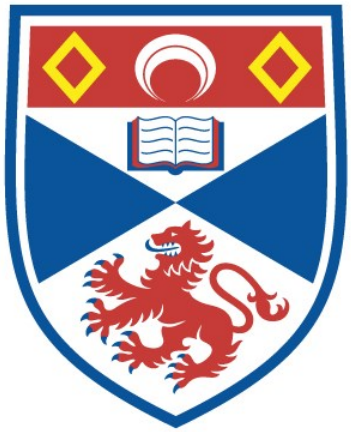
DE NOVO GENOME ASSEMBLY OF *PLASMODIUM KNOWLESI* FROM CONTEMPORARY CLINICAL ISOLATES – A NOVEL SCALABLE RESOURCE TO TAKE FORWARD MALARIA RESEARCH

Damilola Rasheed Oresegun

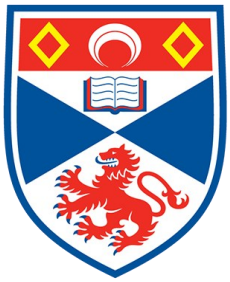
A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2022

De novo genome assembly of *Plasmodium knowlesi* from contemporary clinical isolates - a novel scalable resource to take forward malaria research

Damilola Rasheed Oresegun



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at the University of St Andrews

November 2021

Abstract

Plasmodium knowlesi is a zoonotic malaria parasite of Southeastern macaque monkeys that causes zoonotic malaria in humans. *P. knowlesi* is also an experimental model for malaria, and information on *P. knowlesi* largely stem from experimental lines first isolated >four decades ago, rather than contemporary isolates causing human infections. The experimental lines are laboratory-restricted and have become relatively genetically stagnant and free from the selection pressure that would naturally occur in nature. Within the *P. knowlesi* genome exist the Schizont Infected Cell Agglutination variant antigen (SICAvar) and *Plasmodium knowlesi* interspersed repeat (*kir*) multigene families significant, which are of biological and scientific interest. To provide context using contemporary clinical isolates, this project aimed to generate high-quality genome sequences using long-sequencing from clinical 'wild-type' samples from infected patients. This includes generating new information on variant multigene families in *P. knowlesi* genomes generated from clinical patient whole blood.

The work presented here details a method to deplete leucocytes in thawed *P. knowlesi*-infected patient whole blood samples to generate parasite-enriched DNA for whole-genome sequencing, resulting in >95 % human DNA reduction. The extracted DNA was sequenced with long-read sequencing technology to create *de novo* whole-genome assemblies. From these, two isolate genomes representing the two dimorphic clusters of *P. knowlesi* in clinical samples were analysed. The generated genomes are highly syntenic to the published reference genome, sharing >4500 orthologous clusters with the PKNH reference genome. However, the number of *SICAvar* and *kir* genes present in the dataset deviated from the published reference genomes of *P. knowlesi*. The successful generation and construction of these patient genomes aid further interrogation of the contemporary *P. knowlesi* genome, with a focus on the constituent genes present in comparison to the experimental line.

Acknowledgements

General acknowledgements

Firstly I would like to thank my supervisor Dr Janet Cox-Singh. Thank you for the constant support and sanity checks and keeping me on track. This project has been a difficult time, one made easier by your mentorship and advice on both science and life. Also, thank you for posting this project on Twitter. I will tell you the story of the application process from my end one day!

Secondly and by no means less, I would like to thank my parents, friends and family. You have been a source of motivation, support and love. While I may not have come to thoughts of giving up, your words, companionship and care let me carry on through the tough times. To Daniel, Elena, Eli, Kirsten, Panashe, Rebecca, Sanne and so many other friends and colleagues, you have been my extended support network through these last four years, and I hope I have been able to give even a tiny percentage of what you gave me. To Anna, thank you for supporting me through the most challenging part of this process and not becoming too sick of me during the write-up process. Thank you for being there when I needed you most, including during the pandemic and the difficulty within that period.

I would also like to thank my funding bodies, the Wellcome Trust (via the University of St. Andrews) and Tenovus Scotland. Thank you for your crucial financial support in completing this project.

My thanks go to my PhD committee, Dr Simon Powis, Dr Carolin Kosiol, Dr Terry Smith, and Dr Paul Reynolds, who gave me the confidence to continue on track in my first and second-year meetings. A particular thanks to Simon for his support in solving early problems during the development of crucial laboratory protocols.

Dr Rebecca Mekler, Dr Sophie Turner, Dr Robert Hammond, Dr Ines Nearchou, Chryso (Christie) Ioannou, Dr Silvia Paracchini, Dr Jon Lucocq,

Dr Joseph Ward, In Hwa Um, Karen Ross and Dr Wilbur Sabiiti, thank you. Thank you for the Friday afternoon cakes, the corridor chats, the reassurance, lab laughs and insights. You have been a true benefit to my time in St. Andrews and the School of Medicine.

Finally, to Dr Peter Thorpe. Thank you for your friendship and mentorship. Bouncing ideas off you, your advice and your collaborative spirit were not only welcome but helpful. I have learned a lot from you.

Thank you all, and I wish you all the best for the future.

Funding

This project was funded by the Wellcome Trust ISSF award 204821/Z/16/Z. Bioinformatics and computational biology analyses were supported by the University of St Andrews Bioinformatics Unit (AMD3BIOINF), funded by Wellcome Trust ISSF award 105621/Z/14/Z and 204821/Z/16/Z. The patient sample BioBank was compiled with informed consent (Medical Research Council, www.mrc.ac.uk, grant G0801971). Genome sequencing was supported by Tenovus Scotland (T16/03).

Research Data/Digital Outputs access statement

Outputs and data generated in this project were submitted for publication, with one article published and another prepared for submission. The published article can be found under **https://doi.org/10.3389/fcimb.2021.607686**, while the second article has been accepted and awaiting publication under **https://doi.org/10.3389/fgene.2022.855052**. Data supporting the upcoming submission are deposited in Zenodo under **https://doi.org/10.5281/zenodo.5598264**. Scripts used to generate the data in this project are available on Github: **https://github.com/damioresegun/Pknowlesi_denovo_genome_assembly** and **https://github.com/peterthorpe5/plasmidium_genomes**. Research data underpinning this thesis are available at https://doi.org/10.17630/541e495e-231a-4c43-bd02-904bc3fd6f79

Candidate's declaration

I, Damilola Rasheed Oresegun, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 45,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2017.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 12/05/2022 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 13th May 2022 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Damilola Rasheed Oresegun, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 12/05/2022 Signature of candidate

Date 13th May 2022 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Damilola Rasheed Oresegun, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date 12/05/2022 Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

Embargo on all of electronic files for a period of 1 year on the following ground(s):

- Publication would preclude future publication

Supporting statement for embargo request

A publication is being prepared to be released with the data presented

Date 12/05/2022 Signature of candidate

Date 13th May 2022 Signature of supervisor

For Baba, Mama, Mama Ìbàdàn, Grandpa, Uncle Kolapo, Dipo Ajadi and Dami Adeleye.

For all the little boys and girls with too much responsibility and not enough childhood wonder.

For all black boys and girls.
Don't let them take away your dreams.
Aye alawọdudu pataki

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CODES AND SCRIPTS

ABBREVIATIONS

P. knowlesi *Plasmodium knowlesi*

SICAvar Schizont Infected Cell Agglutination variant antigen

***ip*WB** Infected Patient Whole Blood

***i*RBC** Infected Reb Blood Cells

kir *Plasmodium knowlesi* interspersed repeat

pir *Plasmodium* spp. interspersed repeat

***pk*DNA** *Plasmodium knowlesi* DNA

***si*WB** Simulated-infected Whole Blood

vir *Plasmodium vivax* interspersed repeat

µL microlitre

ABI Applied BioSystems

ANOVA Analysis of Variance

API Apicoplast genome

BAM Binary Alignment Mapping

BED Browser Extensible Data format

BLAST Basic Local Alignment Search Tool

bp base pair

BSA Bovine Serum Albumin

BUSCO Benchmarking Universal Single-Copy Orthologs

BWA Burrows-Wheeler Aliger

CSA Chondroitin sulphate A

CSP Circumsporozoite protein

Ct Cycle threshold

DBG de Bruijn Graph

DBP Duffy-binding protein

ddNTP di-deoxyribonucleotide

DNA Deoxyribonucleic acid

dNTP deoxyribonucleotide

DUP Duplications

EDTA Ethylenediaminetetraacetic acid

FB Flush Buffer

FLT Flush Tether

Gb Gigabases

GFF General (or Genome) Feature Format

GPU Graphical Processing Unit

HAC High-accuracy basecalling model

hDNA human DNA

HLA1 Human Leucocyte Antigen Class 1

HPC High Performance Cluster

HS High Sensitivity

ICAM-1 Intracellular Adhesion Molecule-1

IGV Integrative Genome Viewer

INDEL Insertions and Deletions

INV Inversions

Kb Kilobases

LB Loading Beads

LCA Leucocyte Common Antigen

M.Ct Multiple Cycle Thresholds

Mb Megabases

mg milligram

MIT Mitochondrial genome

mL millilitre

mM Millimolar

MSP Merozoite Surface Protein

N.C Not Calculated

N.V No Value

NCBI National Center for Biotechnology Information

ng nanogram

NGS Next Generation Sequencing

NHP Non-Human Primate (NHP)

nm nanometer

nt NCBI nucleotide database

NTC No Template Control

OCTFTA One code to find them all

OLC Overlap Layout Consensus

ONT Oxford Nanopore Technologies

ONTSV Oxford Nanopore structural variation pipeline

PacBio Pacific Biosciences

PAM Pregnancy-associated malaria

PBS Phosphate-buffered saline

PCR Polymerase Chain Reaction

PEXEL *Plasmodium* export element

PfEMP1 *Plasmodium falciparum* Erythrocyte Membrane Protein 1

Pfvar *Plasmodium falciparum* variable gene

qPCR Real-time quantitave PCR

RAP Rapid Adapter

RBC Red blood cells

RNA Ribonucleic acid

RNN Recurrent Neural Network

ROS Reactive Oxygen Species

SAM Sequence Alignment Mapping

SBS Sequencing by synthesis

SICA Schizont Infected Cell Agglutination

SMRT Single-molecule real time

SNP Single Nucleotide Polymorphism

SNV Single Nucleotide Variation

SPIAP Sporozoite invasion-associated protein

SQB Sequencing Buffer

SV Structural Variant

TE Transposable element

TRA Translocations

VCF Variant Call Format

WBC White blood cell

WHO World Health Organisation (WHO)

ZMW Zero-mode wave guide

GLOSSARY

P. knowlesi *Plasmodium knowlesi*. A zoonotic human malaria found in South East Asia that has recently been identified to be naturally transmitted in the presence of the natural reservoir, vector and susceptible humans. ii, iv, ix–xi, xiii, 24, 28, 29, 51–54, 62, 63, 65–67, 73, 75, 77, 79, 119, 127, 130, 136, 138, 148, 256, 257, 265, 270–272, 274–276, 297

BED The *Browser Extensible Data (BED)* format is plain-text format which is made to hold genomic information such as positions and annotation in the form of co-ordinates. xiii, 103

BLAST *Basic Local Alignment Search Tool.* Algorithm and suite of tools for comparing biological sequence information for similarities. xiii, 110, 114, 128, 130, 173, 175, 195

Ct *Cycle threshold* (Ct), is the number of cycles within a PCR reaction where the fluorescent signal indicates a DNA concentration above the set threshold concentration. xi, xiv, xx, 62–64, 67, 70–74, 76–81, 254–257, 260, 277

DBG *de Bruijn Graph*. A class of *de novo* assembly algorithm that generate *de novo* whole genomes using short reads to assemble the longest sequence possible from k-mer graphs. xiv, 105–108, 118

ddNTP *di-deoxyribonucleotide*. A chemical analogue to the natural deoxyribonucleotides (dNTPs) which make up the backbone of the DNA. ddNTPs possess an inactive hydroxyl group which allows termination of DNA synthesis as part of Sanger sequencing. xiv, 93, 94

FASTA A file containing only the DNA, RNA or amino acid sequence of interest. No quality data can be held here. 103, *Also see abbreviations:* DNA & RNA

FASTQ A file containing DNA or RNA sequence data including additional information about the quality of each nucleotide position. The additional information is not human-readable and thus requires a quality assessment tool to analyse. xix, 103, 137

M.Ct *Multiple cycle thresholds.* Occurs when the qPCR detects the fluorescence curve of a sample crossing the threshold at least twice, resulting in an unambiguous Ct value. xv, 76

N.C *Not Calculated.* Refers to instances where a value could be calculated. Often appears when an average value could not be calculated due to no prior parent values. xv, 76

N.V *No value.* Occurs when the qPCR returns no Ct value for a sample with no clear reason given. It is likely due to insufficient template DNA or the reaction components not being mixed sufficiently. xv, 76

NGS *Next Generation Sequencing*. The colloquial name for the second generation of sequencing technologies. It is known for high-throughput seqencing, resulting in millions of short reads of the DNA of interest. xv, 94, 95, 99, 109, 117, 118

NTC *No Template Control*. A designation given to a sample which is unable to surpass the set threshold percentage in the qPCR experiment. The percentage is relative to the maximum fluorescence change in any sample of the experiment i.e. if the largest fluorescence change is 32, then a threshold of 10 % means no samples with fluorescence <3.2 are considered to surpass this threshold and thus are not considered. xv, 76

OLC *Overlap Layout Consensus*. A class of *de novo* assembly algorithm to generate whole genomes using long read sequence data. Uses overlaps to generate a layout graph to confirm a consensus sequence. xv, 106, 107, 118

ONT *Oxford Nanopore Technologies*. A publicly-traded company producing the MinION platform utilised in this thesis. xv, 96–103, 106, 107, 112, 113, 124, 127, 147, 168, 169, 172, 182, 216, 282, 304, 306

PCR *Polymerase Chain Reaction*. A molecular process where input DNA fragments are continuously replicated, creating millions of copies of the input DNA. xvi, 66, 93, 96, 97, 170, 301

PfEMP1 *Plasmodium falciparum Erythrocyte Membrane Protein 1* are a family of variable proteins found on the surface of *Plasmodium falciparum*-infected red blood cells. They have been implicated with facilitating sequestration in infected red blood cells in the host's blood vessels. xvi, 21–24

RNN *Recurrent Neural Network*. A class of artificial neural networks capable of processing sequential and temporal data; using the previous inputs to influence the current input data over a period of time. xvi, 100, 147

SBS *Sequencing by synthesis*. xvi, 94

SMRT *Single-molecule real time* sequencing patented by Pacific Biosciences. Uses a flowcell (called a SMRTcell) comprised of ZMWs to carry out long read sequencing. xvi, 95, 97, 99, *Also see glossary:* ZMW

WHO *The World Health Organisation*. International body tasked with researching, monitoring and improving international public health. xvii, 8, 9

ZMW *Zero-wave guide*. A small chamber/well in a Pacific Biosciences SMRT cell which is small enough to stop light from completely passing through to allow the detection of fluorescence at the bottom of the ZMW. xvii, 95, *Also see glossary:* SMRT

CHAPTER ONE

INTRODUCTION

Fimí-pamọ́-kí-npa-ọ́ làrùn ńjẹ́ – **A concealed disease is a deadly thing.**

Yorùbá adage

Collectively, mosquitoes have been the greatest source of human death for thousands of years. The chief culprit of their destructive effect on *Homo sapiens* has arguably been in their transmittance of malaria. Malaria is a disease that has been a companion for humanity, stretching from the Neolithic age to the modern age [1]. The first known record of the disease known today as malaria was first described in a clay cuneiform tablet dated for 3200 - 1304 BC [1]. Remaining records from this pre-historic age document the disease, detailing crude but understandable deductions including the association of malarial fevers with splenic enlargement [1, 2]. Additionally, malaria has been described in various cultures and settings ranging from physicians in ancient China to Greek philosophers like Plato, Homer and Aristophanes; including an account from ancient Rome [1]. Indeed, malaria can be argued to be one of the most influential external factors for human development, so much so that some historians have suggested malaria was a factor that influenced the fall of Rome in 79AD [1].

As such, malaria has historically been a disease that caused great fear and confusion in ancient society, often resulting in rapid loss of adults and infants alike, with the only records remaining seen in excavations, centuries after the fact [3]. Truly, malaria is the dark shadow following humanity across both time and space; however, as with everything else, much has come to be understood about the disease, its causative agents, treatment and potential eradication.

1.1 Malaria: The Disease

1.1.1 Background

Malaria is a potentially fatal disease whereby healthy red blood cells (RBC) are invaded by intracellular protozoan parasitic organisms belonging to the genus *Plasmodium*. These *Plasmodium* spp. are mainly transmitted by female Anopheline mosquitoes – and some Culline species – passing from insect vector to vertebrate host [4, 5]. As such, member species of the genus have become adapted to infect animals ranging from birds, lizards and rodents to non-human primates like chimpanzees and macaques [6], and humans. While these causative *Plasmodium* spp. are related, they have evolved to be host-adapted and thus can only infect a particular host species or closely related host species. However, this has come into contention in the last two decades with recent reports of zoonotic malaria in Malaysia [7] and Brazil [8].

1.1.2 Malaria causative agents

As previously stated, multiple species of *Plasmodium* are capable of invading a variety of vertebrates, from the avian (bird) *Plasmodium relictum* to the saurian (lizard) *Plasmodium zonuriae* [5, 9]. In humans, six species have been fully recognised for their ability to elicit an infection: *Plasmodium falciparum*, *Plasmodium malariae*, *Plasmodium vivax*, *Plasmodium ovale* (its two sub-species: *Plasmodium ovale wallikeri* and *Plasmodium ovale curtisi*), and *Plasmodium knowlesi* [10–12].

By far, the largest proportion of malaria global burden and subsequently, the greatest share of research and funding is directed towards *Plasmodium falciparum*. *P. falciparum* is a human-adapted parasite that is present in all continents apart from Europe, though often only endemic in tropical and sub-tropical regions. However, the greatest burden of *P. falciparum* comes from sub-Saharan Africa [13]. *P. falciparum* is arguably the most distinct of the human-adapted *Plasmodium* spp. parasites. Indeed, it is now known that *P. falciparum* evolved from the *Laverania* sub-genus of *Plasmodium*, which is known to infect only chimpanzees and gorillas [6]. As such, *P. falciparum* is genetically

clustered with *Laverania* and *Laverania*-like *Plasmodium* parasites, rather than with other human-adapted organisms [Figure 1.1].

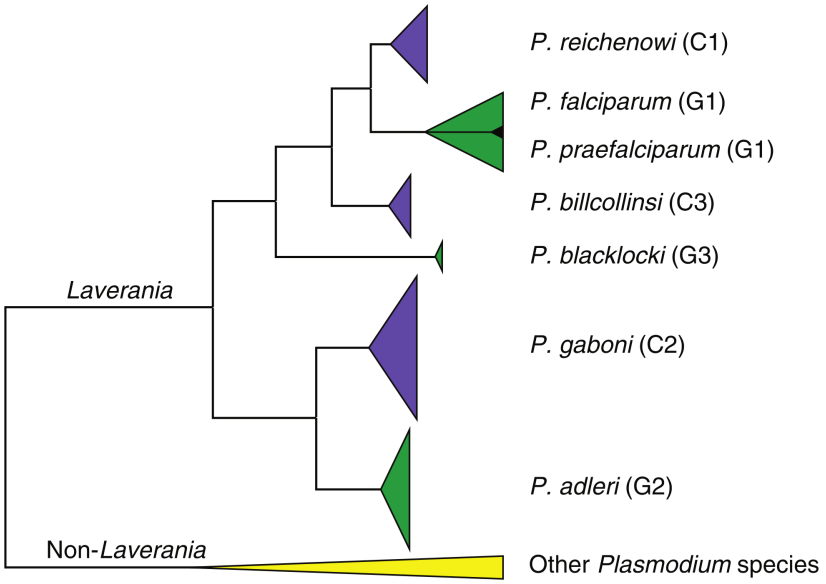


Figure 1.1: Phylogeny of ape-adapted *Plasmodium* species' mitochondral sequences reveals *P. falciparum* genetically clusters with Laverania. Six clades of ape-adapted *Plasmodium* spp. were found; all conforming to the *Laverania* subgenus. Here samples were isolated from gorillas (Green, G1 - G3) and chimpanzees (purple, C1 - C3) and non-Laverania species are presented in yellow. The black triangle represents human *P. falciparum* which form a single clade within the gorilla parasites. Reproduced from: Rayner et al. [6].

