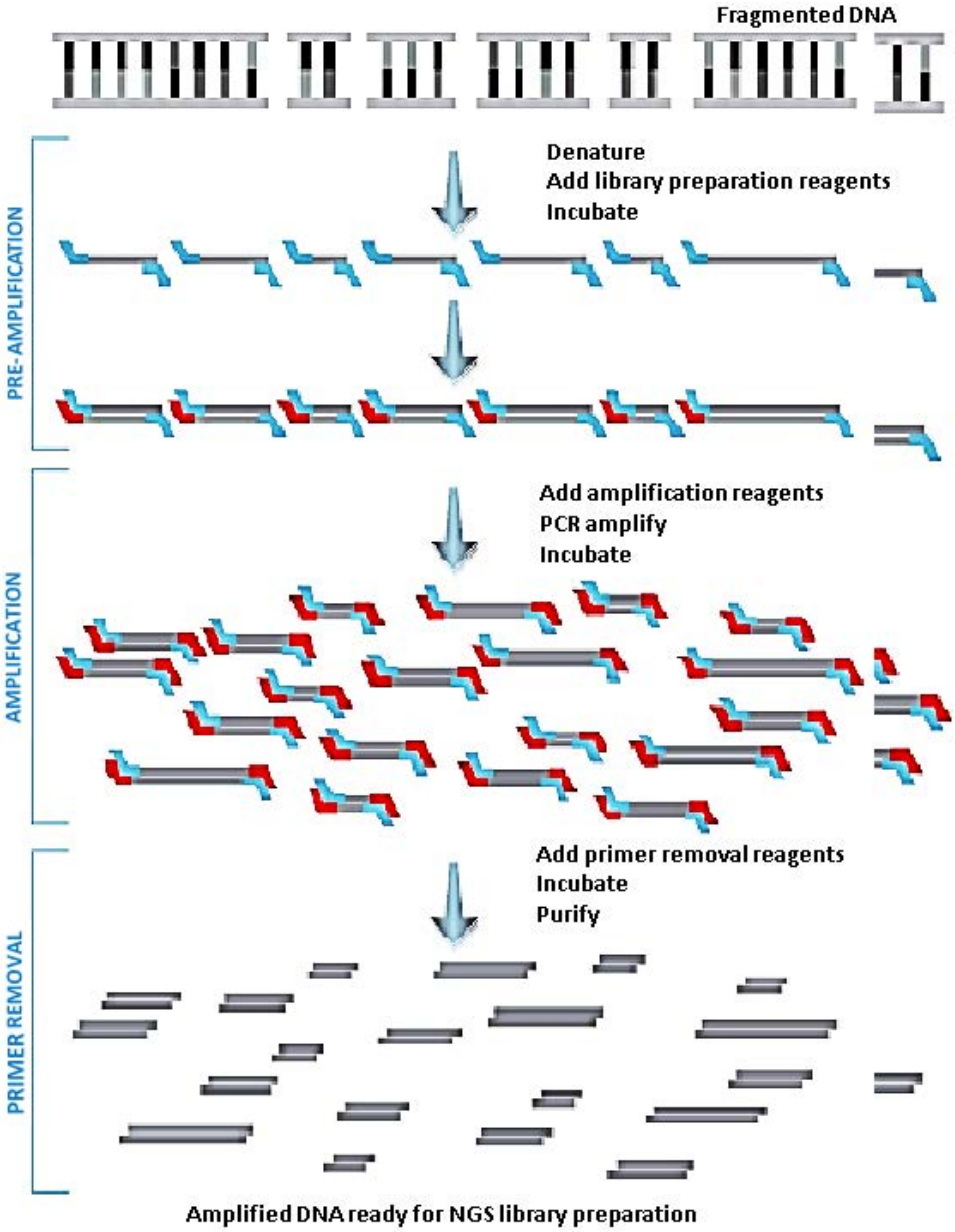


Figure 6.3: SEQXE process workflow

(Reproduced from[180] with permission)

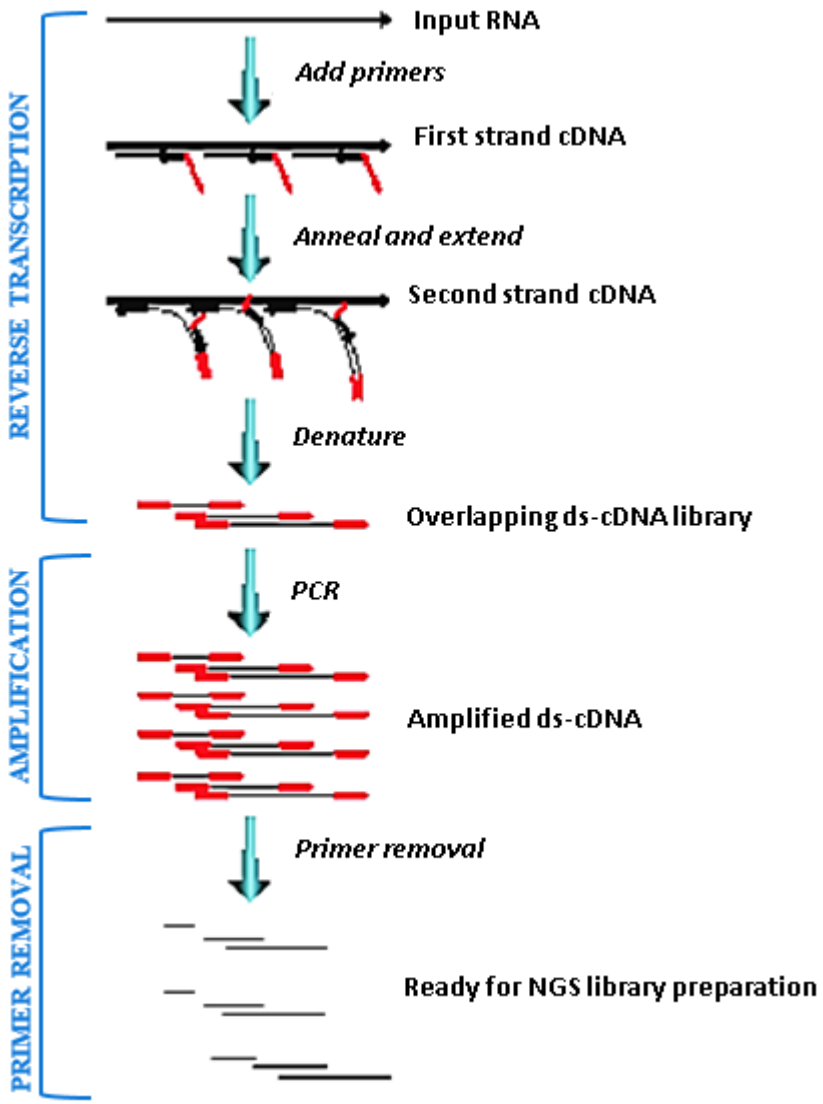


Figure 6.4: SEQR process workflow

(Reproduced from[181] with permission)

GenElute PCR Clean-Up Kit (Sigma-Aldrich, catalogue number: NA1020) was used for the purification of products from the SeqPlex DNA and RNA amplification kits (amplicons). Amplicon purification was performed twice (after amplification and after primer removal steps) to ensure the removal of components from the reactions, such as excess primers, nucleotides, DNA polymerase, oil and salts. The purified amplicons (double-stranded DNA/cDNA) were then stored at -30 °C prior to sequencing.

The quantity and quality of the DNA/cDNA amplicons were determined using gel electrophoresis (1.5% agarose) and a NanoDrop 2000 spectrophotometer (Thermo Scientific), as requested by the commercial sequencing service used for the study: the

AGRF. The protocol for preparing gel electrophoresis was as follows. Briefly, 1.5 g agarose was mixed with 100 ml 1x TBE buffer in a 250 µl Erlenmeyer flask to generate the agar with 1.5% concentration. The mixture was heated in a microwave for 30 seconds, and this was repeated 2–3 times until the agarose was dissolved. After the agarose had cooled, 10 µl of SYBR Green I (Life Technologies) was added into the flask and swirled. Following this, the agar solution was poured into a tray that has been equipped with combs. Once the agar was set, 100 bp ladder (New England Biolabs) and DNA/cDNA samples diluted in loading dye were pipetted into the sample wells. The next step involved placing the tray into a gel tank containing 1x TBE and then running the electrophoresis at 80V voltage. As for the NanoDrop 2000 spectrophotometer, the total absorbance was measured at 260 nm, according to the manufacturer's instruction.

The products of the whole-genome amplification along with the image of gel electrophoresis and the data from the spectrophotometer analysis were sent to AGRF Melbourne via a commercial courier (LabCabs). The AGRF ran quality assessment prior to further processing of the samples. The samples that passed AGRF quality assessment were processed into library preparations using the TruSeq Nano DNA Library Preparation kit protocol (Illumina). Finally, paired-end (PE) 100 bp sequencing was conducted using the Illumina HiSeq 2000 instrument.

6.2.3 Bioinformatics analysis

Bioinformatics analysis to identify pathogens associated with fever was performed on two cloud computing servers: BaseSpace® (Illumina) and CLC Genomics Workbench (Qiagen). Analysis on BaseSpace® was carried out using the Kraken program, an ultrafast and highly accurate program for assigning taxonomic labels to short DNA sequences (reads).[182] During Kraken analysis, the query sequence (read) is computationally chopped into *k-mers* (subsequences of length *k*). Each *k-mer* is then mapped to the nearest taxon in the lowest common ancestor hierarchy of the genomes that contain that *k-mer* in the database. This process generates a taxonomy tree, and subsequently the program puts the taxon and its ancestors into a complete classification tree. A second analysis on the CLC Genomics Workbench was performed to analyse unclassified reads from the first analysis, as well as to validate the findings of the NGS analysis. Finally, the results of the bioinformatics analysis, in conjunction with

supporting clinical data and laboratory findings, were used to inform diagnosis. Figure 6.5 illustrates the workflow for the bioinformatics analyses.

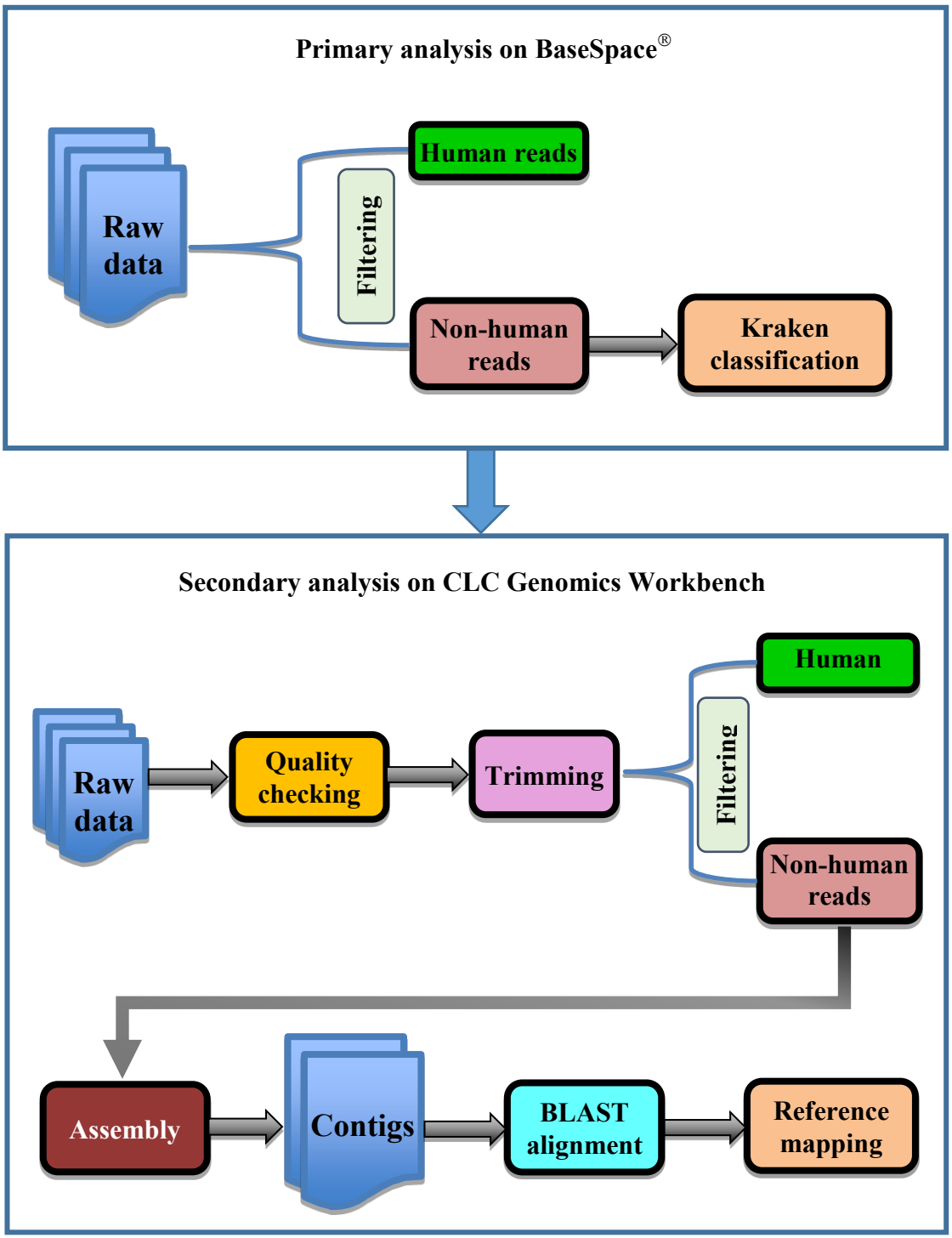


Figure 6.5: Analysis workflow

Primary and secondary analyses were performed sequentially for all samples and performed in parallel for the purpose of validation. Sequential analysis means that all reads were uploaded to the BaseSpace® server (primary analysis), followed by analysis of non-human reads not classified by Kraken on the CLC server (secondary analysis). Validation of the NGS analysis was performed on control samples only by analysing all reads using the BaseSpace® and CLC servers in parallel.

During the first analysis, a straightforward pipeline was carried out to immediately examine microbial communities present in the sample. First, the raw sequence data obtained from AGRF were uploaded onto the BaseSpace® server; then the human sequences were filtered using the program SNAP.[183] Following this, the remaining non-human sequences were classified by Kraken based on their matching organisms (at the lowest common ancestor hierarchy) in the MiniKraken database. The MiniKraken database is a simplified form of the Kraken database, which is constructed from the complete bacterial, archaeal and viral genomes in the National Centre for Biotechnology Information's (NCBI) Reference Sequence database (RefSeq). Kraken's default database requires 70 GB of RAM (random-access memory, a form of computer data storage). In contrast, the MiniKraken database only requires 4 GB of data storage due to the removal of *k-mers* from the database. The results of the analysis in BaseSpace® were presented using the Krona program[184] and examined in an Excel spreadsheet.

The secondary analysis was run in both sequential and parallel modes in relation to the primary analysis. In sequential mode, any non-human reads not classified by Kraken were imported to the CLC Genomics Workbench server. After importing these reads, the quality of the sequence data was examined using the FastQC tool.[185] The reads were then processed using the Trim Sequences tool to remove adapters, low quality bases (using a Phred quality score 33 as the threshold), ambiguous nucleotides, terminal nucleotides (25–35 nucleotides from the 5' end) and short sequences (less than 24 nucleotides). Following this, the next step was filtering to remove from the dataset any remaining human reads that had passed through BaseSpace's® filtering step. The short overlapping sequence reads that did not map to the human genome could now be used as input for the *De Novo* Assembly tool in order to generate longer contiguous sequences (contigs). The last step was the alignment of contigs to reference databases held at the NCBI using the Basic Local Alignment Search Tool (BLAST).[186] This alignment process was more computer-intensive than that used in the primary analysis with Kraken, because in this case, contigs were aligned against several nucleotide databases at NCBI and not just the RefSeq database.[187] BLASTn optimised for highly similar sequences (megaBLAST) was employed to search for homologies between query sequences and reference sequences in the NCBI database. A similarity was considered significant at Expect values (E-values) $\leq 10^{-5}$.

Validation of the bioinformatics analysis was performed by running the secondary analysis in parallel to the primary analysis. This was performed on control samples only, for which the agent causing fever was known. The workflow in the secondary analysis was similar with that for the analysis of non-human reads not classified by Kraken, except that the input data consisted of total reads from the Illumina HiSeq 2000 instead of the Kraken unclassified reads. The pre-processing, filtering, assembly and alignment steps were performed as above. After the alignment step, the resulting contigs of the agent causing fever (if found) were mapped to the reference genome.

6.3 Results

6.3.1 Participants and samples

Database screening at Cairns Hospital identified 241 potential AUF cases. Of these, 197 patients were excluded, either because they met exclusion criteria or because they had been discharged prior to the time of recruitment. Of the remaining 44 patients, four declined to participate in the study.

Informed consent and blood samples were obtained from 40 patients admitted to Cairns Hospital. Of these, two patients were excluded—one had influenza (ID# 036) and another had Creutzfeldt–Jakob syndrome, a progressive motorneuron disease (ID# 041). One patient (ID# 025) was recruited from Cairns Private Hospital after he had been admitted for a few days. However, this patient was excluded because his acute sample could not be obtained retrospectively.

Thus, in total, 38 patients were recruited to the study, 22 male and 16 female, with a mean age of 39.6 (*SD* = 14.9, median = 38.5, interquartile range = 27.5–55). Ten patients had specific diagnoses made, including six patients with dengue fever (ID# 004, 005, 017, 020, 031, 034), one patient with measles (ID# 024), one patient with Hepatitis C (ID# 006), one patient with *Leptospira* (ID# 010) and one patient with a *Streptococcus pyogenes* infection (ID# 032). The other 28 patients were undiagnosed.

A total of 27 plasma samples were obtained directly from the patients, and a further 11 from the Pathology Department of Cairns Hospital. Samples from the Pathology Department consisted of plasma and serum (from patient ID# 011, 014, 016, 019, 027, 028, 032, 033, 034) or serum only (from patient ID# 30 and 37). Despite patients' statements of agreement during the recruitment process to provide both acute

and convalescent samples, compliance with collection of the convalescent sample was below expectation. Paired samples were available from ten patients only (ID# 009, 011, 012, 014, 015, 018, 021, 026, 028, 032).

6.3.2 Sample preparation and sequencing

The Agilent 2100 Bioanalyzer was used for the quantification, sizing and quality control of CNA. The total amounts of DNA and RNA isolated from plasma or serum samples are usually at picogram levels, highly fragmented and highly degraded, as shown by electropherogram and a gel-like image in Figure 6.6. These extremely small quantities of DNA/RNA are much smaller than those required for successful NGS library preparation, with Roche 454 calling for microgram levels of input nucleic acid, while the Illumina Miseq and HiSeq platforms require nanograms of DNA/RNA, as does ABI SoLiD.[115, 188] In this study, the amount of nucleic acid required for Illumina HiSeq sequencing was at least 100 ng to allow AGRF to perform quality control and library preparation prior to sequencing. Thus the amplification step was essential in order to increase the quantity of DNA/RNA.

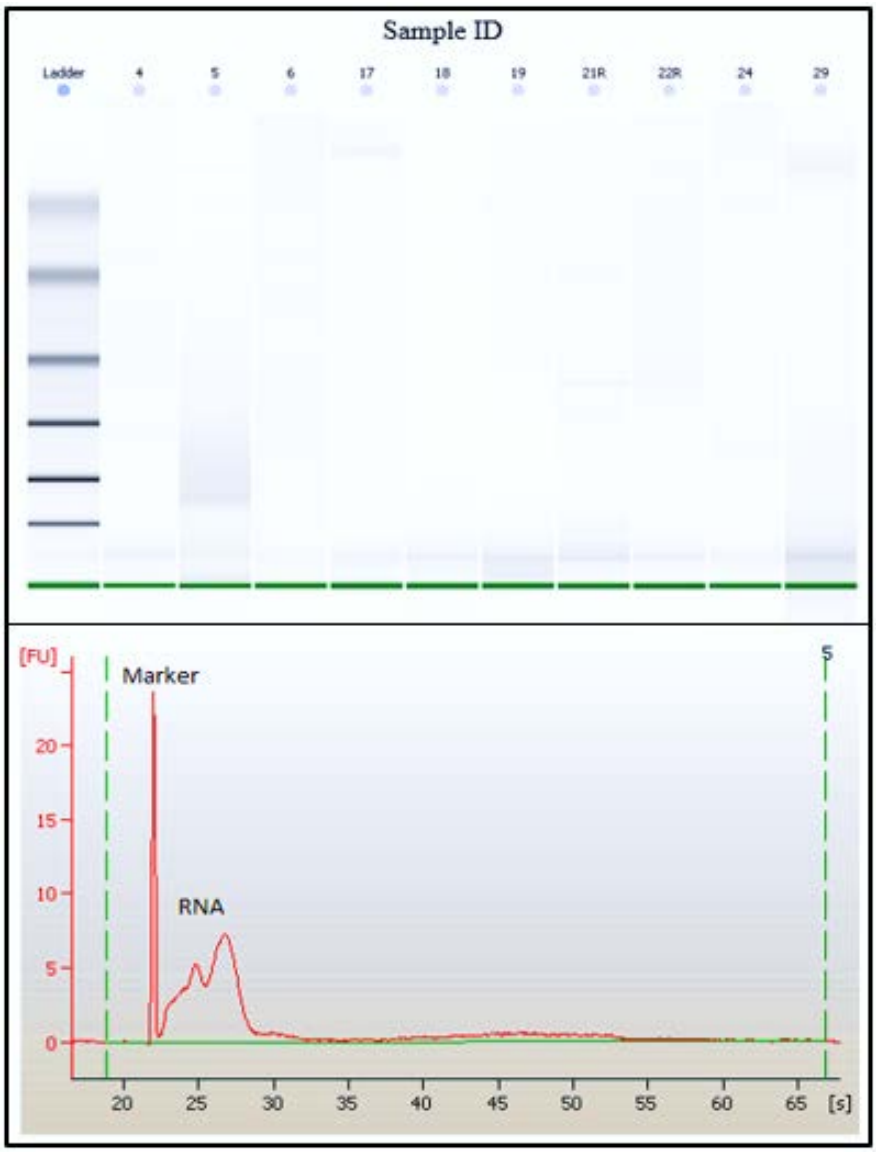


Figure 6.6: Typical quantity and quality of circulating nucleic acids, measured using bioanalyser with detection limit of 200 pg per band

Gel image (above) shows low quantity of highly fragmented nucleic acids, which appear as low-molecular-weight smears. Electropherogram (below) shows the quality of RNA sample isolated from the plasma of one representative subject (ID# 005). RNA of acceptable quality exhibits the 18S and 28S subunits as two distinct bands. This electropherogram shows highly degraded RNA in which the typical 18S and 28S subunits were not detectable, but all of the RNA was grouped around the marker band. In this sample, the RNA integrity number (RIN) is 2.4.

Amplification was performed on 28 samples (14 DNA and 14 cDNA samples) that met the amplification kit's input requirements. These samples originated from 21 patients. Figures 6.7 shows DNA and cDNA amplicons in a 1.5% gel electrophoresis with 100 bp ladder and SYBR Green I dye. Analysis using a NanoDrop 2000

spectrophotometer shows the concentration and purity of the DNA/cDNA (see Table 6.1). Generally, DNA amplicons (ID# 2, 8, 10, 11, 12, 14, 19, 23, 27, 28, 30, 32, 39, 40) were present in higher concentrations and longer fragments than cDNA amplicons (ID# 2c, 5c, 14c, 17c, 18c, 19c, 20c, 21c, 24c, 31c). The average DNA and cDNA concentrations were 44.2 ng/µl (*SD* = 14.96) and 27.62 ng/µl (*SD* = 21.92) respectively. The size of the DNA fragments was ~200 bp, compared to ~100 bp for the cDNA fragments. The purity of DNA and cDNA samples was quite good, as indicated by the value of A 260/280 and A 260/230, which is ≥ 1.8 in the majority of the samples.

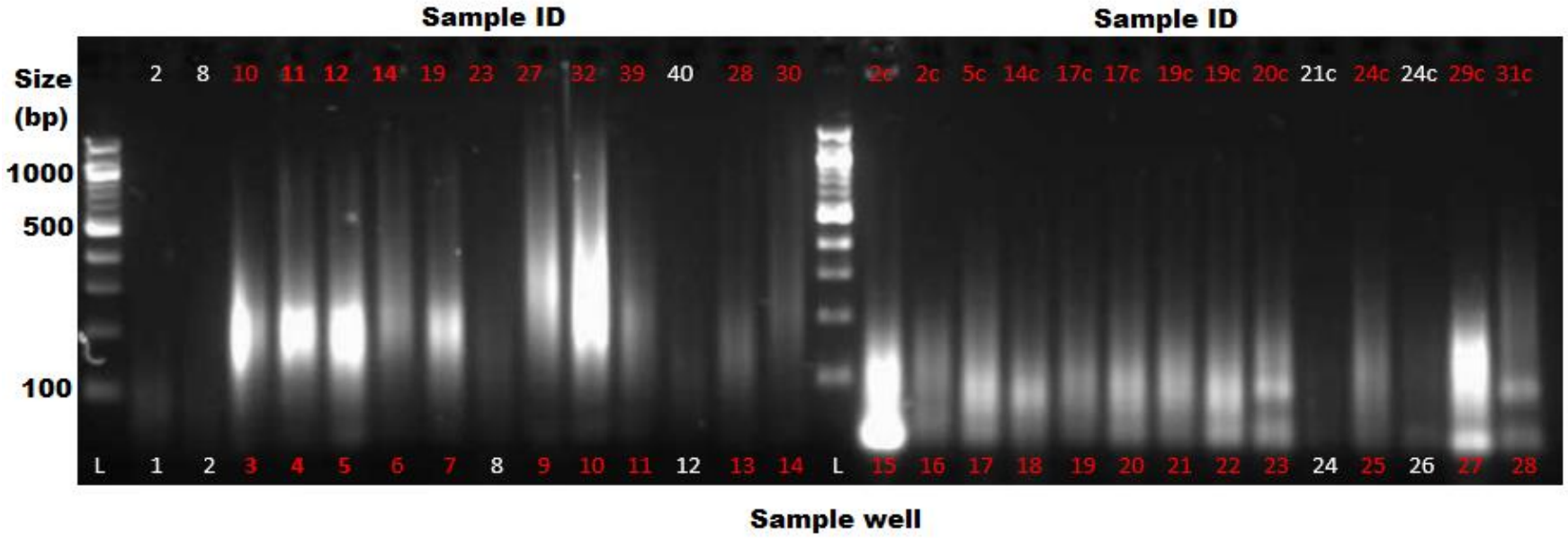


Figure 6.7: DNA and cDNA samples in 1.5% gel electrophoresis pre-casted with SYBR Green I dye

SYBR Green I dye has detection limit of 30–40 pg per band. Identification (ID) number written in red colour indicates high level of DNA/cDNA and thus potential for sequencing.