Meanwhile, *P. vivax*, the second most prevalent human malaria [13], covers a greater regional distribution than *P. falciparum*, although, *P. vivax* is most commonly transmitted in Asia and Latin America [14]. This geographic distribution is largely due to its ability to remain dormant in infected humans, thus allowing for survival in colder climates, which other *Plasmodium* spp. cannot survive [15]. However, it is unable to gain a foothold in sub-Saharan Africa due to the inactivation or absence of the Duffy antigen receptor in indigenous sub-Saharan African genomes [14]. Specifically, *P. vivax* relies on the presence of the 'Duffy antigen receptor for chemokines' on the surface of host erythrocytes [14]. However, 95 - 99 % of people in West/Central Africa are estimated to be missing this antigen, thereby conferring protection to *P. vivax* infections [14, 16]. Although, *P. vivax* infection of Duffy-negative individuals has been reported in

Madagascar where *P. vivax* appears to have successfully surpassed its dependence on the Duffy antigen for human infection [14, 17]. Apart from this, a small population of Duffy-negative individuals with *P. vivax*-specific antibodies have been identified in the Congo; indicating low-level human transmission of the parasite [16, 18]. In comparison to *P. falciparum*, *P. vivax* was thought to be a relatively benign infection [19]. However, *P. vivax* has been implicated of progressing from relatively uncomplicated to severe malaria; much like *P. falciparum* and *P. knowlesi* [15, 19, 20]. Additionally, much like *P. falciparum*, *P. vivax* has a related non-human primate form present in apes [6, 18] although, the hypothesised species is still only '*P. vivax*-like', due to a lack of understanding and research in the organism [6].

All human *Plasmodium* spp. currently in the Americas are thought to the unforeseen exports to the 'New-World' by European colonists. Of these, *Plasmodium malariae* is arguably the least researched; likely due to its reported low prevalence, moderate symptoms and difficulty to grow *in vitro* [21]. While *P. malariae* endemic regions overlap with those for *P. falciparum*, the proportion of *P. malariae* is miniscule in comparison [22]. In such regions, *P. malariae* is found as mixed infections with other *Plasmodium* spp.; most often *P. falciparum* [21]. However, these were based on the use of microscopy for detection and confirmation. A recent change to modern molecular detection reveals *P. malariae* prevalence to be far larger than initially thought; with at least an 8-fold increment on the previous global prevalence levels [21]. However, this does not change its milder infection outlook in susceptible humans, which is likely due to its reduced growth rate in comparison to other *Plasmodium* spp. [21]. Curiously, *P. malariae* forgoes the liver stage entirely and preferentially targets mature RBCs while having low parasitaemia per microlitres (µL) of blood [21].

When it was first discovered, *Plasmodium ovale* was thought to be a variant of *P. vivax* prior to be being fully recognised as a distinct species in its own right [24]. Eventually, *P. ovale* was found to possess two morphologically similar, but genetically distinct forms, *Plasmodium ovale wallekeri* and *Plasmodium ovale curtisi* [24, 25]. Although there is an increasing convention to treat these two forms as two distinct species as no reports of a hybrid of the two forms have occurred, which would be expected if *P. o. wallekeri* and *P. o. curtisi* were indeed forms of the same species [24, 25]. Much like *P. vivax*, *P. ovale* can remain in the liver of the infected human host for a prolonged length of

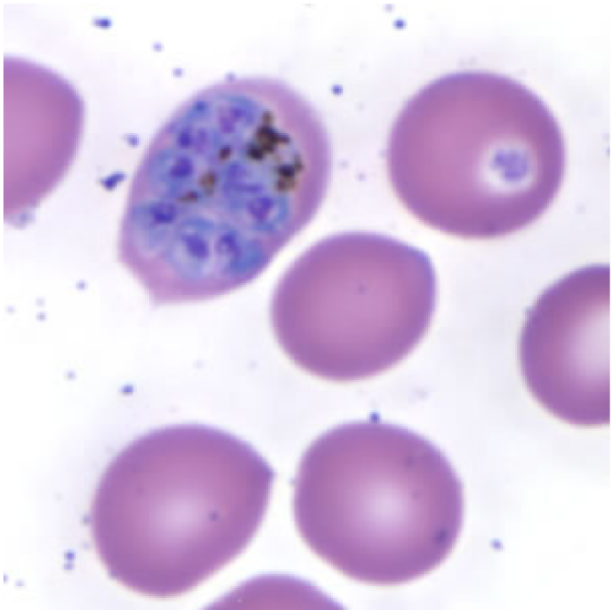


Figure 1.2: *Plasmodium ovale* schizont infected red blood cell. Giemsa stained thin blood film of a *Plasmodium ovale* schizont within an infected human red blood cell, with the characteristic oval shape of the infected red blood cell being shown. Source: Division of Parasitic Diseases and Malaria [23]

time, resulting in a relapse infection months or even years after initial infection [10, 24]. However, *P. ovale* has been neglected in wider global malaria research, likely due to its mild disease outputs and reduced burden around the world [25]. All taken, *P. ovale* represents a fascinating form of *Plasmodium* that appears to be in flux, with a need for greater research investment.

Plasmodium knowlesi is a *Plasmodium* spp. recently recognised by the World Health Organisation (WHO) to be able to infect humans; despite being primarily a malaria of old-world macaques. *P. knowlesi* is the focus of the work being presented within this report and will be expanded upon in section 1.2.

Regardless of the species, in regions where the insect vector, vertebrate host and causative agent exist, malaria can often be found, particularly around the tropical and sub-tropical regions of the world [26]. With 40 % (>3.3 billion) of the world's population residing within these regions [1, 27], it is no surprise that the global toll of human malaria infections appears never-ending, after decades of funding and research.

1.1.3 Global Distribution and Burden of Malaria

In the three decades between 1969 and 2000, malaria appeared to fade into a state of tacit ignorance and abandonment in the global health research community [13]. Perhaps due to the emergence and magnitude of the HIV/AIDs epidemic and other health crises, malaria seemingly fell to the wayside in research and funding. However, this does not discount the many researchers who fervently worked to improve knowledge, treatment and prevention of the disease. Over this period, hundreds of millions became infected, tens of millions died, hundreds of thousands of pregnant women died due to malaria complications and millions of surviving children were left with recurring bouts of malaria infections, seizures and often cognitive impairment [13]. The last two decades (2000 - 2020) has been a veritable age of renewed focus in malaria research, with the World Health Organisation (WHO) and malaria-endemic countries setting lofty aims of eradication, international charitable organisations funding cutting-edge research, and importantly, curiosity and passions for malaria research were rekindled. There has been a remarkable and admittedly impressive reduction in both incidence and fatalities associated with malaria around the world [Figure 1.3a,b]. However, it must be noted that even with such progress, 229 million cases of malaria and 409 thousand deaths were reported globally in 2019; an increase of 1 million cases and a decrease of two thousand deaths from the previous year [13]. As such, the fight for malaria control and elimination continues in the most endemic and often impoverished regions of the world.

As previously mentioned, the global burden of malaria cannot be overlooked due to the sheer proportion of the world's population that lives within malaria-endemic regions. However, it must be noted that the global infection rate for malaria has drastically reduced in the last three decades [1, 13, 27, 28]. This is particularly the case in traditionally malaria-endemic countries and regions that have the resources and have implemented vigorous and expansive elimination programs [27]. Nevertheless, the threat of malaria remains in poor and under-developed nations of the world, many of which have lost countless lives for generations, thus further compounding the effect of the disease and other geo-political hardships. Nowhere is this more evident than sub-Saharan Africa, where *P. falciparum* is consistently implicated to be the *de facto* culprit for 99.7 % of all cases in the region [13]. Indeed, the African region accounts for 94 % of all malaria

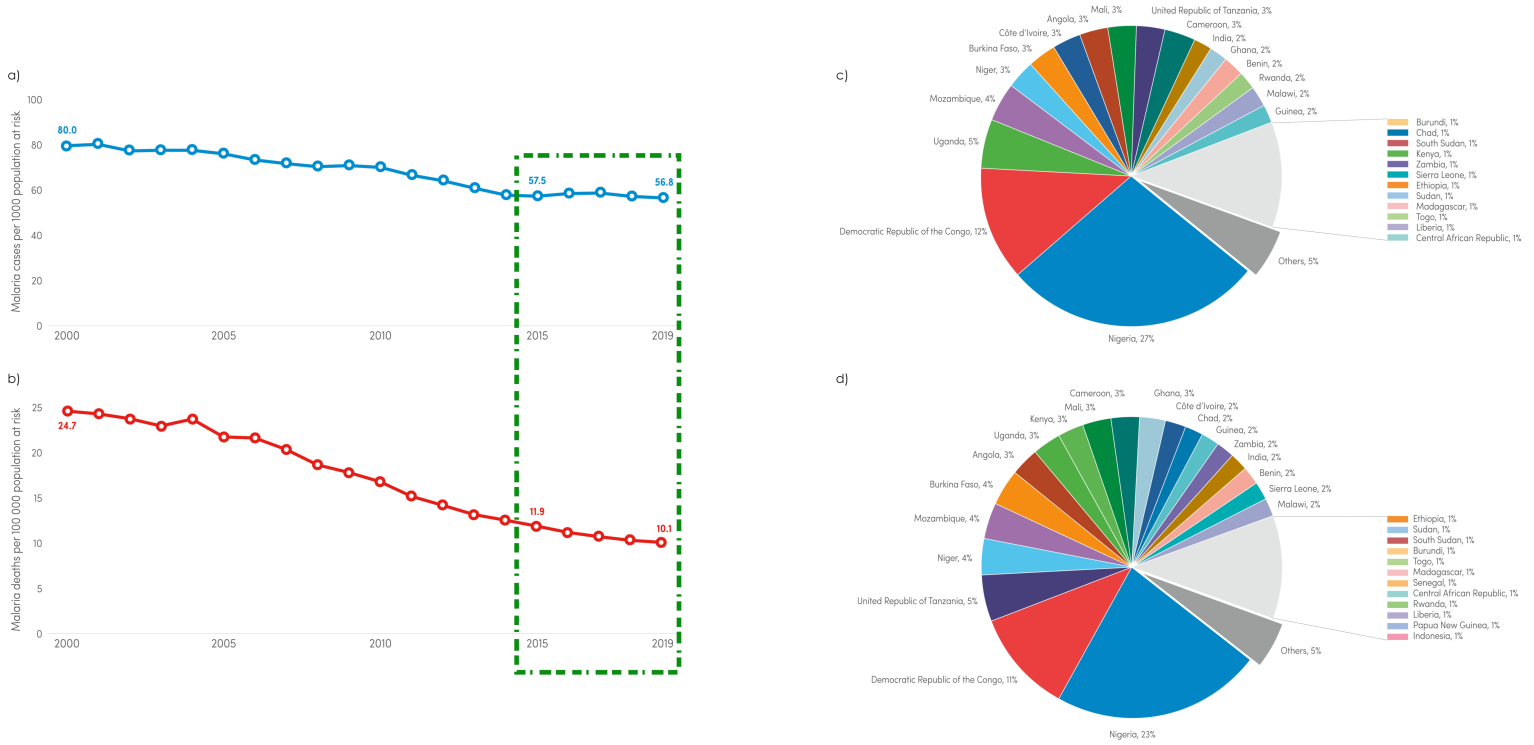


Figure 1.3: Global reports of malaria incidence and deaths over two decades (2000 - 2020). Global trends of malaria incidence and deaths where incidence rate (a) per 1000 population steadily falls between 2000 - 2015 before a stalling of progress from 2015 - 2020 (green box). Death trends per 100,000 population (b) are similar, with very little reduction in the fatality rate since 2015. Within countries, incidence (c) and deaths (d) are mostly represented in sub-Saharan African nations with the largest proportions seen in West Africa. Adapted from World Health Organization [13].

cases in 2019, and five African nations alone account for half of all reported global cases [Figure 1.3] [13, 28]. Apart from the African region, *P. falciparum* was reported to be the cause for 50 %, 71 % and 65 % of reported cases in the Southeast Asian, Eastern Mediterranean and West Pacific regions, respectively [13]; once again solidifying the imminent threat posed by this single species.

Apart from *P. falciparum*, *P. vivax* is the other main contributor to the global impact of malaria. Indeed, *P. vivax* was reported to contribute to ~3 % of malaria cases, which is a decrease from 7 % in 2000 [13]. Geographically, *P. vivax* infections were most prevalent in the Americas, accounting for 75 % of infections reported in the region [13]. Curiously, *P. malariae* and *P. ovale* have not been regularly reported by the WHO in its annual report. It is likely due to these species' low prevalence or their relatively low burden of infection. Finally, malaria caused by *P. knowlesi*, receives a brief mention on the world stage, with the WHO noting a marked increase in *P. knowlesi* associated infections from 1,600 in 2016 to over 4,000 in 2018 [13]; although these were only reported in Malaysia. Other Southeast Asian nations report much lower incidences, such as the 38 cases reported in Vietnam between 2004 and 2010 [29].

While general malaria incidence rates have decreased over time; it is prudent to note that even with a global surveillance program in place, not all malaria cases are reported in endemic countries. Perhaps, due to the infected individual deciding not to go to a healthcare centre, lack of necessary transportation and healthcare infrastructure or any myriad of other factors, only an estimated 19 % of cases were thought to have been reported in 2015 [30]. Truly, with such a staggering under-representation of the incidence, it is likely, the fatality rate would also be misrepresented. Even so, it is undeniable that the fatality rate has reduced worldwide year-on-year with a reported 409,000 deaths in 2019, in comparison to 429,000 in 2016 and 453,000 in 2015 [Figure 1.3] [13, 30].

Importantly, childhood mortality associated with malaria has dropped from 84 % to 67 % between 2000 and 2020, though this is still an unacceptable number of child deaths in these regions [13]. As seen in Figure 1.3, five sub-Saharan nations (Nigeria, the Democratic Republic of the Congo, Uganda, Mozambique and Niger) account for 51 % of all global malaria deaths, with Nigeria (23 %) and the Democratic Republic of the Congo (11 %) reporting the two highest proportions [13]. Away from the African regions, the

Southeast Asia region has also seen a remarkable 74 % reduction in deaths over the two decades (2000 - 2020), and while India reported the most significant drop in the number of case incidence in the region, it still accounted for the most significant proportion of deaths [13]. Similarly, in the East Mediterranean, West Pacific and Americas regions, mortality reduced by at least 50 % between 2000 and 2020, representing an overall trend towards minimised death burdens in the malaria-endemic regions.

Indeed, the goal of these nations and the WHO is the elimination of malaria or minimisation of deaths as a consequence of contracting malaria. However, the lofty aims of malaria eradication, initially agreed upon in the 1950s and again formalised in the 1990s, have been reduced from a global ambition to a country-specific aim or even, in some cases, species-specific elimination [13, 26]. While these nations, particularly in Southeast Asia, have invested a significant amount of effort and financial incentives, they still have malaria cases, which are often reported as non-local transmission and are travel-related [13]. Such strategies and programmes have been ineffective in the sub-Saharan region, although insecticide-treated mosquito nets have proven effective in mosquito and malaria control.

The WHO's stance on the definition of malaria elimination and what is deemed human malaria leaves room for some manoeuvring. A chief example of this being the status of *P. knowlesi* which has been accepted to cause human infection; however, due to being classed as zoonotic malaria, it does affect Malaysia's status for malaria elimination. This appears to be due to no current evidence of human-to-human transmission without a passage through a non-human primate first [13, 31]. Importantly, the question remains if this is the correct means of carrying out reporting to the general public; with the most recent malaria report focussing on *P. falciparum* and *P. vivax* exclusively. With the burden imposed by these two species, it is certainly understandable that these take the fore; however, with no acknowledgement of cases of other human host-adapted *Plasmodium* spp. or indeed, zoonotic malaria species, these cases remain ignored or underestimated in incidence and impact.

1.2 *Plasmodium knowlesi*

As previously mentioned, *Plasmodium* species tend to be host-adapted, thus human adapted *Plasmodium* species tend to be unable to infect other primates or non-human primates. However, with the discovery of a large natural outbreak of *P. knowlesi* (*whose natural animal hosts are old-world macaques*) in human populations of Sarawak Malaysian Borneo [7], host-adaptation has come under some degree of contention. *P. knowlesi* was discovered in the early 19th Century with the extraction of the parasite from the blood of an infected *Macaca fascicularis* (long tailed macaque) and later from a *Macaca nemestrina* (pig tailed macaque) [32]. Early descriptions and observations of experimental infections of *P. knowlesi* revealed the development of a fulminating infection in *Macaca mulatta*, before being shown to be transmitted to humans [32]. Eventually, *P. knowlesi* was adopted as a therapeutic agent in the treatment of neurosyphilis in humans [33, 34]. While varying levels of severity were documented in humans artificially infected with *P. knowlesi*, natural infections were thought to be rare, with only one case being documented prior to 2004 [33, 35–37]. It was the release of the Singh et al. [7] study which has propagated new interest in *P. knowlesi* research [26].

The standout feature of the Singh et al. [7] study was its use of modern molecular analytical approaches, which were not available during the first isolation of *P. knowlesi*. For decades, *Plasmodium* spp. were identified and diagnosed morphologically upon visual observations made via light microscopy [36, 37], whereby features such as the size of the parasite, infected RBCs and subsequent internal structures aid to determine which species is causing the infection [31, 38].

However, subsequent research has revealed that *P. knowlesi* morphologically appears similar to *P. malariae* or *P. falciparum* depending upon the life stage the parasite is observed. As such, early trophozoite *P. knowlesi* parasites morphologically appear as *P. falciparum* while mature trophozoites and schizonts are indistinguishable from *P. malariae* parasites [38, 39]. The life cycle of *P. knowlesi* and by extension the other *Plasmodium* spp. will be discussed in chapter 1 subsection 1.2.1. Lee, Cox-Singh, and Singh [38] report very minor morphological differences such as *P. knowlesi* producing 16 merozoites within schizonts rather than the 12 seen in *P. malariae*.

However, it must be noted that the ability to confidently and consistently identify these features by routine microscopy is doubtful. So much so that a recent review of studies published about *P. knowlesi* has estimated as much as 57 % of the *P. knowlesi* samples had been misdiagnosed as *P. malariae*; while in endemic regions like Sarawak and Sabah, Malaysian Borneo, misdiagnoses can be as high as 87 % and 85 %, respectively [40]. It is this similarity with *P. malariae*, which is suspected to be at the root of *P. knowlesi* seemingly appearing to cause large natural human infections suddenly. The only evident non-molecular signifier for *P. knowlesi* was determined to be a blood-film with morphological features of *P. malariae*, but possessing atypical symptoms and parasitaemia [7, 39, 41].

Natural Vectors and Hosts

Malaria transmission of any kind requires a vertebrate host and insect vector to complete the parasite life cycle and elicit the infection. As mentioned, human malaria is transmitted by female Anopheline mosquitoes, which carry the parasite in their proboscis and deposit parasites into a viable host during a blood meal [4, 5]. Whilst there are ~430 species of Anopheles, only 30 - 40 species are capable of transmitting malaria with *Anopheles gambiae* and *Anopheles funestus* contributing to the greatest malaria burden in the world [42]. However, both species are only found in Africa, and for *P. knowlesi*, transmission occurs via members of the Leucosphyrus group of Anopheline mosquitoes [43–46]. The leucosphyrus group consist of 20 species of diverse mosquitoes that are found in the forests of Southeast Asia [43, 47].

Members such as *An. latens*, *An. dirus*, *An. cracens*, *An. introlatus* and *An. hackeri* have been confirmed to be vectors of *P. knowlesi* across Southeast Asia [29]. However, it appears that the prevalence of a particular *P. knowlesi*-infecting Anopheline species is dependent upon the geographical location. As an example, *P. knowlesi* has been documented, observed and described in the Sabah and Sarawak regions of Malaysian Borneo. However, *An.balabacensis* is the most prevalent leucosphyrus group species in Sabah, and *An. latens* is the most prevalent in Kapit, Sarawak [29, 48]. Importantly, most of the members of this group are known to be exophilic and mostly jungle-restricted, with a few like *An. balabacensis* having rare visits to forest fringes [45, 46, 49]. This

suggests the majority of the insect vectors' activity occurs deeper within the jungle, away from human settlements. This creates a need for a reservoir or host that transports the parasite from within the jungle to the jungle-fringes, where infection and transmission occur in humans.

While *P. knowlesi* is known to cause malaria infections in humans; humans are not the natural hosts of *P. knowlesi*. Rather, the natural host reservoir of *P. knowlesi* are Old-World macaques; chiefly the long-tailed macaque *Macaca fascicularis*, pig-tailed macaque *Macaca nemestrina* and the black-crested Sumatran langur *Presbytis melalophos* [29, 45, 46, 50, 51]. As macaques, these species are arboreal creatures, which naturally dwell deep in the jungle (*some species like the long-tailed macaque live in urban areas, though they have not been documented to confer P. knowlesi*), venturing into forest fringe regions in search of food. Studies into the prevalence of *P. knowlesi* in these species, have identified that wild *M. fascicularis* individuals are most likely to carry *P. knowlesi*; particularly in Sarawak, Malaysian Borneo, where 87 % and 50 % of captured wild *M. fascicularis* and *M. nemestrina* respectively, were positive for *P. knowlesi* [50].

Both the vectors and natural reservoirs live and thrive in forests; hence human transmission only occurs when humans venture into jungle environments or fringe areas where the intersection of Anopheline vector, non-human primate host and healthy human can occur [45, 52]. This interaction is likely the reason for the relatively low *P. knowlesi* incidence reported across the region. However, with the increase in deforestation, human encroachment into jungle spaces coupled with wild macaque troupes venturing closer to human dwellings to scavenge, the causative vectors could potentially accompany the non-human primates (NHP)-hosts to occupy new niches [29]. With this, it is likely the rate of *P. knowlesi* will increase over the coming years. Truly, this is already evident in some rural regions showing elevated incidence rates for *P. knowlesi*, including in children living in these deforested areas of Sabah, Malaysian Borneo [45, 47, 53]. The relationship between the parasite, vector and host are particularly evident given the overlapping geographical range they occupy [Figure 1.4]. However, it must be noted that though the geographical spread of *P. knowlesi* is relatively quite large, the majority of *P. knowlesi* research has occurred in Malaysia, with research in other affected countries lagging, particularly in vector research [51]. With that said, this does pose an avenue

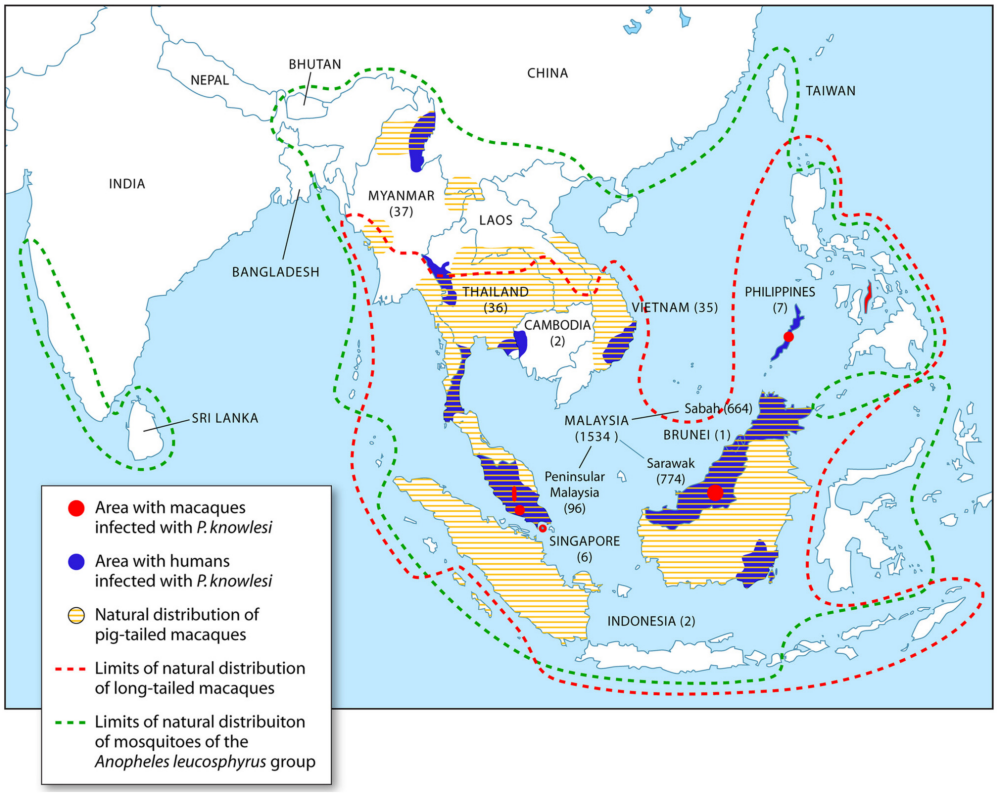


Figure 1.4: Geographical range of *P. knowlesi* vectors and hosts. Regions where *P. knowlesi*, *M. fascicularis* and *M. nemestrina* have been reported across Southeast Asia. Source: Singh and Daneshvar [50].

of research ripe for investigation and discovery, with projects investigating the presence of *P. knowlesi* in Indonesia; including in Sumatra where *P. knowlesi* has been found in humans [54], and Sulawesi, an island where the vector and the natural host are available.

1.2.1 *Plasmodium knowlesi* Life Cycle

Regardless of the Plasmodium species, primate malaria parasites follow a similar life cycle involving an asexual pre-erythrocytic asymptomatic period, followed by an asexual symptomatic erythrocytic period in the liver, and finally, a sexual asymptomatic period; ready to begin the next vector transmission cycle [46, 53]. The life cycle for *P. knowlesi* is presented below with specific mentions provided where differences appear for other

human-adapted *Plasmodium* spp.

Firstly, the asexual pre-erythrocytic asymptomatic period involves the first invasion of the host by the parasites and subsequent propagation of these parasites. Here, the infected female Anopheline mosquito breaks the dermis with its proboscis, laden with sporozoites (the haploid forms of *Plasmodium* spp.) in its salivary glands [46, 55]. Once deposited in the host, the sporozoites are thought to navigate through the blood vessels, ending up at the liver, invading the hepatocytes present in the liver [Figure 1.5] [46, 55]. To succeed, the sporozoites have to traverse the endothelial cells and phagocytic Kupffer cells lining the liver [46, 55, 56]. To achieve this, the sporozoite utilises the well-described circumsporozoite protein to mediate interactions between the parasite and the host Kupffer cells; to gain entry [56]. In truth, this mechanism of invasion is still poorly described, with investigations into the path of invasion required to elucidate interactions between various cytokines of the host as a consequence of the activity of the sporozoite [56].

In one such study, Klotz and Frevert [58] describe the down-regulation of immunogenic Th1 cytokines (Interferon-γ, Interleukin-6) and the up-regulation of Th2 cytokines; allowing the sporozoites relatively unimpeded traversal through the Kupffer cells. This results in an intriguing theory about the effect the sporozoite interactions have on the immunogenic function of the Kupffer cells as there are suggestions the sporozoites monopolise the function of the Kupffer cells [56]. Here, monopolisation includes the inhibition of reactive oxygen species (ROS), thereby effectively removing a key indicator for host cell stress and damage [59]. In any case, once traversal is complete, sporozoites invade the liver hepatocytes [Figure 1.5a]. This begins an asymptomatic period of the parasite's life cycle. Upon entry into the hepatocyte, sporozoites employ further down-regulation and inhibition of key cell pathways such as the mTOR and NF-kβ signalling pathways, thus, encouraging parasite growth and development while suppressing cell processes involving cell death and growth [56].

Additionally, the sporozoite encases itself in an intracellular vacuole as a means of escaping endosomic/lysosomic cell responses [56, 60]. Within the vacuole, the sporozoite matures and replicates into thousands of liver stage merozoites before finally rupturing the hepatocyte, releasing these merozoites into the circulating bloodstream [Figure 1.5b] [55,

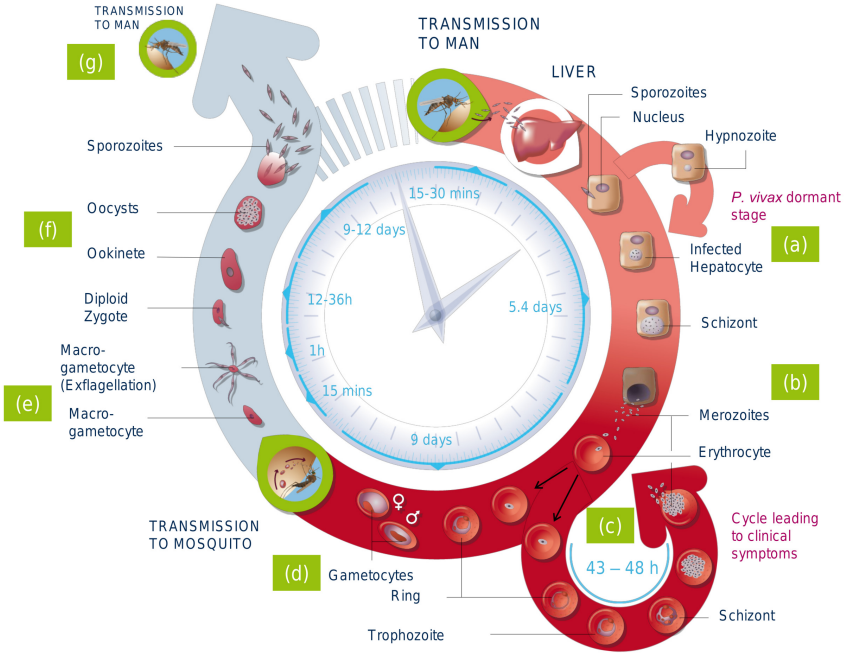


Figure 1.5: The life cycle of *Plasmodium knowlesi*. The stages of development for *Plasmodium* spp. organisms which infect humans with specific differences pointed for aberrant species like *P. vivax*. Injected sporozoites infect liver hepatocytes **(a)**, replicating and maturing and finally releasing merozoites into the host bloodstream **(b)** to invade erythrocytes, triggering an "invade-replicate-rupture" cycle **(c)**. A portion of trophozoites differentiate to become male and female gametocytes **(d)** to be taken up by another female Anopheline mosquito. Within the mosquito, the male gametocytes exflagellates **(e)** to produce smaller gametes to fuse with the female gametocyte. The fusion creates a zygote which develops into an ookinete and finally an oocyst **(f)** containing thousands of sporozoites. The sporozoites are released and migrate to the salivary glands to be transmitted to a new human/primate host **(g)**. Adapted from: Medicines for Malaria Venture [57].

56]. In *P. vivax* and *P. ovale*, a proportion of sporozoites invade hepatocytes and become dormant, allowing for a recurring infection weeks and months after initial infection [55]. Once hepatocytes rupture, the released merozoites are coated in the fragments of the host hepatocyte, subverting host immune recognition to successfully reach the circulating RBCs [55, 56, 61].

It is the invasion of host RBCs that triggers the symptomatic erythrocytic phase of *P. knowlesi* (and *Plasmodium* spp.) malaria in humans. Here, *P. knowlesi*, much like *P. vivax* utilises the Duffy-binding protein (DBP) family to anchor to the host RBC; facilitating ingress [41, 62]. Upon entry, the merozoite replicates and progresses into ring-stage,

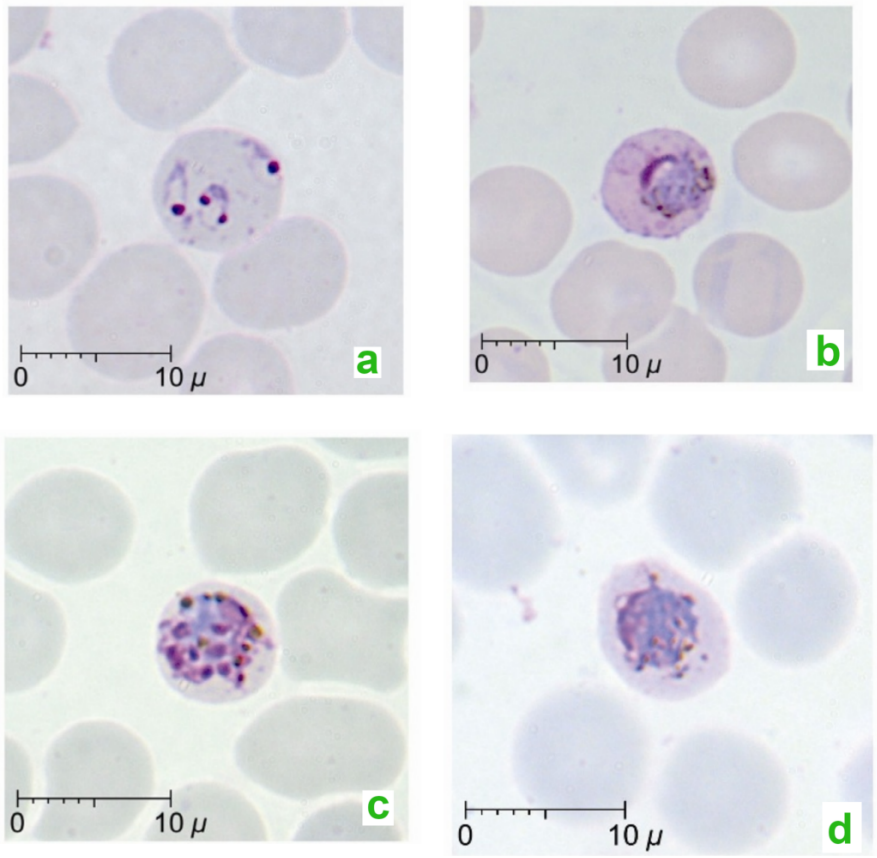


Figure 1.6: *Plasmodium knowlesi* erythrocytic life-cycle stages in human red blood cells Giemsa stained thin blood films of *Plasmodium knowlesi* parasites in infected human hosts at the early trophozoite **(a)**, mature trophozoite **(b)**, schizont **(c)** and gametocyte **(d)** life stages. Adapted from: Lee, Cox-Singh, and Singh [38]

young and mature trophozoites [Figure 1.6a,b]. It is at this point where a striking morphological feature of *P. knowlesi* can be observed. At the early trophozoite life-cycle stage, *P. knowlesi* is morphologically very similar to the trophozoites of *P. falciparum* [41]. Once mature, the trophozoites begin undergoing schizongony [55, 56, 62]. Under this process, a schizont is formed within the infected RBC [Figure 1.6c], culminating in the generation and eventual release of 10 - 16 daughter merozoite clones into the host's circulation, each daughter, ready to invade a new RBC [41, 55, 56, 62]. While other human-adapted *Plasmodium* spp. generate varying numbers of merozoites in their schizonts, here, *P. knowlesi* and *P. malariae* have similar (but not identical) morphological organisation, size and output of schizonts [41]. However, with all similarities between

P. knowlesi and its counterparts, it is unique in the speed at which the cycle of 'invade-replicate-rupture' occurs [Figure 1.5c]. This cycle typically occurs every 48 - 72 hours in *P. falciparum*, *P. vivax*, *P. malariae* and *P. ovale* however, *P. knowlesi* is the only known primate malaria that is quotidian i.e. it possesses a 24-hour cycle [41, 55, 56, 62].

During merozoite release from infected RBCs, it is theorised that a small proportion differentiates further into male and female gametocytes [Figure 1.5d, Figure 1.6d] to be taken up by a feeding mosquito. Here, the immature gametocytes require a short period to become mature [55], although, to our knowledge, little is known about this aspect of *P. knowlesi*'s development. It is possible that *P. knowlesi* gametocytes sequester in the bone marrow for maturation and eventually coalesce within capillaries beneath the skin to foster mosquito uptake, although this has only been documented in *P. falciparum* [63, 64]. Within the mosquito, the male gamete further differentiates and multiplies to fertilise female gametes, forming a zygote, which in turn develops into a motile ookinete [Figure 1.5e] [46, 55, 65]. The ookinete proceeds to penetrate the gut wall and form an oocyst [Figure 1.5f]. Within the oocyst, asexual replication occurs to form thousands of haploid sporozoite progeny [2, 46, 55]. Once formed, the sporozoites rupture the oocyst and are released into the haemolymph of the mosquito before migrating to the acinar cells of the salivary glands [2, 46, 55, 66]. From the salivary glands, up to 100 sporozoites can be injected into a primate host during the mosquito's blood meal [Figure 1.5g] [2, 46, 55].

1.2.2 Pathophysiology and Clinical presentation

The rupturing of the infected RBCs acts as a signal to the host, indicating the presence of a foreign entity. This is mainly due to the release of intracellular structures like haemozoin and other toxins, which are also released into the bloodstream during RBC egress, i.e. the release of parasites from infected red blood cells (*i*RBC) [55]. The host's immune system is signalled, setting off the cascade of symptomatic manifestations described for malaria. Initially, it was thought that *P. knowlesi* infections in humans self-resolved spontaneously two weeks after inoculation [32, 41, 46]. As with other malaria species, *P. knowlesi* infections are characterised by the incitement of a fever (generally between 40 - 41°C), chills, headache, general myalgia and malaise [52]. However, with *P. knowlesi*'s

quotidian nature, the observed fever is renewed every 24 hours.

Indeed, with such clinical features, it is important to provide rapid treatment to the infected individual, as, if left untreated, *Plasmodium* spp. and *P. knowlesi* infections in particular, can become irrevocably damaging and fatal [55, 67]. Other less common symptoms reported for *P. knowlesi* include nausea, abdominal pain, coughs, breathlessness, hyponatremia, diarrhoea, thrombocytopenia and vomiting [67]. Additionally, clinical observations associated with *P. knowlesi* infections include the development of palpable spleen, palpable liver, an increased respiratory rate and tachycardia, with hepatomegaly and splenomegaly seen in 15 - 40 % of patients [67].

Severe malaria is one of the categories into which malaria can be classified. The others, asymptomatic and uncomplicated malaria, represent malaria infections that are relatively common and treatable, with mild complications [68]. Indeed, asymptomatic malaria involves no observable disease manifestations, often due to low immunity levels in the infected individual. On the other hand, uncomplicated malaria is symptomatic, resulting in the classical hallmark symptoms of malaria mentioned above.

Clinical features of Severe malaria

When uncomplicated malaria is left untreated to progress further, severe malaria can develop, resulting in a considerably more complicated outcome for the patient. Here, symptoms transform to become more wide-ranging to include presentations such as respiratory distress, convulsions, organ dysfunction, hyperparasitaemia and coma [Table 1.1] [30]. However, the development of coma is only present in *P. falciparum* and is largely due to its ability to cause sequestration of infected RBCs within the brain, resulting in cerebral malaria [Table 1.1] [69]. While sequestration of infected RBCs has been documented in humans (*post-mortem*), the development of coma has not been observed in *P. knowlesi* infections in humans [70]

Hyperparasitaemia is often thought to be the most indicative symptom of severe malaria, however this is more the case in *P. falciparum*. In *P. knowlesi*, severe malaria is also associated with hyperparasitaemia, however severely ill patients with relatively low parasitaemia of 15,000 parasites/μL have been reported to develop severe malaria [26,

Table 1.1: Complications observed in severe malaria caused by *P. falciparum* and *P. knowlesi*

Complications	*Plasmodium falciparum*	*Plasmodium knowlesi*
Hyperparasitemia	✓	✓
Anaemia	✓	✓
Acute renal injury	✓	✓
Respiratory distress	✓	✓
Hypotension	✓	✓
Jaundice	✓	✓
Hyperlactaemia	✓	✓
Coma	✓	×

Described symptoms of severe malaria in infected patients of *P. falciparum* and *P. knowlesi*, showing the only difference between the two species is *P. falciparum*'s ability to incite a coma, which *P. knowlesi* is unable to do. Source: Cox-Singh and Culleton [45]

71]. Willmann et al. [72] concluded that patients recorded with parasitaemias >20,000 parasites/μL are at an increased risk of developing complicated *P. knowlesi* malaria. The low parasitaemia threshold observed in *P. knowlesi* severe infections may contribute to 9 - 29 % of *P. knowlesi* patients developing severe malaria symptoms, with a fatality rate of 0 - 1.9 % [19, 52, 67, 73]. It must be noted that hyperparasitaemia in *P. falciparum* is higher than *P. knowlesi* [49]; however, Barber et al. [73] suggest *P. knowlesi* has a 3-fold greater chance of developing into severe malaria in comparison to *P. falciparum* with parasitaemia and schizontaemia of >10 % being independent predictors of severe malaria observed in *P. knowlesi*.

Sequestration in *P. falciparum* and *P. knowlesi*

Hyperparasitaemia is an important identifier in the propensity for a *P. falciparum* infection to proceed from uncomplicated to complicated and eventually severe malaria. Here, hyperparasitaemia refers to a parasite load of >100,000 parasites/μL in non-endemic/non-immune individuals and >250,000 parasites/μL in regions of high transmission [74]. Such a high parasite load is likely due in part to sequestration, a process by which infected RBCs are removed from circulation by localising in the microvasculature of the

host [75].

Sequestration in *P. falciparum* occurs through cytoadherence, a feature that appears to be unique to *P. falciparum* in human host-adapted *Plasmodium* spp. [70]. Localisation within the microvasculature results in obstruction of these vessels, culminating in localised hypoxia and potential tissue damage [75]. Sequestration and cytoadherence to the vascular endothelium of different tissues result in different forms of *P. falciparum* infections; as an example, sequestration in the capillaries of the brain is thought to result in cerebral malaria [75]. On the other hand, sequestration of *P. falciparum*-infected RBCs in the endothelium of the placenta gives rise to pregnancy-associated malaria (PAM), a leading cause of death in pregnant women of malaria-endemic nations [13].