Table 6.1: Concentrations of DNA and cDNA in total volume of 20 µl per sample, measured by NanoDrop 2000 spectrophotometer with detection limit of 2 ng/µl

Patient ID	Sample ID	Sample type	NanoDrop measurement		
			Concentration (ng/µl)	A 260/280	A 260/230
002	2	Plasma DNA	40.0	1.81	1.55
	2c1 (duplicate 1)	Plasma cDNA	18.4	2.00	2.14
	2c2 (duplicate 2)	Plasma cDNA	16.2	1.88	1.91
005	5c	Plasma cDNA	26.3	1.76	1.17
008	8	Plasma DNA	27.4	1.85	1.78
010	10	Plasma DNA	46.4	1.85	2.18
011	11	Plasma DNA	40.5	1.86	2.18
012	12	Plasma DNA	49.7	1.84	1.89
014	14	Plasma DNA	34.0	1.83	1.86
	14c	Plasma cDNA	36.3	1.72	0.91
017	17c1 (duplicate 1)	Plasma cDNA	15.3	1.89	1.16
	17c2 (duplicate 2)	Plasma cDNA	17.9	1.90	1.5
019	19	Plasma DNA	44.0	1.88	2.03
	19c1 (duplicate 1)	Plasma cDNA	14.1	1.91	1.66
	19c2 (duplicate 2)	Plasma cDNA	91.7	1.59	0.66
020	20c	Plasma cDNA	47.6	1.70	0.74
021	21c	Plasma cDNA	7.3	1.85	1.32
023	23	Plasma DNA	25.4	1.85	2.15
024	24c (duplicate 1)	Plasma cDNA	18.8	1.90	1.54
	24c (duplicate 2)	Plasma cDNA	16.3	1.80	0.74
027	27	Plasma DNA	46.8	1.85	2.25
028	28	Plasma DNA	61.2	1.71	0.9
029	29c	Plasma cDNA	44.2	1.87	2.27
030	30	Serum DNA	41.6	1.92	1.95
031	31c	Plasma cDNA	16.3	1.79	1.19
032	32	Plasma DNA	85.0	1.86	2.16
039	39	Plasma DNA	32.6	1.97	2.21
040	40	Plasma DNA	44.2	1.81	1.27

Of the 28 DNA/cDNA samples that were prepared, only 22 samples passed the AGRF's quality assessment. The acceptable samples originated from 17 patients (ID# 002, 005, 010, 011, 012, 014, 017, 019, 020, 024, 027, 028, 029, 030, 031, 032, 039). Library preparation was successfully performed on all 22 samples, and these were pooled in a single sequencing lane. AGRF guaranteed 150 million reads (~35 Gb of data) per lane. Despite the success of the library preparation, AGRF could only provide ~20 Gb of sequencing data, so AGRF repeated the sequencing on the same library to produce another ~20 Gb of data. Upon the completion of sequencing, AGRF checked the quality of the ~40 Gb of data prior to providing the results. Sequencing data from the Illumina HiSeq platform were presented in FASTQ format.[189] This format records both a nucleotide sequence and its corresponding quality score with each base position. Both the sequence letter and quality score are encoded with a single American Standard Code for Information Interchange (ASCII) character for brevity.

6.3.3 Summary of bioinformatics analysis

Figure 6.8 shows a summary of the primary analysis using Kraken for all 22 DNA/cDNA samples sequenced by the AGRF. Deep sequencing generated between 4 and 20 million reads per sample. As expected, the majority of reads (43.67–94.38% of the total reads in each sample) were of human origin. Only a small proportion (8.31–43.55%) of non-host reads could be classified into particular taxa, except in sample ID 14c, where the majority (76.64%) of non-host reads could be classified by Kraken. The number of viral, bacterial and archaeal species reported by Kraken varied considerably, from 146 to 505 species per sample.

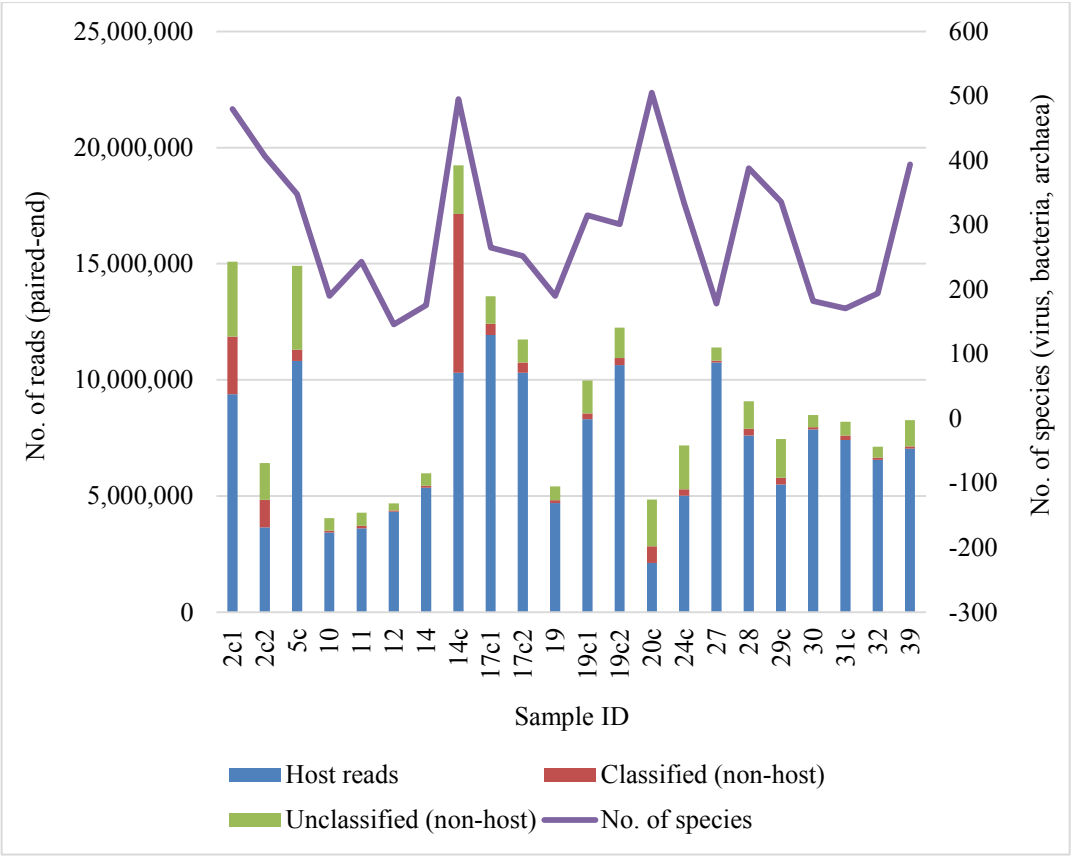


Figure 6.8: Kraken analysis on reads generated by Illumina HiSeq 2000

cDNA samples from patient ID# 002, 017 and 019 were prepared and sequenced in duplicates (sample ID# 2c1, 2c2, 17c1, 17c2, 19c1, 19c2).

The secondary analysis using the CLC Genomics Workbench facilitated further classification of reads that were previously unclassified by Kraken. Analysis with the CLC Genomics Workbench revealed that the Kraken unclassified reads still contained human sequences, accounting for 18.92–81.71% of total contigs in each sample. Non-host contigs were detected, including viruses, bacteria and other organisms such as archaea (e.g., *Sulfolobus* sp.), fungi (e.g., *Saccharomyces* sp., *Cryptococcus* sp., *Penicillium* sp.), algae (e.g., *Navicula gregaria*), plant (e.g., rice, tomato, grain, tobacco), protozoa (e.g., *Toxoplasma gondii*, *Plasmodium berghei*), human parasites (e.g., roundworm, tapeworm, pinworm, *Schistosoma mansoni*) and larger animals (e.g., snail, fish, rat, monkey, orangutan, gorilla). At the end of analysis, a small proportion (2.71–29.57%) of total contigs in each sample remained unclassified (see Figure 6.9).

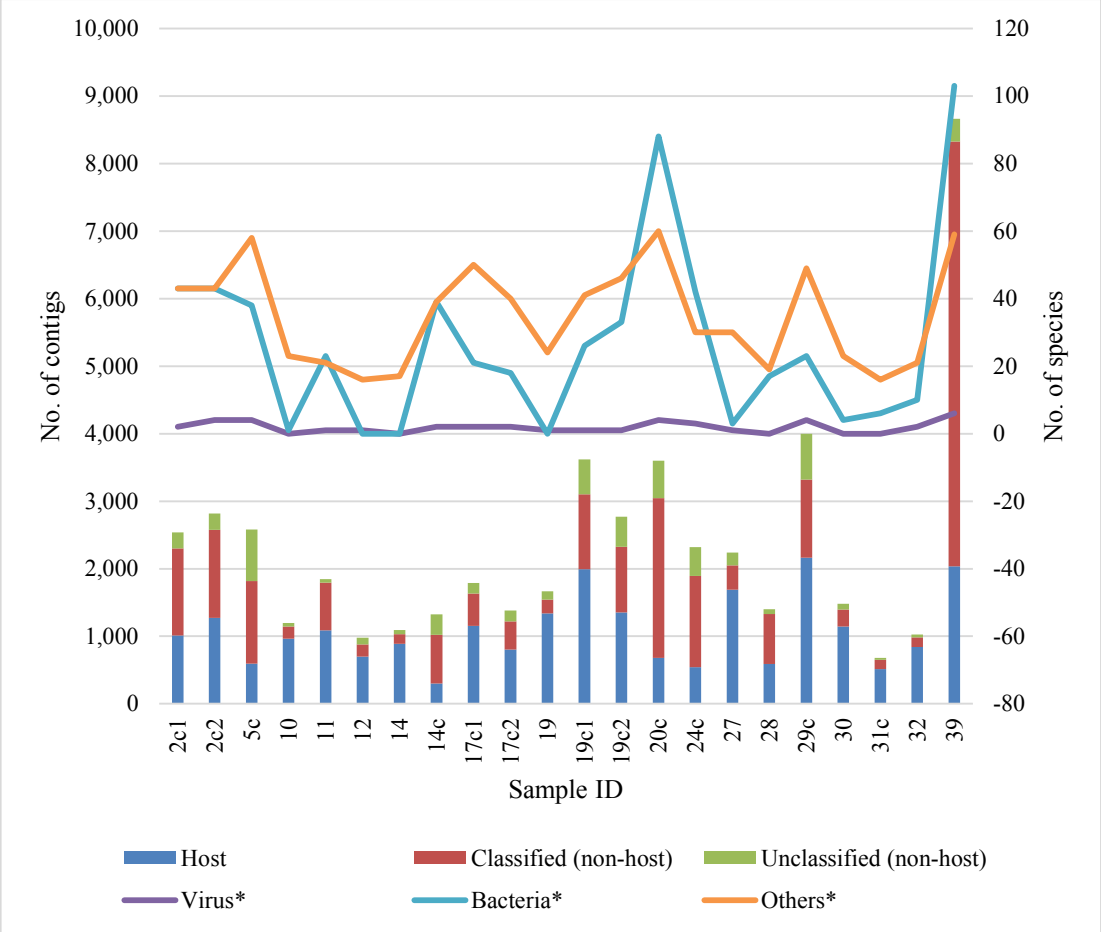


Figure 6.9: CLC Genomics Workbench analysis on Kraken unclassified reads

* BLASTn optimised for highly similar sequences (megaBLAST) was performed to search the homology between query sequences and reference sequences in database. If multiple significant similarities matched with a single species, only the highest scoring hit was included in the table. A similarity was considered significant at E-value $\leq 10^{-5}$.

Validation of NGS diagnosis was successfully performed in two patients with dengue (ID# 005, 017) in whom the secondary analysis on the CLC Genomics Workbench server detected the agent causing fever (see Table 6.2). In the first dengue case (ID# 005), the CLC Genomics analysis successfully constructed 8 contigs of dengue virus 1 from the NGS dataset of this sample. These contigs were between 166 and 1,328 bp in length. In the second dengue case (ID# 017), only 1 contig of dengue virus 1 was available, with a length of 217 bp. Figures 6.10 and 6.11 show the positions of the dengue contigs obtained from plasma cDNA samples of patient IDs# 005 and 017 respectively, against the complete genome of dengue virus 1 obtained from NCBI Reference Sequence (NC_001477.1).

Table 6.2: Validation of diagnosis in positive control samples

Sample ID	Diagnosis	Results from hospital diagnostics	Time of sample collection for NGS study	Detection of pathogen in NGS analysis	
				Kraken	CLC
5c	Dengue	PCR+, NS1+, IgM-	Fever day 4	Yes—105,738 reads	Yes
10	Leptospirosis	IgM+	Fever day 9, antibiotics day-3	Yes—1 read	No
17c1	Dengue	NS1+, IgM+	Fever day 8	Yes—64 reads	Yes
17c2	Dengue	NS1+, IgM+	Fever day 8	Yes—4 reads	No
20c	Dengue	NS1+, IgM+	Fever day 6	Yes—25 reads	No
24c	Measles	PCR+, IgM+	Fever day 4	No	No
31c	Dengue	PCR-, NS1+, IgM+	Fever day 8	Yes—6 reads	No
32	*Streptococcus pyogenes*	Blood culture +	Fever day 2, antibiotics day 1	Yes—1 read	No

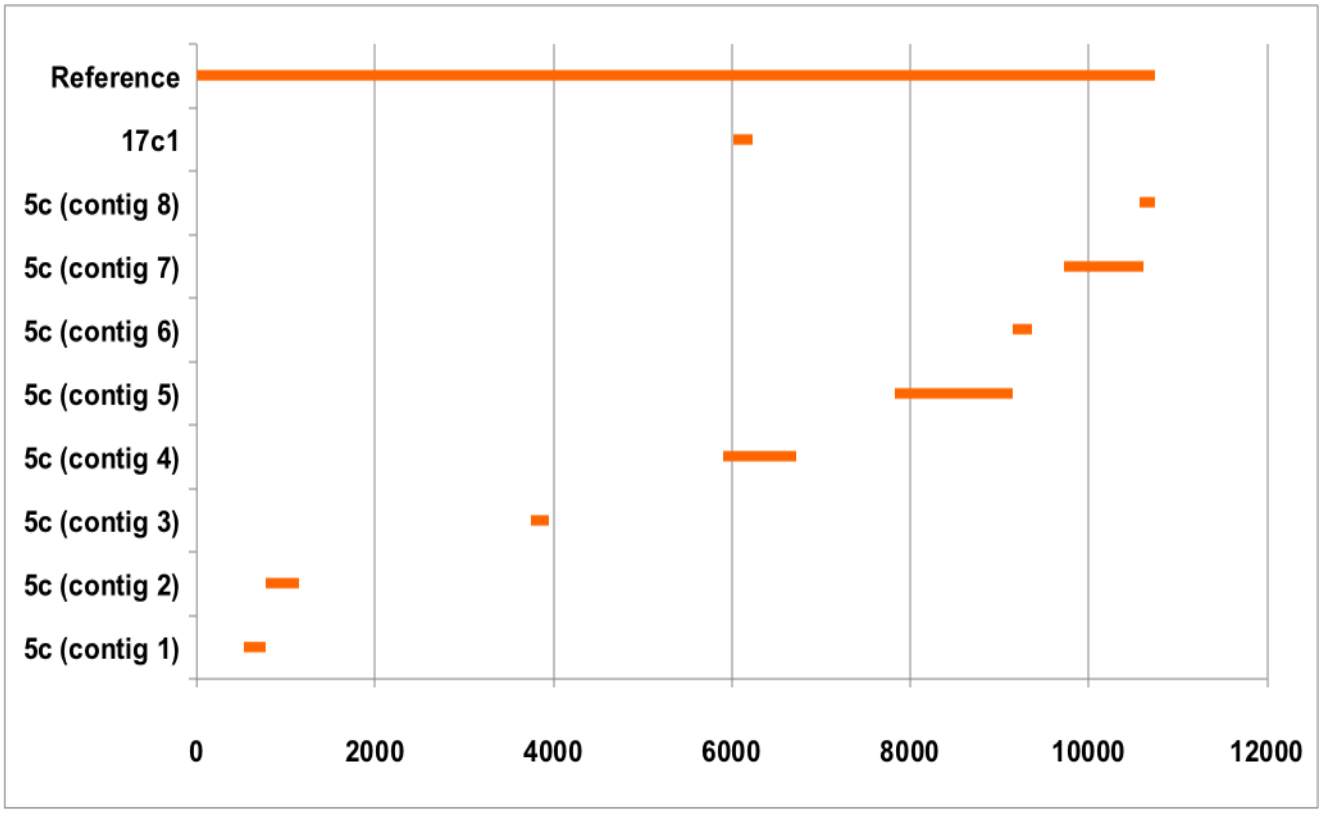


Figure 6.10: Mapping of dengue virus 1 contigs from sample ID# 5c and 17c1 against dengue virus 1 complete genome of 10,735 bp genomic DNA (NCBI Reference Sequence: NC_001477.1)

The X-axis represents the size and position of the genome/contigs in base pair (bp); the Y-axis represents the dengue virus 1 reference genome and the dengue virus 1 contigs found in the sample ID# 5c and 17c1, both are cDNA samples from patient ID# 005 and 017, respectively.

6.3.4 Control subjects

Control subjects consisted of patients with dengue (ID# 005, 017, 020, 031), *Leptospira* (ID# 010), measles (ID# 024) and *S. pyogenes* (sample ID# 032) infection (see Table 6.2).

The first control subject (patient ID# 005) was a 20-year-old man with dengue fever. He presented to Cairns Hospital with a 4-day history of fever, headache, muscle pain and joint pain. He stated that his brother had been recently admitted to Cairns Hospital with dengue fever. The patient was diagnosed with dengue fever on the basis of positive NS1, and PCR was positive for Den-1. His dengue IgM was non-reactive at the time of diagnosis. His condition improved with supportive therapy and he was discharged the following day. Sequencing was performed on plasma cDNA and

generated ~15 million PE reads. Of these, 10,814,715 (72.51%) of reads were of host origin and were removed from the NGS dataset. Kraken analysis classified 2.58% (105,738/4,100,716) of the non-human reads as dengue virus 1. Secondary analysis using the CLC Genomics Workbench validated the findings of the primary analysis by reporting 8 contigs corresponding to dengue virus 1 (see Table 6.2 and Figure 6.10).

The second control subject (patient ID# 010) was an Indigenous 22-year-old man transferred from Cooktown hospital with haemoptysis associated with a 9-day history of fever. The doctor suspected *Leptospira* infection following exposure to and consumption of creek water after becoming lost. There was no consolidation on chest X-ray and no growth of the organism on blood culture. The serum level of CRP was elevated, at 246 mg per litre. Urinalysis showed increased excretion of leukocytes, at 80 x 10^9 cells per litre, and erythrocytes at 20 x 10^9 cells per litre. In Cooktown, the patient had been started on intravenous (IV) ampicillin and gentamicin, and switched to IV benzylpenicillin and doxycycline on the following day. The infectious diseases team at Cairns Hospital recommended that he continue doxycycline 100 mg twice per day for a total of 14 days. The haemoptysis stopped and he exhibited improvement during the 4 days of his admission. Serologic assay showed IgM reactivity to *Leptospira interrogans* serovar Australis. The blood sample for sequencing was collected on the third day of antibiotics administration. Sequencing was performed on plasma DNA and generated ~4 million PE reads. Kraken analysis identified 1 read corresponding to *L. interrogans* serovar Copenhageni, but analysis on the CLC Genomics Workbench did not reveal the presence of a *Leptospira* sequence.

The third control (ID# 017) was a 57-year-old woman with fever, myalgia and rash (see Figure 6.12). She was diagnosed with dengue on the basis of positive IgM and NS1. A blood sample was collected on day 8 of illness, from which cDNA amplicons were prepared in duplicate. Deep sequencing generated ~13.5 million and ~11.5 million PE reads. The majority of reads (87.71% and 87.75%) were of human origin. Kraken detected 64 and 4 reads of dengue virus 1 in the first and second duplicate analyses of the sample, respectively. Dengue virus 1 was also detected on both duplicates analysed on CLC Genomics Workbench.

Figure 6.11: Patient ID# 017 with dengue rash over her trunk and extremities

The fourth control (ID# 020) also had dengue, evident by positive dengue IgM and NS1. Her Flavivirus IgM was positive and serotyping showed Den-1. She was an 18-year-old girl who presented with fever, retro-orbital headache, myalgia, nausea, lethargy, abdominal pain and facial rash (see Figure 6.13). The blood sample was collected on day 6 of fever and sequencing was performed on plasma cDNA to generate ~5 million PE reads. Host reads accounted for 43.67% of total reads. Analysis with Kraken detected 25 reads of dengue virus 1, but this virus could not be identified on CLC Genomics analysis.

Figure 6.12: Patient ID# 020 with dengue rash on her face

The fifth control (ID# 024) was a 23-year-old man with fever and rash after a recent trip to Thailand. He had had a tattoo done in Thailand, but denied the use of intravenous drugs or unprotected sex. He was well in Thailand, but developed the rash upon return to Australia. His rash was maculopapular, beginning on his forehead and spreading to involve the majority of the body (see Figure 6.14). He also had tender cervical lymphadenopathy. At Cairns Hospital, blood culture was performed along with other laboratory investigations, including respiratory viruses PCR, dengue PCR, measles PCR and serology screening for Hepatitis (A, B and C), HIV, syphilis, herpes, CMV, arboviruses (Flavivirus, dengue, Ross River, Barmah Forest, Alphavirus, Sindbis and Chikungunya virus), rubella and measles. His dengue IgM was reactive, dengue IgG was non-reactive, dengue NS1 antigen was non-reactive and dengue PCR was negative. He was diagnosed with measles based on positive PCR and IgM (his measles IgG was non-reactive). The blood sample for sequencing was collected on day 4 of fever. Sequencing was performed on plasma cDNA and generated ~7 million PE reads. Measles virus could not be identified, either by Kraken or CLC Genomics analysis. However, Kraken analysis detected 11 reads of dengue virus 1.

Figure 6.13: Patient ID# 024 with measles rash on his face, torso and extremities

The sixth control (ID# 031) was a 27-year-old female with an 8-day history of feeling unwell, myalgia, fevers, sore throat, light sensitivity, nausea and vomiting. She stated that she had a mosquito bite prior to illness. She saw her general practitioner (GP) after 3 days of sickness and was placed on Augmentin Duo Forte (875 mg of amoxycillin and 125 mg of clavulanic acid) twice a day. Three days later, the pain became worse and she was not improving. Her GP ordered blood tests, including

dengue serology. Later on, she was informed by Public Health that she had dengue fever. She presented back to her GP and, due to ongoing sore throat, reduced appetite and general malaise, she was referred to Cairns Hospital for further management. At the hospital, it was noted that throat ulcers were present. Her Flavivirus IgM was positive with unspecified dengue serotype. Dengue IgM and NS1 were positive. However, the dengue universal PCR was negative. Sequencing was performed on plasma cDNA and generated ~8 million PE reads. Kraken analysis detected 6 reads of dengue virus 1, but this virus was not detected by secondary analysis on the CLC Genomics Workbench.

The final control (ID# 032) was a 56-year-old man with viraemic symptoms including 2 days fever, chills, headache, neck pain, muscle pain, joint pain, back pain, weakness, fatigue, shortness of breath and abdominal pain. Blood cultures taken from a private laboratory showed 3/3 positive blood cultures for *Streptococcus pyogenes*. The blood sample for sequencing was collected after the administration of antibiotics. Sequencing was performed on plasma DNA and generated ~7 million PE reads. Kraken analysis detected 1 read of *S. pyogenes*, but this bacterium was not detected on CLC Genomics analysis.

6.3.5 Undiagnosed subjects

In most cases, simultaneous sequencing of the DNA and cDNA samples was unable to be performed and this may affect the interpretation of the results. The presence of bacteria might not be detected if deep sequencing was performed on cDNA sample because cDNA (from RNA) contains only the coding portion of the genome (i.e., the genes) whereas the bacterial genome contains the coding and non-coding portions. In this case, the bacterial non-coding portions could not be detected in cDNA samples which reduce the chance of bacteria identification. Compared to bacteria, viral DNA is gene dense, with fewer non-coding regions, allowing for high read identification using cDNA (RNA) samples. A strategy was developed to determine the most likely cause(s) of patients' illness. This strategy involved screening the results of the primary and secondary analyses for pathogens that were known to cause prominent symptoms in infected patients. When the sequencing was performed on duplicate samples (from DNA or cDNA samples) originating from the same subject, only the results from the sample that produced the largest sequencing data would be reported. If bioinformatics results from both DNA and cDNA are available, the results from cDNA sample would be screen first to allow the detection of viral RNA, viral DNA and

bacteria. Following this, the results from DNA sample would be screened for further detection of bacteria. Finally, the probable causative agents were listed according to the number of reads (from largest to smallest) obtained from the primary analysis (Kraken). E-value would be reported if the pathogen was detected in the secondary analysis (CLC). The clinical data and the outcomes from conventional investigation as well as from NGS testing in patients with undiagnosed fevers are summarized in Table 6.3.

Table 6.3: Plaussible NGS diagnoses in patients with undiagnosed fevers

Patient (sample) ID	Clinical data	Pathology and radiology findings from conventional investigation at Cairns Hospital	Plaussible NGS diagnosis
002 (2c1, 2c2)[a]	A 59-year-old Indigenous man with diarrhoea and fevers. On examination, his respiratory rate was 26x/minute, temperature was 38.8 ^{0}C, and there was dual heart sound with systolic murmur. Other active problems: ischaemic heart disease, pulmonary hypertension, atrial flutter, type 2 diabetes mellitus, hypertension and dyslipidemia. During admission, the patient commenced gentle rehydration, ceftriaxone and doxycycline. Ongoing diarrhoea, dehydration, and low blood pressure led to patient transfer to intensive care on day 4. Ascitic tap was performed at ICU and haemofiltration was commenced. The clinical diagnosis was viral gastroenteritis and acute renal failure, possibly due to dehydration. He was given poor prognosis and transferred back to the ward for comfort measures. He died on day 18 due to multi-organ failure.	• Serum urea: 51.7 mmol/l (↑), creatinine: 1000 µmol/l (↑), ALT: 414 U/l (↑), AST: 1290 U/l (↑), CRP: 102 mg/l (↑) • Platelet: 60 x 10^9/l (↓), WBC: 23.2 x 10^9/l (↑) with neutrophil predominant (20.69 x 10^9/l) • Antinuclear antibody (ANA): (+) • Blood culture: no growth after 5 days incubation in aerobic and anaerobic condition • Dengue and hepatitis C serology: (-) • Faecal analysis: normal faecal flora were absent on culture, *Clostridium difficile* screen (-), no evidence of parasitic infection on microscopic examination, and , *Shigella*, *Yersinia* or *Campylobacter* grown on culture. • Chest X-ray: cardiomegaly, no consolidation in lungs; chest CT: small pleural effusion; abdomen CT: ascites, fatty liver, hepatomegaly and splenomegaly • Ascites analysis (chemistry and microbiology): normal	Results of the NGS analysis were screened for agents causing diarrhoea including bacteria (*V. cholerae*, *C. difficile*, *Shigella* sp., *Salmonella* sp., *E. coli, Yersinia enterocolica*, *Campylobacter jejuni*), viruses (Rotavirus, Norovirus, Astrovirus), parasites (*Giardia lamblia* and *Entamoeba histolytica*).[190, 191] The most likely cause of illness was *E. coli* (57,657 reads, E-value 4.3E-49). In addition to the finding of agents causing diarrhoea, NGS also detected *Acinetobacter baumannii* (556 reads, E-value 1.15E-22), an opportunistic pathogen that often causes nosocomial infection.
011 (11)[b]	A 40-year-old man with a 3-day history of rash, fever, myalgia and nausea, following recent travel to the Torres Strait Islands and possible contact with mites. Erythematous blanching papules	• Serum ALT: 177 U/l (↑), AST: 164 U/l (↑), CRP: 97 mg/l (↑). • Platelet: 108 x 10^9/l (↓), WBC: 3.4 x 10^9/l (↓). • Urinalysis: no abnormality detected. • Blood culture: (-)	Results of the NGS analysis were screened for *Rickettsia*, a bacterial agent transmitted to human bloodstream via an infected tick bite. Analysis with Kraken detected 11

	rash were found on trunk, chest and extremities (see Figure 6.14). There was an eschar on his upper left arm. He showed marked clinical improvement with doxycycline.	• Malaria screening: (-) • Flavivirus, Q fever and *Leptospira* serology: (-) • Spotted fever group (SFG) *Rickettsia* and scrub typhus PCR: (-)	reads corresponding to *O. tsutsugamushi*, family of Rickettsiaceae, the causative organism of scrub typhus. Even though the CLC analysis did not detect *O. tsutsugamushi* sequence, the clinical data, laboratory findings, deep sequencing results and positive response to doxycycline supported the diagnosis of scrub typhus.
012 (12)[b]	A 31-year-old female from the Atherton Tablelands presenting with abdominal pain, nausea, vomiting and constipation on a background of inflammatory bowel syndrome. The provisional clinical diagnosis was acute Hepatitis, likely caused by leptospirosis. The patient was given analgesia and was advised to return to the emergency department if the pain returned.	• Serum ALT: 545 U/l (↑), AST: 428 U/l (↑) • ANA: (+) • Blood culture: (-) • Hepatitis B and C serology: (-) • Serology and PCR for CMV, *Leptospira* and Q fever: (-) • Chest X-ray: no abnormality detected • Abdomen CT: splenomegaly	The results of NGS analysis did not support the reported clinical diagnosis, as neither primary nor secondary analysis detected *Leptospira* sequences. In fact, there was insufficient evidence to establish a causal relationship between agents detected on NGS and the illness. There was a possibility that the illness was caused by a non-infectious condition, such as inflammatory disease or autoimmune disease, as evidenced by the positive ANA test.
014 (14, 14c)[a,b]	A 55-year-old man from Babinda with fever, myalgia and blanching maculopapular rash (see Figure 6.15) with the history of mosquito bites 2 weeks prior to presentation. The patient's wife (ID# 019) concomitantly presented with a more severe variant of the same illness. The patient was	• Serum ALT: 63 U/l (↑), AST: 56 U/l (↑) • Platelet: 122 x 10^9/l (↓). • Blood culture: (-) • Serology for EBV, *R. rickettsia, O. tsutsugamushi, Mycoplasma pneumoniae, Leptospira*, Flavivirus: (-) • PCR for *Rickettsia* (SFG and thypus group),	Results of the NGS analysis were screened for bacteria and viruses causing maculopapular rash and fever, such as *Rickettsia*, *Leptospira*, *Mycoplasma* sp., *S. pneumoniae*, *S. aureus*, *N. gonorrhoeae, N. meningitidis*, measles, rubella and EBV (human herpesvirus 4), human

	diagnosed with probable scrub typhus. He clinically improved with ceftriaxone and doxycycline.	*Streptococcus pneumoniae, Neisseria meningitidis* and *Leptospira*: (-) • *S. pneumoniae* and *Legionella* antigens in urine: (-)	herspesvirus 6, human parvovirus B19, enteroviruses and dengue fever.[192, 193] The most likely aetiology was dengue virus (25 reads).
019 (19, 19c1, 19c2)[a,b]	A 57-year-old female transferred from Babinda hospital and was the spouse of patient ID# 014. She had an 8 day history of fever, rash, retro-orbital headache, myalgia, malaise, vomiting and diarrhoea. Patient was hypotensive at initial presentation to Babinda hospital and was given ceftriaxone, vancomycin and doxycycline intravenously prior to transfer. Patient had evidence of multi-organ failure at transfer to Cairns Hospital and was intubated. She died with a diagnosis of septic shock.	• Serum urea: 43.7 mmol/l (↑), creatinine: 839 µmol/l (↑), ALT: 1170 U/l (↑), AST: 6160 U/l (↑) • Platelet: 14 x 10^9/l (↓), WBC: 35.1 x 10^9/l (↑) with neutrophil predominance (30.57 x 10^9/l). • Blood culture: (-) • Serology for *Leptospira*, *S. pneumoniae,* EBV, CMV, Flavivirus, Q fever, *R. typhi, R.ricketsii, O. tsutsugamushi,* SFG and typhus group *Rickettsia*: (-) • PCR for *Leptospira*, Dengue, scrub typhus, *S.pneumoniae* and *N. meningitidis*: (-) • Malaria screen: (-)	Results of the NGS analysis were screened as in patient ID# 014, and the following agents were detected and suspected as the cause of fever/death: 1. EBV (7,023 reads) 2. Dengue virus (19 reads) 3. *R. conorii* (1 read) 4. *A. baumannii* (2,514 reads, E-value 2.0E-68) 5. *Listeria monocytogenes* (13 reads). *A. baumannii* and *L. monocytogenes* are known to cause fatality in immunocompromised patients.
027 (27)[b]	A 29-year-old man presented with fever, generalised weakness, nausea, vomiting, diarrhoea, dry cough, and headaches after returning from Thailand, where he had spent some time in the jungle. The provisional diagnosis was likely arbovirus infection. He was given doxycycline to cover leptospirosis and clinically improved.	• Serum urea: 11.1 mmol/l (↑), creatinine: 181 µmol/l (↑), ALT: 94 U/l (↑), AST: 68 U/l (↑), CRP: 487 mg/l (↑) • Platelet: 90 x 10^9/l (↓), WBC: 15.3 x 10^9/l (↑) • Blood culture: (-) • Faeces culture: no *Salmonella, Shigella, Yersinia, Campylobacter* • Malaria screening: (-) • *Leptospira* and Dengue serology: (-) • Respiratory viruses PCR: (-)	Results of the NGS analysis were screened for arboviruses, *Rickettsia*, and agents causing gastroenteritis. The most likely aetiologies were dengue virus (11 reads) and *Salmonella enterica* (3 reads).

028 (28)[b]	A 64-year-old man with 4-day fever, chills, sore eyes and ulcerated mouth (see Figure 6.16). He had a tick bite recently and had been on a cruise from Vanuatu through the Solomon Islands to Papua New Guinea. He was taking several medications for high blood pressure and type 2 diabetes mellitus.	• Serum ALT: 64 U/l (↑), AST: 43 U/l (↑) • WBC: 0.8 x 10^9/l (↓), neutrophils 0.02 x 10^9/l (↓), platelet: 147 x 10^9/l (↓) • Screening for malaria, Hepatitis B, Hepatitis C and HIV: (-) • Serology for CMV, EBV, *R. rickettsii*, *Leptospira*, Flavivirus, Q fever, *Brucella, Cryptococcus* and *M. pneumoniae*: (-) • Serology for *O. tsutsugamushi*: weakly positive (total immunoglobulin: 128) • PCR for HSV 1 and 2: (-) • Chest X-ray: clear • Blood and urine cultures: (-)	The patient's medical note was reviewed and it was found that the patient had prolonged neutropenia (up to 4 weeks). It was likely that the patient had drug-induced febrile neutropenia and the results of NGS analysis were not considered.
029 (29c)[a]	A 38-year-old female admitted with possible measles. Initially, patient had back pain and fevers without urinary symptoms. She was started on antibiotics for suspected urinary tract infection. Subsequently, she developed a rash, which began on her face and spread to her torso and limbs (see Figure 6.17). She worked as a housekeeper and had no exposure to potential allergens or new chemicals. Patient was admitted with strict respiratory isolation. On day 1 of admission, the doctor suspected a small bite on the patient's back. This lesion was reviewed by infectious diseases team and the impression was rickettsial in nature. Fevers were improving with	• Serum ALT: 448 U/l (↑), AST: 519 U/l (↑) • Platelet: 99 x 10^9/l (↓) • Serology for CMV, EBV, Dengue, Ross River, Alphavirus, Barmah Forest, Sindbis, Chikungunya, Q fever, *Leptospira, Brucella, R. rickettsii, O. tsutsugamushi,* Rubella and Measles: (-) • Measles PCR: (-)	Kraken analysis detected Jingmen tick virus (52 reads) and *R. africae* (10 reads) but these organisms were not detected on the secondary analysis with the CLC server. The detection of theses organisms in Kraken analysis was likely insignificant, possibly due to spurious alignment and cross contamination (see discussion about these organisms in page 137).

	doxycyline.		
030 (30)[b]	An 18-year-old man with fever, chills, headache, muscle pain, joint pain, back pain, cough, sore throat, nausea and rash. He had history of a tick bite on the upper right thigh when camping at Tinaroo Dam, Atherton Tablelands. He had been taking doxycycline for approximately 2 days as prescribed by a 24-hour medical center. On examination, the doctor found 1 cm firm mobile tender inguinal lymph node on the right inguinal region and 1 cm eschar over upper right thigh. Patient continued on doxycycline for two weeks.	• CRP: 37 mg/l (↑) • Patelet: 125 x 10^9/l (↓) • *R. rickettsii* and *O. tsutsagamushi* serology: (-)	Kraken analysis detected Jingmen tick virus (32 reads) and *R. africae* (1 read) but these agents were not detected on the secondary analysis with the CLC Genomic Workbench. The detection of theses organisms in Kraken analysis was likely insignificant, possibly due to spurious alignment and cross contamination (see discussion about these organisms in page 137).
039 (39)[b]	A 57-year-old man with gradual onset of malaise, vomiting, myalgia, fevers and white productive cough. The provisional diagnosis was viral infection with possible acute renal failure secondary to viral illness. He was treated with doxycycline and Augmentin.	• Serum creatinine: 157 U/l (↑), AST: 40 U/l (↑) • Chest X-ray was normal • *Leptospira* and Flavivirus serology: (-)	The most likely aetiology was dengue virus (27 reads).

[a] Sequencing was performed on cDNA samples

[b] Sequencing was performed on DNA samples

Figure 6.14: Patient ID# 011 with eschar (upper left) and rash on his extremities (bottom left) and trunk (right)

Figure 6.15: Patient ID# 014 with rash on his body and extremities

Figure 6.16: Patient ID# 028 with mouth ulcer (left) and tick bite (right)

Figure 6.17: Patient ID# 029 with rash on her face, torso and extremities

Photo on the upper right shows attempt to identify Koplik spots, a pathognomonic sign of measles.

6.4 Discussion

6.4.1 Participants and samples

The success of pathogen detection using a deep sequencing approach highly depends on the burden of the pathogen in a particular sample. Obtaining optimal samples for deep sequencing is challenging when conducting a study at a tertiary hospital. Most patients had already presented to their GP and had been administered medications that may have interfered with pathogen load. GPs usually expect that AUFs resolve spontaneously, and only severe cases are referred to a tertiary hospital. When patients presented late to hospital (day 5 or beyond), or after taking antibiotics, the level of pathogen in blood may have been inadequate to generate sufficient reads for attaining diagnosis. Thus, deep sequencing is not a recommended approach in this case.

Ideally, samples should be snap frozen in liquid nitrogen to preserve CNA and to prevent further degradation of RNA in particular. However, if a deep sequencing approach is to be used for routine diagnosis of AUF, the use of liquid nitrogen is impractical, costly and hazardous. An alternative way to preserve RNA is by using RNAlater® solutionl however, this solution is not really suitable for cell-free liquid samples (plasma or serum). If RNAlater® were used in liquid sample, the ratio would be 15:1 of RNAlater® to liquid sample (as advised by Life Technologies, an RNAlater® manufacturer). Trizol® LS reagent can be added to the sample before or after the sample is stored in the freezer in order to maintain the integrity of RNA by inhibiting RNase activity during sample homogenisation. A study conducted by Cerkovnik et al[194] demonstrated no significant difference in the quality of RNA isolated from plasma with the addition of TRIzol® LS reagent before or after the samples were frozen.

6.4.2 Sample preparation

Previous studies have used deep sequencing approaches to facilitate the discovery of novel viruses in plasma/serum samples.[15, 127–130] However, the use of plasma/serum samples in the present study posed a challenge to the downstream sequencing analysis, because the researchers did not have in-house access to a high-throughput sequencing platform, necessitating the employment of a sequencing company. DNA and RNA in plasma/serum samples were present in low quantity and low quality, which restrained sequencing companies from using the samples as NGS

input. The researchers contacted several research centres and sequencing companies in Australia, including the AGRF, Ramaciotti, Micromon, Australian National University and Garvan Institute of Medical Research, and overseas, including New Zealand Genomics Limited, Macrogen (Korea), Beijing Genomics Institute (BGI), Yourgene BioScience (Taiwan), LC Sciences and Whitehead Institute (USA) and Exiqon (Denmark) to inquire about sequencing costs and NGS input requirements. All companies requested high quantity (nanograms to micrograms), high quality (RIN $\geq$ 8) and high purity (A 260/280 $\geq$ 1.8) DNA/RNA samples, conditions that are impossible to attain from plasma/serum samples. The eventual solution was to send the samples to AGRF, the sequencing company that required the smallest amount of NGS input.

The quantities of DNA and RNA that can be isolated from human plasma/serum are very low, and frequently represent the limiting factor for metagenomic research. Previous studies were reviewed to select the best method for isolating DNA and RNA from cell-free samples. There were limited choices regarding commercial kits for the isolation of circulating DNA. Currently available kits on the market are mostly developed with a focus on medium- and high-molecular weight DNA, and are thus ineffective for the isolation of short fragments of DNA from body fluids.

A comparison of four different kits by van der Vaart and Poterious[195] showed that the QIAamp® DNA Blood Mini Kit and the MagNA Pure Compact System (Roche) produced lower yields than isolating DNA by either salting-out or extracting with phenol-chloroform. Board et al[154] showed that the QIAamp® Viral Spin Kit (Qiagen) was more effective than the QIAamp® DNA Mini Blood Kit (Qiagen), Agencourt® Genfind Blood and Serum Genomic DNA Isolation Kit (Agencourt Bioscience Corporation) and ChargeSwitch® gDNA 1 mL Serum Kit (Invitrogen). A study by Fleischhacker et al[196] showed that the MagNA Pure isolation system (Roche) produced higher DNA yields than the NucleoSpin® Kit (Macherey-Nagel) or QIAamp® Blood Midi kit (Qiagen). NucleoSpin® Plasma XS kit (Macherey-Nagel) was shown by Kirsch et al[197] to produce all fragment sizes as opposed to the QIAamp® DNA Blood Mini Kit (Qiagen), which failed to produce $\leq$ 50-bp fragments and also yielded lower amounts of the other sized fragments. This is because the NucleoSpin® Plasma XS kit was developed for the isolation of low-molecular weight DNA, with a special emphasis on the high recovery of DNA fragments < 200 bp. For the purpose of this study, however, small-sized fragments are not informative for pathogen identification. The DNA or RNA fragments must be long enough to unambiguously identify the presence

of pathogen. In the present study, I compared DNA yield isolated from two plasma samples using the NucleoSpin® Plasma XS and the QIAamp® DNA Mini kit (data not shown). The QIAamp® DNA Mini kit produced a higher level of DNA, so it was decided that the DNA isolation would be performed using the QIAamp® DNA Mini kit for the rest of the samples.

Commercially available kits often contain poly(A) carrier RNA to facilitate RNA isolation. In this study, carrier RNA was not desired, as this would swamp the sequencing data. To optimise the isolation of circulating RNA, a number of commercial kits and reagents were tested, including Quick-RNA MiniPrep and Direct-zol RNA MiniPrep (Zymo Research), RNeasy® Mini Kit (Qiagen), TRI Reagent® (Sigma-Aldrich), TRIzol® and TRIzol® LS reagents (Life Technologies) on a subset of sample. It was found that TRIzol® LS reagent was the most efficient reagent for isolating plasma and serum RNA (data not shown). The reagent is a monophasic solution of phenol, guanidine isothiocyanate and other proprietary components that facilitate the isolation of a variety of RNA species of large or small molecular size. More importantly, this reagent is free from carrier RNA. Previous research comparing various RNA extraction methods has also shown that plasma RNA can be isolated most efficiently by guanidium-phenol extraction followed by precipitation.[157]

Several techniques can be employed in the quantitative analysis of DNA and RNA. The most common technique is to determine absorbance at 260 nm (A_{260}) with a UV spectrophotometer. This detection method has long been a standard in DNA and RNA quantitation, which is largely due to its ease of sample preparation, requiring no additional mixing of reagents and resulting in good reproducibility. UV analysis is very stable because no injection or separation takes place. One major disadvantage to using UV analysis during RNA quantitation is the impact of sample contaminants such as genomic DNA or phenol, which also absorb at 260 nm, thereby giving false quantitation readings. UV spectrophotometer analysis (e.g., by using NanoDrop) cannot discriminate between RNA and genomic DNA contaminants, so concentration measurements may be affected. Additionally, contaminants such as phenol can yield irreproducible data. In this study, the main factor that makes UV spectrophotometer analysis prohibitive during the first QC assessment is the low concentration of DNA and RNA in plasma and serum samples. The concentration of this DNA and RNA is often below the level of detection of a UV spectrophotometer demanding a more

sensitive method to assess the quantity and quality of CNA. Some CNA quantitation methods that have been employed in previous studies include quantitative real-time PCR (qRT-PCR),[169] fluorometer[150, 158, 163] and bioanalyser.[157]

The Agilent 2100 Bioanalyzer is a microfluidics-based platform for sizing, quantification and quality control of DNA and RNA samples. It uses a laser for the excitation of intercalating fluorescent dyes, thereby achieving a high level of sensitivity. Data is presented in an easy-to-read format consisting of electropherograms, a gel-like image and tabular results. There are several important advantages of the 2100 Bioanalyzer over UV. First, the 2100 Bioanalyzer assesses the quality and the quantity of DNA/RNA in one step, combining two traditional techniques (slab gel electrophoresis and UV measurement) all on the same platform. This saves a considerable amount of time and resources. Second, only a very small amount of sample is consumed, saving precious DNA/RNA samples. Third, DNA/RNA quantitation using the 2100 Bioanalyzer is far more independent from sample contaminants than UV measurements, which are strongly influenced by low- and high-molecular weight genomic DNA and phenol. Low-molecular weight DNA contamination is shown as a distinctive baseline 'hump' that can be observed in the electropherograms of bioanalyser. Contamination with high-molecular weight DNA results in clogging channels so that the DNA is no longer injected into the separation channel. Therefore, it goes undetected and does not have influence on the RNA concentration measurement. As with phenol, the reagent does not interact with the fluorescent dye used in the bioanalyser, so the DNA/RNA concentrations remain stable in the presence of phenol contamination.

In the second QC assessment, UV spectrophotometry was used to determine the quantity and purity of the amplification products (amplicons). At this stage, the concentrations of DNA and cDNA were adequate for UV spectrophotometry analysis. The size of DNA and cDNA amplicons in a gel electrophoresis were also checked, as requested by AGRF. The data from UV spectrophotometry and the gel image were sent to AGRF together with the samples.

While the essence of metagenomics research is to directly sequence the DNA/RNA in the sample, the task is challenging; in the present study, it was necessary to provide at least 100 ng of DNA/RNA per sample. The attempt to meet the minimum input requirement for sequencing necessitated amplification, which involved five additional steps (see Figure 6.2). Every step carried out during sample preparation is a

potential source of contamination, and might cause further degradation of the nucleic acids. Amplification itself could introduce biases during sequencing.

NGS is so sensitive that it detects organisms that, though present in the sample, are not actually responsible for infection. These contaminants could be minimised if sample processing is minimal and is performed in a committed laboratory. The problem is that hospitals, and other clinical institutions, often lack facilities for preparing samples for sequencing. In this study, plasma and serum separation were performed in the hospital, but the other protocols were mostly performed at a university laboratory. Finally, the library preparation and sequencing had to be outsourced. Processing of samples at three different sites increased chances of contamination and sample degradation.

A high number of contaminating agents could overshadow the true cause of fever. Moreover, the use of plasma and serum as the starting material may have inadvertently excluded intracelluler microbes causing the infection as their DNA/RNA would have been mainly within the celluler fraction of the blood and thus discarded. Thus, in a fever case where the NGS diagnosis is ambiguous, an expert clinician or an infectious disease specialist is needed to interpret NGS findings.

6.4.3 Bioinformatics analysis

With the increasing speed and quantity of data generated from sequencing platforms, bioinformatics analysis has been found to be the most complex and time consuming element of a deep sequencing approach. In fact, the time taken for data analysis was far longer than that taken for data generation.

In this study, the data generated from NGS were analysed in two stages. The primary analysis (perfromed using Kraken program) involved the alignment of the sequencing reads to a database of known viruses, bacteria and archaea. Pathogen identification was inferred from the resulting frequency of aligned reads. Secondary analysis in the CLC Genomics Workbench server included *de novo* assembly of the reads and subsequent BLAST analysis of resulting contigs to identify viruses and bacteria causing fever. The secondary analysis was performed to analyse reads that were unclassified by Kraken, as well as to validate the findings of the Kraken analysis on positive control samples. For users with no medical background, the screening of pathogens associated with fever can be performed quickly by transferring the outputs of Kraken and CLC Genomics Workbench to an Excel spreadsheet, and then using various

tools in Excel (such as Find, Sort and Filter) in conjunction with the lists of medically important bacteria (http://www.tostepharmd.net/pharm/clinical/bacteria.html) and viruses infecting humans (http://viralzone.expasy.org/all_by_species/656.html).

In clinical settings, diagnostic tools that can provide rapid results are favoured because of the need to administer the appropriate treatment promptly to improve patient outcomes, that is, to maximise the chances of patient recovery and minimise the occurrence of clinical complications. Analysis of sequencing data using Kraken on a cloud computing service has several advantages. First, analysing ~2 Gb of raw sequence data could be completed in 1 hour. This speed far exceeds the duration of a conventional BLAST search, which may take more than 24 hours to complete. Instead of assembling short reads into longer contigs, Kraken directly classifies the reads into taxonomic labels, thus shortening the analysis time. Second, Kraken uses an exact alignment strategy for assigning taxonomic labels to DNA sequences. Hence, Kraken's precision is higher than that of BLAST, which tolerates slight mismatches in order to achieve high sensitivity. Therefore, in this study, a BLAST search was performed after the assembly of short sequences into contigs in order to achieve high precision. Third, the feasibility of running bioinformatics analysis in a simple pipeline permits the use of Kraken by novices. Traditionally, analysis for pathogen detection requires immense bioinformatics support and a customised pipeline. When using Kraken and many other tools, BaseSpace® provides step-by-step guidance and also technical support for users, all free of charge.