While sequestration in *P. falciparum* is well documented; this is not the case for *P. knowlesi*. Very little is known about sequestration in *P. knowlesi*, however some indication of partial sequestration has been described by Cox-Singh et al. [70] and Miller et al. [76]. Although sequestration of infected RBCs in the vasculature of the cerebrum, heart and kidney was observed, Cox-Singh et al. [70] describe this within a single fatal case of severe *P. knowlesi*. Thus, it is currently unknown if these observations translate to be a part of the wider *P. knowlesi* repertoire. Furthermore, the observations made in *P. knowlesi* differ from *P. falciparum*, particularly in the other features of the infected RBCs, whereby no platelet clumping was observed [70]; a hallmark of malaria sequestration [76]. Importantly, a lack of up-regulation in the concentration of Intracellular Adhesion Molecule-1 (ICAM-1) within the brain of the host suggests that *P. knowlesi* utilises a different mode of action for sequestration in comparison to *P. falciparum*, where ICAM-1 is the main target on the cerebral endothelium [70, 77]. As such, the mechanisms for sequestration and thus host immune evasion or subversion by *P. knowlesi* requires further study.

1.3 Multiple gene families in malaria parasites

While sequestration in *P. knowlesi* is still not as characterised as in *P. falciparum*, it is thought that they have similar characteristics. This is especially the case due to the historical links within the research of the two species. Historically, *P. knowlesi* acted

as a malaria model for *P. falciparum*; due to its high infectivity in non-human primate models as well as its relative ease of control and expedient yet consistent symptomatic manifestations [46]. Many *Plasmodium* spp. mechanisms such as antigenic variation were first hypothesised or indeed discovered in *P. knowlesi* [62]. Using antigenic variation, *P. falciparum* can evade host immune response, thus, prolonging the infection. However, antigenic variation was first described in *P. knowlesi* through the agglutination of infected RBCs facilitated by Schizont Infected Cell Agglutination (SICA) proteins [78]. By demonstrating the expression of a family of variant genes in consecutive waves of parasitaemia, Brown and Brown [78] were able to show variant-specific interactions of infected RBCs to mediate host evasion. A similar family of genes showing these variations were later discovered to be present in *P. falciparum*, prompting questions as to the similarity of their loci, function and mechanism of action.

1.3.1 The *P. falciparum* variant (var) genes

Plasmodium spp. employ a wide variety of genes to infect and propagate in a given host environment however, the genes encoding host evasion are arguably the most important for infection prolongation and survivability in hostile environments. In *P. falciparum*, evasion mainly occurs due to sequestration (*see chapter 1 subsubsection 1.2.2*), which is facilitated by the *var* multigene family. The *Plasmodium falciparum* variable gene (*Pfvar*) multigene family are a group of approximately 60 variable genes dispersed across the 14 chromosomes of the *P. falciparum* genome [79, 80]. Each member of the *var* genes family encodes a variant of the *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (PfEMP1) protein that is expressed on the surface of infected RBCs, with some variants better suited for specific host environments [Figure 1.7] [81].

However, not all variant forms of PfEMP1 can be expressed at the same time. Indeed, only one major variant is translated and expressed at a time for each infected RBC; although, due to the nature of the protein acting as both an antigen and adhesin, the host's immune system is triggered. By targeting and binding to surface receptors on the endothelium of the human host's microvasculature, PfEMP1 allows the infected RBCs to evade clearance by the spleen [Figure 1.7] [81]. In the event where PfEMP1 is inactive, absent or lacking targets in the local vasculature, infected RBCs are filtered through the

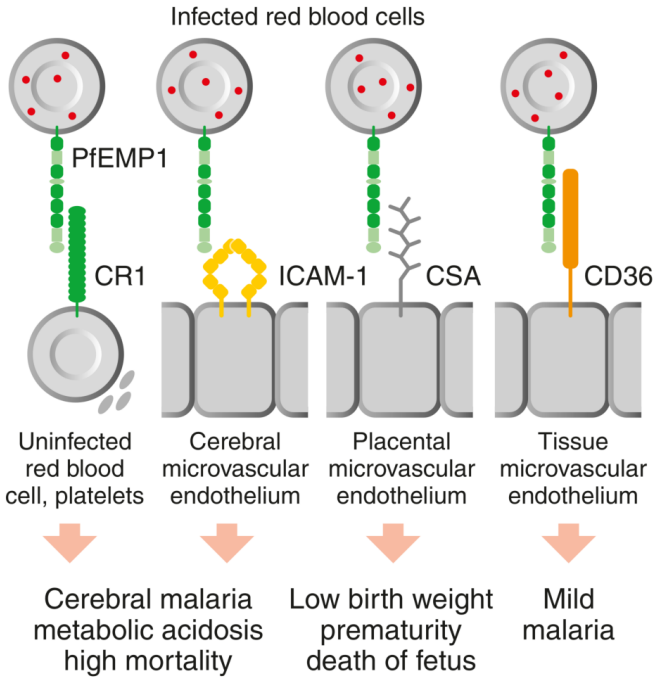


Figure 1.7: Binding sites present on different human host endothelial environments. Utilising *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (PfEMP1), *P. falciparum* can bind to various host endothelia. Using different variants of *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (PfEMP1), infected RBCs can bind to the adhesion molecules of different tissues, which have vastly different configurations. Differential symptoms are observed depending on the tissue the parasites bind to and subsequently the type of malaria incited. Source: Penman and Gupta [77].

spleen and targeted for destruction. Were multiple variants of *Pfvar* genes (encoding several variants of PfEMP1) be expressed at a time, this would expose the host to the entirety of the parasite's repertoire, potentially reducing the effectiveness of antigenic variation [82, 83]. It is suggested that this is epigenetically controlled as a means of ensuring switching between *Pfvar* genes occurs at a low rate to avoid exhaustion of the parasite's *Pfvar* repertoire [82, 83]. The process of silencing variant PfEMP1 genes from being transcribed is allelic exclusion, and the subsequent process of switching the variant form is antigenic variation [82].

The parasite expresses and constructs PfEMP1 after successfully invading a RBC before the protein is transported to the surface of the infected RBC. At the surface, if a

corresponding endothelial adhesin is present, PfEMP1 initiates cytoadherence, a process where the infected RBC binds to the vascular wall to stop circulating through the host [Figure 1.7] [81, 83]. Where a suitable endothelial adhesin is not present, the infected RBC continues to circulate through the host and is eventually cleared within the red pulp of the spleen by splenic macrophages and monocytes, which may, in turn, induce a cytokine storm [84–86].

However, before splenic clearance or when cytoadherence occurs, circulating and sequestered parasitised RBCs can induce malaria of differing severity, depending on the location within the human host [Figure 1.7]. This results in differential observable disease outcomes for the patient and a potentially different constellation of symptoms [Figure 1.7]. As an example, cerebral malaria and pregnancy-associated malaria (PAM) are two clinical outcomes of *P. falciparum* associated with sequestration of the parasite and infected RBCs in the cerebrum and placenta, respectively. In particular, pregnancy-associated malaria (PAM) is a form of *P. falciparum* malaria in pregnant women of malaria-endemic regions that is characterised by maternal anaemia, low birth weight and high mortality [30, 81]. PAM is facilitated by the sequestration of infected RBCs in the placenta of the mother via the action of VAR2CSA, a specific variant of PfEMP1 encoded by the *var2csa* gene [56]. Once expressed on the surface of the infected RBC, VAR2CSA targets and binds chondroitin sulphate A (CSA), a surface protein readily expressed on the surface of placental endothelial cells [Figure 1.7]. Interestingly, fully functioning copies of the gene *var2csa* has been shown to be distributed across the *P. falciparum* genome [80, 87]. The impact of multiple copies of this gene remains a point of study; however, these are chiefly being carried out on clinical patient samples (venous, cord and placental blood) of *P. falciparum* infected pregnant women [80, 88–90].

1.3.2 The *Schizont-Infected Cell Agglutination variant* (SICAvar) genes

As previously stated, *P. knowlesi* acted as a *de facto* model for human *Plasmodium* spp. infections; specifically *P. falciparum* malaria [91]. So much so that antigenic variation and the *Pfvar* genes associated with the process were first discovered in *P. knowlesi* [91,

92]. Here, Howard, Barnwell, and Kao [92] described a link between proteins expressed on the surface of infected RBCs and observed immune evasion in laboratory infected rhesus macaques. Further studies by Barnwell et al. [93] proceeded to suggest that the presence of these surface proteins could facilitate prolongment of the infection within the infected host. Over the years, further work has revealed that Schizont Infected Cell Agglutination variant antigen (SICAvar) proteins are analogous and evolutionarily related to the *Pfvar* genes [91, 94, 95].

Currently, the *SICAvar*s are the subject of much interest, partly due to their similarity to the *Pfvar* genes and the growing importance of *P. knowlesi* in Southeast Asia. However, the true role of the proteins being encoded by this family of genes remains poorly characterised. Brown and Brown [78] and Howard, Barnwell, and Kao [92] show evidence of their involvement in the agglutination of *P. knowlesi*-infected RBCs; however, it remains to be seen if this agglutination acts as a form of sequestration or if it serves another purpose (see *chapter 1 subsubsection 1.2.2*) Indeed, both gene families share the same characteristics which have been described, with the only exception to date being the association of PfEMP1 with cytoadherence and the *SICAvar*'s lack of evidence for this feature [Table 1.2].

It is now understood that Schizont Infected Cell Agglutination (SICA) proteins (products of the *SICAvar* genes) are relatively large proteins between 180 - 220 kDa expressed on the surface of trophozoites and early schizont-infected RBCs [91, 98]. They are widespread across all 14 chromosomes of *P. knowlesi* and unlike *P. falciparum* which possesses a well-characterised *Pfvar* gene family, *P. knowlesi*'s *SICAvar*s requires considerably more research. An example of this being the increase in the number of *SICAvar*s found in the *P. knowlesi* whole genome with successive research [99, 100].

The difficulty in quantifying the number and precise location of the *SICAvar* genes lies in their variable nature. Additionally, due to the relatively recent confirmation of naturally occurring *P. knowlesi* infection in humans, little is known about *SICAvar*s in wild-type parasites. *SICAvar* variability lies in the low complexity yet highly repetitive sequences in large portions of the gene, including some *SICAvar* variants with very large introns. Nevertheless, the first *P. knowlesi* sequence genome (Pain et al. [99]) attempted quantification, resulting in a combined 107 full-length and fragmented *SICAvar* genes.

Table 1.2: Features described to be present in encoded proteins of the *SICAvar* and Pfvar multigene families

Variant antigen characteristic	SICA protein	PfEMP1
Large parasite-encoded proteins	✓	✓
Exposed at infected RBC surface	✓	✓
Extractable only by ionic detergents	✓	✓
Undergo antigenic variation	✓	✓
Associated with virulence	✓	✓
Encoded by large multigene family	✓	✓
Contain cysteine-rich domains	✓	✓
Subject to much diversity	✓	✓
Genes on all chromosomes	✓	✓
In situ transcription	✓	✓
Abundant transcription in ring stage	✓	✓
Translation in trophozoite stage	✓	✓
Antisense long non-coding RNAs in gene regulation	✓	✓
Epigenetic gene control	✓	✓
Role for spleen in expression control	✓	✓
Associated with typical cytoadherence/sequestration	×[a]	✓

Features previously described to be present in either the *Pfvar* or *SICAvar* encoded proteins. Previously reviewed by Korir and Galinski [91] and Corredor et al. [96]. Source: Galinski et al. [97].
a - Some evidence of sequestration has been seen in *P. knowlesi* as described by Cox-Singh et al. [70] and Miller et al. [76]

However, due to the limitations of the Sanger sequencing technology used, Pain et al. [99] were unable to resolve all of the *SICAvar* proteins. Recently, Lapp et al. [100] carried out another *P. knowlesi* whole genome sequencing, using Pacific Biosciences (PacBio) long-read sequencing. This sequencing technology will be further explored in chapter 3 subsection 3.1.4.

From this, Lapp et al. [100] were able to describe 136 full-length *SICAvar* genes within *P. knowlesi*. As with other organisms, including *P. falciparum*, long-read sequencing technology allowed the generation of reads that could span regions of low complexity or highly repetitive DNA (to varying levels of accuracy). Additionally, Lapp et al. [100] further classified the *SICAvar*'s two subtypes; based on the number of exons the gene

possesses [99]. From their generated experimental line-derived whole-genome assembly (*Malayan strain*), *SICAvar* type I genes were reported to possess 5 - 16 exons with large introns while type II genes possess 3 - 4 exons [100]. Thus, of the 136 total complete *SICAvar* genes detailed by Lapp et al. [100], 117 and 19 were classified as *SICAvar* type I and type II respectively. However, the genome was released in an incomplete form, with only pseudochromosomes rather than chromosomes.

Notwithstanding, the true purpose and means of action of the SICAvar protein is still poorly described, although a considerable amount of research has been carried out over the years. However, it must be noted that this research has primarily only occurred on the experimental lines, which were first isolated over four decades ago. Hence queries arise as to the viability of potentially evasion protein-encoding genes (*SICAvar*s) which have been effectively removed from selective pressure. Thus, the *SICAvar* genes and proteins must be investigated within contemporary patient samples.

1.3.3 *Plasmodium knowlesi* interspersed repeat (KIR) genes

The *kir* genes are the second largest multiple gene variant family currently identified within *P. knowlesi*, encoding proteins of varying sizes (36 - 97 kDa) that have been observed to be expressed on the surface of infected RBCs [99, 101]. *kir* genes are also members of the *Plasmodium* spp. interspersed repeat (*pir*) gene family, the largest *Plasmodium* spp. variant protein family [102–104]. Hence, while the *Pfvar* and *SICAvar* genes are only found in *P. falciparum* and *P. knowlesi* respectively, members of the *pir* gene family are ubiquitous in all *Plasmodium* spp. sequenced so far apart from parasites of the *Laverania* sub-species [102, 104].

Indeed, recent studies have determined that the *pir* superfamily, which consists of *cir*, *yir*, *bir*, *vir* and *kir* from *P. chabaudi*, *P. yoelii*, *P. berghei*, *P. vivax* and *P. knowlesi* respectively, may have evolved after *Laverania* and bird-infective *Plasmodia* had diverged from the rodent and primate-infective *Plasmodia* [102]. This would support recent suggestions that *rif*/*stevor* genes from *P. falciparum* are not members of the *pir* superfamily even though, they had initially been considered to be similar enough to be classed together [102, 103]. Recent investigations into the structures of *pir*s have shown only a 1 % likelihood that

rif/*stevor* gene architecture are similar to other members of the *pir* superfamily [102].

However, even though the *pir* superfamily is so widespread, very little is known about their function or mechanism of action [102, 103]. Considering their abundance in their respective genomes – where *P. knowlesi* has 68, *P. vivax* has 350 and *P. yoelii* has 800 –, they may carry a key function within the parasite [104]. It must be said that much like the *SICAvar*s, the *Kir*s have very little direct research but rather much of what is known of them, is inferred from the research of other *Plasmodium* spp.; in this case, *Plasmodium vivax interspersed repeat (vir)* genes, which occupy up to 10 % of the*P. vivax* coding genome [101]. Although *vir* genes are the most diverse group in the superfamily, they are the most similar to the *kir* genes [103, 105].

Nevertheless, what is known about the *pir* supergene family and, by extension, the *kir*s suggests they are involved in multiple processes throughout the development of the parasite, with *cir* genes in *P. chabaudi* being detected as early as the ring stage of the parasite life cycle [104]. Other studies have associated PIR proteins with antigenic variation, cytoadherence and even host cell invasion [102, 104, 105]. With this, it is clear that the expression of *pir* genes is less restrictive than that of *SICAvar*s that may also be controlled by antigenic variation, however, given the large copy numbers of *pir*s, it is unlikely that they are expressed at the same time. Some evidence to support this has been observed where less than 50 % of *vir*s possess the *Plasmodium* export element (PEXEL) motif essential for transporting proteins to the extracellular surface of infected RBCs [101, 104]. This suggests that although part of the same family, there remains considerable variation within the *vir* genes, which may, in turn, be present in the *kir* genes of *P. knowlesi*.

Although potentially variant within the multigene family, Pain et al. [99] described an unusual but definitive feature of the *kir* genes. The KIR amino acid sequence was found to contain stretches of up to 36 amino acids which perfectly matched portions of the extracellular domain of the human CD99 molecule [99]. Such a development is undoubtedly a source of intrigue as *Plasmodium* spp. are known to modulate the host immune system; however, this type of structural similarity appears to be the first example of a direct interaction of parasite-derived host peptides and the host immune system [99]. With this hypothesis, Pain et al. [99] suggest that KIR proteins function to facilitate

host immune response reduction by inhibiting T-cell activity through competition for activation.

1.4 Previous work within the Cox-Singh group

The Cox-Singh group has largely focussed on the research and understanding of zoonotic malaria caused by *Plasmodium knowlesi*. Over time, this has resulted in multiple projects furthering the knowledge of this little known yet important organism. Recently the group described a genetic dimorphism within one of the two copies of the *P. knowlesi* normocyte binding protein gene (*Pknbp*) (Xa and Xb) [47]. From this initial discovery, the group expanded the confirmation of dimorphism genome-wide. Here, six frozen patient isolates from a *P. knowlesi* clinical patient biobank were sequenced using Illumina HiSeq and MiSeq technology [106]. Whole-genome assembly was carried out on the resulting sequence data and subsequently, full-length *PknbpXa* in which dimorphism was shown [106]. From this, Pinheiro et al. [106] determined that $>50\,\%$ of the *P. knowlesi* genome is dimorphic, encompassing all chromosomes and genes, thus suggesting two forms of *P. knowlesi* are causing disease in and around the study site (Sarawak, Malaysian Borneo) [106]. This was corroborated by Divis et al. [107] and Divis et al. [108], revealing the two intra-species clusters present in *P. knowlesi*. Importantly, Pinheiro et al. [106] proved frozen archival patient samples are amenable for generating high-quality genome sequence data to further *P. knowlesi* research. Concurrently, Millar [109] was able to successfully transfect an artificial *PknbpXa* gene identified by Pinheiro et al. [106] into a laboratory-restricted experimental strain of *P. knowlesi* (Millar [109]:unpublished thesis). In achieving this, Millar [109] was able to expand and optimise the purification method developed by Pinheiro et al. [106] and demonstrate the ability of an experimental line to express clinically relevant alleles. Thus, allowing for functional, controllable and reproducible analysis of clinically relevant mutations expressed in experimental lines of *P. knowlesi*, reconstituted without loss of precious clinical material [109].

1.5 Project Motivation, Hypothesis, Aims and Objectives for this project

While the group successfully purified patient isolates for whole-genome assemblies using Illumina short-read sequencing [106, 109], the limitations of short-read sequencing to study large genetic variations and variable multigene families persisted. Additionally, substantial losses of the precious parasite nuclei acid were made during the extraction process. With this, the next step was to develop a means of reducing or eliminating the loss of genetic material in a bid to describe these genetic variations and resolve the multigene loci within the *P. knowlesi* genome. A question was posed: can we investigate variant multiple gene families within the *Plasmodium knowlesi* genome using clinical patient isolates?

To achieve this, we aim to employ third-generation sequencing in the form of Oxford Nanopore long-read sequencing. The benefit of utilising long-read sequencing lies in the length of the sequence reads produced. The use of PacBio long-read sequencing was not viable within this project due to insufficient whole blood volume the procedure would require – further supporting our use of Oxford Nanopore's platform.

To generate high-quality *P. knowlesi* genome sequences, parasite DNA of adequate volumes and high quality are to be extracted from the archived frozen whole blood in our biobank. To achieve this, a method to deplete human leucocytic content from the whole blood using a relatively simple antibody-antigen interaction will be developed. The method is to be optimised to determine the most appropriate usage in depleting human leucocytes and the isolation/retention of parasites within the remnant blood. Using a ready-made antibody, the method will aim to standardise the depletion process whilst minimising the introduction of adverse chemicals into the blood. Furthermore, depleting leucocytes before genome sequencing will allow the generation of a purified, enriched parasite DNA eluate, facilitating better yield from the long-read sequencing platform.

Once depletion and extraction is sufficiently optimised, *de novo* whole-genome assemblies of *P. knowlesi* infected patient isolates will be implemented. This would involve the development of a robust *in silico* pipeline capable of carrying out this process with

minimal interaction from the researcher. In order to ensure adequate development, assessments of tools and methods are to be implemented throughout the project and at each 'major' step of the pipeline. As such, the pipeline will be able to reproduce the analyses carried out for the same sequence data to be generated in this project or for new *P. knowlesi* sequence data.

Finally, using the generated *de novo* whole genomes, an investigation into the multiple gene families that are present within the generated complete genomes will be carried out. Here, we use custom scripts and analyses to determine their abundance and loci as well as other information for both the Schizont Infected Cell Agglutination variant antigen (SICAvar) and *Plasmodium knowlesi* interspersed repeat (*kir*) multiple gene families used and recorded for reproducibility.