Despite the usefulness of Kraken, it was not possible to recover measles virus in the plasma sample of patient ID# 024 with ~7 million PE reads generated from NGS. This is probably because the measles virus presents for only a short period in the circulation, resulting in a low level of viremia. The virus is a negative, single-strand RNA virus that has a unique mechanism of replication: its genome has to be transcribed as soon as the virus enters the host in order to carry out viral replication. Thus, the ratio of this virus will be very low compared to other microbes that circulate for longer than a negative strand RNA virus. A deeper level of sequencing with high coverage is necessary to detect this virus in a sample with a high titre of measles virus. Grard et al[129] performed an ultra-deep sequencing, generating ~140 million reads, to identify a negative-sense single-strand RNA virus (Bas-Congo virus) at a concentration of 1.09×10^6 RNA copies/ml. Although patient ID# 024 had positive dengue IgM, it was

unlikely that he had dengue infection because Kraken analysis only detected 11 reads of dengue virus 1 in the sample collected on day 4 of fever (see Section 6.3.4 'Control subjects: patient ID# 024'), a time when dengue primary infection typically reaches its peak of viraemia.[198–200]

The sensitivity of deep sequencing highly depends on the progress of the illness and time of sample collection, which may exceed the sensitivity of PCR. For example, in positive control subject ID# 031, deep sequencing permitted the detection of dengue virus in a PCR-negative sample when the patient presented on day 8 of illness. Similarly, in undiagnosed subject ID# 011 with clinical features suggestive for scrub typhus infection, Kraken analysis on the sample colected on day 3 of illness detected 11 reads of *O. tsutsugamushi*, a rickettsial agent causing scrub typhus, which escaped PCR detection.

When a patient presents at a late stage of leptospirosis and after taking antibiotics, as in the case of patient ID# 10, deep sequencing could detect only 1 read of *Leptospira interrogans* serovar Copenhageni in the sample taken on day 9 of fever. In contrast, the serologic assay at Cairns Hospital identified IgM reactivity to *Leptospira interrogans* serovar Australis. A possible explanation for this difference is that there are 18 different species of *Leptospira* and more than 200 serovars,[201] so cross-reactivity could occur in a serologic assay. The bacteria are only present in the blood for a few days after the onset of symptoms,[202] thus a low number of bacterial reads, with supportive clinical and laboratory findings, is certainly indicative of an attained diagnosis when sequencing is performed on cell-free samples.

It is unreasonable, however, to determine infectious cause of fever based on deep sequencing results solely because Kraken has reported an excessive amount (hundreds) of viruses and bacteria in a single sample (see Figure 6.8), yielding a non-specific answer. In addition to sequences from known pathogens, sequences derived from commensal bacteria and/or laboratory contamination were common in all NGS data sets. There is a possibility that ambiguous reads might map to multiple taxa in the MiniKraken database, resulting in the detection of organisms that were not actually present in the sample (false positive).

Performing more complex analysis does not guarantee reliability of results, and may indeed result in failure to detect the true cause of fever (false negative). The secondary analysis, which involved pre-processing and assembly prior to the BLAST search, failed to detect the true pathogen in 75% (6/8) of positive control samples (see

Table 6.2). Thus, it can be argued that the success of the deep sequencing approach is more likely to be determined by the sample condition (course of illness, antibiotics treatment) and the dataset (sequencing depth) rather than the analysis pipeline. It is important to collect the samples during the highest level of viraemia or bacteraemia to achieve optimum sensitivity of detection. In the sample taken from ID# 005, dengue virus 1 was constantly detected on the top of the list of pathogens reported by Kraken and the CLC Genomics Workbench server. Sequences of dengue virus 1 were clearly identified when the sample was collected during the acute stage of dengue viraemia (day 4 of fever) and when sufficient sequence data (14,915,431 PE reads, equal to 2.2 Gb) were available to permit detection. As a result, Kraken was able to classify 2.58% (105,738/4,100,716) of non-host reads into the dengue virus 1 taxon.

The failure of secondary analysis to detect the cause of fever in the majority (75%, 6/8) of positive control samples was likely due to the limited number of pathogen reads that can be assembled into contigs. This server, however, was able to construct contigs from relatively abundant reads in the samples with a high level of dengue viraemia (as shown in the sample ID# 5c taken from patient ID# 005; see Figure 6.10). Moreover, the Workbench permitted the detection of other organisms, such as those from the kingdoms Animalia, Plantae and Fungi, which were not classified by Kraken because the MiniKraken database only consists of viruses, bacteria and archaeal reference genome.

Last but not least, interpretation of the metagenomics results was challenging. It was difficult to determine causal relationships in the context of AUF or to justify the finding of microbes in normally 'sterile' samples such as blood, plasma or serum. Bioinformatic analysis reported numerous viruses and bacteria in each sample, and it was not possible to distinguish pathogenic microbes from commensal organisms or contaminants based on the number of reads, as shown in control samples where the true pathogens were often less abundant than contaminants. Reliable clinical information and the results of other investigations (pathology, radiology) are thus critical to formulating a diagnosis or to propose a differential diagnosis.

Previous studies using the deep sequencing approach for the diagnosis of fever have experienced similar challenges. A study conducted by Yozwiak et al in 2012[132] detected that many samples contained sequences resembling viruses with no well-established link to human disease, including GB virus C (GBV-C), African swine fever virus (ASFV), Torque teno virus (TTV) and circovirus. Naccache et al[203] employed a

different algorithm of analysis and found various organisms in human blood samples spiked with a particular virus. For example, a spiked HIV NGS dataset could be 'contaminated' by various bacteriophages, TTV, SEN virus, GBV-C, human herpesvirus 1, Trichoderma hypovirus, Megavirus, Acanthamoeba castellanii mamavirus, Simian-Human immunodeficiency virus, and murine leukemia-related retrovirus. It was not clear whether these 'contaminant' viruses were truly present in the sample or had been 'falsely' identified during the analysis.

Despite the identification of the pathogens and contaminants, some reads remained unclassified even after the implementation of further analysis in CLC Genomics Workbench. These unclassified reads might have originated from the remaining Illumina adapters and some primers used in the amplification step that could not be completely removed. Alternatively, these unclassified reads might indicate that there are microbes for which sequences are not homologous to anything known in the NCBI database. In 2012, Mokili et al[80] reported that in previous metagenomics studies, unknown sequences varied between 60% and 99% depending on sample type, read length, homology search method, similarity threshold and database used. The two-tiered approach used in this study to analyse NGS data demonstrates a methodological analysis for identifying viruses and bacteria associated with fever as well as for illustrating microbial diversities present in the sample. Further, this approach was able to classify most of the sequences at the species level, leaving less than 30% contigs unknown (see Figure 6.9).

6.4.4 Microbial diversity in human blood

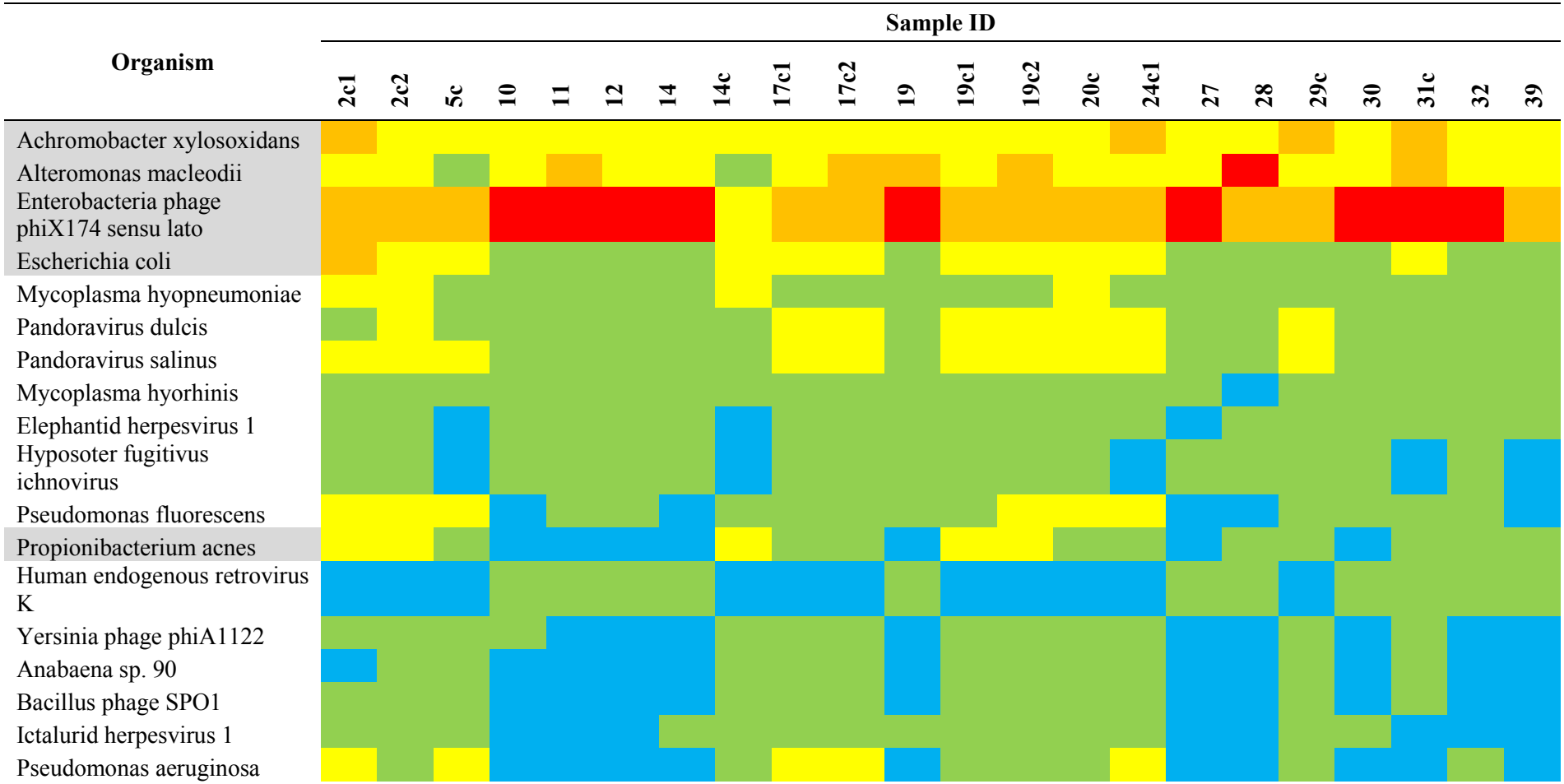
Although the primary aim of this study was to evaluate the use of NGS in detecting pathogens associated with AUF, the present study's data provide an insight into the diversity of the microbiome in human blood. It has been contended that only a few (1%) organisms on earth can be cultured,[204, 205] and it has generally been believed that under 'normal' conditions, human blood is 'sterile'. Culture-independent methods are clearly required in order to extend our knowledge of microbial diversity.

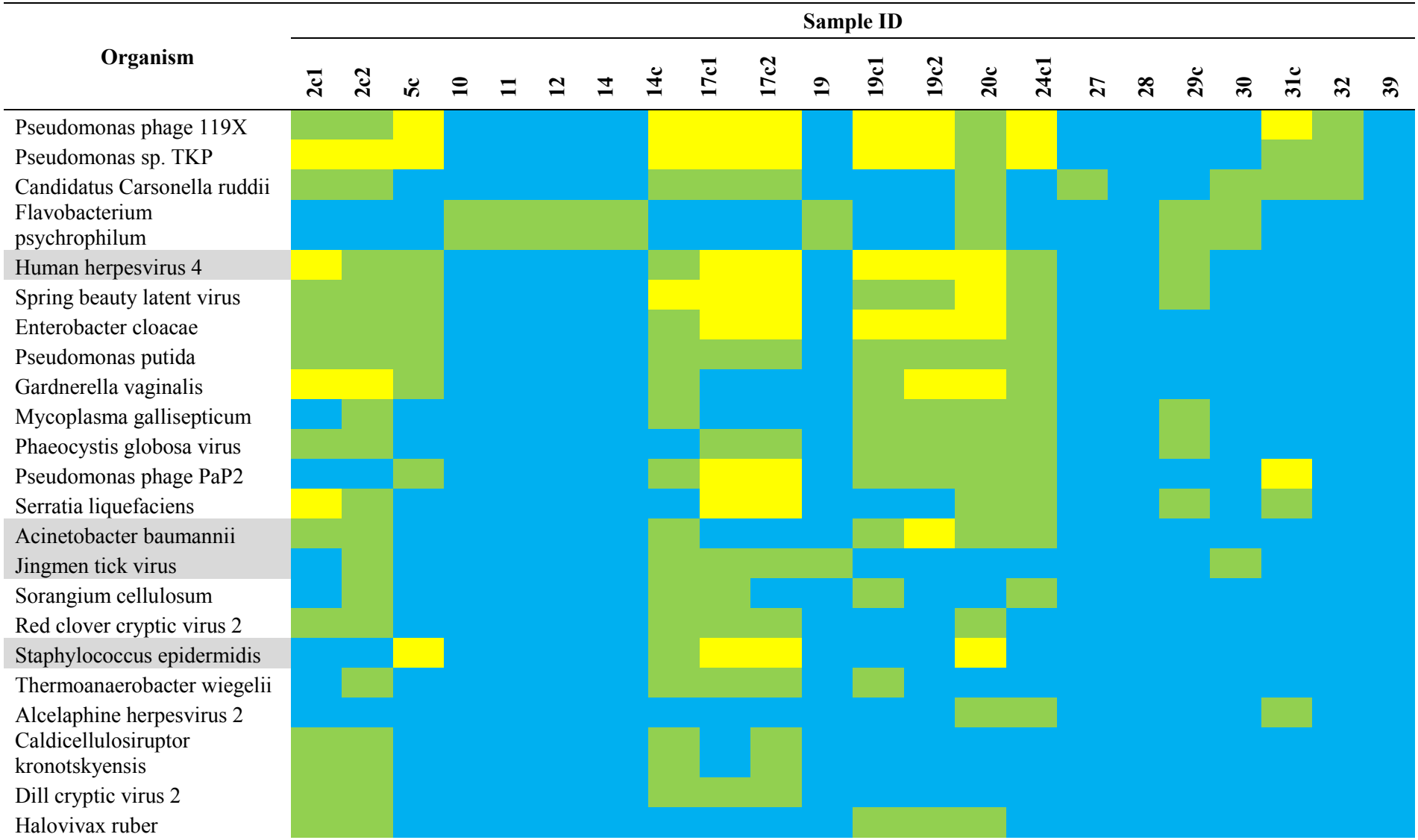
Previous research conducted by Nikkari et al[206] successfully detected bacterial 16S ribosomal DNA in healthy human blood using real-time PCR and traditional sequencing on an ABI PRISM platform. Phylogenetic analysis inferred similarity between the 16S rDNA sequences found in the samples and *Riemerella anatipestifer*, *Pseudomonas fluorescens*, *Propionibacterium acnes*, *Microbacterium schleiferi*,

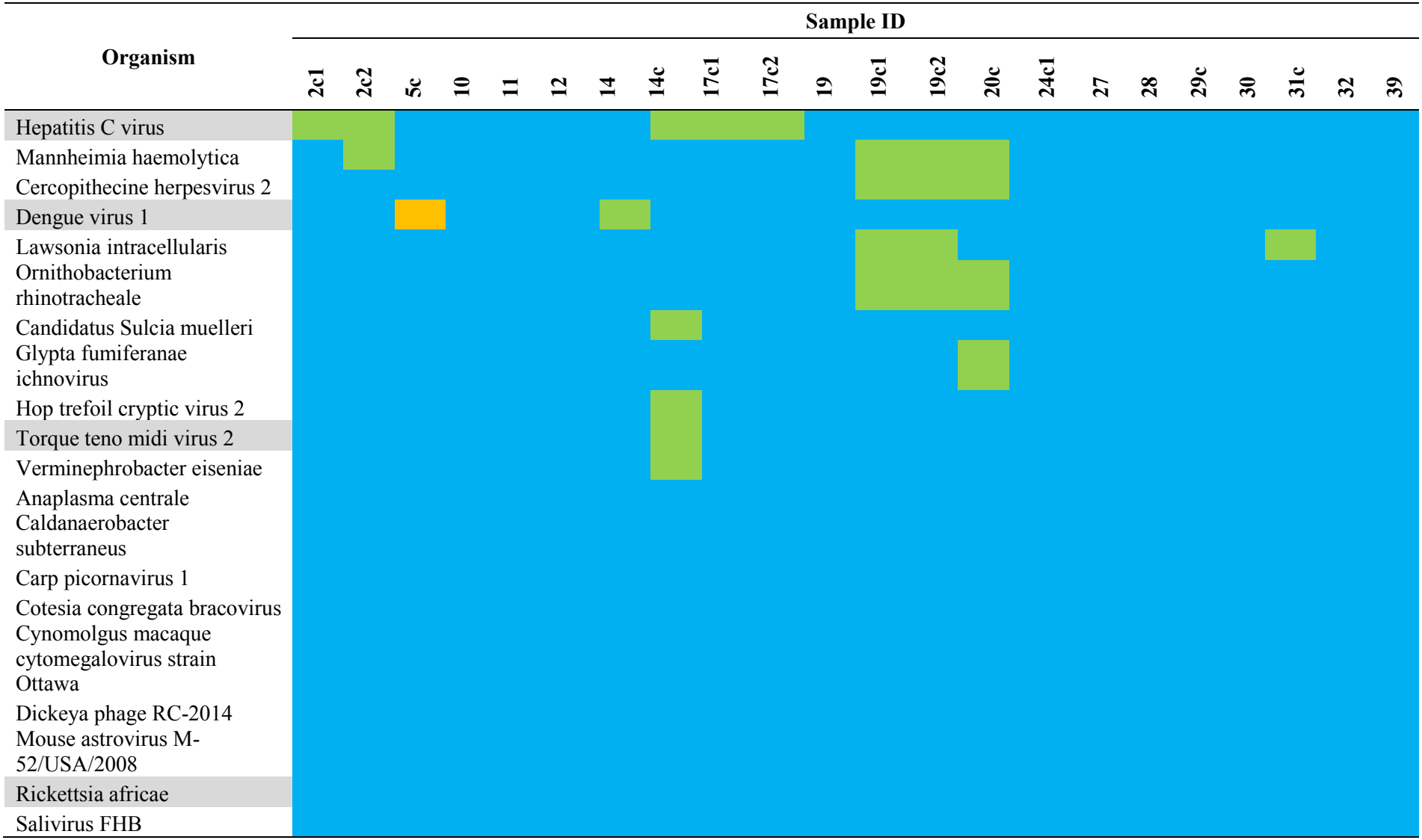
Stenotrophomonas and *Pseudomonas putida.* The authors presumed that the origins of these sequences were either from experimental reagents, from the skin during phlebotomy or from the blood itself. Nevertheless, the findings of the study raised the possibility that there is a 'normal' population of bacterial DNA sequences in blood that has previously been considered sterile. The use of NGS in the present study facilitated the identification of organisms that might well escape cultivation because of their low burden in the blood or simply because they are unculturable. Table 6.3 shows 61 organisms that were present in all samples and their relative abundance. Some of these organisms (those highlighted grey in Table 6.3) and their clinical relevance are discussed further in this chapter.

Table 6.4: Organisms present in all samples, detected by Kraken analysis

Relative abundance (proportion of organism read among non-host read) is shown by graded colour with 'blue' indicates < 0.01%, 'green' indicates 0.01%–< 0.1%, 'yellow' indicates 0.1%–< 1%, 'orange' indicates 1%–< 10% and 'red' indicates 10%–18.48%.




Organism
Sample ID
2c1
2c2
5c
10
11
12
14
14c
17c1
17c2
19
19c1
19c2
20c
24c1
27
28
29c
30
31c
32
39
Pseudomonas phage 119X
Pseudomonas sp. TKP
Candidatus Carsonella ruddii
Flavobacterium psychrophilum
Human herpesvirus 4
Spring beauty latent virus
Enterobacter cloacae
Pseudomonas putida
Gardnerella vaginalis
Mycoplasma gallisepticum
Phaeocystis globosa virus
Pseudomonas phage PaP2
Serratia liquefaciens
Acinetobacter baumannii
Jingmen tick virus
Sorangium cellulosum
Red clover cryptic virus 2
Staphylococcus epidermidis
Thermoanaerobacter wiegelii
Alcelaphine herpesvirus 2
Caldicellulosiruptor kronotskyensis
Dill cryptic virus 2
Halovivax ruber


Organism
Sample ID
2c1
2c2
5c
10
11
12
14
14c
17c1
17c2
19
19c1
19c2
20c
24c1
27
28
29c
30
31c
32
39
Hepatitis C virus
Mannheimia haemolytica
Cercopithecine herpesvirus 2
Dengue virus 1
Lawsonia intracellularis
Ornithobacterium rhinotracheale
Candidatus Sulcia muelleri
Glypta fumiferanae ichnovirus
Hop trefoil cryptic virus 2
Torque teno midi virus 2
Verminephrobacter eiseniae
Anaplasma centrale
Caldanaerobacter subterraneus
Carp picornavirus 1
Cotesia congregata bracovirus
Cynomolgus macaque cytomegalovirus strain Ottawa
Dickeya phage RC-2014
Mouse astrovirus M-52/USA/2008
Rickettsia africae
Salivirus FHB

6.4.4.1 Achromobacter xylosoxidans and Alteromonas macleodii

Achromobacter xylosoxidans is a Gram-negative bacterium with flagella. It can be found in water environments and has been isolated from both immunocompetent and immunocompromised patients with bacteraemia, chronic otitis media, meningitis, UTIs, abscesses, osteomyelitis, corneal ulcers, prosthetic valve endocarditis, peritonitis and pneumonia.[207, 208] These bacteria are not a typical component of human flora and have low virulence.[209] Infection with *A. xylosoxidans* is widely considered to be opportunistic, and the source of infection is usually found to be a contaminated solution.[207] The bacteria can survive in aqueous environments with minimal nutrients, so it is likely that the relatively high abundance of these bacteria (0.16–1.38% of non-host reads) indicates contamination from the water or reagent used during sample preparation for sequencing.

Alteromonas macleodii is a marine bacterium. These bacteria are commonly found in temperate or tropical sea waters.[210, 211] The presence of these Gram-negative bacteria in humans has not been reported. The present study detected high levels of *A. macleodii* reads across the samples, accounting for 0.51–11.35% of non-host reads. It is suspected that this organism is a contaminant, and its presence in the NGS dataset can be disregarded.

6.4.4.2 Enterobacteria phage phiX174 sensu lato

A bacteriophage (phage) is a viral predator that can infect and replicate within a bacterium. As *Enterobacter* is part of normal gut flora, there is obviously abundant enterobacteria phage in the human body. Assuming that the phage truly originated from the patients' samples, the question is how this phage can escape the gut–blood barrier. Sequencing phage genomes is an interesting field of research, with potential uses for phages as antimicrobials and biocontrol agents for food production.[212] In fact, the phiX174 bacteriophage was the first DNA-based genome to be sequenced, dating back to 1977.[213] There is the possibility that enterobacteria phage phiX174 from another research group was introduced to the samples during library preparation and sequencing at AGRF. Previous research[214] has also detected enterobacteria phage phiX174 sensu lato as a common contaminant in NGS analysis performed on blood samples.

6.4.4.3 Escherichia coli

E. coli is a Gram-negative bacterium that frequently causes bacteraemia and is the most common organism associated with nosocomial infection.[215, 216] The highest level of *E. coli* in the present study (1.01% of non-host reads) was found in sample 2c1, and it is possible that patient ID# 002 had diarrhoea caused by *E. coli*, although the burden of this organism in the blood did not permit detection by blood culture (see Section 6.3.5 'Undiagnosed subjects: patient ID# 002').

6.4.4.4 Human herpesvirus 4

Human herpes virus 4 (HHV-4) or Epstein-Barr virus (EBV) is one of the most common viruses in humans. This virus is an agent causing infectious mononucleosis (glandular fever);[217] it is also associated with certain cancers, such as Hodgkin's[218] or non-Hodgkin's[219] lymphomas and nasopharyngeal carcinoma,[220] and autoimmune diseases, such as systemic lupus erythematosus[221] and multiple sclerosis.[222] This virus is widely spread around the globe, and around 95% of the human population is infected with EBV.[223] Therefore, the presence of low levels (< 1% of non-host reads) of EBV in all samples is not surprising.

6.4.4.5 Propionibacterium acnes and Staphylococcus epidermidis

Propionibacterium acnes and *S. epidermidis* are Gram-positive bacteria that are commonly found in human skin. These organisms are likely sample contaminants that were introduced during blood collection. *P. acnes* and *S. epidermidis* were also detected on previous metagenomics and metatranscriptomics studies on human-derived samples.[206, 224]

6.4.4.6 Acinetobacter baumannii

Acinetobacter baumannii is a Gram-negative bacterium that is frequently reported as a cause of fatal nosocomial infection due to its multi-drug-resistant properties.[225] Resistant antibiotype and mechanical ventilation were found to be potential independent risk factors for mortality.[226] However, this organism is ubiquitous in the environment, so it is difficult to differentiate true infection of *A. baumannii* and contamination based on sequencing results. Sample 19c2 had significantly higher reads for these bacteria compared to other samples. It is likely that the cause of fatality in

patient ID# 019 was *A. baumannii* (see Section 6.3.5 'Undiagnosed subjects: patient ID# 019').