1.6 References

[1] **INSTITUTE OF MEDICINE (US) COMMITTEE ON THE ECONOMICS OF ANTIMALARIAL DRUGS**. *A Brief History of Malaria*. Ed. by **K. J. ARROW**, **C. PANOSIAN**, and **H. GELBAND**. National Academies Press (US), 2004 (see pp. 1, 5, 6)

[2] **F. E. COX**. "History of the Discovery of the Malaria Parasites and Their Vectors". In: *Parasites & Vectors* **3**:1 (Feb. 2010), 5. DOI: 10.1186/1756-3305-3-5 (see pp. 1, 17)

[3] **D. SOREN**. "Can Archaeologists Excavate Evidence of Malaria?" In: *World Archaeology* **35**:2 (2003), 193–209 (see p. 1)

[4] **M. IURESCIA**, **F. ROMITI**, **C. COCUMELLI**, **E. L. DIACONU**, **F. STRAVINO**, **R. ONORATI**, **P. ALBA**, **K. G. FRIEDRICH**, **F. MAGGI**, **A. MAGLIANO**, **A. ERMENEGILDI**, **V. CARFORA**, **A. CAPRIOLI**, **C. DE LIBERATO**, and **A. BATTISTI**. "Plasmodium Matutinum Transmitted by Culex Pipiens as a Cause of Avian Malaria in Captive African Penguins (Spheniscus Demersus) in Italy". In: *Frontiers in Veterinary Science* **8**: (2021), 198. DOI: 10.3389/fvets.2021.621974 (see pp. 2, 11)

[5] **R. GUTIÉRREZ-LÓPEZ**, **J. M.-D. LA PUENTE**, **L. GANGOSO**, **R. SORIGUER**, and **J. FIGUEROLA**. "Plasmodium Transmission Differs between Mosquito Species and Parasite Lineages". In: *Parasitology* **147**:4 (Apr. 2020), 441–447. DOI: 10.1017/S0031182020000062 (see pp. 2, 11)

[6] **J. C. RAYNER**, **W. LIU**, **M. PEETERS**, **P. M. SHARP**, and **B. H. HAHN**. "A Plethora of Plasmodium Species in Wild Apes: A Source of Human Infection?" In: *Trends in Parasitology* **27**:5 (May 2011), 222–229. DOI: 10.1016/j.pt.2011.01.006 (see pp. 2–4)

[7] **B. SINGH**, **L. K. SUNG**, **A. MATUSOP**, **A. RADHAKRISHNAN**, **S. S. SHAMSUL**, **J. COX-SINGH**, **A. THOMAS**, and **D. J. CONWAY**. "A Large Focus of Naturally Acquired Plasmodium Knowlesi Infections in Human Beings". In: *The Lancet* **363**:9414 (2004), 1017–1024 (see pp. 2, 10, 11)

[8] **P. BRASIL**, **M. G. ZALIS**, **A. DE PINA-COSTA**, **A. M. SIQUEIRA**, **C. B. JÚNIOR**, **S. SILVA**, **A. L. L. AREAS**, **M. PELAJO-MACHADO**, **D. A. M. DE ALVARENGA**, **A. C. F. D. S. SANTELLI**, **H. G. ALBUQUERQUE**, **P. CRAVO**, **F. V. S. DE ABREU**, **C. L. PETERKA**, **G. M. ZANINI**, **M. C. S. MUTIS**, **A. PISSINATTI**, **R. LOURENÇO-DE-OLIVEIRA**, **C. F. A. DE BRITO**, **M. D. F. FERREIRA-DA-CRUZ**, **R. CULLETON**, and **C. T. DANIEL-RIBEIRO**. "Outbreak of Human Malaria Caused by Plasmodium Simium in the Atlantic Forest in Rio de Janeiro: A Molecular Epidemiological Investigation". In: *The Lancet Global Health* **5**:10 (Oct. 2017), e1038–e1046. DOI: 10.1016/S2214-109X(17)30333-9 (see p. 2)

[9] **J. VAN AS**, **C. A. COOK**, **E. C. NETHERLANDS**, and **N. J. SMIT**. "A New Lizard Malaria Parasite Plasmodium Intabazwe n. Sp. (Apicomplexa: Haemospororida: Plasmodiidae) in the Afromontane Pseudocordylus Melanotus (Sauria: Cordylidae) with a Review of African Saurian Malaria Parasites". In: *Parasites & Vectors* **9**: (Aug. 2016), 437. DOI: 10.1186/s13071-016-1702-3 (see p. 2)

[10] **H. R. ANSARI**, **T. J. TEMPLETON**, **A. K. SUBUDHI**, **A. RAMAPRASAD**, **J. TANG**, **F. LU**, **R. NAEEM**, **Y. HASHISH**, **M. C. OGUIKE**, **E. D. BENAVENTE**, **T. G. CLARK**, **C. J. SUTHERLAND**, **J. W. BARNWELL**, **R. CULLETON**, **J. CAO**, and **A. PAIN**. "Genome-Scale Comparison of Expanded Gene Families in

Plasmodium Ovale Wallikeri and Plasmodium Ovale Curtisi with Plasmodium Malariae and with Other Plasmodium Species". In: *International Journal for Parasitology* **46**:11 (Oct. 2016), 685–696. DOI: 10.1016/j.ijpara.2016.05.009 (see pp. 2, 5)

[11] **U. FREVERT**. "Sneaking in through the Back Entrance: The Biology of Malaria Liver Stages". In: *Trends in Parasitology* **20**:9 (Sept. 2004), 417–424. DOI: 10.1016/j.pt.2004.07.007 (see p. 2)

[12] **J. COX-SINGH**, **T. M. E. DAVIS**, **K.-S. LEE**, **S. S. G. SHAMSUL**, **A. MATUSOP**, **S. RATNAM**, **H. A. RAHMAN**, **D. J. CONWAY**, and **B. SINGH**. "Plasmodium Knowlesi Malaria in Humans Is Widely Distributed and Potentially Life Threatening". In: *Clinical Infectious Diseases* **46**:2 (Jan. 2008), 165–171. DOI: 10.1086/524888 (see p. 2)

[13] **WORLD HEALTH ORGANIZATION**. *World Malaria Report 2020*. Global Health. World Health Organisation, 2020, 300 (see pp. 2, 3, 6–9, 20)

[14] **W. LIU**, **Y. LI**, **K. S. SHAW**, **G. H. LEARN**, **L. J. PLENDERLEITH**, **J. A. MALENKE**, **S. A. SUNDARARAMAN**, **M. A. RAMIREZ**, **P. A. CRYSTAL**, **A. G. SMITH**, **F. BIBOLLET-RUCHE**, **A. AYOUBA**, **S. LOCATELLI**, **A. ESTEBAN**, **F. MOUACHA**, **E. GUICHET**, **C. BUTEL**, **S. AHUKA-MUNDEKE**, **B.-I. INOGWABINI**, **J.-B. N. NDJANGO**, **S. SPEEDE**, **C. M. SANZ**, **D. B. MORGAN**, **M. K. GONDER**, **P. J. KRANZUSCH**, **P. D. WALSH**, **A. V. GEORGIEV**, **M. N. MULLER**, **A. K. PIEL**, **F. A. STEWART**, **M. L. WILSON**, **A. E. PUSEY**, **L. CUI**, **Z. WANG**, **A. FÄRNERT**, **C. J. SUTHERLAND**, **D. NOLDER**, **J. A. HART**, **T. B. HART**, **P. BERTOLANI**, **A. GILLIS**, **M. LEBRETON**, **B. TAFON**, **J. KIYANG**, **C. F. DJOKO**, **B. S. SCHNEIDER**, **N. D. WOLFE**, **E. MPOUDI-NGOLE**, **E. DELAPORTE**, **R. CARTER**, **R. L. CULLETON**, **G. M. SHAW**, **J. C. RAYNER**, **M. PEETERS**, **B. H. HAHN**, and **P. M. SHARP**. "African Origin of the Malaria Parasite Plasmodium Vivax". In: *Nature Communications* **5**:1 (May 2014), 3346. DOI: 10.1038/ncomms4346 (see pp. 3, 4)

[15] **L. MENKIN-SMITH** and **W. T. WINDERS**. "Plasmodium Vivax Malaria". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2021 (see pp. 3, 4)

[16] **R. CULLETON**, **M. NDOUNGA**, **F. Y. ZEYREK**, **C. COBAN**, **P. N. CASIMIRO**, **S. TAKEO**, **T. TSUBOI**, **A. YADAVA**, **R. CARTER**, and **K. TANABE**. "Evidence for the Transmission of Plasmodium Vivax in the Republic of the Congo, West Central Africa". In: *The Journal of Infectious Diseases* **200**:9 (Oct. 2009), 1465–1469. DOI: 10.1086/644510 (see pp. 3, 4)

[17] **D. MÉNARD**, **C. BARNADAS**, **C. BOUCHIER**, **C. HENRY-HALLDIN**, **L. R. GRAY**, **A. RATSIMBASOA**, **V. THONIER**, **J.-F. CAROD**, **O. DOMARLE**, **Y. COLIN**, **O. BERTRAND**, **J. PICOT**, **C. L. KING**, **B. T. GRIMBERG**, **O. MERCEREAU-PUIJALON**, and **P. A. ZIMMERMAN**. "Plasmodium Vivax Clinical Malaria Is Commonly Observed in Duffy-Negative Malagasy People". In: *Proceedings of the National Academy of Sciences of the United States of America* **107**:13 (Mar. 2010), 5967–5971. DOI: 10.1073/pnas.0912496107 (see p. 4)

[18] **S. A. SUNDARARAMAN**, **W. LIU**, **B. F. KEELE**, **G. H. LEARN**, **K. BITTINGER**, **F. MOUACHA**, **S. AHUKA-MUNDEKE**, **M. MANSKE**, **S. SHERRILL-MIX**, **Y. LI**, **J. A. MALENKE**, **E. DELAPORTE**, **C. LAURENT**, **E. M. NGOLE**, **D. P. KWIATKOWSKI**, **G. M. SHAW**, **J. C. RAYNER**, **M. PEETERS**, **P. M. SHARP**, **F. D. BUSHMAN**, and **B. H. HAHN**. "Plasmodium Falciparum-like Parasites Infecting Wild Apes in Southern Cameroon Do Not Represent a Recurrent Source of Human Malaria". In: *Proceedings of the National Academy of Sciences* **110**:17 (Apr. 2013), 7020–7025. DOI: 10.1073/pnas.1305201110 (see p. 4)

[19] **B. A. RAHIMI**, **A. THAKKINSTIAN**, **N. J. WHITE**, **C. SIRIVICHAYAKUL**, **A. M. DONDORP**, and **W. CHOKEJINDACHAI**. "Severe Vivax Malaria: A Systematic Review and Meta-Analysis of Clinical Studies since 1900". In: *Malaria Journal* **13**:1 (Dec. 2014), 481. DOI: 10.1186/1475-2875-13-481 (see pp. 4, 19)

[20] **J. K. BAIRD**. "Evidence and Implications of Mortality Associated with Acute Plasmodium Vivax Malaria". In: *Clinical Microbiology Reviews* **26**:1 (Jan. 2013), 36–57. DOI: 10.1128/CMR.00074-12 (see p. 4)

[21] **M. C. BRUCE**, **A. MACHESO**, **M. R. GALINSKI**, and **J. W. BARNWELL**. "Characterization and Application of Multiple Genetic Markers for *Plasmodium*

Malariae". In: *Parasitology* **134**:5 (May 2006), 637–650. DOI: 10.1017/S0031182006001958 (see p. 4)

[22] **W. E. COLLINS** and **G. M. JEFFERY**. "Plasmodium Malariae: Parasite and Disease". In: *Clinical Microbiology Reviews* **20**:4 (Oct. 2007), 579–592. DOI: 10.1128/CMR.00027-07 (see p. 4)

[23] **DIVISION OF PARASITIC DISEASES AND MALARIA**. *Plasmodium Ovale Bench Aid*. 2020 (see p. 5)

[24] **W. E. COLLINS** and **G. M. JEFFERY**. "Plasmodium Ovale: Parasite and Disease". In: *Clinical Microbiology Reviews* **18**:3 (July 2005), 570–581. DOI: 10.1128/CMR.18.3.570-581.2005 (see pp. 4, 5)

[25] **J. A. GARRIDO-CARDENAS**, **L. GONZÁLEZ-CERÓN**, **F. MANZANO-AGUGLIARO**, and **C. MESA-VALLE**. "Plasmodium Genomics: An Approach for Learning about and Ending Human Malaria". In: *Parasitology Research* **118**:1 (Jan. 2019), 1–27. DOI: 10.1007/s00436-018-6127-9 (see pp. 4, 5)

[26] **D. R. ORESEGUN**, **C. DANESHVAR**, and **J. COX-SINGH**. "Plasmodium Knowlesi – Clinical Isolate Genome Sequencing to Inform Translational Same-Species Model System for Severe Malaria". In: *Frontiers in Cellular and Infection Microbiology* **11**: (2021). DOI: 10.3389/fcimb.2021.607686 (see pp. 5, 9, 10, 18)

[27] **A. Z. CHIN**, **M. C. M. MALUDA**, **J. JELIP**, **M. S. B. JEFFREE**, **R. CULLETON**, and **K. AHMED**. "Malaria Elimination in Malaysia and the Rising Threat of Plasmodium Knowlesi". In: *Journal of Physiological Anthropology* **39**:1 (Nov. 2020), 36. DOI: 10.1186/s40101-020-00247-5 (see pp. 5, 6)

[28] **WORLD HEALTH ORGANISATION**. *WHO — Malaria*. http://www.who.int/ith/diseases/malaria 2018 (see pp. 6, 8)

[29] **N. K. JEYAPRAKASAM**, **J. W. K. LIEW**, **V. L. LOW**, **W.-Y. WAN-SULAIMAN**, and **I. VYTHILINGAM**. "Plasmodium Knowlesi Infecting Humans in Southeast Asia: What's next?" In: *PLOS Neglected Tropical Diseases* **14**:12 (Dec. 2020), e0008900. DOI: 10.1371/journal.pntd.0008900 (see pp. 8, 11, 12)

[30] **WORLD HEALTH ORGANISATION**. *World Malaria Report 2016: Summary*. Tech. rep. World Health Organisation, 2017 (see pp. 8, 18, 23)

[31] **DIVISION OF PARASITIC DISEASES AND MALARIA**. *Malaria*. https://www.cdc.gov/dpdx/malaria/ind Health Information. Oct. 2020 (see pp. 9, 10)

[32] **J. SINTON** and **H. MULLIGAN**. "Plasmodium Knowlesi". In: *The Primate Malarias*. Ed. by **A. GALLEGO**, **W. COLLINS**, **M. WARREN**, and **P. CONTACOS**. Second. Division of Parasitic Diseases, 1971, 317–333 (see pp. 10, 17)

[33] **C. E. VAN ROOYEN** and **G. R. PILE**. "Observations on Infection by Plasmodium Knowlesi (Ape Malaria) in the Treatment of General Paralysis of the Insane". In: *British medical journal* **2**:3901 (1935), 662 (see p. 10)

[34] **R. N. CHOPRA** and **A. S. B. D. GUPTA**. "A Preliminary Note on the Treatment of Neuro-Syphilis with Monkey Malaria". In: *The Indian Medical Gazette* **71**:4 (1936), 187 (see p. 10)

[35] **W. CHIN**, **P. G. CONTACOS**, **G. R. COATNEY**, and **H. R. KIMBALL**. "A Naturally Acquired Quotidian-Type Malaria in Man Transferable to Monkeys". In: *Science* **149**:3686 (Aug. 1965), 865–865. DOI: 10.1126/science.149.3686.865 (see p. 10)

[36] **N. J. WHITE**. "Plasmodium Knowlesi: The Fifth Human Malaria Parasite". In: *Clinical Infectious Diseases* **46**:2 (Jan. 2008), 172–173. DOI: 10.1086/524889 (see p. 10)

[37] **W. E. COLLINS** and **J. W. BARNWELL**. "Plasmodium Knowlesi: Finally Being Recognized". In: *The Journal of Infectious Diseases* **199**:8 (Apr. 2009), 1107–1108. DOI: 10.1086/597415 (see p. 10)

[38] **K.-S. LEE**, **J. COX-SINGH**, and **B. SINGH**. "Morphological Features and Differential Counts of Plasmodium Knowles i Parasites in Naturally Acquired Human Infections". In: *Malaria Journal* **8**:1 (Apr. 2009), 73. DOI: 10.1186/1475-2875-8-73 (see pp. 10, 16)

[39] **B. E. BARBER**, **M. J. GRIGG**, **T. WILLIAM**, **T. W. YEO**, and **N. M. ANSTEY**. "The Treatment of Plasmodium Knowlesi Malaria". In: *Trends in Parasitology* **33**:3 (Mar. 2017), 242–253. DOI: 10.1016/j.pt.2016.09.002 (see pp. 10, 11)

[40] **A. MAHITTIKORN**, **F. R. MASANGKAY**, **K. U. KOTEPUI**, **G. D. J. MILANEZ**, and **M. KOTEPUI**. "Quantification of the Misidentification of Plasmodium Knowlesi as Plasmodium Malariae by Microscopy: An Analysis of 1569 P. Knowlesi Cases". In: *Malaria Journal* **20**: (Apr. 2021), 179. DOI: 10.1186/s12936-021-03714-1 (see p. 11)

[41] **J. K. BAIRD**. "Malaria Zoonoses". In: *Travel Medicine and Infectious Disease* **7**:5 (Sept. 2009), 269–277. DOI: 10.1016/j.tmaid.2009.06.004 (see pp. 11, 15–17)

[42] **CENTERS FOR DISEASE CONTROL AND PREVENTION**. *Malaria*. https://www.cdc.gov/dpdx/m Dec. 2017 (see p. 11)

[43] **M. A. M. SALLUM**, **P. G. FOSTER**, **C. LI**, **R. SITHIPRASASNA**, and **R. C. WILKERSON**. "Phylogeny of the Leucosphyrus Group of *Anopheles* (*Cellia*) (Diptera: Culicidae) Based on Mitochondrial Gene Sequences". In: *Annals of the Entomological Society of America* **100**:1 (Jan. 2007), 27–35. DOI: 10.1603/0013-8746(2007)100[27:POTLGO]2.0.CO;2 (see p. 11)

[44] **M. A. M. SALLUM**, **E. L. PEYTON**, and **R. C. WILKERSON**. "Six New Species of the Anopheles Leucosphyrus Group, Reinterpretation of An. Elegans and Vector Implications". In: *Medical and Veterinary Entomology* **19**:2 (June 2005), 158–199. DOI: 10.1111/j.0269-283X.2005.00551.x (see p. 11)

[45] **J. COX-SINGH** and **R. CULLETON**. "Plasmodium Knowlesi: From Severe Zoonosis to Animal Model". In: *Trends in Parasitology* **31**:6 (June 2015), 232–238. DOI: 10.1016/j.pt.2015.03.003 (see pp. 11, 12, 19)

[46] **W. E. COLLINS**. "Plasmodium Knowlesi: A Malaria Parasite of Monkeys and Humans". In: *Annual Review of Entomology* **57**:1 (2012), 107–121. DOI: 10.1146/annurev-ento-121510-133540 (see pp. 11–14, 17, 21)

[47] **A. M. AHMED**, **M. M. PINHEIRO**, **P. C. DIVIS**, **A. SINER**, **R. ZAINUDIN**, **I. T. WONG**, **C. W. LU**, **S. K. SINGH-KHAIRA**, **S. B. MILLAR**, **S. LYNCH**, **M. WILLMANN**, **B. SINGH**, **S. KRISHNA**, and **J. COX-SINGH**. "Disease Progression in Plasmodium Knowlesi Malaria Is Linked to Variation in Invasion Gene Family Members". In: *PLoS Neglected Tropical Diseases* **8**:8 (Aug. 2014).