6.4.4.7 Hepatitis C virus and dengue virus 1

The presence of these viruses in all samples is evidence of cross-contamination from one sample to another. One sample was collected that was positive for hepatitis C virus (patient ID# 006); however, this sample was not sequenced due to insufficient amounts of nucleic acids to enter the amplification step. Accordingly, it was presumed that contamination occurred during DNA/RNA isolation. The low abundance (< 0.1%) of hepatitis C virus reads in all samples suggests that this organism is likely present as a contaminant. As for dengue virus, although it was detected in all samples, its presence in high titre (2.58%) in sample ID# 5c indicates a true infection. Dengue virus reads were also present in relatively higher proportion (0.01%) in sample ID# 014 compared to the rest of the samples. It is possible that patient ID# 014 had a dengue infection (see Section 6.3.5 'Undiagnosed subjects: patient ID# 014').

6.4.4.8 Jingmen tick virus

Jingmen tick virus is a recently discovered RNA virus.[227] This virus is widely distributed in tick populations across China, and its genome has similarity to those of Flavivirus and *Toxocara canis*, a roundworm found in dogs.[228] The detection of Jingmen tick virus in all samples could be the result of spurious alignment during bioinformatics analysis. This virus sequence might be inferred from dengue virus (a Flavivirus) sequence, which was also detected in all samples.

6.4.4.9 Rickettsia africae

Rickettsia africae has been reported as the common cause of African tick-bite fever, which is commonly found in people travelling to sub-Saharan Africa.[229] The symptoms of illness are usually mild, and may include influenza-like syndrome; rash is not always present.[230] One of the patients recruited for the study (ID# 026) was of South African nationality, but that patient's sample did not undergo deep sequencing due to a low DNA/RNA yield. She was in Australia in 2007–2009, returned to South Africa in 2009–2011, and stayed in Australia from 2011 onwards. This patient was a 37-year-old female who presented to Cairns Hospital on 12 June 2013 with fever, sore throat, headache, myalgia, arthralgia, abdominal pain and loin pain. She stated that the

fever was associated with the feeling of hot and cold, whole body rigors and sweats. She reported several previous episodes dating back to 2007. The fever episode was experienced approximately once per year for the duration of 2007–2012, mainly presenting as whole body rigors and sweats. Her last episode had occurred in April 2013, lasting for one day, and she did not receive medical treatment for it. At Cairns Hospital, she was diagnosed with possible arbovirus infection. Pathology results showed elevated WBC count (12.7 x 10^9/l) with neutrophil predominance and increased CRP (30 mg/l). Arboviral serology showed IgM reactivity to Barmah Forest virus. The attending doctor did not order a test for *Rickettsia*. The *R. africae* detected in study samples might have originated from this patient.

6.4.4.10 Torque teno midi virus

Torque teno midi virus (TTMDV) is a member of the genus *Gammatorquevirus* in the family *Annelloviridae*. The virus was first isolated in 2005 from plasma samples of individuals with high-risk behaviour leading to HIV-1 acquisition.[231] While other annelloviruses (such as torque teno mini virus [TTMV] and torque teno virus [TTV]) are frequently detected in healthy and diseased humans as well as in non-human primates,[232–234] TTMDV hosts are likely limited to humans.[235] This virus has been found in various body fluids, including saliva and nasopharyngeal aspirates, serum, urine and stool collected from children with acute respiratory disease.[236] The frequent detection of TTMDV in our study is consistent with a previous metagenomics study conducted by Li et al in 2012,[237] which reported that TTMDV constituted the second largest viral community after TTV in the plasma of healthy adults.

6.5 Chapter summary

A prospective study was conducted in a hospital setting recruiting patients with AUF for whom a diagnosis was not immediately obvious. The subjects of this study included individuals with specific diagnoses made during the course of routine investigations at the hospital and those who remained undiagnosed after a series of investigations ordered by the treating doctors. This study applied a deep sequencing approach to the investigation of the agent causing fever in both groups of patients. Deep sequencing results (with Kraken analysis) concurred with diagnoses obtained by other

means in 87.5% (7/8) of positive control samples, or 85.7% (6/7) of positive control subjects.

As for the patients with undiagnosed illnesses, deep sequencing identified some plausible causes of fever in 60% (6/10) of subjects (ID# 002, 011, 014, 019, 027, 039). The cause of fever was likely non-infectious in 2 subjects (ID# 012, 028). It was not possible to determine the cause of fever in 2 patients (ID# 029, 030); the detection of Jingmen tick virus and *Rickettsia africae* in these patients was likely insignificant. In addition to confirming and facilitating diagnosis, this study provides important information on microbial diversity in a 'sterile' environment, that is, human blood. Some findings from the NGS analysis, including medically important viruses and bacteria, were discussed to illustrate the high sensitivity of the deep sequencing approach. Unfortunately, this sensitivity may interfere with diagnostic appraisal due to sample contamination or spurious alignment during bioinformatics analysis.

The next chapter presents a general discussion integrating the topics of the previous chapters, practical challenges in metagenomics research, the present study's strengths and limitations, and suggestions for implementation of the results. The chapter concludes by revisiting the research questions and offering reflections.

Chapter 7: General Discussion and Conclusions

7.1 The deep sequencing approach to fever investigation

In previous research, deep sequencing has been initially used to reveal microbial diversity in environmental samples through direct isolation of genomic DNA.[238] This approach relies on the genomic analysis of a population of microorganisms and circumvents the process of culturing. The application of this approach has since been extended to clinical practice for the identification of markers of genetic diseases[239] and tumours.[240] With the invention of high-throughput next generation sequencing machines, referred to as NGS in this thesis, it is now possible to produce large amounts of sequencing data within days in order to obtain scarce genetic information of interest. This research project has evaluated the use of NGS as a broad generalist tool for characterising pathogens associated with AUF.

Little is known about the aetiology and investigation of AUF, as previous research has largely focused on FUO, which is clearly defined by the WHO and involves fevers of longer duration than AUF. A common method of investigation in AUF involves the evaluation of one or several agent(s) of interest, and does not measure the epidemiology of undiagnosed cases as a result of the comprehensive investigation of infectious diseases. While the majority of AUFs resolve spontaneously, some cases become prolonged and cause significant morbidity and mortality. It is important to address this disease burden and to find a new method for investigating the causes of AUF.

Although it is common in clinical practice, the extent of AUF in Far North Queensland is poorly understood. The first phase of this study sought to determine the epidemiology of AUF in Far North Queensland, to measure proportions of undiagnosed cases, and to gather information for the subsequent metagenomics study of patients with AUF. The findings of this initial study supported the hypothesis that AUFs are common in the population of Far North Queensland. A total of 340 AUFs presented to Cairns Hospital over a 3-year period (around two cases per week). The findings also supported the hypothesis that a significant proportion of AUFs do not result in a specific diagnosis. Among the 340 AUFs identified in this initial study, about half of all cases remained undiagnosed.

The high proportion of undiagnosed cases supported the hypothesis that UUDFs occur frequently enough at Cairns Hospital to justify a subsequent NGS study looking for unpredicted and unknown (novel) infectious agents associated with UUDF at this site. Further, a robust definition of UUDF could be developed based on the study's findings. An implication of this is that future research could compare incidences of this condition across different geographical sites. Since a global epidemiology of UUDF has not yet been achieved, the criteria proposed in this study for the definition of UUDF could be used for the surveillance of this condition in any part of the world, from well-developed countries to rural areas. For example, an unusually high incidence of UUDF in a particular area could be an indication of a new emerging disease; clinicians working in that area could report this via a program for monitoring emerging diseases, such as ProMED-mail (promedmail.org). Such reports serve to increase global awareness and encourage further efforts to characterise and contain the emerging disease. These findings also urge improvement in the diagnosis of AUF so that the number of cases of UUDF and FUO can decrease in the future.

To date, there has been little research undertaken to investigate the use of NGS in clinical practice, due to its high cost and complex methods of sample preparation and data analysis. One way to reduce the sequencing costs involved in NGS is by requesting the minimum amount of sequencing data that can be produced without compromising the sensitivity of detection of the pathogen. This can be achieved by reducing human DNA contamination to maximise the yield of pathogen sequences.

The second phase of the study investigated amounts of human DNA in plasma and serum collected from healthy volunteers. The study further investigated the effects of different techniques for blood collection on the concentration of human DNA in plasma and serum. Plasma or serum were chosen for the metagenomics study because of the significant costs of deep sequencing, and the need to avoid swamping the data with excessive human sequences derived from whole blood. The findings of this study support the hypothesis that plasma and serum contain different amounts of human DNA. It was shown that DNA concentrations are significantly lower in plasma than those found in serum. However, the study findings do not support the hypothesis that DNA concentrations in plasma and serum are affected by the method of blood collection. No differences were found in the DNA concentrations of plasma and serum collected either by vacuum system or by standard syringe and needles. There were also no differences found in the DNA concentrations of plasma and serum collected

according to whether or not a tourniquet was used. Following the implications of this study, it was decided that plasma samples would be collected for the subsequent metagenomics study. If a sample should be collected retrospectively and plasma was not available, serum samples would be used.

In the main study, 22 samples originating from 17 patients with AUF were deep sequenced. The findings of the main study suggest that deep sequencing represents a highly sensitive but not specific method for identifying pathogens associated with fever. The study findings support the hypothesis that the use of NGS technology can identify the same pathogens in patients with fever as can diagnoses achieved using contemporary diagnostic methods; this finding was achieved from the control participants, who had already received a definite diagnosis. Deep sequencing (with Kraken analysis) was capable of detecting the pathogen causing fever in 7/8 of the control samples. The lengthy analysis of reads unclassified by Kraken made it difficult to support the hypothesis that the use of NGS technology in patients with UUDF (study subjects) can inform diagnosis by detecting genome sequence(s) of previously known (unpredicted) or unknown (novel) pathogens. However, it was possible to propose the most likely causes of fever in some undiagnosed subjects. The implication of this study is that if the deep sequencing approach is to be used for routine investigation of fever, it is necessary to address several challenges, as discussed below.

7.2 Practical challenges in the study

Ethics and governance approval was the first obstacle that had to be overcome before starting this project. It took about one and year to gain full approval, mainly because of the unforeseen delay in the process of obtaining governance approval. There was a 7-month delay in obtaining permission to access retrospective samples stored at the Pathology Department of Cairns Hospital. This permission had to be granted by the CaSS located in Brisbane. Further delay was experienced during the process of SSA approval (see Chapter 3, Section 3.3.3, Figure 3.3). The SSA application form was detained by the financial officer for 4 months so the process of SSA approval could not move forward. This slow administrative process was unusual and unexpected and had affected a number of ongoing studies at that time. As for ethics approval, it was anticipated that the HREC would express their concerns with regards to the use and

storage of human genetic data. Ethical queries and their resolutions were discussed in Chapter 3 (see Section 3.3.3–3.5).

The next challenge arose during patient recruitment. Most of the patients were not familiar with genomics studies, and this lack of understanding often made them reluctant to participate. This problem was discussed in Chapter 3, including approaches for resolution (see Section 3.4). The recruitment strategy resulted in the majority (41 out of 45) of patients approached giving their consent to participate in the study; 40 patients were recruited from Cairns Hospital and one patient from Cairns Private Hospital. Another problem relating to participation was patients' commitment. While they were sick, all 41 participants had consented to attending a follow-up appointment and donating convalescent samples. However, most of the patients did not comply. As a result, it was only possible to collect 10 paired samples from the total of 41 participants. If the results of deep sequencing had been available within a 2–3 week time frame, most patients probably would have been happy to attend follow-up appointments and discuss the results of tests conducted on their blood.

Sample preparation for deep sequencing was the major problem encountered in the course of the research. Despite achieving low-level human DNA contamination, which is necessary to keep sequencing costs down, the plasma and serum samples had very low quantity and quality of nucleic acids. It was necessary to process the samples carefully to prevent further degradation and loss of the pathogen nucleic acids. Low-input samples are challenging for library preparation, and more so when the fragment lengths are small. At the time of preparing the sample for sequencing, the available NGS library preparation kits (e.g., the TruSeq and the Nextera kits from Illumina) required hundreds of nanograms of total DNA/RNA input and lengths of at least 300 bp, necessitating whole-genome or whole-transcriptome amplification (WGA and WTA) of low quantities of DNA/RNA isolated from the cell-free samples.

It is important to note that, although several WGA and WTA kits are available on the market, their use is limited by the input requirement and downstream application. For example, Qiagen released the Repli-g® WGA kit, which requires > 10 ng of starting DNA at a size of >2kb in length,[241] and the QuantiTect® WTA kit requires a starting amount of 10 ng of intact RNA at a size of > 500 nucleotides.[242] While the Repli-g® WGA kit could be used for downstream applications in NGS analysis, the QuantiTect® WTA kit is intended for use in real-time PCR, and has not been tested for NGS (this information was obtained through personal correspondence

with Qiagen technical support). The Ovation® RNA-Seq kit (Nugen) requires 500 pg–50 ng of total RNA as starting material,[243] much less than the amount required by the Qiagen amplification kits. However, the Ovation® kit selectively amplifies polyA RNA species,[244] meaning that non-polyA species, such as dengue virus, might not be amplified for NGS analysis.

Bioinformatics analysis was an enormous challenge in this study because, while many tools are available for analysing sequence data, they require expert users to put them together into an effective workflow. The most widely used strategy is computational subtraction: reads are first sequentially aligned to reference databases to filter out sequences corresponding to host background. Sequences derived from microbes are then typically identified by nucleotide or translated amino acid alignments using BLAST.[106] It is a challenging task to handle multi-gigabyte files, remove the reads that map to human DNA, perform *de novo* assembly on the remaining reads, and then investigate those assembled reads (contigs) to see what they match up to in the databases. Conventional BLAST alignment is a highly computer-intensive process that maps millions of contigs to the NCBI non-redundant nucleotide database.

The available metagenomics analysis methods are generic and identify all organisms present in the sample but provide no guidance in terms of how to interpret what is pathogenic and what is not, or what is most likely contributing to the AUF symptoms. We used Kraken, an ultrafast program for assigning short DNA sequences and performed secondary analysis involving BLAST search on reads that were unclassified by Kraken (see Chapter 6, Section 6.2.3). On completing the analysis, it was found that it was difficult to interpret the results based on sequence data only. Detection of organisms may or may not be meaningful; it may be due to spurious alignment or contamination. There is also a possible sequencing bias introduced by the WGA step, as reported in previous studies,[245] that the current amplification methodologies (including multiple displacement amplification, primer extension pre-amplification and degenerate oligonucleotide primed PCR) induced statistically significant bias relative to unamplified control. On the other hand, lack of pathogen detection can be caused by lack of coverage or unmatched sequences to any known pathogen in the database. The sensitivity of NGS itself has made it difficult to maintain the reliability of findings. Every sample was different, even among duplicates processed and analysed in the same manner. The 'incidental' agents detected on the NGS datasets complicated data interpretation and diagnosis appraisal.

It is of key importance to estimate the sequencing depth required to effectively identify the pathogen present in the sample. At the beginning of the study, it was unknown how 'deep' the sequence analysis would need to be in order to demonstrate clinical utility. If data were insufficient, any given pathogen could have been missed, while too much data would be costly to generate and analyse in sufficient depth, and would hinder the implementation of the approach in routine clinical investigation. As there was wide variation in the DNA/RNA yields, the amounts of sequencing data and the organisms identified in the samples, it is not possible to rely on the results of a pilot sequencing. This meant that running a pilot sequencing effort using samples spiked with known quantities of virus/bacteria might not always be useful to assess the detection sensitivity and the interpretation of findings.

The amount of data produced in this study was determined pragmatically; the target was that the cost of deep sequencing would not exceed the cost of routine investigation of fever. In this study, AUD $9,570 was paid to AGRF for library preparation and sequencing. Additional costs incurred during sample preparation included AUD $2,500 for purchasing kits for nucleic acid isolation and amplification. Laboratory consumables were not included in this cost; these are estimated at AUD $500–1,000. As an estimate, a deep sequencing approach is feasible for investigating cause of fever at a cost of AUD $600 per sample. The ability to detect the pathogens causing fever depends on the pathogen burden in the sample and the depth of sequencing.

7.3 Study strengths and limitations

To the author's knowledge, this was the first study assessing the practicality of the deep sequencing approach for routine investigation of AUF. Although deep sequencing had been used in previous studies on AUF,[129, 132, 203] those studies were performed at their academic centres with an in-house sequencing machine, and did not disclose the cost-effectiveness of the deep sequencing approach nor the challenges involved in working with cell-free samples and lengthy bioinformatic analysis. The researchers' experience with the deep sequencing approach showed that the technology is not yet ready for routine clinical investigation. The development of NGS technology should be directed at reducing sample input. Improvement in bioinformatics is also urgently needed for the adoption of this approach in clinical settings.

This study provided important information with regards to the sensitivity and cost-effectiveness of deep sequencing on an Illumina HiSeq instrument for fever investigation. It was demonstrated that, with an optimum sample such as that from patient ID# 005, ~15 million PE reads (~2 Gb of data) are sufficient to facilitate a diagnosis of dengue virus. When the virus load was sufficiently high, *de novo* assembly permitted the generation of several contigs corresponding to a nearly full-length genome, thus confirming the diagnosis of dengue fever. Identification of pathogens was a fundamental goal in this study, regardless of whether or not complete genome sequences can be assembled. The two-tier bioinformatics analysis was useful in accomplishing the study aim in the majority of cases and resulting in only a small proportion (< 30%) of contigs remaining unclassified. These unclassified contigs (reads) might have originated from a highly divergent group of species that did not align well with the reference sequences. Further work could be performed to analyse these unclassified reads; however, this work is beyond the scope of this PhD project for the following reasons.

The individual reads generated in this study were short, typically only ~100 nucleotides (nt) in length. Such sparse reads do not overlap sufficiently to permit *de novo* assembly into longer sequences. For the construction of high-quality genome(s) for new pathogen(s), re-sequencing is likely required, followed by complicated gap-closing procedures. In addition, the search for novel microorganisms is a huge challenge in bioinformatics analysis, because the divergent genomes of these organisms are not adequately represented in existing reference databases. These novel microorganisms often can only be identified on the basis of remote amino acid homology; for example, by using protein similarity search tools such as BLASTx[186] and RAPSearch.[246] This is followed by phylogenetic analysis for inferring or estimating the evolutionary relationship between the novel pathogen candidate and its ancestors, as well as for studying the divergence between the novel pathogen candidate and the other species within the same taxonomic group. Finally, validation tests are required prior to reporting a new organism. The extra time required for pathogen discovery did not fit into the timeline of this PhD.

Another limitation of this study was that the bioinformatics findings were not cross-validated using conventional methods of investigation such as culture, serology or PCR. The analysis using Kraken and the CLC Genomics Workbench reported an exhaustive list of microbes, and it would have been time consuming and costly to

perform separate validation tests on each (or even the most likely) pathogen associated with fever. In addition, some agents were detected in a small number of reads, representing a low pathogen burden in blood samples that is potentially insufficient for detection by culture or PCR. Among the undiagnosed subjects tested for deep sequencing, convalescent samples were available in four patients (ID# 011, 012, 014, 028), but serology was not performed on these samples for the following reasons. The results of deep sequencing supported the provisional clinical diagnosis of scrub typhus in patient ID# 011. The causes of fever in patient ID# 012 and 028 were likely non-infectious. In patient ID #014, the most likely aetiologies detected on deep sequencing were EBV, dengue virus, Jingmen tick virus, *Rickettsia africae* and *Streptococcus pneumoniae.* This patient had been investigated for EBV, dengue virus, *Rickettsia* and *S. pneumoniae* by means of serology and PCR during his admission at hospital, and all results were negative (see Section 6.3.5 'Undiagnosed subjects: patient ID# 014'). Therefore, no validations of the deep sequencing results were performed using the same methods that had been performed by the hospital. As for Jingmen tick virus, this virus was detected in all samples, and it is believed that the detection of this virus in deep sequencing was incidental due to the similarity of the Jingmen tick virus genome and the Flavivirus genome (see Section 6.4.4.8 'Microbial diversity in human blood: Jingmen tick virus').

7.4 Suggestions for implementation

Although deep sequencing is not yet ready to be utilised as a routine diagnostic tool, this technology has developed at a rapid pace over the last decade, so it is likely that in the near future deep sequencing will become commonplace in clinical settings. This section discusses several aspects of the present study that can contribute to the success of fever investigations using a deep sequencing approach.

Searching for scarce pathogen nucleic acids on an abundant background of human nucleic acids is a daunting challenge. Therefore, samples should be collected from patients in acute stages of illness and prior to administration of antibiotics to increase the sensitivity of detection of the pathogen. Study participation should be limited to patients with fever of less than a one-week duration, and should exclude patients who have taken antibiotics. Patients with these characteristics are more likely to present to primary health care rather than a tertiary hospital. Regardless of the study

setting, a grading system is proposed to determine the potential significance of undifferentiated fever for deep sequencing analysis (see Chapter 4, Section 4.4.5, Table 4.2). In this system, disease severity is scored based on fever duration, length of hospitalisation, thoroughness of investigation and laboratory abnormalities. This scoring system may assist in selecting the most clinically significant samples for more intensive investigations. We propose that patients scoring 5 or more points should be considered for investigation using deep sequencing.

Once the ideal subject for deep sequencing analysis has been identified, sample collection should be performed aseptically. Plasma is an ideal sample for pathogen detection and discovery, as it has the lowest level of human nucleic acids compared to serum or whole blood (see Chapter 5, Section 5.3, Figure 5.2). On the downside, the concentration of DNA/RNA in plasma varies from undetectable up to hundreds of nanograms, making it frequently inadequate for metagenomics analyses using current NGS platforms. While it is possible to obtain micrograms of DNA/RNA from whole blood, excessive amounts of human nucleic acids from leukocytes and other immune cells would likely swamp the sequence data, therefore requiring a higher volume of sequencing data in order to gain adequate coverage of the pathogen nucleic acids in the sample. After collection, samples should be snap frozen or stored at -70 °C, or in buffers designed to preserve nucleic acid integrity and to avoid further degradation of nucleic acids. It is important to note that delays in separating the plasma or serum at room temperature can increase the concentration of human genomic DNA from leukocyte lysis.[145, 154]

A sample preparation for deep sequencing should be performed quickly but carefully, because the quantity and quality of input material is one of the most important determinants of a successful sequencing library. Working in a fume hood facility with equipment specifically designated for a deep sequencing project is necessary to prevent contamination from the environment, and using different gloves to process samples from different individuals is recommended to prevent cross-contamination. Other important sources of viral and bacterial contamination can include reagents and laboratory equipment, as reported previously.[247–249] Further, a strict washing procedure to clean the Illumina flow cell should be applied to prevent cross-contamination caused by carryover. Alternatively, each sample may be sequenced in different lanes, but that would dramatically increase the cost and time of sequencing. Overall, it is important to keep the sample processing steps to a minimum. A good

example for this found in a study conducted by Bzhalava et al,[250] in which the authors attempted to recover Human Papilloma virus (HPV) from skin lesions. The authors reported that the separation of viral DNA from human DNA before WGA and sequencing was less successful for detecting viral reads than directly subjecting samples to WGA and sequencing.

Because the cost of sequencing is significant, sample processing must be carefully designed to bring down the costs. There are several methods that can be used to pre-treat samples, and research needs to be done to determine the most effective one. The methods of sample preparation for sequencing can be grouped into two strategies: depletion of human nucleic acids and enrichment of pathogen nucleic acids. Depletion of human RNA can be conducted using kits specifically designed for the removal of human ubiquitous ribosomal RNA (rRNA), such as the Ribo-Zero rRNA removal kit from Epicentre and the rRNA removal kit from Arraystar. Removal of human DNA can be accomplished using methylation-dependent restriction endonucleases to remove methylated DNA[251] or by performing sample filtration and DNase treatment.[252, 253] It should be noted, however, that every additional step will reduce the DNA/RNA yield and increase the possibility of contamination. Enrichment of pathogen nucleic acids is not applicable if we aim to use deep sequencing as a broad generalist diagnostic tool, because current methods of nucleic acid enrichment rely on the use of PCR, which will carry out preferential amplification based on the primers used. However, if deep sequencing is to be used for the study of selective organisms of interest, enrichment of the desired subsets can be performed using a method called suppression subtractive hybridisation.[121] Other strategies for target DNA/RNA enrichment, as well as selecting an appropriate NGS platform for a particular metagenomics project, have been discussed previously.[119]

The choice of NGS platforms on the market today for pathogen detection and discovery is driven by two main parameters: read length and read depth. The maximum capacity of the 454 platform (currently the average read length generated from the Roche 454 GS-FLX platform is 700 bp) provides the longest sequencing reads but the lowest throughput, while other NGS platforms give much higher throughput but very short reads.[254] The long reads produced by the 454 increase the specificity of pathogen identification by facilitating the discrimination of pathogen reads from hosts or endogenous flora.

Greninger et al[255] reported that long reads from 454 pyro-sequencing are more suitable for the identification of bacteria compared to Illumina short reads. Although Illumina platforms generate short reads (currently up to 2 x 150 bp for the HiSeq and Genome Analyzer II and 2 x 300 bp for the MiSeq), this platform can generate sequencing data much faster than the 454 platform, thus providing sufficient read depth or number of sequence reads generated per run to detect pathogens with a high degree of sensitivity. Recent Illumina NGS sequencing platforms (GAIIx^{TM}, HiSeq^{TM} and MiSeq^{TM}) can generate 10–1000-fold improved read depth relative to the 454, which makes Illumina sequencing an ideal platform for viral discovery.[256] For the purpose of identifying scarce amounts of pathogen nucleic acids among abundant human nucleic acids, paired-end sequencing (instead of single-end sequencing) should be performed to double the overall amount of data generated by the NGS and thus increase the sensitivity of pathogen detection. In addition, the use of paired-end sequencing can be particularly useful for pathogen discovery given that the forward and reverse reads can facilitate the design of PCR primers to confirm the existence of novel microbes. All things considered, in my opinion, it is more appropriate to use the Illumina MiSeq rather than the Illumina HiSeq for small-scale investigation such as that in a clinical setting.