Ed. by **K. HIRAYAMA**, e3086. DOI: 10.1371/journal.pntd.0003086 (see pp. 11, 12, 28)

[48] **J. X. DE ANG**, **K. YAMAN**, **K. A. KADIR**, **A. MATUSOP**, and **B. SINGH**. "New Vectors That Are Early Feeders for Plasmodium Knowlesi and Other Simian Malaria Parasites in Sarawak, Malaysian Borneo". In: *Scientific Reports* **11**:1 (Apr. 2021), 7739. DOI: 10.1038/s41598-021-86107-3 (see p. 11)

[49] **D. R. ABEYASINGHE**. *Outcomes from the Evidence Review Group on Plasmodium Knowlesi*. Mar. 2017 (see pp. 11, 19)

[50] **B. SINGH** and **C. DANESHVAR**. "Human Infections and Detection of Plasmodium Knowlesi". In: *Clinical Microbiology Reviews* **26**:2 (Apr. 2013), 165–184. DOI: 10.1128/CMR.00079-12 (see pp. 12, 13)

[51] **I. VYTHILINGAM**. "Plasmodium Knowlesi and Wuchereria Bancrofti: Their Vectors and Challenges for the Future". In: *Frontiers in Physiology* **3**: (2012), 115. DOI: 10.3389/fphys.2012.00115 (see p. 12)

[52] **C. DANESHVAR**, **T. M. E. DAVIS**, **J. COX-SINGH**, **M. Z. RAFA'EE**, **S. K. ZAKARIA**, **P. C. S. DIVIS**, and **B. SINGH**. "Clinical and Laboratory Features of Human *Plasmodium Knowlesi* Infection". In: *Clinical Infectious Diseases* **49**:6 (15 Septemer 2009), 852–860. DOI: 10.1086/605439 (see pp. 12, 17, 19)

[53] **B. E. BARBER**, **T. WILLIAM**, **M. JIKAL**, **J. JILIP**, **P. DHARARAJ**, **J. MENON**, **T. W. YEO**, and **N. M. ANSTEY**. "Plasmodium Knowlesi Malaria in Children". In: *Emerging Infectious Diseases* **17**:5 (May 2011), 814–820. DOI: 10.3201/eid1705.101489 (see pp. 12, 13)

[54] **I. N. D. LUBIS**, **H. WIJAYA**, **M. LUBIS**, **C. P. LUBIS**, **P. C. S. DIVIS**, **K. B. BESHIR**, and **C. J. SUTHERLAND**. "Contribution of Plasmodium Knowlesi to Multispecies Human Malaria Infections in North Sumatera, Indonesia". In: *The Journal of Infectious Diseases* **215**:7 (Apr. 2017), 1148–1155. DOI: 10.1093/infdis/jix091 (see p. 13)

[55] **M. R. GALINSKI** and **J. W. BARNWELL**. "Chapter 5 - Nonhuman Primate Models for Human Malaria Research". In: *Nonhuman Primates in Biomedical Research (Second Edition)*. Ed. by **C. R. ABEE**, **K. MANSFIELD**, **S. TARDIF**,

and **T. MORRIS**. Boston: Academic Press, 2012, 299–323. DOI: 10.1016/B978-0-12-381366-4.00005-5 (see pp. 14–18)

[56] **P. S. GOMES**, **J. BHARDWAJ**, **J. RIVERA-CORREA**, **C. G. FREIRE-DE-LIMA**, and **A. MORROT**. "Immune Escape Strategies of Malaria Parasites". In: *Frontiers in Microbiology* **7**: (Oct. 2016). DOI: 10.3389/fmicb.2016.01617 (see pp. 14–17, 23)

[57] **MEDICINES FOR MALARIA VENTURE**. *Parasite Life Cycle*. https://www.mmv.org/malaria-medicines/parasite-lifecycle. Research. 2016 (see p. 15)

[58] **C. KLOTZ** and **U. FREVERT**. "Plasmodium Yoelii Sporozoites Modulate Cytokine Profile and Induce Apoptosis in Murine Kupffer Cells". In: *International journal for parasitology* **38**:14 (Dec. 2008), 1639–1650. DOI: 10.1016/j.ijpara.2008.05.018 (see p. 14)

[59] **M. IKARASHI**, **H. NAKASHIMA**, **M. KINOSHITA**, **A. SATO**, **M. NAKASHIMA**, **H. MIYAZAKI**, **K. NISHIYAMA**, **J. YAMAMOTO**, and **S. SEKI**. "Distinct Development and Functions of Resident and Recruited Liver Kupffer Cells/-Macrophages". In: *Journal of Leukocyte Biology* **94**:6 (2013), 1325–1336. DOI: 10.1189/jlb.0313144 (see p. 14)

[60] **C. THIELEKE-MATOS**, **M. L. DA SILVA**, **L. CABRITA-SANTOS**, **M. D. PORTAL**, **I. P. RODRIGUES**, **V. ZUZARTE-LUIS**, **J. S. RAMALHO**, **C. E. FUTTER**, **M. M. MOTA**, **D. C. BARRAL**, and **M. C. SEABRA**. "Host Cell Autophagy Contributes to Plasmodium Liver Development". In: *Cellular Microbiology* **18**:3 (2016), 437–450. DOI: 10.1111/cmi.12524 (see p. 14)

[61] **A. STURM**, **R. AMINO**, **C. VAN DE SAND**, **T. REGEN**, **S. RETZLAFF**, **A. RENNENBERG**, **A. KRUEGER**, **J.-M. POLLOK**, **R. MENARD**, and **V. T. HEUSSLER**. "Manipulation of Host Hepatocytes by the Malaria Parasite for Delivery into Liver Sinusoids". In: *Science* **313**:5791 (Sept. 2006), 1287–1290. DOI: 10.1126/science.1129720 (see p. 15)

[62] **S. ANTINORI**, **L. GALIMBERTI**, **L. MILAZZO**, and **M. CORBELLINO**. "Plasmodium Knowlesi: The Emerging Zoonotic Malaria Parasite". In: *Acta Tropica* **125**:2 (Feb. 2013), 191–201. DOI: 10.1016/j.actatropica.2012.10.008 (see pp. 15–17, 21)

[63] **A. M. TALMAN**, **D. T. D. OUOLOGUEM**, **K. LOVE**, **V. M. HOWICK**, **C. MULAMBA**, **A. HAIDARA**, **N. DARA**, **D. SYLLA**, **A. SACKO**, **M. M. COULIBALY**, **F. DAO**, **C. P. O. SANGARE**, **A. DJIMDE**, and **M. K. N. LAWNICZAK**. "Uptake of Plasmodium Falciparum Gametocytes During Mosquito Bloodmeal by Direct and Membrane Feeding". In: *Frontiers in Microbiology* **11**: (Mar. 2020), 246. DOI: 10.3389/fmicb.2020.00246 (see p. 17)

[64] **V. MESSINA**, **M. VALTIERI**, **M. RUBIO**, **M. FALCHI**, **F. MANCINI**, **A. MAYOR**, **P. ALANO**, and **F. SILVESTRINI**. "Gametocytes of the Malaria Parasite Plasmodium Falciparum Interact With and Stimulate Bone Marrow Mesenchymal Cells to Secrete Angiogenetic Factors". In: *Frontiers in Cellular and Infection Microbiology* **8**: (2018), 50. DOI: 10.3389/fcimb.2018.00050 (see p. 17)

[65] **K. VENUGOPAL**, **F. HENTZSCHEL**, **G. VALKIŪNAS**, and **M. MARTI**. "Plasmodium Asexual Growth and Sexual Development in the Haematopoietic Niche of the Host". In: *Nature Reviews Microbiology* **18**:3 (Mar. 2020), 177–189. DOI: 10.1038/s41579-019-0306-2 (see p. 17)

[66] **I. W. SHERMAN**. *Malaria: Parasite Biology, Pathogenesis, and Protection*. ASM Press, 1998 (see p. 17)

[67] **C. DANESHVAR**, **T. WILLIAM**, and **T. M. E. DAVIS**. "Clinical Features and Management of Plasmodium Knowlesi Infections in Humans". In: *Parasitology* **145**:1 (Jan. 2018), 18–31. DOI: 10.1017/S0031182016002638 (see pp. 18, 19)

[68] **D. D. LAISHRAM**, **P. L. SUTTON**, **N. NANDA**, **V. L. SHARMA**, **R. C. SOBTI**, **J. M. CARLTON**, and **H. JOSHI**. "The Complexities of Malaria Disease Manifestations with a Focus on Asymptomatic Malaria". In: *Malaria Journal* **11**:1 (Jan. 2012), 29. DOI: 10.1186/1475-2875-11-29 (see p. 18)

[69] **N. SRIBOONVORAKUL**, **A. GHOSE**, **M. M. U. HASSAN**, **M. A. HOSSAIN**, **M. A. FAIZ**, **S. PUKRITTAYAKAMEE**, **K. CHOTIVANICH**, **Y. SUKTHANA**, **S. J. LEOPOLD**, **K. PLEWES**, **N. P. J. DAY**, **N. J. WHITE**, **J. TARNING**, and **A. M. DONDORP**. "Acidosis and Acute Kidney Injury in Severe Malaria". In: *Malaria Journal* **17**: (Mar. 2018), 128. DOI: 10.1186/s12936-018-2274-9 (see p. 18)

[70] **J. COX-SINGH**, **J. HIU**, **S. B. LUCAS**, **P. C. DIVIS**, **M. ZULKARNAEN**, **P. CHANDRAN**, **K. T. WONG**, **P. ADEM**, **S. R. ZAKI**, **B. SINGH**, et al. "Severe Malaria - A Case of Fatal Plasmodium Knowlesi Infection with Post-Mortem Findings: A Case Report". In: *Malaria journal* **9**:1 (2010), 10 (see pp. 18, 20, 25)

[71] **D. J. COOPER**, **G. S. RAJAHRAM**, **T. WILLIAM**, **J. JELIP**, **R. MOHAMMAD**, **J. BENEDICT**, **D. A. ALAZA**, **E. MALACOVA**, **T. W. YEO**, **M. J. GRIGG**, **N. M. ANSTEY**, and **B. E. BARBER**. "*Plasmodium Knowlesi* Malaria in Sabah, Malaysia, 2015–2017: Ongoing Increase in Incidence Despite Near-Elimination of the Human-Only *Plasmodium* Species". In: *Clinical Infectious Diseases* **70**:3 (Jan. 2020), 361–367. DOI: 10.1093/cid/ciz237 (see p. 18)

[72] **M. WILLMANN**, **A. AHMED**, **A. SINER**, **I. T. WONG**, **L. C. WOON**, **B. SINGH**, **S. KRISHNA**, and **J. COX-SINGH**. "Laboratory Markers of Disease Severity in Plasmodium Knowlesi Infection: A Case Control Study". In: *Malaria Journal* **11**: (Oct. 2012), 363. DOI: 10.1186/1475-2875-11-363 (see p. 19)

[73] **B. E. BARBER**, **T. WILLIAM**, **M. J. GRIGG**, **J. MENON**, **S. AUBURN**, **J. MARFURT**, **N. M. ANSTEY**, and **T. W. YEO**. "A Prospective Comparative Study of Knowlesi, Falciparum, and Vivax Malaria in Sabah, Malaysia: High Proportion With Severe Disease From Plasmodium Knowlesi and Plasmodium Vivax But No Mortality With Early Referral and Artesunate Therapy". In: *Clinical Infectious Diseases* **56**:3 (Feb. 2013), 383–397. DOI: 10.1093/cid/cis902 (see p. 19)

[74] **WORLD HEALTH ORGANIZATION**. *Management of Severe and Complicated Malaria: A Practical Handbook*. Geneva: World Health Organization, 2012 (see p. 19)

[75] **K. PLEWES**, **G. D. TURNER**, and **A. M. DONDORP**. "Pathophysiology, Clinical Presentation, and Treatment of Coma and Acute Kidney Injury Complicating Falciparum Malaria". In: *Current Opinion in Infectious Diseases* **31**:1 (Feb. 2018), 69–77. DOI: 10.1097/QCO.0000000000000419 (see p. 20)

[76] **L. H. MILLER**, **D. I. BARUCH**, **K. MARSH**, and **O. K. DOUMBO**. "The Pathogenic Basis of Malaria". In: *Nature* **415**:6872 (Feb. 2002), 673–679. DOI: 10.1038/415673a (see pp. 20, 25)

[77] **B. PENMAN** and **S. GUPTA**. "Evolution of Virulence in Malaria". In: *Journal of Biology* **7**:6 (Aug. 2008), 22. DOI: 10.1186/jbiol83 (see pp. 20, 22)

[78] **K. N. BROWN** and **I. N. BROWN**. "Immunity to Malaria: Antigenic Variation in Chronic Infections of Plasmodium Knowlesi". In: *Nature* **208**: (1965), 1286–1288 (see pp. 21, 24)

[79] **S. CAMPINO**, **E. D. BENAVENTE**, **S. ASSEFA**, **E. THOMPSON**, **L. G. DROUGHT**, **C. J. TAYLOR**, **Z. GORVETT**, **C. K. CARRET**, **C. FLUECK**, **A. C. IVENS**, **D. P. KWIATKOWSKI**, **P. ALANO**, **D. A. BAKER**, and **T. G. CLARK**. "Genomic Variation in Two Gametocyte Non-Producing Plasmodium Falciparum Clonal Lines". In: *Malaria Journal* **15**: (Apr. 2016), 229. DOI: 10.1186/s12936-016-1254-1 (see p. 21)

[80] **A. F. SANDER**, **A. SALANTI**, **T. LAVSTSEN**, **M. A. NIELSEN**, **P. MAGISTRADO**, **J. LUSINGU**, **N. T. NDAM**, and **D. E. ARNOT**. "Multiple Var2csa-Type PfEMP1 Genes Located at Different Chromosomal Loci Occur in Many Plasmodium Falciparum Isolates". In: *PLoS ONE* **4**:8 (Aug. 2009). Ed. by **D. J. DIEMERT**, e6667. DOI: 10.1371/journal.pone.0006667 (see pp. 21, 23)

[81] **C. BANCELLS** and **K. W. DEITSCH**. "A Molecular Switch in the Efficiency of Translation Reinitiation Controls Expression of *Var2csa* , a Gene Implicated in Pregnancy-Associated Malaria: *Var2csa* Is Expressed by Translation Reinitiation". In: *Molecular Microbiology* **90**:3 (Nov. 2013), 472–488. DOI: 10.1111/mmi.12379 (see pp. 21, 23)

[82] **M. S. CALDERWOOD**, **L. GANNOUN-ZAKI**, **T. E. WELLEMS**, and **K. W. DEITSCH**. "Plasmodium Falciparum Var Genes Are Regulated by Two Regions with Separate Promoters, One Upstream of the Coding Region and a Second within the Intron". In: *Journal of Biological Chemistry* **278**:36 (Sept. 2003), 34125–34132. DOI: 10.1074/jbc.M213065200 (see p. 22)

[83] **N. D. PASTERNAK** and **R. DZIKOWSKI**. "PfEMP1: An Antigen That Plays a Key Role in the Pathogenicity and Immune Evasion of the Malaria Parasite Plasmodium Falciparum". In: *The International Journal of Biochemistry & Cell Biology* **41**:7 (July 2009), 1463–1466. DOI: 10.1016/j.biocel.2008.12.012 (see pp. 22, 23)

[84] **P. A. BUFFET**, **I. SAFEUKUI**, **G. DEPLAINE**, **V. BROUSSE**, **V. PRENDKI**, **M. THELLIER**, **G. D. TURNER**, and **O. MERCEREAU-PUIJALON**. "The Pathogenesis of Plasmodium Falciparum Malaria in Humans: Insights from Splenic Physiology". In: *Blood* **117**:2 (Jan. 2011), 381–392. DOI: 10.1182/blood-2010-04-202911 (see p. 23)

[85] **J. KRÜCKEN**, **L. I. MEHNERT**, **M. A. DKHIL**, **M. EL-KHADRAGY**, **W. P. M. BENTEN**, **H. MOSSMANN**, and **F. WUNDERLICH**. "Massive Destruction of Malaria-Parasitized Red Blood Cells despite Spleen Closure". In: *Infection and Immunity* **73**:10 (Oct. 2005), 6390–6398. DOI: 10.1128/IAI.73.10.6390-6398.2005 (see p. 23)

[86] **R. T. GAZZINELLI**, **P. KALANTARI**, **K. A. FITZGERALD**, and **D. T. GOLENBOCK**. "Innate Sensing of Malaria Parasites". In: *Nature Reviews Immunology* **14**:11 (Nov. 2014), 744–757. DOI: 10.1038/nri3742 (see p. 23)

[87] **E. D. BENAVENTE**, **D. R. ORESEGUN**, **P. F. DE SESSIONS**, **E. M. WALKER**, **C. ROPER**, **J. G. DOMBROWSKI**, **R. M. DE SOUZA**, **C. R. F. MARINHO**, **C. J. SUTHERLAND**, **M. L. HIBBERD**, **F. MOHAREB**, **D. A. BAKER**, **T. G. CLARK**, and **S. CAMPINO**. "Global Genetic Diversity of Var2csa in Plasmodium Falciparum with Implications for Malaria in Pregnancy and Vaccine Development". In: *Scientific Reports* **8**: (Oct. 2018). DOI: 10.1038/s41598-018-33767-3 (see p. 23)

[88] **D. K. DOSOO**, **D. CHANDRAMOHAN**, **D. ATIBILLA**, **F. B. OPPONG**, **L. ANKRAH**, **K. KAYAN**, **V. AGYEMANG**, **D. ADU-GYASI**, **M. TWUMASI**, **S. AMENGA-ETEGO**, **J. BRUCE**, **K. P. ASANTE**, **B. GREENWOOD**, and **S. OWUSU-AGYEI**. "Epidemiology of Malaria among Pregnant Women during Their First Antenatal Clinic Visit in the Middle Belt of Ghana: A Cross Sectional Study". In: *Malaria Journal* **19**:1 (Oct. 2020), 381. DOI: 10.1186/s12936-020-03457-5 (see p. 23)

[89] **M. OFORI**, **E. ANSAH**, **I. AGYEPONG**, **D. OFORI-ADJEI**, **L. HVIID**, and **B. AKANMORI**. "Pregnancy-Associated Malaria in a Rural Community of Ghana". In: *Ghana Medical Journal* **43**:1 (Mar. 2009), 13–18 (see p. 23)

[90] **P. I. MAYENGUE**, **H. RIETH**, **A. KHATTAB**, **S. ISSIFOU**, **P. G. KREMSNER**, **M.-Q. KLINKERT**, and **F. NTOUMI**. "Submicroscopic Plasmodium Falciparum Infections and Multiplicity of Infection in Matched Peripheral, Placental and Umbilical Cord Blood Samples from Gabonese Women". In: *Tropical medicine & international health: TM & IH* **9**:9 (Sept. 2004), 949–958. DOI: 10.1111/j.1365-3156.2004.01294.x (see p. 23)

[91] **C. C. KORIR** and **M. R. GALINSKI**. "Proteomic Studies of Plasmodium Knowlesi SICA Variant Antigens Demonstrate Their Relationship with P. Falciparum EMP1". In: *Infection, Genetics and Evolution* **6**:1 (Jan. 2006), 75–79. DOI: 10.1016/j.meegid.2005.01.003 (see pp. 23–25)

[92] **R. J. HOWARD**, **J. W. BARNWELL**, and **V. KAO**. "Antigenic Variation of Plasmodium Knowlesi Malaria: Identification of the Variant Antigen on Infected Erythrocytes." In: *Proceedings of the National Academy of Sciences of the United States of America* **80**:13 (July 1983), 4129–4133 (see pp. 23, 24)

[93] **J. W. BARNWELL**, **R. J. HOWARD**, **H. G. COON**, and **L. H. MILLER**. "Splenic Requirement for Antigenic Variation and Expression of the Variant Antigen on the Erythrocyte Membrane in Cloned Plasmodium Knowlesi Malaria." In: *Infection and Immunity* **40**:3 (June 1983), 985–994 (see p. 24)

[94] **S. A. LAPP**, **S. MOK**, **L. ZHU**, **H. WU**, **P. R. PREISER**, **Z. BOZDECH**, and **M. R. GALINSKI**. "Plasmodium Knowlesi Gene Expression Differs in Ex Vivo Compared to in Vitro Blood-Stage Cultures". In: *Malaria Journal* **14**:1 (Mar. 2015), 110. DOI: 10.1186/s12936-015-0612-8 (see p. 24)

[95] **C. FRECH** and **N. CHEN**. "Variant Surface Antigens of Malaria Parasites: Functional and Evolutionary Insights from Comparative Gene Family Classification and Analysis". In: *BMC Genomics* **14**: (June 2013), 427. DOI: 10.1186/1471-2164-14-427 (see p. 24)

[96] **V. CORREDOR**, **E. V. S. MEYER**, **S. LAPP**, **C. CORREDOR-MEDINA**, **C. S. HUBER**, **A. G. EVANS**, **J. W. BARNWELL**, and **M. R. GALINSKI**. "A SICAvar Switching Event in Plasmodium Knowlesi Is Associated with the DNA Rearrangement of Conserved 3′ Non-Coding Sequences". In: *Molecular*

and Biochemical Parasitology **138**:1 (Nov. 2004), 37–49. DOI: 10.1016/j.molbiopara.2004.05.017 (see p. 25)

[97] **M. R. GALINSKI**, **S. A. LAPP**, **M. S. PETERSON**, **F. AY**, **C. J. JOYNER**, **K. G. LE ROCH**, **L. L. FONSECA**, **E. O. VOIT**, and **THE MAHPIC CONSORTIUM**. "Plasmodium Knowlesi: A Superb in Vivo Nonhuman Primate Model of Antigenic Variation in Malaria". In: *Parasitology* (July 2017), 1–16. DOI: 10.1017/S0031182017001135 (see p. 25)

[98] **S. A. LAPP**, **C. C. KORIR**, and **M. R. GALINSKI**. "Redefining the Expressed Prototype SICAvar Gene Involved in Plasmodium Knowlesi Antigenic Variation". In: *Malaria Journal* **8**:1 (July 2009), 181. DOI: 10.1186/1475-2875-8-181 (see p. 24)

[99] **A. PAIN**, **U. BÖHME**, **A. E. BERRY**, **K. MUNGALL**, **R. D. FINN**, **A. P. JACKSON**, **T. MOURIER**, **J. MISTRY**, **E. M. PASINI**, **M. A. ASLETT**, **S. BALASUBRAMMANIAM**, **K. BORGWARDT**, **K. BROOKS**, **C. CARRET**, **T. J. CARVER**, **I. CHEREVACH**, **T. CHILLINGWORTH**, **T. G. CLARK**, **M. R. GALINSKI**, **N. HALL**, **D. HARPER**, **D. HARRIS**, **H. HAUSER**, **A. IVENS**, **C. S. JANSSEN**, **T. KEANE**, **N. LARKE**, **S. LAPP**, **M. MARTI**, **S. MOULE**, **I. M. MEYER**, **D. ORMOND**, **N. PETERS**, **M. SANDERS**, **S. SANDERS**, **T. J. SARGEANT**, **M. SIMMONDS**, **F. SMITH**, **R. SQUARES**, **S. THURSTON**, **A. R. TIVEY**, **D. WALKER**, **B. WHITE**, **E. ZUIDERWIJK**, **C. CHURCHER**, **M. A. QUAIL**, **A. F. COWMAN**, **C. M. R. TURNER**, **M. A. RAJANDREAM**, **C. H. M. KOCKEN**, **A. W. THOMAS**, **C. I. NEWBOLD**, **B. G. BARRELL**, and **M. BERRIMAN**. "The Genome of the Simian and Human Malaria Parasite Plasmodium Knowlesi". In: *Nature* **455**:7214 (Oct. 2008), 799–803. DOI: 10.1038/nature07306 (see pp. 24–27)

[100] **S. A. LAPP**, **J. A. GERALDO**, **J.-T. CHIEN**, **F. AY**, **S. B. PAKALA**, **G. BATUGEDARA**, **J. HUMPHREY**, **THE MAHPIC CONSORTIUM**, **J. D. DEBARRY**, **K. G. LE ROCH**, **M. R. GALINSKI**, and **J. C. KISSINGER**. "PacBio Assembly of a Plasmodium Knowlesi Genome Sequence with Hi-C Correction and Manual Annotation of the SICAvar Gene Family". In: *Parasitology* (July 2017), 1–14. DOI: 10.1017/S0031182017001329 (see pp. 24–26)

[101] **E. F. MERINO**, **C. FERNANDEZ-BECERRA**, **A. M. DURHAM**, **J. E. FERREIRA**, **V. F. TUMILASCI**, **J. D'ARC-NEVES**, **M. DA SILVA-NUNES**, **M. U. FERREIRA**, **T. WICKRAMARACHCHI**, **P. UDAGAMA-RANDENIYA**, **S. M. HANDUNNETTI**, and **H. A. DEL PORTILLO**. "Multi-Character Population Study of the Vir Subtelomeric Multigene Superfamily of Plasmodium Vivax, a Major Human Malaria Parasite". In: *Molecular and Biochemical Parasitology* **149**:1 (Sept. 2006), 10–16. DOI: 10.1016/j.molbiopara.2006.04.002 (see pp. 26, 27)

[102] **T. E. HARRISON**, **A. J. REID**, **D. CUNNINGHAM**, **J. LANGHORNE**, and **M. K. HIGGINS**. "Structure of the Plasmodium-Interspersed Repeat Proteins of the Malaria Parasite". In: *Proceedings of the National Academy of Sciences* **117**:50 (Dec. 2020), 32098–32104. DOI: 10.1073/pnas.2016775117 (see pp. 26, 27)

[103] **C. S. JANSSEN**, **R. S. PHILLIPS**, **C. M. R. TURNER**, and **M. P. BARRETT**. "Plasmodium Interspersed Repeats: The Major Multigene Superfamily of Malaria Parasites". In: *Nucleic Acids Research* **32**:19 (Oct. 2004), 5712–5720. DOI: 10.1093/nar/gkh907 (see pp. 26, 27)

[104] **X. Y. YAM**, **T. BRUGAT**, **A. SIAU**, **J. LAWTON**, **D. S. WONG**, **A. FARAH**, **J. S. TWANG**, **X. GAO**, **J. LANGHORNE**, and **P. R. PREISER**. "Characterization of the Plasmodium Interspersed Repeats (PIR) Proteins of Plasmodium Chabaudi Indicates Functional Diversity". In: *Scientific Reports* **6**:1 (Mar. 2016), 23449. DOI: 10.1038/srep23449 (see pp. 26, 27)

[105] **V. SINGH**, **P. GUPTA**, and **V. PANDE**. "Revisiting the Multigene Families: Plasmodium Var and Vir Genes". In: *Journal of Vector Borne Diseases* **51**: (June 2014), 75–81 (see p. 27)

[106] **M. M. PINHEIRO**, **M. A. AHMED**, **S. B. MILLAR**, **T. SANDERSON**, **T. D. OTTO**, **W. C. LU**, **S. KRISHNA**, **J. C. RAYNER**, and **J. COX-SINGH**. "Plasmodium Knowlesi Genome Sequences from Clinical Isolates Reveal Extensive Genomic Dimorphism". In: *PLOS ONE* **10**:4 (Apr. 2015). Ed. by **O. KANEKO**, e0121303. DOI: 10.1371/journal.pone.0121303 (see pp. 28, 29)

[107] **P. C. S. DIVIS**, **C. W. DUFFY**, **K. A. KADIR**, **B. SINGH**, and **D. J. CONWAY**. "Genome-Wide Mosaicism in Divergence between Zoonotic Malaria Parasite

Subpopulations with Separate Sympatric Transmission Cycles". In: *Molecular Ecology* **27**:4 (2018), 860–870. DOI: 10.1111/mec.14477 (see p. 28)

[108] **P. C. S. DIVIS**, **L. C. LIN**, **J. J. ROVIE-RYAN**, **K. A. KADIR**, **F. ANDERIOS**, **S. HISAM**, **R. S. K. SHARMA**, **B. SINGH**, and **D. J. CONWAY**. "Three Divergent Subpopulations of the Malaria Parasite Plasmodium Knowlesi". In: *Emerging Infectious Diseases* **23**:4 (2017). DOI: 10.3201/eid2304.161738 (see p. 28)

[109] **S. B. MILLAR**. "Gene Knock-in as a Tool to Phenotype Clinically Relevant Variant Alleles for Studies on Malaria Pathobiology: Proof of Concept Using the Plasmodium Knowlesi Normocyte Binding Protein Xa Gene". Thesis. St Andrews, Scotland: University of St Andrews, 2017 (see pp. 28, 29)

CHAPTER TWO

DEPLETION OF HUMAN LEUCOCYTES FROM THAWED INFECTED WHOLE BLOOD USING CD45 DYNABEADS

A kì í fi ojúbọ́rọ́ gba ọmọ lọ́wọ́ èkùrọ́ – **One does not easily or casually take the child from the palm-nut.**

Yorùbá adage

Malaria research is entirely dependent upon the whole blood that is either extracted from a willing infected patient or used to experimentally cultivate *Plasmodium spp.* laboratory lines. As such, it is necessary to ensure that the whole blood being used is carefully collected and stored. Subsequent analyses carried out both via microscopy and molecular DNA analyses, may be affected by variation between the host patients in the form of white blood cell composition and abundance as well as other idiosyncrasies [1]. Additionally, the location of venipuncture –subject to parasite localisation– can also play a role in the eventual estimation of parasitaemia and DNA quantification [1]. As such, it becomes prudent to ensure uniformity during blood collection to provide a representative parasite yield for the isolate.

2.1 Introduction

Genome sequencing of pathogenic organisms can often be complex when using field or clinical samples that are 'contaminated' with the host DNA. For example, in humans infected with malaria, the host human DNA content within an extracted blood sample can be orders of magnitude larger than the Plasmodium DNA. Such disproportionality creates challenges in generating laboratory culture lines for controlled studies as well as increasing the difficulty of investigating contemporary *in vivo* strains undergoing continuous selective pressure. Depending on the organism of choice, various methods can be implemented to either deplete the host DNA content or isolate the organism of interest's DNA. As such, at the fundamental level, most genome sequencing investigations begin with an isolation or depletion step to ensure enrichment of the desired organism's genetic content.

Genome sequences are available for all the human malaria pathogens (*Plasmodium falciparum, Plasmodium malariae, Plasmodium knowlesi, Plasmodium vivax* and *Plasmodium ovale*) [2–5], however only *P. falciparum* and *P. knowlesi* have been laboratory-adapted. Hence, the reference sequences generated for these two species had access to relatively large supplies of these experimental lines to accomplish sequencing with sufficient genome coverage. Infected patient whole blood (*ip*WB) collected from malaria patients contains a mixture of healthy host erythrocytes, parasite-infected erythrocytes, host leucocytes, free soluble DNA, various soluble host molecules and also free, circulating parasites. Thus, upon direct DNA extraction of *ip*WB, *Plasmodium* parasite DNA can make up a tiny proportion of the DNA extracted from the whole blood.

Furthermore, there are ~8,000 white blood cells (WBC)/µL of whole blood while there can be $<10^6$ parasites/microlitres (µL), with each human cell containing a diploid genome with 3.2 - 3.5 Gb of nucleotide bases. On the other hand, *Plasmodium spp.* possess a miniscule 21 - 25 Mb of DNA; signifying that the human genome is >110-fold the size of a *Plasmodium* parasite ([6]; *Janet Cox-Singh, personal communications*). It is this difference in WBC abundance and subsequent genome size that is the main reason for removing or depleting WBCs from infected whole blood. Indeed, an individual WBC contains the molecular weight equivalent to 200 merozoite parasites [1]. Due to this,

Plasmodium genome sequencing begins with a depletion of the human host DNA content and other free, soluble DNA through the means of depleting host cell content in the extracted sample. However host RBC pose little difficulty in removal due to containing no DNA. Indeed the challenge to attain adequate parasite DNA lies in the depletion of host WBC content.

With this in mind, many techniques have been developed as a means of reducing the leucocyte content of an infected whole blood isolate. Depletion techniques such as separation by filtration using varieties of filtration systems [1, 5, 7, 8], anti-HLA1 DynaBeads separation [1, 9, 10], flow cytometry [11] and various density gradient separation systems [1, 10], have been described and developed to achieve parasite enrichment of a particular sample. The techniques described in the literature above have varying ranges of success, however, such methods were developed for cultured samples or fresh, non-frozen whole blood isolates. In truth, while leucocyte depletion is necessary, many sub-zero archival biobanks of infected *Plasmodium* spp. isolates still possess leucocytes to be depleted prior to any downstream preparation for whole genome sequencing [12].

To our knowledge, only two methods successfully carried out leucocyte depletion of thawed infected patient whole blood for whole-genome sequencing. The first [12] utilised a Whatman paper filtration system to generate six high-quality Illumina short genomes investigating dimorphism in *P. knowlesi*. The second [13], expanded Pinheiro et al. [12]'s method further, adding a Plasmodipur step as a means to facilitate better *P. knowlesi* parasite retention. While both methods succeeded in extracting *P. knowlesi* DNA, their efficiency and yield required further optimisation to generate a reference genome from thawed clinical samples [14]. As such, a third method was developed and is described below in this chapter. All three methods assume that a certain proportion of leucocytes survive the freeze-thaw process rather than being ruptured by the ice crystals formed [12].

Additionally, it was hypothesised that while some leucocytes, RBCs and *i*RBC may rupture during the freeze/thaw process and the parasites contained within remained intact. Thus, the aim of such a method would be the purification of the intact parasite from the unnecessary human cell debris. One early example of this involves the use

of the commercially available LD (MACS) separation system, which was applied to successfully deplete the unwanted human cells from whole blood [10]. However, the LD (MACS) require the presence of haemozoin in the whole blood in order to be effective [1]. Haemozoin is a by-product of the parasite's digestion of the host RBC's haemoglobin; thus, it is not present in early-stage, non-blood malaria parasites [15]. Other methods implemented involve the use of antibody-mediated isolation and separation techniques, chiefly through the use of the anti-HLA1 antibody [1, 9].

Rather than the Human Leucocyte Antigen Class 1 (HLA1) protein family, in the method described below, the leucocyte common antigen (LCA) or CD45 that is ubiquitous in all nucleated haemopoietic cells was selected as a means of targetting intact and lysed host white blood cells (WBC) from the thawed whole blood samples available to study [16–18]. Indeed, CD45 is densely populated on the surface of WBCs; commanding up to 10 % of the cell surface with 10^6 antigens per cell [16–18]. Thus, we hypothesise that a CD45 antibody can mediate an isolation of WBCs via this interaction. As a member of the protein tyrosine phosphatase family, CD45 has been linked with various functions in the regulation of the adaptive immune response [17]. Chief of these being its activity in regulating signal transduction in T-cells and B-cells as part of the cell's development [16–19]. Apart from this, CD45 has been shown to be associated with histamine deregulation in mast cells, modulation of chemokine-induced signalling in neutrophils and perhaps regulation of myelopoiesis [17]. As such, CD45 has been a prospective target in various therapies and studies involving Alzheimer's Disease, rheumatoid arthritis and HIV [17].

Initial work into the leucocyte depletion from the frozen samples in our biobank had been started by Millar [13]. The outputs of this work were used in this study, in part, to generate the simulated-infected whole blood (*si*WB), where cultured PkA1-H.1 samples were used to seed healthy donor whole blood as a means of providing a comparative medium to the precious frozen patient samples. While Millar [13] utilised the Whatman-Plasmodipur leucocyte depletion method, the method incurred parasite material loss and hDNA contamination. In a bid to overcome this, a conceptually simpler method of antibody-antigen interaction-mediated depletion was deemed promising. Here, the simulated-infected whole blood (*si*WB) would prove helpful in first determining the appropriateness and eventual effectiveness of the hypothesis. Additionally, the qPCR

protocols and conditions had been developed and extensively described by Millar [13], with this study expanding upon the protocol as necessary.

2.2 Chapter 2: Aim and Rationale

Through the ubiquity of CD45 on the surface of leucocyte cells and the superparamagnetic capability of the DynaBeads, we aimed to exploit the capabilities of the commercially available CD45 DynaBeads™[Invitrogen]. Through this, we aimed to develop and optimise a depletion method to remove surviving human leucocyte content within freeze/thawed infected patient whole blood, whilst retaining an enriched population of *P. knowlesi* parasites for DNA extraction and subsequent whole-genome sequencing.

2.3 Equipment and Reagents

The full complement of equipment and reagents utilised in developing this method is presented in Appendix section A.1.

2.4 Methods

2.4.1 Whole Blood Sample Collection

Infected patient whole blood (*ip*WB) was collected from consenting adult patients infected with *P. knowlesi* as part of a non-interventional study carried out by Ahmed et al. [20] and later utilised by Pinheiro et al. [12]. The infected whole blood samples were archived and stored in -80°C storage at the School of Medicine at the University of St. Andrews. The *ip*WB were not leucocyte depleted before being frozen, although their parasitaemia was recorded and stored with the relevant patient information.

Due to the precious nature of the *ip*WB, simulated infected whole blood was required for the development and optimisation of the leucocyte depletion method for this

project. Here, ten millilitres (mL) of healthy donor blood was collected by venipuncture into anticoagulant Ethylenediaminetetraacetic acid (EDTA) tubes by Dr Fiona Cooke according to the University of St. Andrews Teaching and Research Ethics Committee guidelines and approval. The donor whole blood was aliquoted into 300 µL volumes, which were in turn stored in -30°C for varying lengths of time. It was ensured that each aliquot was fully frozen before use in order to simulate the 'freeze-thaw' process which the *ip*WB would undergo. The donor whole blood was not pretreated, to ensure that leucocytes remained in the blood sample.

On the other hand, positive *P. knowlesi* control blood was prepared using packed cell volumes of the *P. knowlesi* PkA1-H.1 *in vitro* clone strain (*also PkA1H1*) which had been cultured as part of a previous study by Millar [13]. These were stored at -30°C with an estimated parasitaemia of 2 % in 200 µL aliquots. Additionally, as part of a separate study, Dr Fauzi Muh produced cultured stocks of PkA1-H.1 with 5 - 7 % parasitaemia to be used in the latter stages of this study.

Simulated infected whole blood was produced through a proportional mixture of the healthy whole blood with the *P. knowlesi* PkA1-H.1 experimental line cultured in washed human red blood cells. Cultured parasites (2 % parasitaemia) and healthy whole blood were mixed at a 2:1 ratio by volume to give a final parasitaemia of 1 - 1.5 %. The *si*WB was aliquoted into 200 µL volumes and stored in -30°C.

2.4.2 Measuring human and parasite DNA using Quantitative Real-time PCR to generate a calibration curve

To ensure parity across numerous experiments and to determine a relative baseline for both human and parasite DNA concentration, qPCR was implemented to construct a calibration curve. For this, a 10 ng/µL TaqMan hDNA standard solution [ABI] was used for the human standard however, a *pk*DNA standard was not available. Hence, to compensate for this, previously extracted DNA from the cultured *P. knowlesi* PkA1-H.1 experimental line was used. Using a NanoDrop 2000 spectrophotometer [ThermoFisher], the cultured *P. knowlesi* standard was measured to be 6.4 ng/µL. Following this, a 1:5 by concentration dilution series was carried out on both the hDNA and *P. knowlesi* cultured

standards.

qPCR was performed with the 2x TaqMan Universal Mastermix [ABI], forward and reverse hDNA primers (*specific for the human β-globin gene*), the hDNA probe [21], *P. knowlesi* primers (*specific for the P. knowlesi 18S rRNA* [22]) and *P. knowlesi* probes [23]. Additionally the DNA template and nuclease-free water was added to make a final reaction volume of 20 µL [Table 2.1]. The primers and probes were obtained from Eurofins Genomics and their sequences are presented in Table 2.2.

Table 2.1: Mastermix preparation volumes used for calibration curve

Material	**Stock Concentration (µM)**	**Volume in** 20 µL	**Volume x34 (µL)**	**Final concentration (µM)**
Klaassen hDNA primer Forward	10	0.5	17	250
Klaassen hDNA primer Reverse	10	0.5	17	250
Plasmo 1	10	0.4	13.6	250
Plasmo 2	10	0.4	13.6	250
Human DNA Probe	10	0.2	6.8	100
P. knowlesi Probe	10	0.16	5.44	80
TaqMan Universal Mastermix	2x	10	340	1x
Nuclease-Free Water	-	2.84	94.56	-
Total	-	15	510	-
DNA Template	-	5	N.A	-

The primers and probes were sourced from Eurofins based on sequences described by Klaassen et al. [21], Rougemont et al. [22], and Divis et al. [23].

The qPCR was carried out in triplicate on the Rotor-Gene Q-Series platform [QIAGEN] with the following conditions: 10 minutes @ 95°C, followed by 55 cycles of 95°C for 15 seconds, and 60°C for 60 seconds, resulting in ~143 minutes runtime. The results are assessed on the Rotor-Gene Q-Series software [QIAGEN] using two different absorbance wavelength channels [Table 2.3], with the Green channel detecting *pk*DNA

Table 2.2: Probe and primer sequences for qPCR

Probes/Primers	Sequence (5'>3')
Klaassen hDNA primer Forward	GGGCAACGTGCTGGTCTG
Klaassen hDNA primer Reverse	AGGCAGCCTGCACTGGT
Plasmo 1	GTTAAGGGAGTGAAGACGATCAGA
Plasmo 2	AACCCAAAGACTTTGATTTCTCATAA
Human DNA Probe	YAKYE*-CTGGCCCATCACTTTGGCAAAGAA-BHQ1**
P. knowlesi Probe	FAM*-CTCTCCGGAGATTAGAACTCTTAGATTGCT-BHQ1**

Sequences for primers and probes used in carrying out the quantitative real-time PCR. The sequences were sourced from Eurofins based on sequences described by Klaassen et al. [21], Rougemont et al. [22], and Divis et al. [23].
* *- 5' manufacturer's modification*
** *- 3' manufacturer's modification*

and the Yellow channel detecting hDNA. A standard calibration curve was constructed for both human and *pk*DNA content based on the fluorescence of each DNA at the yellow and green channels respectively. Reports were generated for both channels including calculated concentrations based on the input concentration of the pure standards.

Table 2.3: qPCR channel gain wavelengths

Channels	Source	Detector	Gain
Green	470 nm	510 nm	4.67
Yellow	530 nm	555 nm	5

Values for channel gain wavelengths on the Rotor-Gene Q software to detect human and *P. knowlesi* DNA.