It is certainly more convenient if a sequencing machine is available in-house. CNAs are often present at picogram levels, and the typical fragment length is less than 200 bp (DNA) or < 200 nt (RNA).[154, 257, 258] Having ready access to a sequencing machine will provide the flexibility to modify standard protocols for library preparation and sequencing without being tied to the use of a certain level of DNA or RNA as a pre-requirement for proceeding to NGS libraries. In addition, data acquisition can be obtained in a few hours or days, as opposed to a 6-week (or longer, depending on the sample queue) turn-around-time when sequencing is performed by a commercial company.

The availability of cheaper, faster, and more sensitive sequencing technologies will be greatly advantageous for outbreak investigation and pathogen discovery. For facilitating diagnosis of infectious diseases, specificity is the key of importance. Ideally, a NGS platform that provides long reads will generate a more specific diagnosis and will facilitate a reliable bioinformatics assembly of new pathogens where a reference sequence is not available. In the long run, the ideal platform should also be high-throughput, but economical, to be used for routine diagnosis purposes. The

development of single-molecule sequencing (also known as third-generation sequencing) via technologies such as nanopore sequencing or MinION (Oxford Nanopore Technologies) is highly promising; this pocket-sized genome sequencer can generate longer reads (tens of kilobases) at a cost comparable to that of currently available NGS instruments.[259] Another advantage of this portable DNA sequencer over NGS is its ability to perform real-time sequence analysis, which is highly valuable for providing results rapidly. The downside of nanopore sequencing is that the technology currently requires micrograms of DNA/cDNA input. Two recent studies[260, 261] reported disadvantages of MinION sequencing, including higher error rates (10–30%) and relatively lower throughput (< 100,000 reads per cell) than second-generation sequencing (NGS).

Access to a high-performance computer enables researchers to experiment with various analysis pipelines and to choose the most suitable pipeline and programs for metagenomics projects. Various programs for conducting taxonomic analyses of microbial communities in samples have been discussed previously.[262] The efficiency of taxonomic analyses can be assessed using four parameters: accuracy, specificity, execution time and computing power requirement. The accuracy of the analysis is good if the reads can be assigned to any taxon that lies in the taxonomic lineage of the source organism in the read. Specificity relates to the assignment of reads to a specific taxonomic level (strain, species, genus, family, order, class, phylum, kingdom and domain) that corresponds to the source organism in query. Requirements for computing power and analysis time are also significant parameters, especially in clinical diagnostics.

Bazinet and Cummings[263] contended that the performance of sequence classification programs can be evaluated through two main areas: assignment accuracy (sensitivity and precision) and resource requirements. Sensitivity of analysis was defined as the number of correct assignments divided by the number of sequences in the data set. Precision can be assessed by calculating the number of correct assignments divided by the number of assignments made. Resource requirements may relate to processing time, RAM and disk requirements. Bioinformatics programs can vary by orders of magnitude in terms of their computational resource requirements, so researchers should choose programs appropriately depending on the available computing resources, the amount of data to analyse and the particular bioinformatics application.

In the present study, sequencing data were analysed in the cloud environment to eliminate the need for a standalone supercomputer. Cloud-based analysis allowed for high volumes of data to be processed within reasonable requirements of computing power and within a clinically relevant timeframe. It is noted that there was a trade-off between the accuracy and specificity of a method and its time and computing power requirements. A recent study[264] described that, despite its impressive execution time, the Kraken program reported high numbers of false positives. Therefore, the usefulness of Kraken in the present study relates to the rapid screening of pathogens associated with fever. The second analysis pipeline, using the CLC Genomics Workbench, applied more stringent parameters than the Kraken analysis, and was useful for validating the Kraken findings. The downside of the CLC pipeline was the alignment step using BLAST, which completely dominates the runtime for the analysis, and the sequences of hitherto unknown organisms remained unidentified. Analysis on CLC can use a large amount of disk space and RAM if analysing a large number of sequences.

Taking everything into account, bioinformatics analysis for fever investigation should be developed for non-specialist users with the aim of providing a powerful tool for rapid pathogen detection in mixed microbial communities. The use of deep sequencing to identify pathogens associated with fever in the current environment requires communication and collaboration among clinicians, sequencing service providers, bioinformaticians and other experts such as computer scientists and infectious disease specialists.

7.5 Revisiting research questions

The findings from the first study provided answers to the following research questions: How common is AUF in the population of Far North Queensland, Australia? How is this condition investigated? What are the frequent diagnoses? What is the proportion of undiagnosed cases and what information exists with regards to this condition? Over a three-year period, 340 cases of AUF were detected, meaning that on average, two cases of AUF presented to Cairns Hospital per week. The real prevalence of AUF in the population of Far North Queensland is obviously higher than two cases per week, as most AUFs are treated by GPs, and only severe/prolonged cases are referred to the hospital.

Common investigations for AUF included routine blood tests, chest X-ray to rule out pneumonia and pulmonary tuberculosis, and specific tests for detecting infectious agents that are locally prevalent and endemic, such as arboviruses, *Leptospira* and *Rickettsia*. Dengue fever was frequently found as the cause of AUF in this region. Despite the availability of advanced medical tests and resident experts at Cairns Hospital, around half of AUFs were undiagnosed. This undiagnosed illness was referred to as UUDF in this thesis. It was found that the patients with UUDF presented with non-specific symptoms, predominantly constitutional and gastrointestinal symptoms, and were admitted for shorter periods compared to those who had specific diagnoses made. The results of routine blood tests were normal in the majority of patients with UUDF. Some notable abnormalities included elevated levels of hepatic aminotransferases and CRP, thrombocytopenia and leukocytosis. The results of urinalysis and blood culture were predominantly negative, ruling out UTI and bacteraemia. The reports of chest X-rays were inconclusive for diagnosis appraisal.

Based on the information gleaned from this study, a case definition for UUDF was proposed, including five criteria: 1) a fever of $\geq$ 38.0 °C or symptoms suggestive of fever; 2) a duration of fever of $\leq$ 21 days; 3) a failure to reach a diagnosis after performing clinical evaluation and laboratory investigations, including complete blood count, serum biochemistry, urinalysis, blood culture or chest X-ray; 4) a request by the clinician for specific tests for at least one infectious agent; and 5) a failure to make a specific diagnosis.

The findings from the second study provided answers to the following research questions: What type of blood specimen contains the least human DNA? Does blood sampling technique affect the concentration of circulating DNA? It was found that plasma has the lowest level of human DNA compared to serum and whole blood. Levels of human DNA in samples were not significantly different despite variations in technique during blood collection.

The findings from the third study provided answers to the following research questions: Is a deep sequencing approach using NGS technology a reliable method for diagnosing infectious diseases? Is deep sequencing a practical approach for identifying unknown pathogens in human blood? It was found that deep sequencing is a reliable and vigorous approach for microbe hunting, as it has a high degree of sensitivity in detecting any nucleic acid sequence present in a particular sample. However, for the purpose of diagnosing infectious diseases, results of bioinformatics analysis should be

interpreted carefully in conjunction with clinical data and the results of pathology or radiology tests.

Currently, the complexity of sample preparation and bioinformatics analysis makes the deep sequencing approach impractical for routine diagnosis, but it is believed that significant progress in sample preparation methods, NGS technology and bioinformatics tools will be accomplished in the near future. For instance, Parkinson et al[188] reported a modified transpososome-mediated fragmentation technique for NGS library preparation in Illumina sequencing platforms (Illumina GAII Genome Analyser and Illumina HiSeq 2000) using picogram quantities of DNA. If this technique can be adopted by commercial sequencing companies, input requirement will decrease significantly until amplification of DNA/RNA is no longer needed.

Recent developments in NGS technology have included the initiation of single molecular sequencing (third-generation sequencing). Third-generation sequencing platforms produce longer nucleotide sequences and higher volumes of data compared to the NGS platform used in the present study. These features will make the diagnosis of infectious diseases more reliable and more rapid. It is anticipated that integration of single molecular sequencing with fast bioinformatics tools will facilitate diagnosis appraisal in a real-time manner. A recent study by Greninger et al[265] reported the use of MinION (a third-generation sequencing platform) and MetaPORE (a newly developed program) to confirm diagnosis of Chikungunya, Ebola and Hepatitis C infections within 6 hours of sample receipt. Nonetheless, with the rapid decreases in sequencing costs, where $1,000 genomes and $700 exomes are now available,[266] NGS opens a frontier to personalised genomic medicine.

7.6 Reflections

I learnt some valuable lessons from my PhD project. The combination of clinical, molecular and bioinformatics work done represents an initial metagenomics project using clinical specimens in North Queensland. The molecular aspect of this research would have been much easier if the project had been conducted at a major laboratory with well-established sequencing equipment, such as those in Europe or America. It was a challenge to convince the sequencing suppliers to make an exception for the DNA/RNA levels given their pre-requirements for proceeding to NGS library-building and sequencing. Their hesitations in processing our samples are quoted below:

> Some samples look good and some samples looked not so good. Yes, I think it would be good idea to perform trial version first, to see if they are promising. To be honest with you, finding the possible pathogens can be quite challenging. I have discussed with our technical expert at our HQ and he also agrees the BI analysis can be a challenge, we are thinking to perform BLAST against NCBI.—Macrogen
>
> As for sequencing, some core facilities are not used to working with samples such as plasma/serum where the RNA is bound to be smaller/fragmented in length. Most RNA in plasma is microRNA of 200bp and under so we always recommend using facilities that have experience with these more specialized samples.—Norgen Biotek
>
> Looking at the gel, I would largely agree with Tri's assessment for lanes 1–14, however lanes 15 and onwards are very small (~100 bp), so I'm not sure about the suitability of these. Is the bionalyzer the same set of samples? If going by this I'd say ~10 of these (the ones with a distinct smear) are ok.—AGRF
>
> The samples for Nextera XT needs to be submitted at 5 ng/ul (total amount 100 ng). The ratios need to be ~2 for both 260/230 and 260/280. The samples that are under 100 bp will not work in the library prep.—Ramaciotti
>
> Your application is not appropriate for PacBio sequencing which is a single-molecule long-read sequencing technology requiring large quantities of DNA as starting material.—PacBio

Despite these difficult negotiations with sequencing companies, performing this study in Cairns provided a better opportunity to recruit participants with tropical infectious diseases. I also learnt that metagenomics projects require high levels of bioinformatics support. We underestimated the complexity of the bioinformatics analysis, which set us back when two bioinformaticians left JCU during the course of the project. All in all, investigating the causes of fever using a deep sequencing approach is a challenging task requiring immense intellectual and infrastructural support.

By the conclusion of the study, new library preparation kits requiring low levels of DNA input had been released to the market. These low-input library preparation kits include the NEBNext® Ultra™ DNA Library Prep Kit (New England BioLabs® Inc.), with input amounts as low as 5 ng of DNA, and the ThruPLEX® DNA-seq Kit (Rubicon Genomics), which can generate DNA libraries from as little as 50 pg of DNA. Within the last 2 years (2014–2015), the sequencing platform itself has been rapidly upgraded into Illumina HiSeq 2500/3000/4000, HiSeq X Five/Ten and NextSeq 500/550 systems. These new Illumina platforms run faster than the HiSeq 2000 used in

this study, enabling the generation of up to ~1,000 Gb of sequencing data in a couple of days. With the rapid development of sequencing technologies and simplified sample preparation techniques, deep sequencing approaches will become more likely to be adopted in clinical settings, particularly for the investigation of undiagnosed fever.

7.7 Conclusions

A study was performed to investigate the aetiology of AUF using a deep sequencing approach. Information on pathogens associated with fever could be obtained from ~2 Gb of data at a cost of ~AUD $600 per sample. Challenges were identified in conducting a deep sequencing approach, and resolutions were proposed. The success of the deep sequencing approach for fever investigation is influenced by appropriate timing of the sample collection, optimum sample preparation and adequate coverage of pathogen sequences. Future improvements in sequencing platforms are needed to provide longer reads and enable sequencing from smaller amounts of input material. The development of bioinformatics tools should be directed towards user-friendly options and the means to provide answers in clinically relevant timeframes (e.g., within hours of sample receipt). Recent advancements in sequencing technologies and bioinformatics analyses in the recent past provide a positive outlook for the application of the deep sequencing approach to facilitate the diagnosis of human febrile illness.

References

1. Ogoina D. Fever, fever patterns and diseases called 'fever'—a review. *J Infect Public Health*. 2011;4(3):108–24.
2. Reid JV. Pyrexia of unknown origin; study of a series of cases. *Br Med J*. 1956;2(4983):23–25.
3. Petersdorf RG, Beeson PB. Fever of unexplained origin: report on 100 cases. Medicine (Baltimore). 1961;40:1–30.
4. Durack DT, Street AC. Fever of unknown origin—reexamined and redefined. *Clin Top Infect Dis*. 1991;11:35–51.
5. World Health Organization. ICD-10 Version: 2015. http://apps.who.int/classifications/icd10/browse/2015/en#/R50. Updated 1 December 2014. Accessed 18 December 2014.
6. Susilawati TN, McBride WJ. Acute undifferentiated fever in Asia: a review of the literature. *Southeast Asian J Trop Med Public Health*. 2014;45(3):719–726.
7. Parra Ruiz J, Pena Monje A, Tomas Jimenez C, et al. Clinical spectrum of fever of intermediate duration in the south of Spain. *Eur J Clin Microbiol Infect Dis*. 2008;27(10):993–995.
8. Susilawati TN, McBride WJ. Undiagnosed undifferentiated fever in Far North Queensland, Australia: a retrospective study. *Int J Infect Dis*. 2014;27:59–64.
9. Kircher M, Kelso J. High-throughput DNA sequencing—concepts and limitations. *Bioessays*. 2010;32(6):524–536.
10. Loman NJ, Misra RV, Dallman TJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*. 2012;30(5):434–439.
11. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. *Clin Chem*. 2009;55(4):641–658.
12. Hayden EC. Technology: the $1,000 genome. *Nature*. 2014;507(7492):294–295.
13. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol*. 2005;13(6):278–284.
14. Gahan PB, Swaminathan R. Circulating nucleic acids in plasma and serum. Recent developments. *Ann N Y Acad Sci*. 2008;1137:1–6.

15. Palacios G, Druce J, Du L, et al. A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med*. 2008;358(10):991–998.
16. Beck J, Urnovitz HB, Riggert J, Clerici M, Schutz E. Profile of the circulating DNA in apparently healthy individuals. *Clin Chem*. 2009;55(4):730–738.
17. Ergonul O, Willke A, Azap A, Tekeli E. Revised definition of 'fever of unknown origin': limitations and opportunities. *J Infect*. 2005;50(1):1–5.
18. Efstathiou SP, Pefanis AV, Tsiakou AG, et al. Fever of unknown origin: discrimination between infectious and non-infectious causes. *Eur J Intern Med*. 2010;21(2):137–143.
19. Bleeker-Rovers CP, van der Meer JWM, Beeching NJ. Fever. *Medicine*. 2009;37(1):28–34.
20. Henker R, Kramer D, Rogers S. Fever. *AACN Clin Issues*. 1997;8(3):351–367; quiz 505–506.
21. Atkins E. Fever: its history, cause, and function. *Yale J Biol Med*. 1982;55(3–4):283–289.
22. Mackowiak PA. Concepts of fever. *Arch Intern Med*. 1998;158(17):1870–1881.
23. Leelarasamee A, Chupaprawan C, Chenchittikul M, Udompanthurat S. Etiologies of acute undifferentiated febrile illness in Thailand. *J Med Assoc Thai*. 2004;87(5):464–472.
24. Suttinont C, Losuwanaluk K, Niwatayakul K, et al. Causes of acute, undifferentiated, febrile illness in rural Thailand: results of a prospective observational study. *Ann Trop Med Parasitol*. 2006;100(4):363–370.
25. Ellis RD, Fukuda MM, McDaniel P, et al. Causes of fever in adults on the Thai-Myanmar border. *Am J Trop Med Hyg*. 2006;74(1):108–113.
26. Watt G, Jongsakul K. Acute undifferentiated fever caused by infection with Japanese encephalitis virus. *Am J Trop Med Hyg*. 2003;68(6):704–706.
27. Phuong HL, de Vries PJ, Nagelkerke N, et al. Acute undifferentiated fever in Binh Thuan province, Vietnam: imprecise clinical diagnosis and irrational pharmaco-therapy. *Trop Med Int Health*. 2006;11(6):869–879.
28. Thai KT, Phuong HL, Thanh Nga TT, et al. Clinical, epidemiological and virological features of Dengue virus infections in Vietnamese patients presenting to primary care facilities with acute undifferentiated fever. *J Infect*. 2010;60(3):229–237.

29. Nga TT, de Vries PJ, Abdoel TH, Smits HL. Brucellosis is not a major cause of febrile illness in patients at public health care facilities in Binh Thuan Province, Vietnam. *J Infect*. 2006;53(1):12–15.
30. Joshi R, Colford JM, Jr., Reingold AL, Kalantri S. Nonmalarial acute undifferentiated fever in a rural hospital in central India: diagnostic uncertainty and overtreatment with antimalarial agents. *Am J Trop Med Hyg*. 2008;78(3):393–399.
31. Chrispal A, Boorugu H, Gopinath KG, et al. Acute undifferentiated febrile illness in adult hospitalized patients: the disease spectrum and diagnostic predictors—an experience from a tertiary care hospital in South India. *Trop Doct*. 2010;40(4):230–234.
32. Thangarasu S, Natarajan P, Rajavelu P, Rajagopalan A, Seelinger Devey J. A protocol for the emergency department management of acute undifferentiated febrile illness in India. *Int J Emerg Med*. 2011;4(1):57.
33. Chandy S, Yoshimatsu K, Boorugu HK, et al. Acute febrile illness caused by hantavirus: serological and molecular evidence from India. *Trans R Soc Trop Med Hyg*. 2009;103(4):407–412.
34. Murdoch DR, Woods CW, Zimmerman MD, et al. The etiology of febrile illness in adults presenting to Patan hospital in Kathmandu, Nepal. *Am J Trop Med Hyg*. 2004;70(6):670–675.
35. Premaratna R, Rajapakse RP, Chandrasena TG, et al. Contribution of rickettsioses in Sri Lankan patients with fever who responded to empirical doxycycline treatment. *Trans R Soc Trop Med Hyg*. 2010;104(5):368–370.
36. Phongmany S, Rolain JM, Phetsouvanh R, et al. Rickettsial infections and fever, Vientiane, Laos. *Emerg Infect Dis*. 2006;12(2):256–262.
37. Gasem MH, Wagenaar JF, Goris MG, et al. Murine typhus and leptospirosis as causes of acute undifferentiated fever, Indonesia. *Emerg Infect Dis*. 2009;15(6):975–977.
38. Low JG, Ong A, Tan LK, et al. The early clinical features of dengue in adults: challenges for early clinical diagnosis. *PLoS Negl Trop Dis*. 2011;5(5):e1191.
39. Ndip LM, Labruna M, Ndip RN, Walker DH, McBride JW. Molecular and clinical evidence of *Ehrlichia chaffeensis* infection in Cameroonian patients with undifferentiated febrile illness. *Ann Trop Med Parasitol*. 2009;103(8):719–725.

40. Jentes ES, Robinson J, Johnson BW, et al. Acute arboviral infections in Guinea, West Africa, 2006. *Am J Trop Med Hyg*. 2010;83(2):388–394.
41. Prabhu M, Nicholson WL, Roche AJ, et al. Q fever, spotted fever group, and typhus group rickettsioses among hospitalized febrile patients in northern Tanzania. *Clin Infect Dis*. 2011;53(4):e8–15.
42. Forshey BM, Guevara C, Laguna-Torres VA, et al. Arboviral etiologies of acute febrile illnesses in Western South America, 2000–2007. *PLoS Negl Trop Dis*. 2010;4(8):e787.
43. Manock SR, Jacobsen KH, Bravo NBd, et al. Etiology of Acute Undifferentiated Febrile Illness in the Amazon Basin of Ecuador. *Am J Trop Med Hyg*. 2009;81(1):146–151.
44. Silva AD, Evangelista Mdo S. Syndromic surveillance: etiologic study of acute febrile illness in dengue suspicious cases with negative serology. Brazil, Federal District, 2008. *Rev Inst Med Trop Sao Paulo*. 2010;52(5):237–242.
45. Figueiredo RM, Naveca FG, Oliveira CM, et al. Co-infection of Dengue virus by serotypes 3 and 4 in patients from Amazonas, Brazil. *Rev Inst Med Trop Sao Paulo*. 2011;53(6):321–323.
46. King LA, Goirand L, Tissot-Dupont H, et al. Outbreak of Q fever, Florac, Southern France, Spring 2007. *Vector Borne Zoonotic Dis*. 2011;11(4):341–347.
47. Siikamaki HM, Kivela PS, Sipila PN, et al. Fever in travelers returning from malaria-endemic areas: don't look for malaria only. *J Travel Med*. 2011;18(4):239–244.
48. Australian Institute of Health and Welfare. National health priority areas. Digital Media and Communications Unit, Australian Institute of Health and Welfare, Australian Government. http://www.aihw.gov.au/national-health-priority-areas/. Published 2015. Accessed 14 May 2015.
49. Quinn HE, Gatton ML, Hall G, Young M, Ryan PA. Analysis of Barmah Forest virus disease activity in Queensland, Australia, 1993–2003: identification of a large, isolated outbreak of disease. *J Med Entomol*. 2005;42(5):882–890.
50. Kelly-Hope LA, Kay BH, Purdie DM, Williams GM. The risk of Ross River and Barmah Forest virus disease in Queensland: implications for New Zealand. *Aust N Z J Public Health*. 2002;26(1):69–77.

51. Hossain I, Tambyah PA, Wilder-Smith A. Ross River virus disease in a traveler to Australia. *J Travel Med*. 2009;16(6):420–423.
52. Hanna JN, Ritchie SA, Richards AR, et al. Dengue in north Queensland, 2005–2008. *Commun Dis Intell Q Rep*. 2009;33(2):198–203.
53. Hanna JN, Ritchie SA, Eisen DP, et al. An outbreak of *Plasmodium vivax* malaria in Far North Queensland, 2002. *Med J Aust*. 2004;180(1):24–28.
54. Hanna JN, Brookes DL, Ritchie SA, van den Hurk AF, Loewenthal MR. Malaria and its implications for public health in Far North Queensland: a prospective study. *Aust N Z J Public Health*. 1998;22(2):196–199.
55. Mannestal Johansson C, McBride WJ, Engstrom K, Mills J. Who brings dengue into North Queensland? A descriptive, exploratory study. *Aust J Rural Health*. 2012;20(3):150–155.
56. McBride WJ. Infections in travellers arriving from Australia. *Trans R Soc Trop Med Hyg*. 2008;102(4):312–313.
57. McBride WJ, Hanson JP, Miller R, Wenck D. Severe spotted fever group rickettsiosis, Australia. *Emerg Infect Dis*. 2007;13(11):1742–1744.
58. De Paula SO, Fonseca BA. Dengue: a review of the laboratory tests a clinician must know to achieve a correct diagnosis. *Braz J Infect Dis*. 2004;8(6):390–398.
59. Kasper MR, Blair PJ, Touch S, et al. Infectious etiologies of acute febrile illness among patients seeking health care in south-central Cambodia. *Am J Trop Med Hyg*. 2012;86(2):246–253.
60. Brown GW, Shirai A, Jegathesan M, et al. Febrile illness in Malaysia—an analysis of 1,629 hospitalized patients. *Am J Trop Med Hyg*. 1984;33(2):311–315.
61. Anderson KE, Joseph SW, Nasution R, et al. Febrile illnesses resulting in hospital admission: a bacteriological and serological study in Jakarta, Indonesia. *Am J Trop Med Hyg*. 1976;25(1):116–121.
62. Wollman AJ, Nudd R, Hedlund EG, Leake MC. From Animaculum to single molecules: 300 years of the light microscope. *Open Biol*. 2015;5(4):150019.
63. Burbelo PD, Ching KH, Bush ER, Han BL, Iadarola MJ. Antibody-profiling technologies for studying humoral responses to infectious agents. *Expert Rev Vaccines*. 2010;9(6):567–578.

64. Vigil A, Davies DH, Felgner PL. Defining the humoral immune response to infectious agents using high-density protein microarrays. *Future Microbiol*. 2010;5(2):241–251.
65. Huse SM, Ye Y, Zhou Y, Fodor AA. A core human microbiome as viewed through 16S rRNA sequence clusters. *PLoS One*. 2012;7(6):e34242.
66. Li K, Bihan M, Yooseph S, Methe BA. Analyses of the microbial diversity across the human microbiome. *PLoS One*. 2012;7(6):e32118.
67. Koch R. The aetiology of tuberculosis (translation of Die Aetiologie der Tuberculose [1882]). In: Clark D, editor. Source Book of Medical History. New York: Dover Publications, Inc.; 1942. p. 392–406.
68. Fredricks DN, Relman DA. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin Microbiol Rev*. 1996;9(1):18–33.
69. Manso AS, Chai MH, Atack JM, et al. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat Commun*. 2014;5:5055.
70. Pandrea I, Silvestri G, Apetrei C. AIDS in African nonhuman primate hosts of SIVs: a new paradigm of SIV infection. *Curr HIV Res*. 2009;7(1):57–72.
71. Haigwood NL. Update on animal models for HIV research. *Eur J Immunol*. 2009;39(8):1994–1999.
72. Evans AS. Limitation of Koch's postulates. *Lancet*. 1977;2(8051):1277–1278.
73. Rivers TM. Viruses and Koch's Postulates. *J Bacteriol*. 1937;33(1):1–12.
74. Huebner RJ. Criteria for etiologic association of prevalent viruses with prevalent diseases; the virologist's dilemma. *Ann N Y Acad Sci*. 1957;67(8):430–438.
75. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295–300.
76. Evans AS. Causation and disease: the Henle-Koch postulates revisited. *Yale J Biol Med*. 1976;49(2):175–195.
77. Henle G, Henle W, Diehl V. Relation of Burkitt's tumor-associated herpes-ytpe virus to infectious mononucleosis. *Proc Natl Acad Sci U S A*. 1968;59(1):94–101.
78. Walker L, Levine H, Jucker M. Koch's postulates and infectious proteins. *Acta neuropathologica*. 2006;112(1):1–4.

79. Lipkin WI. Microbe hunting. *Microbiol Mol Biol Rev*. 2010;74(3):363–377.
80. Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol*. 2012;2(1):63–77.
81. Fenner F, World Health Organisation. *Smallpox and its eradication*. Geneva: World Health Organization; 1988.
82. Lustig A, Levine AJ. One hundred years of virology. *J Virol*. 1992;66(8):4629–4631.
83. Bos L. Beijerinck's work on tobacco mosaic virus: historical context and legacy. *Philos Trans R Soc Lond B Biol Sci*. 1999;354(1383):675–685.
84. Reed W, Carroll J, Agramonte A. The etiology of yellow fever: an additional note. *JAMA*. 1901;36:431–440.
85. Woolhouse ME, Howey R, Gaunt E, et al. Temporal trends in the discovery of human viruses. *Proc Biol Sci*. 2008;275(1647):2111–2115.
86. Goodpasture EW, Woodruff AM, Buddingh GJ. The cultivation of vaccine and other viruses in the chorioallantoic membrane of chick embryos. *Science*. 1931;74(1919):371–372.
87. Hao W, Bernard K, Patel N, et al. Infection and propagation of human rhinovirus C in human airway epithelial cells. *J Virol*. 2012;86(24):13524–13532.
88. Wilson GK, Stamataki Z. In vitro systems for the study of hepatitis C virus infection. *Int J Hepatol*. 2012;2012:292591.
89. Williams JV. Deja vu all over again: Koch's postulates and virology in the 21st century. *J Infect Dis*. 2010;201(11):1611-1614.
90. Delwart EL. Viral metagenomics. *Rev Med Virol*. 2007;17(2):115–131.
91. Mullis KB. The unusual origin of the polymerase chain reaction. *Sci Am*. 1990;262(4):56–61, 4–5.
92. Staples JE, Monath TP. Yellow Fever: 100 years of discovery. *JAMA*. 2008;300(8):960–962.
93. Frierson JG. The yellow fever vaccine: a history. *Yale J Biol Med*. 2010;83(2):77–85.
94. Tiroumourougane SV, Raghava P, Srinivasan S. Japanese viral encephalitis. *Postgrad Med J*. 2002;78(918):205–215.

95. Smithburn KC, Hughes TP, Burke AW, Paul JH. A neurotropic virus isolated from the blood of a native of Uganda. *Am J Trop Med Hyg* (0002–9637). 1940;1(4):471–492.
96. Harley D, Sleigh A, Ritchie S. Ross River virus transmission, infection, and disease: a cross-disciplinary review. Clin Microbiol Rev. 2001;14(4):909-32.
97. Feinstone SM, Kapikian AZ, Purcell RH, Alter HJ, Holland PV. Transfusion-associated hepatitis not due to viral hepatitis type A or B. *New Engl J Med*. 1975;292(15):767–770.
98. Alter HJ, Holland PV, Morrow AG, Purcell RH, Feinstone SM, Moritsugu Y. Clinical and serological analysis of transfusion-associated hepatitis. *Lancet*. 1975;2(7940):838–841.
99. Alter MJ, Gerety RJ, Smallwood LA, et al. Sporadic non-A, non-B hepatitis: frequency and epidemiology in an urban U.S. population. *J Infect Dis*. 1982;145(6):886–893.
100. Bukh J. Animal models for the study of hepatitis C virus infection and related liver disease. *Gastroenterology*. 2012;142(6):1279–1287 e3.
101. Houghton M. Discovery of the hepatitis C virus. *Liver Int*. 2009;29 Suppl 1:82–88.
102. Groneberg DA, Zhang L, Welte T, et al. Severe acute respiratory syndrome: global initiatives for disease diagnosis. *QJM*. 2003;96(11):845–852.
103. Drosten C, Gunther S, Preiser W, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med*. 2003;348(20):1967–1976.
104. Ksiazek TG, Erdman D, Goldsmith CS, et al. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med*. 2003;348(20):1953–1966.
105. Palacios G, Quan PL, Jabado OJ, et al. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg Infect Dis*. 2007;13(1):73–81.
106. Chiu CY. Viral pathogen discovery. *Curr Opin Microbiol*. 2013;16(4):468–478.
107. Hoffmann B, Tappe D, Hoper D, et al. A Variegated squirrel bornavirus associated with fatal human encephalitis. *N Engl J Med*. 2015;373(2):154–162.
108. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463–5467.

109. Smith LM, Sanders JZ, Kaiser RJ, et al. Fluorescence detection in automated DNA sequence analysis. *Nature*. 1986;321(6071):674–679.
110. Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods*. 2008;5(1):16–18.
111. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. *J Genet Genomics*. 2011;38(3):95–109.
112. Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinform*. 2009;10(4):354–366.
113. MacLean D, Jones JDG, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Micro*. 2009;7(4):287–296.
114. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet*. 2010;11(1):31–46.
115. Liu L, Li Y, Li S, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol*. 2012:251364.
116. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*. 2008;9:387–402.
117. Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol*. 2012;10(9):618–630.
118. Desai AN, Jere A. Next-generation sequencing: ready for the clinics? *Clin Genet*. 2012;81(6):503–510.
119. Shendure J, Lieberman Aiden E. The expanding scope of DNA sequencing. *Nat Biotechnol*. 2012;30(11):1084–1094.
120. Munroe DJ, Harris TJ. Third-generation sequencing fireworks at Marco Island. *Nat Biotechnol*. 2010;28(5):426–428.
121. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 1998;5:245–249.
122. Baker BJ, Dick GJ. Omic approaches in microbial ecology: Charting the unknown. *Microbe*. 2013;8:353–360.
123. Breitbart M, Salamon P, Andresen B, et al. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A*. 2002;99(22):14250–14255.

124. Breitbart M, Hewson I, Felts B, et al. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol*. 2003;185(20):6220–6223.

125. Zhang T, Breitbart M, Lee WH, et al. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol*. 2006;4(1):e3.

126. Poza M, Gayoso C, Gomez MJ, et al. Exploring bacterial diversity in hospital environments by GS-FLX Titanium pyrosequencing. *PLoS One*. 2012;7(8):e44105.

127. Xu B, Liu L, Huang X, et al. Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. PLoS Pathog. 2011;7(11):e1002369.

128. McMullan LK, Folk SM, Kelly AJ, et al. A new phlebovirus associated with severe febrile illness in Missouri. *N Engl J Med*. 2012;367(9):834–841.

129. Grard G, Fair JN, Lee D, et al. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS Pathog*. 2012;8(9):e1002924.

130. Briese T, Paweska JT, McMullan LK, et al. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever–associated arenavirus from southern Africa. *PLoS Pathog*. 2009;5(5):e1000455.

131. Hang J, Forshey BM, Kochel TJ, et al. Random amplification and pyrosequencing for identification of novel viral genome sequences. *J Biomol Tech*. 2012;23(1):4–10.

132. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis*. 2012;6(2):e1485.

133. Australian Bureau of Statistics. 3218.0—Regional Population Growth, Australia, 2010–11. http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3218.02010-11. Updated 30 July 2012. Accessed 29 August 2012.

134. Queensland Government. About us. Cairns and Hinterland Hospital and Health Service. Queensland Health. http://www.health.qld.gov.au/cairns_hinterland/html/about_us.asp. Updated 21 June 2012. Accessed 24 July 2012.

135. Australian Government. National Statement on Ethical Conduct in Human Research. http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf.

Published 2007. Accessed 27 May 2015.

136. Office of Health and Medical Research, Queensland Health. Site Specific Assessment Form Guidance. http://www.health.qld.gov.au/ohmr/documents/regu/ssa_guide_v1.pdf. Published 2010. Accessed 27 May 2015.
137. Nyika A. Ethical and practical challenges surrounding genetic and genomic research in developing countries. *Acta Trop*. 2009;112 Suppl 1:S21–31.
138. Tindana P, Bull S, Amenga-Etego L, et al. Seeking consent to genetic and genomic research in a rural Ghanaian setting: a qualitative study of the MalariaGEN experience. BMC medical ethics. 2012;13:15.
139. Manasco PK. Ethical and legal aspects of applied genomic technologies: practical solutions. *Curr Mol Med*. 2005;5(1):23–28.
140. Goto M, Koyama H, Takahashi O, Fukui T. A retrospective review of 226 hospitalized patients with fever. *Intern Med*. 2007;46(1):17–22.
141. Australian Government Bureau of Meteorology. Monthly rainfall, Cairns Aero. http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=139&p_display_type=dataFile&p_stn_num=031011. Accessed 14 February 2013.
142. McGready R, Ashley EA, Wuthiekanun V, et al. Arthropod borne disease: the leading cause of fever in pregnancy on the Thai–Burmese border. *PLoS Negl Trop Dis*. 2010;4(11):e888.
143. Blacksell SD, Sharma NP, Phumratanaprapin W, et al. Serological and blood culture investigations of Nepalese fever patients. *Trans R Soc Trop Med Hyg*. 2007;101(7):686–690.
144. Espinosa N, Canas E, Bernabeu-Wittel M, Martin A, Viciana P, Pachon J. The changing etiology of fever of intermediate duration. *Enferm Infecc Microbiol Clin*. 2010;28(7):416–420.
145. Xue X, Teare MD, Holen I, Zhu YM, Woll PJ. Optimizing the yield and utility of circulating cell-free DNA from plasma and serum. *Clin Chim Acta*. 2009;404(2):100–104.
146. Elshimali YI, Khaddour H, Sarkissyan M, Wu Y, Vadgama JV. The clinical utilization of circulating cell free DNA (CCFDNA) in blood of cancer patients. *Int J Mol Sci*. 2013;14(9):18925–18958.

147. Huang ZH, Li LH, Hua D. Quantitative analysis of plasma circulating DNA at diagnosis and during follow-up of breast cancer patients. *Cancer Lett*. 2006;243(1):64–70.
148. Sozzi G, Conte D, Leon M, et al. Quantification of free circulating DNA as a diagnostic marker in lung cancer. *J Clin Oncol*. 2003;21(21):3902–3908.
149. Zhang P, Ren J, Shen Z. A new quantitative method for circulating DNA level in human serum by capillary zone electrophoresis with laser-induced fluorescence detection. *Electrophoresis*. 2004;25(12):1823–1828.
150. Ha TT, Huy NT, Murao LA, et al. Elevated levels of cell-free circulating DNA in patients with acute dengue virus infection. *PLoS One*. 2011;6(10):e25969.
151. Machado AS, Da Silva Robaina MC, Magalhaes De Rezende LM, et al. Circulating cell-free and Epstein-Barr virus DNA in pediatric B-non-Hodgkin lymphomas. *Leuk Lymphoma*. 2010;51(6):1020–1027.
152. Catarino R, Ferreira MM, Rodrigues H, et al. Quantification of free circulating tumor DNA as a diagnostic marker for breast cancer. *DNA Cell Biol*. 2008;27(8):415–421.
153. Lo YM, Chiu RW. Next-generation sequencing of plasma/serum DNA: an emerging research and molecular diagnostic tool. *Clin Chem*. 2009;55(4):607–608.
154. Board RE, Williams VS, Knight L, et al. Isolation and extraction of circulating tumor DNA from patients with small cell lung cancer. *Ann N Y Acad Sci*. 2008;1137:98–107.
155. Xia P, Radpour R, Zachariah R, et al. Simultaneous quantitative assessment of circulating cell-free mitochondrial and nuclear DNA by multiplex real-time PCR. *Genet Mol Biol*. 2009;32(1):20–24.
156. Ong ME, Chan YH, Lim CS. Observational study to determine factors associated with blood sample haemolysis in the emergency department. *Ann Acad Med Singapore*. 2008;37(9):745–748.
157. El-Hefnawy T, Raja S, Kelly L, et al. Characterization of amplifiable, circulating RNA in plasma and its potential as a tool for cancer diagnostics. *Clin Chem*. 2004;50(3):564–573.
158. Rengarajan K, Cristol SM, Mehta M, Nickerson JM. Quantifying DNA concentrations using fluorometry: a comparison of fluorophores. *Mol Vis*. 2002;8:416–421.

159. Szpechcinski A, Struniawska R, Zaleska J, et al. Evaluation of fluorescence-based methods for total vs. amplifiable DNA quantification in plasma of lung cancer patients. *J Physiol Pharmacol*. 2008;59 Suppl 6:675–681.

160. Sambrook J, Russell DW. Molecular cloning: a laboratory manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory; 2001.

161. Wilfinger WW, Mackey K, Chomczynski P. Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *BioTechniques*. 1997;22(3):474–476, 478–481.

162. Bolger R, Lenoch F, Allen E, Meiklejohn B, Burke T. Fluorescent dye assay for detection of DNA in recombinant protein products. *BioTechniques*. 1997;23(3):532–537.

163. Leggate J, Allain R, Isaac L, Blais BW. Microplate fluorescence assay for the quantification of double stranded DNA using SYBR Green I dye. *Biotechnol Lett*. 2006;28(19):1587–1594.

164. Molecular Probes Inc. Molecular Probes Handbook. ThermoFisher Scientific website. https://www.thermofisher.com/au/en/home/references/molecular-probes-the-handbook.html. Published 2010. Accessed 3 July 2015.

165. Klein RC. Ultraviolet light hazards from transilluminators. *Health Physics*. 2000;78(5 Suppl):S48–50.

166. Cariello NF, Keohavong P, Sanderson BJ, Thilly WG. DNA damage produced by ethidium bromide staining and exposure to ultraviolet light. *Nucleic Acids Res*. 1988;16(9):4157.

167. Singer VL, Jones LJ, Yue ST, Haugland RP. Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation. *Anal Biochem*. 1997;249(2):228–238.

168. Vitzthum F, Geiger G, Bisswanger H, Brunner H, Bernhagen J. A quantitative fluorescence-based microplate assay for the determination of double-stranded DNA using SYBR Green I and a standard ultraviolet transilluminator gel imaging system. *Anal Biochem*. 1999;276(1):59–64.

169. Horlitz M, Lucas A, Sprenger-Haussels M. Optimized quantification of fragmented, free circulating DNA in human blood plasma using a calibrated duplex real-time PCR. *PLoS One*. 2009;4(9):e7207.

170. Gal S, Fidler C, Lo YM, et al. Quantitation of circulating DNA in the serum of breast cancer patients by real-time PCR. Br J Cancer. 2004;90(6):1211–1215.

171. Cunha BA. Fever of unknown origin: focused diagnostic approach based on clinical clues from the history, physical examination, and laboratory tests. *Infect Dis Clin North Am*. 2007;21(4):1137–1187, xi.

172. Abrahamsen SK, Haugen CN, Rupali P, et al. Fever in the tropics: aetiology and case-fatality—a prospective observational study in a tertiary care hospital in South India. *BMC Infect Dis*. 2013;13:355.

173. Sint D, Raso L, Traugott M. Advances in multiplex PCR: balancing primer efficiencies and improving detection success. *Methods Ecol Evol*. 2012;3(5):898–905.

174. Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE. Multiplex PCR: optimization and application in diagnostic virology. *Clin Microbiol Rev*. 2000;13(4):559–570.

175. Hsu CC, Tokarz R, Briese T, Tsai HC, Quan PL, Lipkin WI. Use of staged molecular analysis to determine causes of unexplained central nervous system infections. *Emerg Infect Dis*. 2013;19(9):1470–1477.

176. McLoughlin KS. Microarrays for pathogen detection and analysis. *Brief Funct Genomics*. 2011;10(6):342–353.

177. Cao B, Wang S, Tian Z, Hu P, Feng L, Wang L. DNA microarray characterization of pathogens associated with sexually transmitted diseases. *PloS One*. 2015; 10(7):e0133927.

178. Calistri A, Palù G. Unbiased next-generation sequencing and new pathogen discovery: Undeniable advantages and still-existing drawbacks. *Clin Infect Dis*. 2015;60(6):889–891.

179. Perlejewski K, Popiel M, Laskus T, et al. Next-generation sequencing (NGS) in the identification of encephalitis-causing viruses: Unexpected detection of human herpesvirus 1 while searching for RNA pathogens. *J Virol Methods*. 2015;226:1–6.

180. Sigma-Aldrich. *Product Information SeqPlexTM Enhanced DNA Amplification Kit*. St Louis, MO: Sigma-Aldrich; 2014.

181. Sigma-Aldrich. *Product Information SeqPlexTM RNA Amplification Kit 2014*. St Louis, MO: Sigma-Aldrich; 2014.

182. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 2014;15(3):R46.

183. Zaharia M, Bolosky WJ, Curtis K, et al. Faster and more accurate sequence alignment with SNAP. *arXiv*. 2011;arXiv(1111.5572v1).

184. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*. 2011;12(1):385.

185. Andrews S. FastQC: a quality control tool for high-throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Published 2010. Accessed 3 November 2015.

186. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–410.

187. CLC Bio. *CLC Genomics Workbench User Manual*. Denmark; 2015. http://www.clcbio.com/files/usermanuals/CLC_Genomics_Workbench_User_Manual.pdf.

188. Parkinson NJ, Maslau S, Ferneyhough B, et al. Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res*. 2012;22(1):125–133.

189. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. 2010;38(6):1767–1771.

190. Hodges K, Gill R. Infectious diarrhoea: Cellular and molecular mechanisms. *Gut Microbes*. 2010;1(1):4–21.

191. Gorschlüter M, Hahn C, Ziske C, et al. Low frequency of enteric infections by salmonella, shigella, yersinia and campylobacter in patients with acute leukemia. *Infection*. 2002;30(1):22–25.

192. Tabak F, Murtezaoglu A, Tabak O, et al. Clinical features and etiology of adult patients with fever and rash. *Ann Dermatol*. 2012;24(4):420–425.

193. Kang JH. Febrile illness with skin rashes. *Infect Chemother*. 2015;47(3):155–166.

194. Cerkovnik P, Perhavec A, Zgajnar J, Novakovic S. Optimization of an RNA isolation procedure from plasma samples. *Int J Mol Med*. 2007;20(3):293–300.

195. van der Vaart M, Pretorius PJ. Circulating DNA. Its origin and fluctuation. *Ann N Y Acad Sci*. 2008;1137:18–26.

196. Fleischhacker M, Schmidt B, Weickmann S, et al. Methods for isolation of cell-free plasma DNA strongly affect DNA yield. *Clin Chim Acta*. 2011;412(23-24):2085–2088.

197. Kirsch C, Weickmann S, Schmidt B, Fleischhacker M. An improved method for the isolation of free-circulating plasma DNA and cell-free DNA from other body fluids. *Ann N Y Acad Sci*. 2008;1137:135–139.

198. Tricou V, Minh NN, Farrar J, Tran HT, Simmons CP. Kinetics of viremia and NS1 antigenemia are shaped by immune status and virus serotype in adults with dengue. *PLoS Negl Trop Dis*. 2011;5(9):e1309.

199. Vaughn DW, Green S, Kalayanarooj S, et al. Dengue viremia titer, antibody response pattern, and virus serotype correlate with disease severity. *J Infect Dis*. 2000;181(1):2–9.

200. Gubler DJ, Suharyono W, Tan R, Abidin M, Sie A. Viraemia in patients with naturally acquired dengue infection. *Bull World Health Organ*. 1981;59(4):623–630.

201. Levett PN, Morey RE, Galloway RL, Steigerwalt AG. Leptospira broomii sp. nov., isolated from humans with leptospirosis. *Int J Syst Evol Microbiol*. 2006;56(Pt 3):671–673.

202. Stoddard RA, Gee JE, Wilkins PP, McCaustland K, Hoffmaster AR. Detection of pathogenic Leptospira spp. through TaqMan polymerase chain reaction targeting the LipL32 gene. *Diagn Microbiol Infect Dis*. 2009;64(3):247–255.

203. Naccache SN, Federman S, Veeraraghavan N, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res*. 2014;24(7):1180–1192.

204. Hugon P, Dufour JC, Colson P, Fournier PE, Sallah K, Raoult D. A comprehensive repertoire of prokaryotic species identified in human beings. *Lancet Infect Dis*. 2015;15(10):1211–1219.

205. Vartoukian SR, Palmer RM, Wade WG. Strategies for culture of 'unculturable' bacteria. *FEMS Microbiol Lett*. 2010;309(1):1–7.

206. Nikkari S, McLaughlin IJ, Bi W, Dodge DE, Relman DA. Does blood of healthy subjects contain bacterial ribosomal DNA? *J Clin Microbiol*. 2001;39(5):1956–1959.

207. Goldstein SJ. Achromobacter xylosoxidans bacteremia: Report of four cases and review of the literature. *Clin Infect Dis*. 1996;23(3):569–576.

208. Igra-Siegman Y, Chmel H, Cobbs C. Clinical and laboratory characteristics of Achromobacter xylosoxidans infection. *J Clin Microbiol*. 1980;11(2):141–145.

209. Claassen SL, Reese JM, Mysliwiec V, Mahlen SD. Achromobacter xylosoxidans infection presenting as a pulmonary nodule mimicking cancer. *J Clin Microbiol*. 2011;49(7):2751–2754.

210. López-Pérez M, Gonzaga A, Martin-Cuadrado A-B, et al. Genomes of surface isolates of Alteromonas macleodii: the life of a widespread marine opportunistic copiotroph. *Scientific Reports*. 2012;2:696.

211. Ivars-Martinez E, Martin-Cuadrado AB, D'Auria G, et al. Comparative genomics of two ecotypes of the marine planktonic copiotroph Alteromonas macleodii suggests alternative lifestyles associated with different kinds of particulate organic matter. *Isme J*. 2008;2(12):1194–1212.

212. Klumpp J, Fouts DE, Sozhamannan S. Next generation sequencing technologies and the changing landscape of phage genomics. *Bacteriophage*. 2012;2(3):190–199.

213. Sanger F, Air GM, Barrell BG, et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*. 1977;265(5596):687–695.

214. Lei H, Li T, Hung GC, Li B, Tsai S, Lo SC. Identification and characterization of EBV genomes in spontaneously immortalized human peripheral blood B lymphocytes by NGS technology. *BMC Genomics*. 2013;14:804.

215. Maslow JN, Mulligan ME, Arbeit RD. Recurrent Escherichia coli bacteremia. *J Clin Microbiol*. 1994;32(3):710–714.

216. Peleg AY, Hooper DC. Hospital-acquired infections due to Gram-negative bacteria. *N Engl J Med*. 2010;362(19):1804–1813.

217. Hislop AD. Early virological and immunological events in Epstein-Barr virus infection. *Curr Op Virol*. 2015;15:75–79.

218. Krueger GR, Huetter ML, Rojo J, Romero M, Cruz-Ortiz H. Human herpesviruses HHV-4 (EBV) and HHV-6 in Hodgkin's and Kikuchi's diseases and their relation to proliferation and apoptosis. *Anticancer Res*. 2001;21(3C):2155–2161.

219. Suarez F, Lecuit M. Infection-associated non-Hodgkin lymphomas. *Clin Microbiol Infect*. 2015;21(11):991–997.

220. Chua ML, Wee JT, Hui EP, Chan AT. Nasopharyngeal carcinoma. *Lancet*. 2015;S0140–6736(15):00055-0.

221. Rigante D, Esposito S. Infections and systemic lupus erythematosus: binding or sparring partners? *Int J Mol Sci*. 2015;16(8):17331–17343.

222. McKay KA, Kwan V, Duggan T, Tremlett H. Risk factors associated with the onset of relapsing-remitting and primary progressive multiple sclerosis: a systematic review. *Biomed Res Int*. 2015:817238.

223. Walling DM, Shebib N, Weaver SC, Nichols CM, Flaitz CM, Webster-Cyriaque J. The molecular epidemiology and evolution of Epstein-Barr virus: sequence variation and genetic recombination in the latent membrane protein-1 gene. *J Infect Dis*. 1999;179(4):763–774.

224. Strong MJ, Xu GR, Morici L, et al. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog*. 2014;10(11):e1004437.

225. Vahdani P, Yaghoubi T, Aminzadeh Z. Hospital acquired antibiotic-resistant Acinetobacter baumannii infections in a 400-bed hospital in Tehran, Iran. *Int J Prev Med*. 2011;2(3):127–130.

226. Prashanth K BS. Nosocomial infections due to Acinetobacter species: Clinical findings, risk and prognostic factors. *Indian J Med Microbiol*. 2006;24(1):39–44.

227. Bennett HM. Genome watch split reality for novel tick virus. Nat Rev Microbiol. 2014;12(7):464.

228. Qin X-C, Shi M, Tian J-H, et al. A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors. *Proc Nat Acad Sci*. 2014;111(18):6744–6749.

229. Stephany D, Buffet P, Rolain JM, Raoult D, Consigny PH. Rickettsia africae infection in man after travel to Ethiopia. *Emerg Infect Dis*. 2009;15(11):1867–1869.

230. Raoult D, Fournier PE, Fenollar F, et al. Rickettsia africae, a tick-borne pathogen in travelers to sub-Saharan Africa. *N Engl J Med*. 2001;344(20):1504–1510.

231. Jones MS, Kapoor A, Lukashov VV, Simmonds P, Hecht F, Delwart E. New DNA viruses identified in patients with acute viral infection syndrome. *J Virol*. 2005;79(13):8230–8236.

232. Thom K, Morrison C, Lewis JCM, Simmonds P. Distribution of TT virus (TTV), TTV-like minivirus, and related viruses in humans and nonhuman primates. *Virology*. 2003;306(2):324–333.

233. Leary TP, Erker JC, Chalmers ML, Desai SM, Mushahwar IK. Improved detection systems for TT virus reveal high prevalence in humans, non-human primates and farm animals. *J Gen Virol*. 1999;80(Pt 8):2115–2120.

234. Desai SM, Muerhoff AS, Leary TP, et al. Prevalence of TT virus infection in US blood donors and populations at risk for acquiring parenterally transmitted viruses. *J Infect Dis*. 1999;179(5):1242–1244.

235. Ninomiya M, Takahashi M, Hoshino Y, Ichiyama K, Simmonds P, Okamoto H. Analysis of the entire genomes of torque teno midi virus variants in chimpanzees: infrequent cross-species infection between humans and chimpanzees. *J Gen Virol*. 2009;90(Pt 2):347–358.

236. Burian Z, Szabo H, Szekely G, et al. Detection and follow-up of torque teno midi virus ('small anelloviruses') in nasopharyngeal aspirates and three other human body fluids in children. *Arch Virol*. 2011;156(9):1537–1541.

237. Li SK, Leung RKK, Guo HX, et al. Detection and identification of plasma bacterial and viral elements in HIV/AIDS patients in comparison to healthy adults. *Clin Microbiol Infect*. 2012;18(11):1126–1133.

238. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 2004;68(4):669–685.

239. Saunders CJ, Miller NA, Soden SE, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med*. 2012;4(154):154ra35.

240. Meldrum C, Doyle MA, Tothill RW. Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin Biochem Rev*. 2011;32(4):177–195.

241. Qiagen. Repli-g® Mini/Midi Handbook. Qiagen website. https://www.qiagen.com/au/resources/resourcedetail?id=843654e0-2ccb-474b-b4b8-8744453ed5cb&lang=en. Published July, 2011. Accessed November 10, 2015.

242. Qiagen. QuantiTect® Whole Transcriptome Handbook. Qiagen website. https://www.qiagen.com/au/resources/resourcedetail?id=6910f167-e4ae-43a2-b87d-26a74c55a75c&lang=en. Published July, 2011. Accessed November 10, 2015.

243. NuGEN. Ovation® RNA-Seq System V2 Data Sheet. http://www.nugen.com/sites/default/files/M01254_v4%20-

%20Data%20Sheet,%20Ovation%20RNA-Seq%20V2.pdf. Published 2011. Accessed 10 November 2015.

244. Tariq MA, Kim HJ, Jejelowo O, Pourmand N. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res*. 2011;39(18):e120.

245. Pinard R, de Winter A, Sarkis GJ, et al. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*. 2006;7:216.

246. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*. 2012;28(1):125–126.

247. Naccache SN, Greninger AL, Lee D, et al. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol*. 2013;87(22):11966–11977.

248. Evans GE, Murdoch DR, Anderson TP, Potter HC, George PM, Chambers ST. Contamination of Qiagen DNA extraction kits with Legionella DNA. *J Clin Microbiol*. 2003;41(7):3452–3453.

249. Tilburg JJ, Nabuurs-Franssen MH, van Hannen EJ, Horrevorts AM, Melchers WJ, Klaassen CH. Contamination of commercial PCR master mix with DNA from Coxiella burnetii. *J Clin Microbiol*. 2010;48(12):4634–4635.

250. Bzhalava D, Johansson H, Ekstrom J, et al. Unbiased approach for virus detection in skin lesions. *PLoS One*. 2013;8(6):e65953.

251. Oyola SO, Gu Y, Manske M, et al. Efficient depletion of host DNA contamination in malaria clinical sequencing. *J Clin Microbiol*. 2013;51(3):745–751.

252. Law J, Jovel J, Patterson J, et al. Identification of hepatotropic viruses from plasma using deep sequencing: a next generation diagnostic tool. *PLoS One*. 2013;8(4):e60595.

253. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci U S A*. 2001;98(20):11609–11614.

254. Fournier PE, Dubourg G, Raoult D. Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Med*. 2014;6(11):114.

255. Greninger AL, Chen EC, Sittler T, et al. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS One*. 2010;5(10):e13381.

256. Cheval J, Sauvage V, Frangeul L, et al. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. *J Clin Microbiol*. 2011;49(9):3268–3275.

257. Pinzani P, Salvianti F, Zaccara S, et al. Circulating cell-free DNA in plasma of melanoma patients: qualitative and quantitative considerations. *Clin Chim Acta*. 2011;412(23–24):2141–2145.

258. Spornraft M, Kirchner B, Haase B, Benes V, Pfaffl MW, Riedmaier I. Optimization of extraction of circulating RNAs from plasma—enabling small RNA sequencing. *PLoS One*. 2014;9(9):e107259.

259. Laver T, Harrison J, O'Neill PA, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. *BQD*. 2015;3:1–8.

260. Ashton PM, Nair S, Dallman T, et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol*. 2015;33(3):296–300.

261. Kilianski A, Haas JL, Corriveau EJ, et al. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *Gigascience*. 2015;4:12.

262. Mande SS, Mohammed MH, Ghosh TS. Classification of metagenomic sequences: methods and challenges. *Brief Bioinform*. 2012;13(6):669–681.

263. Bazinet AL, Cummings MP. A comparative evaluation of sequence classification programs. *BMC Bioinformatics*. 2012;13:92.

264. Dodson S, Ricke DO, Kepner J, Chiu N, Shcherbina A. Rapid sequence identification of potential pathogens using techniques from sparse linear algebra. 2015. https://www.researchgate.net/publication/271218635_Rapid_Sequence_Identification_of_Potential_Pathogens_Using_Techniques_from_Sparse_Linear_Algebra. Accessed 18 November 2015.

265. Greninger AL, Naccache SN, Federman S, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med*. 2015;7(1):99.

266. Kingsmore SF, Petrikin J, Willig LK, Guest E. Emergency medical genomes: a breakthrough application of precision medicine. *Genome Med*. 2015;7(1):82.

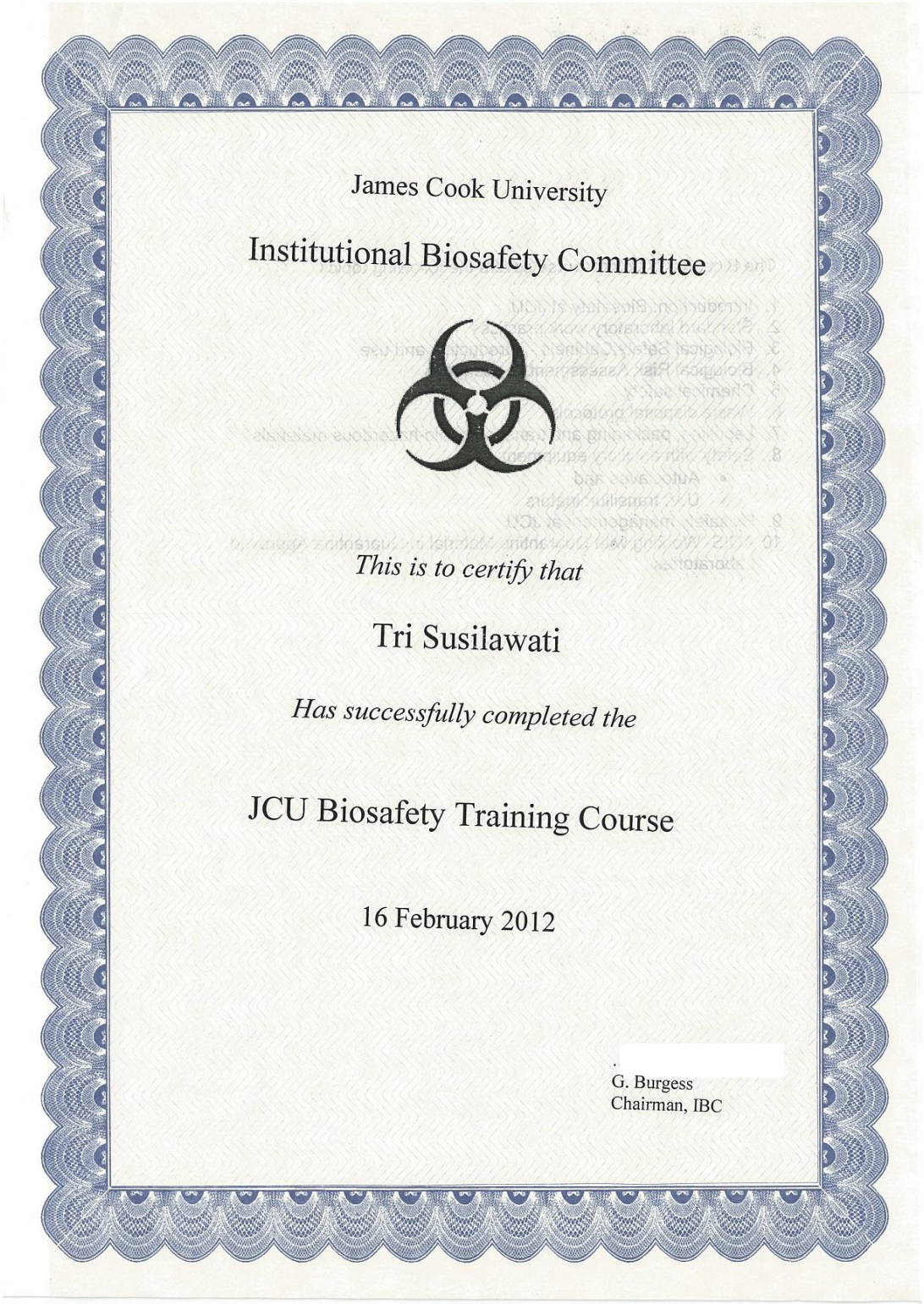
Appendix A. External Approval from Queensland Government for Study 1

This administrative form has been removed

Appendix B. JCU Ethics Approval for Study 1

This administrative form has been removed

Appendix C. Certificate of Biosafety Training Course



James Cook University

Institutional Biosafety Committee

This is to certify that

Tri Susilawati

Has successfully completed the

JCU Biosafety Training Course

16 February 2012

G. Burgess
Chairman, IBC

Appendix D. JCU Ethics Approval for Study 2

This administrative form has been removed

Appendix E. External Approval for Conducting Study 3 at Cairns (Base) Hospital

This administrative form has been removed

Appendix F. JCU Ethics Approval for Study 3

This administrative form has been removed

Appendix G. External Approval for Conducting Study 3 at Cairns Private Hospital

This administrative form has been removed


This administrative form
has been removed

Appendix H. Participant Information Sheet for Study 2

JAMES COOK UNIVERSITY
AUSTRALIA

Cairns Campus
PO Box 6811 Cairns Qld 4870 Australia
Telephone (07) 4042 1111
International +61 7 4042 1111
www.jcu.edu.au

PARTICIPANT INFORMATION SHEET

Project Title: Quantitative analysis of nucleic acid in serum and plasma

This participant information sheet is 2 pages long; please make sure you have all the pages.

Who are we?
We are a study group investigating the variability of amount of nucleic acid in various blood specimens.

This project will be conducted as part of a PhD research project by:

Dr Tri Susilawati (MD, MMed)
PhD candidate, James Cook University
E: tri.susilawati@jcu.edu.au
Tel: 0412036695

Under supervision of the following James Cook University staff:

Prof. John McBride (MBBS, DTM&H, FRACP, FRCPA, PhD)
E: john.mcbride@jcu.edu.au
Tel: (07) 42266530

Prof. Alex Loukas (BSc Hons, PhD)
E: alex.loukas@jcu.edu.au
Tel: (07) 4042 1608

Dr. Jason Mulvenna (PhD)
E: jason.mulvenna@jcu.edu.au
Tel: (07) 40421866

Why are we doing this?
It is commonly taught that a human's DNA is found within the nucleus of a cell, and RNA is confined within a cell. We now know that even in samples of blood from which cells have been removed will contain some DNA and RNA, but the amount has never been accurately quantified. When someone has an infection, a small amount of the pathogen's nucleic acid can be found in blood circulation together with human nucleic acid.

The development of next-generation sequencing (NGS) technology offers possibility to identify previously unknown microorganisms in humans' blood circulation by producing genomic information of the pathogens. The massively parallel nature of NGS technology produces large volumes of sequence data within short period. The proposed research project is important to gather information that will be used in the planning of a prospective study investigating the infectious causes of human febrile illness in far north Queensland. The output of this research will provide information on what is the most appropriate blood sampling technique and what kind of blood specimen contains least human nucleic acid. This study will be conducted in 4 month period (1 March 2012 – 31 June 2012) to answer this question: 'Is there any difference in the amount of human nucleic acid in different blood specimens and different sampling technique?' The aim of this study is:

1. to quantify nucleic acid obtained from whole blood, serum and plasma from healthy individuals with and without application of tourniquet during specimen collection
2. to find correlation between the type of specimen and the amount of human nucleic acid

What will we ask you to do?

1. If you consent to participate in this study, there will be a short interview with the investigators to check your eligibility for the study.
2. If you are eligible, you will be assigned a specific code/ID number.
3. Your blood will be collected by venous puncture of both of your arms. Blood collection will be conducted at James Cook University Cairns and should only take 5-10 minutes to complete.

Who can participate?

Those who are:

1. 18-50 years of age
2. do not have chronic illness
3. do not have acute infection
4. willing to provide 25 ml of blood sample
5. able to provide informed consent

What if I change my mind about participating?

Your involvement in the research is entirely voluntary. You are free to withdraw from the research at anytime without penalty. You may, if you wish also withdraw your consent for the use of your data in the study.

What are the possible benefits and risks of taking part?

You will not benefit personally by participating in the study. There will be some discomfort associated with the use of a needle to collect blood, and bruising is a potential complication. If you feel distressed or experience continuing complication, you can contact **Prof. John McBride (MBBS, DTM&H, FRACP, FRCPA, PhD)** by email or phone to have free advice and medical care. Please find his contact details in the front page of this document.

Possible benefits to general community

The results of the research will benefits to the society in terms of providing valuable information for future research investigating the extent of human and pathogen nucleic acid in blood samples. This study will benefit clinicians in improving management of fever of unknown origin which in turn benefit the patients. The study also will benefit any future research in establishing the infectious cause of undiagnosed fever and defining regional pathogens responsible for undifferentiated fever. Furthermore, this research will contribute to the advancement of medical knowledge and better management of patients with fever in the future.

Ethics review and complaints

This study has been reviewed by the Human Research Ethics Committee (HREC) of James Cook University. If you are not happy with the way this research has been conducted, you can contact:

Human Ethics, Research Office
James Cook University, Townsville, Qld, 4811
Phone: (07) 4781 5011
***Email: ethics*@jcu.edu.au**

Before you make your decision, a member of the research team will be available to answer any questions you have about the research project. You can ask for any information you want. Sign the consent form only after you have had a chance to ask your questions and received satisfactory answers.

If you require further information concerning this project, you can contact the principal investigator or the supervisory team listed above.

Thank you for your interest in this study.

Appendix I. Consent Form for Study 2

This administrative form has been removed

Appendix J. Data Collection Form for Study 3

Queensland Government

JAMES COOK UNIVERSITY AUSTRALIA

Fever study in Cairns Base Hospital

Data Collection Form

Principal Investigator: Dr Tri Susilawati

For Investigator use only
UR number:
Tests requested:

Diagnosis:

Inclusion criteria

1. Are you 16 -65 years old? Yes / No
2. Highest body temperature: _____^{0}C
3. Have you been feeling cold or shivering? Yes / No
4. When did you start having fever? _____/_____/_____
5. Are you willing to provide 10 ml of blood samples now and 2-3 weeks later? Yes / No

Demographic Data

1. Name: _____
2. Date of Birth: _____/_____/_____
3. Sex: ☐ Male ☐ Female
4. Citizenship: _____
5. Occupation: _____
6. Mailing address: _____
7. Email address: _____
8. Home phone:_____
9. Mobile phone:_____

Infection contact

1. Do you live in urban or rural area? _____
2. Were you exposed to animals in the last 3 weeks? Yes / No
   Example: cats, dogs, cattle, horses, goat, sheep, pigs, monkeys, rodents, chickens, etc
   If Yes, please specify_____
   _____
3. Is there any of your family or co-worker who has similar symptoms with you? Yes / No
   If Yes, please specify _____
   _____
4. Did you get mosquito bite or tick bite within 3 weeks prior to the onset of fever? Yes / No
   If Yes, please specify _____
   _____
5. Did you do any outdoor activities in the period of 3 weeks before the onset of fever? Yes / No
   Example: gardening, bush walking, jungle trips, animal hunting, camping, etc
   If Yes, please specify _____
   _____
6. Did you travel overseas in the last 6 months? Yes / No
   If Yes, please specify _____
   _____

Queensland Government

JAMES COOK UNIVERSITY AUSTRALIA

Clinical Data

Over the past 3 weeks have you experienced the following symptoms? (please circle)

1. Headache
2. Neck pain
3. Neck stiffness
4. Muscle pain
5. Joint pain
6. Back pain
7. Weakness
8. Fatigue
9. Eyes sore
10. Light sensitivity
11. Cough
12. Sore throat
13. Short of breath
14. Chest pain
15. Abdominal pain
16. Nausea
17. Vomiting
18. Diarrhea
19. Rash
20. Others: _____

21. Have you seen a GP before presenting to Cairns Base Hospital? Yes / No

If Yes, please specify any tests that the GP ordered for investigating the cause of your illness

_____
_____
_____
_____

22. Do you have any current medical conditions (e.g. High Blood Pressure, Diabetes, etc)? Yes / No

If Yes, please list

_____
_____
_____
_____

23. Are you on any current medications: Yes / No

If Yes, please list:

_____
_____

For Investigator use only

Date of admission:

Date of enrolment:

Researcher signature:

Appendix K. Participant Information Sheet for Study 3

PARTICIPANT INFORMATION SHEET

Project Title: Evaluation of next-generation sequencing (NGS) technology in determining infectious causes of human febrile illness

---

This information sheet is being provided to you to describe a study we are doing to identify causes of fever. You have been identified as someone who may be interested in participating in this study. Alternatively you may be given this document because you are authorised to provide consent on behalf of the potential participant.

Who are we?
We are a study group investigating the infectious cause of fever by using the next-generation sequencing technology.

This project will be conducted as part of a PhD research project by:

Dr. Tri Susilawati (MD, MMed)
PhD candidate, James Cook University
E: tri.susilawati@jcu.edu.au
Tel: (07) 4226 6996

Under supervision of the following James Cook University and Melbourne University staff:

Prof. John McBride (MBBS, DTM&H, FRACP, FRCPA, PhD)
E: john.mcbride@jcu.edu.au
Tel: (07) 42266530

Prof. Alex Loukas (BSc Hons, PhD)
E: alex.loukas@jcu.edu.au
Tel: (07) 4042 1608

Aaron Jex, PhD
E: ajex@unimelb.edu.au
Tel: (03) 9731 2294

The Problem – Why are we doing this?

- An audit at the Cairns Base Hospital (CBH) has found that around 1-3 people per week present to CBH with undifferentiated fever, a febrile illness for which the diagnosis is not immediately obvious.
- This audit revealed that after a series of investigations, the majority of undifferentiated fevers are not diagnosed.

What is this research project?

- We are conducting research using "next-generation sequencing (NGS)" technology which has the potential to identify any infectious cause of fever.
  The primary advantage of NGS platforms over conventional methods is that doctors do not need to specify what infection they are looking for. All lifeforms rely on molecules called DNA to determine their characteristics – the sequence of this DNA is a unique "fingerprint" which distinguishes one lifeform from another. NGS identifies all the "fingerprints" in any sample it is used on. A blood sample will contain the "fingerprints" of both the infectious agent and that contained in your own human DNA. As a consequence of this NGS will not only generate data about an infectious agent, but about YOU.
- In this study we only intend to analyse non-human DNA. However, human illness due to infection often involves an interaction between an infectious organism and the host, and in the future your genetic information may help to provide insights into this interaction. You can choose whether or not you want your genetic information to be stored for further study. The genetic information we obtain from you will NOT be analysed – rather stored as raw data. It would only be analysed in the future, after approval by a research ethics committee.
- Research on the application of NGS technology in clinical practice is needed to validate the use of this technology for routine diagnosis of infectious diseases. The aim of this study is both to re-identify the same pathogens that are detected by conventional methods utilizing NGS technology, and to discover genetic information of previously known and/or new pathogens in humans with undiagnosed fever.

What do we want to do?

- We want to enrol patients with undifferentiated fever into this study.
- Consent to participate in this study will be sought from the patient him/her-self. Consent from other people authorised to provide consent will be sought only if the patient is unable to consent due to their illness (eg the person is in intensive care).
- We will collect 10 ml (about 2 teaspoons) of blood by venous puncture on presentation (acute sample) and after 2-4 weeks (convalescent sample). Samples will be assigned a specific code/ID number.
- We would like your permission to use leftover blood that is collected from you by pathology laboratories for the investigation of fever, once the diagnostic laboratory has completed tests that your doctor has ordered.
- We will ask you whether you agree for us to store your genetic information for future study or to delete this genetic information. For participants who are under 18 years old or unable to give their consent, genetic information will be deleted.
- We will ask your permission to take a picture of abnormal physical findings related to your illness (eg.rash)

Who can participate? Those who are:

1. 16-65 years of age
2. fever for 21 days or less
3. documented temperature is at least 38^0C or history of fever with feeling cold or shivering
4. no obvious cause of fever after initial investigations
5. undergoing diagnostic tests for at least 1 specific infectious agent (can include rapid serological tests and malaria parasite screening)
6. willing to provide acute and convalescent serum samples

What if I change my mind about participating?

Involvement in the research is entirely voluntary. Participants are free to withdraw from the research at any time without penalty. You may, if you wish, also withdraw your consent for the use of information generated in the study. A decision to not participate in this study will NOT affect medical care.

What are the possible benefits and risks of taking part?

Possible benefits and risks to participants

There will be no direct benefits from participation in this research to participants. Participants may experience mild discomfort during blood collection. It is possible that we will detect a potential cause for fever in your blood. If you indicate on the consent form that you wish to know the results of tests done on you we will arrange for an interview to discuss the meaning of the results. The results will not be available until completion of the study, estimated to be in 2014. We do not guarantee that we will analyse every specimen we collect.

Possible benefits to general community

The results of the research will benefit the community in terms of providing valuable information on the reliability of NGS technology for investigating infectious cause of undiagnosed fever. Furthermore, this research may discover previously unknown pathogens in Far North Queensland which will contribute to the advancement of medical knowledge and better management of patients with fever in the future.

Ethics review and complaints

This study has been reviewed by the Human Research Ethics Committee (HREC) of Cairns and Hinterland Health Service District. If you are not happy with the way this research has been conducted, you can contact:

Chair of HREC of Cairns and Hinterland Health Service District
***P**: (07) 422 65312 **F**: (07) 422 65352*
***E**:Cairns_Ethics@health.qld.gov.au*
***Office**: 88 Abbott Street, Cairns QLD 4870*
***Postal:** PO Box 902, Cairns Q 4870*

Before you make your decision, a member of the research team will be available to answer any questions you have about the research project. You can ask for any information you want. Sign the consent form only after you have had a chance to ask your questions and received satisfactory answers. If you require further information concerning this project, you can contact the principal investigator or the supervisory team listed above.

Thank you for your interest in this study.

Appendix L. Consent Form for Participants ≥ 18 Years (Study 3)

This administrative form has been removed

Appendix M. Assent Form for Participants 16-18 Years (Study 3)

This administrative form has been removed

Appendix N. Consent Form for Incapacitated Participants (Study 3)

This administrative form has been removed

Appendix O. Photography Consent for Participants with Visual Abnormal Physical Findings (Study 3)

This administrative form has been removed

Appendix P. Publication and Presentation Arise from This Thesis

Thesis Chapter	Publication or Presentation Type	Year	Name of Journal, Conference, etc.	Title of Publication or Presentation
2	Review	2014	Southeast Asian Journal of Tropical Medicine and Public Health	Acute undifferentiated fever in Asia" a review of the literature
3	Oral presentation	2013	North Queensland Festival of Life Sciences	Pathogen detection: how deep can you go?
4	Poster presentation	2012	Australian Society for Microbiology Annual Scientific Meeting	Diagnostic approach to investigate the aetiology of fever during a dengue outbreak: don't look for dengue only
	Research article	2014	International Journal of Infectious Diseases	Undiagnosed undifferentiated fever in Far North Queensland, Australia: a retrospective study
5	Poster presentation	2013	Townsville Health Research Week	Quantification of circulating DNA in healthy volunteers
	Abstract		Annals of the Australasian College of Tropical Medicine	
6	Oral presentation	2015	7th Annual New Zealand NGS Conference	Detection of Dengue virus, *Leptospira* and *Rickettsia* using next-generation sequencing
	Oral presentation	2015	Cairns Hospital Grand Rounds	Undiagnosed fever – a next generation approach?

Appendix Q. Statements from Copyright Owners to Use Material Reproduced in the Thesis

1. Permission to reproduce material in Figure 2.1

This administrative form has been removed

2. Permission to reproduce material in Figure 2.2

3. Permission to reproduce material in Figure 6.3 and 6.4