2.4.3 *P. knowlesi* parasite purification using the Saponin Method

A method described by Hyde and Read [24] to lyse RBCs with Saponin, but leave malaria parasites intact was used. Briefly 10 mL of a 10 % Saponin [SIGMA] solution

was prepared by dissolving one gram of white saponin into 9.9 mL of phosphate-buffered saline (PBS) and 100 µL EDTA. From this, a 1 % Saponin solution was made and a 10^{th} volume added to the thawed donor blood to give a final saponin concentration of 0.1 % in the blood sample. The mixture was incubated at room temperature for four minutes before centrifuging at 2.500 rpm for five minutes. The supernatant was discarded and the parasite enriched pellet resuspended in 1 mL of PBS and centrifuged at 13.000 rpm for 1 minute. These washing and resuspension steps were repeated three more times, finally suspending the recovered pellet in 200 µL of PBS for DNA extraction (see *chapter 2 subsection 2.4.6, DNA Extraction*).

2.4.4 CD45 DynaBeads Method Development

Proof of Concept for CD45 DynaBeads

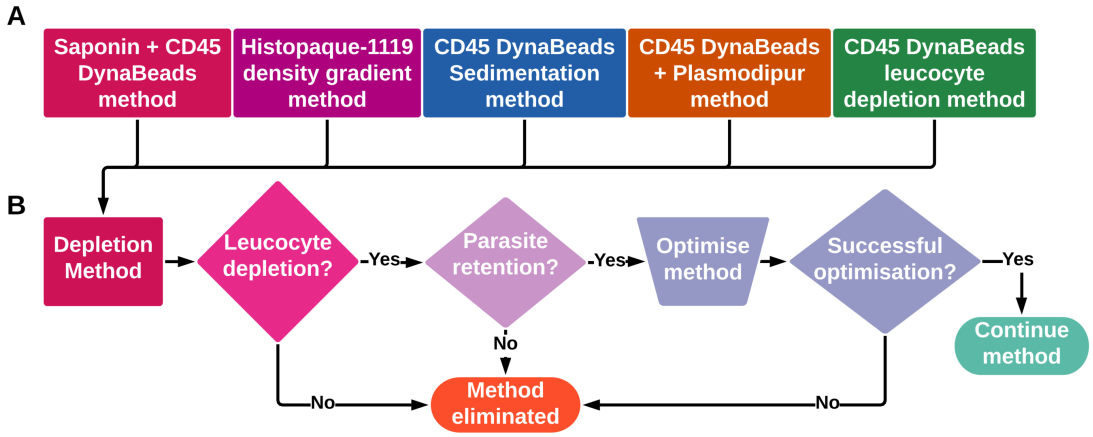


Figure 2.1: Representation of the assessment steps employed to develop leucocyte depletion methods. (A) After a proof of concept for CD45 DynaBeads (*not shown*), multiple methods for leucocyte depletion in simulated infected patient whole blood were developed to take through the assessment steps detailed. (B) Each method was assessed for leucocyte depletion, and if depletion was sufficient, parasite retention was assessed. If the method efficiently retains parasite DNA, the method is further optimised and assessed for appropriateness. Where leucocyte depletion or parasite retention fails, the method is eliminated.

As a proof of concept, the efficiency of CD45 DynaBeads [Invitrogen] to remove human leucocytes from whole blood was first assessed using thawed uninfected and healthy

donor blood. In the first instance, the saponin lysis method described in chapter 2 subsection 2.4.3 (*Saponin Lysis Method*) was implemented. Following this, 50 µL of anti-CD45 DynaBeads [Invitrogen] was added to the pellet containing unlysed WBCs and mixed for 30 minutes at 4°C. The solution was placed in a magnetic stand for two minutes to remove any WBC bound to the DynaBeads from the solution. The leucocyte-depleted solution was transferred into a new 1.5 mL microfuge tube to take forward for DNA extraction and subsequent hDNA qPCR. Following this, positive *pk*DNA and hDNA were added as assay controls.

Saponin and CD45 DynaBeads Method

Using simulated-infected whole blood (*si*WB), Saponin lysis of infected RBCs and subsequent leucocyte depletion of surviving WBCs by CD45 DynaBeads was carried out as described in chapter 2 subsubsection 2.4.4, (*Proof of Concept for CD45 DynaBeads*; Figure 2.1). However, rather than the use of healthy whole blood, the prepared simulated-infected whole blood was used.

Histopaque-1119 Density Gradient Separation Method

Histopaque-1119 is a solution mixture of a polysaccharide and sodium diatrizoate adjusted to a density of 1.119 g/mL [25]. The use of Histopaque-1119 allows for the separation of granulocytic cells from blood plasma and packed red cells. It is often used in combination with Histopaque-1077 in order to also facilitate the separation of mononuclear cells [25]. Here, using two 200 µL aliquots of simulated-infected whole blood (*si*WB) isolates, two different density volumes of Histopaque-1119 [SIGMA] were investigated for their ability to deplete leucocyte content in simulated malaria-infected blood [Figure 2.1]. For this, the two 200 µL aliquots of *si*WB were first taken through the Saponin lysis method (see *chapter 2 subsection 2.4.3*). The resulting pellets containing parasites and leucocytes were treated differently. To one, an equal volume of Histopaque-1119 [SIGMA] was carefully layered onto the solution creating two distinct layers. On the other aliquot, a 1:5 ratio by volume of Saponin output to Histopaque-1119 solution was also carefully layered. Both tubes were centrifuged at 700 *x* g for 30 minutes at

room temperature. After this, the top layer for both aliquots was carefully removed and discarded. Washing began with the addition of 1 mL PBS to the remaining solution before subsequent centrifugation for 10 minutes at 200 *x* g to recover the parasite pellet. The resulting supernatant was removed, and washing steps were repeated twice. The washed parasite pellets were resuspended in 200 µL of PBS ready for DNA extraction.

CD45 DynaBeads Sedimentation Method

Here, a simulated-infected whole blood (*si*WB) isolate was taken through the Saponin/CD45 DynaBeads method [*chapter 2 subsubsection 2.4.4*, Figure 2.1]. However, after adding the DynaBeads and subsequent incubation for 30 minutes, the solution was not placed in a magnetic stand. The solution was left to stand for 15 minutes, leaving the suspended DynaBeads to settle at the tube base. The supernatant was transferred into a new 1.5 mL microfuge tube, and the remaining settled DynaBeads resuspended in 200 µL PBS. Both the leucocyte-free solution and resuspended DynaBeads pellet were taken forward for DNA extraction.

Inverse Saponin and CD45 DynaBeads method

Here, patient infected whole blood isolates (sks265 and sks078) were utilised in an inversion of the order of the Saponin and DynaBeads methods described in chapter 2 subsubsection 2.4.4 (*Saponin and CD45 DynaBeads method*) i.e. rather than Saponin lysis followed by CD45 depletion, CD45 depletion was carried out first, followed by Saponin lysis. Briefly, 100 µL CD45 DynaBeads was added to the thawed *ip*WB and the solution taken through the DynaBeads method as stated in chapter 2 subsubsection 2.4.4 (*Proof of Concept for CD45 DynaBeads*); after which the Saponin Lysis method was carried out. After adding 1 % Saponin to the leucocyte-depleted blood, the tubes were incubated at room temperature for four minutes with gentle inversion. The solution was then centrifuged for five minutes at 800 *x* g, after which the supernatant was discarded and the pellet resuspended in 1 mL of PBS. The solution was mixed well and centrifuged at 10,000 *x* g for one minute, the supernatant discarded and the pellet resuspended in 1 mL PBS. Wash steps were repeated three times, and the resulting pellet was finally

suspended in 200 µL of PBS, ready for DNA extraction.

CD45 DynaBeads and Plasmodipur Method

Firstly 550 µL of the thawed patient isolate sks269 [Table 2.8] was diluted at a 1:50 ratio by volume with cold PBS before centrifugation at 2021 *x* g for 20 minutes at 4°C. The supernatant was removed into a new tube and spun again at the same speed and duration. The supernatant was again removed and transferred into a new tube, which was then centrifuged at 2,558 *x* g for 20 minutes at 4°C. The supernatant was discarded; the formed pellets were combined and then resuspended in cold PBS and made up to 1.2 mL. To this, 100 µL of CD45 DynaBeads was added before incubation at 4°C for 30 minutes with gentle tilting and rotation. Following this, the mixture was placed in a magnetic stand to separate the magnetic DynaBeads, after which; the remaining supernatant was transferred into a new tube. To proceed, the Plasmodipur filters [Europroxima] were pre-wet with 10 mL of cold PBS before diluting the leucocyte-free 1.2 mL solution with a further 10 mL of cold PBS. The 11.2 mL elution was loaded in a syringe with the plasmodipur filter attached and passed through the filter into a 50 mL centrifuge tube. The filters were removed and washed with 10 mL of PBS three times with each PBS wash passed through into a fresh tube. All eluates and PBS washes were then centrifuged at 2,000 *x* g at 4°C for 20 minutes. The resulting pellet was washed three times with 1 mL PBS and then centrifuged at 14,000 *x* g for 2 minutes at 4°C. The final pellet was then resuspended in 200 µL of PBS for DNA extraction.

2.4.5 CD45 DynaBeads leucocyte depletion of infected patient Whole Blood (ipWB)

Firstly, a wash buffer was prepared per the DynaBeads' manufacturer's instructions. For the wash buffer, 50 mL PBS and 0.2 mL EDTA were combined to create a stock background solution. Subsequently 10 milligram (mg) Bovine Serum Albumin (BSA) was added to 10 mL of the background solution to create the wash buffer. For each infected patient isolate to be processed, 100 µL CD45 DynaBeads were aliquoted into a 1.5 mL microfuge tube with 1 mL of the wash buffer added and mixed by inversion.

The tube was placed in a magnetic stand for one minute, and the clear supernatant was discarded. The tube was removed from the stand, and the beads were resuspended in 100 µL of the wash buffer.

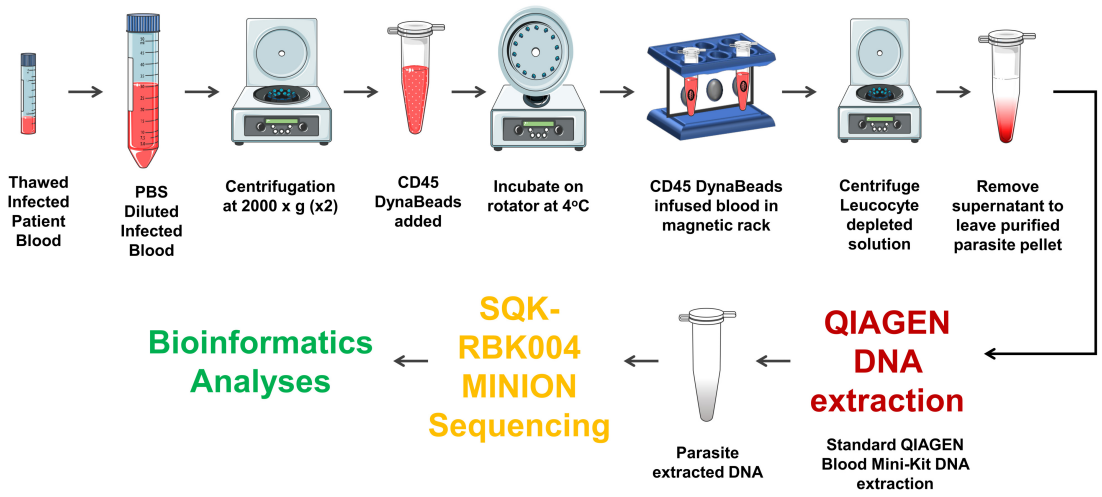


Figure 2.2: The Leucocyte Depletion Protocol. The leucocyte depletion method using CD45 DynaBeads including subsequent DNA extraction, sequencing and analyses. Patient whole blood had to be thawed before carrying out the leucocyte depletion with the effect of depletion determined via qPCR for every experiment. DNA concentration was quantified using either the NanoDrop 2000 or Qubit quantification methods.

Following this, infected patient isolates were thawed in a lukewarm water bath with gentle agitation before being immediately placed in ice. Using a wide-bore pipette tip, the infected whole blood was gently mixed before dispensing 10 µL of whole blood into a new 1.5 mL microfuge tube to act as an untreated experiment and isolate control. The control tube was returned to a -30°C freezer, and the volume of the remaining whole blood was measured before being transferred to a sterile 50 mL skirted centrifuge tube. The whole blood was then diluted at 1:50 ratio by volume with cold PBS and then centrifuged for 2000 *x* g for 20 minutes at 4°C [Figure 2.2]. The resulting supernatant was decanted into a new 50 mL centrifuge tube while the pellet was placed in ice. The supernatant tube was further centrifuged at 2558 *x* g for 20 minutes at 4°C, with the resulting supernatant discarded. Pellets from both tubes were combined in a 1.5 mL microfuge tube and resuspended in cold PBS to make a final volume of 1.2 mL.

The washed 100 µL DynaBeads suspensions were then transferred to the resuspended packed cell pellet and subsequently incubated for 30 minutes at 4°C at a low-medium speed on a rotator [Figure 2.2]. Following this, the tube was placed in a magnetic stand for two minutes, and the leucocyte-depleted supernatant was transferred into a clean 1.5 mL microfuge tube. The remaining DynaBeads pellet held by the magnetic field was resuspended in 200 µL PBS and placed in ice.

The leucocyte-depleted supernatant tube was then centrifuged at 2000 *x* g for 20 minutes to pull down any suspended RBCs, *i*RBCs and parasites. The supernatant was removed, leaving 200 µL containing the parasite-enriched pellet [Figure 2.2]. The parasite-enriched pellet tube was then placed in ice. Following this, the control tube containing 10 µL of untreated infected whole blood was removed from the freezer, and 190 µL PBS added. The control tube, CD45-treated parasite-enriched pellet tube (*henceforth CD45-treated*) and DynaBeads pellet tube (containing the magnetic beads bound to host leucocytes) – each made up to 200 µL – were all taken forward for DNA extraction.

2.4.6 DNA Extraction

DNA extraction was carried out using the QIAamp Blood Mini kit [QIAGEN]. Twenty µL of proteinase K [QIAGEN], 4 µL RNAse A and 200 µL Buffer AL were added to each tube (*Control, CD45-treated pellet and DynaBeads pellet*). Each tube was pulse vortexed for 15 seconds before being placed in a heating block for 10 minutes at 56°C. On completion, the tubes were briefly centrifuged to remove droplets and then 200 µL of fresh ethanol (96 - 100 %) was added before the vortex and centrifuge steps were repeated. Using wide-bore pipette tips, the content of each tube was transferred into individual QIAGEN spin columns, which were then centrifuged for 1 minute at 6,000 *x* g passing the DNA mixture through the filter, binding the DNA to the filter. The passed eluent was collected in the collection tube and discarded. The spin-column with bound DNA was transferred into a clean collection tube. The column was washed with 500 µL buffer AW1, and the tubes were centrifuged at 6000 *x* g for 1 minute. Once again, the spin column was transferred into a clean collection tube, and the wash step repeated with 500 µL buffer AW2 and centrifuged at 20,000 *x* g for 3 minutes. The spin-column was then transferred into a clean 1.5 mL microfuge tube and centrifuged for one minute at 20,000

x g to remove any residual buffer AW2 from the spin column. Following this, the spin column was finally transferred to a 1.5 mL Lo-Bind microfuge tube [Ambion]. 150 µL of buffer AE was added to the spin column and left to incubate at room temperature for 5 minutes to elute the DNA from the filter column. DNA was recovered by centrifugation at 6,000 *x* g for 1 minute. The resulting eluted parasite-enriched DNA was ready for quantification, whole-genome sequencing or storage at -30°C for later use.

2.4.7 DNA Quantification

Nanodrop 2000

Once extracted and eluted, DNA was quantified using the NanoDrop 2000 spectrophotometer blanked with water. One microlitre of the eluted DNA was loaded onto the NanoDrop. Each DNA template was measured twice, with the concentration recorded in ng/µL and the 260/280 DNA purity also recorded.

Qubit Quantification

For high quality, reliable DNA quantification, Qubit quantification was used. Briefly, a High Sensitivity (HS) working solution was generated for each DNA sample to be tested at a 1:200 ratio by volume by combining the HS Reagent and HS Buffer, respectively. High and low standard solutions were prepared using 190 µL of the working solution and 10 µL of the low and high standard solutions. Following this, 1 µL of DNA template samples were added to 199 µL of the working solution and vortexed for 2 - 3 seconds before incubation for 2 minutes at room temperature. Using the "dsDNA" setting, the low standard was measured, followed by the high standard concentration. With this set, each sample had its DNA content quantified and converted to ng/µL to be recorded.

2.4.8 Real-time Quantitative Polymerase Chain Reaction (qPCR)

To quantify the concentration of *P. knowlesi* content in the extracted DNA, a method similar to that described in chapter 2 subsection 2.4.2 (*P. knowlesi and Human DNA Calibration Curve*) was used. However, the qPCR mastermix was prepared using the volumes stated in Table 2.4. Each isolate was carried out in duplicate and run on the RotorGene Q-series platform to achieve maximum cycle threshold (Ct), for ~143 minutes. Quantification was detected and achieved using the channel gains described in Table 2.3.

Table 2.4: qPCR Mastermix Volumes to determine leucocyte depletion

Materials	**Stock Concentration (µM)**	**Volume in** 20 µL	**Final concentration (µM)**
Klaassan hDNA Forward	10	0.5	250
Klaassan hDNA Reverse	10	0.5	250
Plasmo 1	10	0.5	250
Plasmo 2	10	0.5	250
Human DNA Probe	10	0.2	100
***P. knowlesi* Probe**	10	0.16	80
TaqMan Universal Mastermix	2x	10	1x
Nuclease-Free Water	-	2.64	-
Total	-	15	-
DNA Template	-	5	-

Mastermix preparation volumes used for carrying out real-time qPCR of leucocyte-depleted parasite-enriched DNA. An increase in the volume of Plasmo 1 and 2 *P. knowlesi* primers in comparison to the volume used for the calibration curve [Table 2.1] is observed. However, the same final concentration is maintained overall.

Outputs were analysed on the Rotor-Gene Q Series Software. Outliers showing less than 10 % change in total fluorescence were removed from consideration as part of the quality control. Other quality control steps involved the use of the 'Dynamic Tube' and 'Slope Correct' settings in the Rotor-Gene Q Series Software.

2.4.9 Cycle threshold normalisation and calculations

Upon completing the qPCR run, the Rotor-Gene Q Series software was utilised to analyse the dataset and generate reports. Here, each channel carried out quantitation analysis of the fluorescence of the processed isolates using the generated calibration curves [*chapter 2 subsection 2.4.2*]. The Rotor-Gene Q software initially set calibration thresholds for both channels; however, these were manually changed to fit the model better. With this, DNA concentration for each duplicate sample was calculated; however, due to the lack of a true *P. knowlesi* standard DNA solution during the generation of the calibration curve, the calculated concentration from the qPCR was not used. Rather, the Ct value which is the true measured value by the RotorGene platform, provided a better comparison between isolates and experiments. However, the control sample in each experiment often had smaller starting volumes; hence normalisation of the Ct values was required to achieve parity. To achieve this, "a percentage of input" calculation was implemented. Firstly an adjustment factor was calculated based on the starting volumes of the control and CD45-treated samples according to this equation:

$$log((\frac{Control\ starting\ vol.}{CD45treated\ starting\ vol.}),2)$$

Once calculated, the adjustment factor is then subtracted from the average Ct for the control sample. This generates the normalised average Ct value. Once normalised, a delta (δ) change value can be calculated using this equation:

$$average\ CD45treated\ Ct - average\ normalised\ control\ Ct$$

A working example is shown below using values present in Table 2.5.

$$\begin{aligned} Adjustment\ factor: &\ \log((\frac{25}{650}),2) = 4.70 \\ Normalised\ Control\ Ct\ value: &\ 21.73 - 4.70 = 17.03 \\ CD45treated\ Delta\ change: &\ 17.03 - 21.37 = -4.34 \end{aligned}$$

Table 2.5: Normalising measured Ct values from a leucocyte depleted isolate

Starting volume		Measured Cycle times (Ct)		
Control	***CD45-treated***	***Control***	***CD45-treated***	***DynaBeads Pellet***
25	650	21.73	21.37	18.47

Volumes were manually measured by hand by the same researcher using a standard pipette Measured Ct values from a patient isolate qPCR are shown. The control and CD45-treated starting volumes are used for normalisation.

2.4.10 Exploratory experiments

Assessing the effect of Saponin concentration on the *pk*DNA yield

As the method development and optimisation progressed, the observed results from whole blood saponinisation as well as other reported evidence from published methods prompted a need to determine the effect of altering the saponin concentration on the eventual *pk*DNA yield. For this, simulated-infected whole blood (*si*WB) isolates were generated by combining 400 µL cultured PkA1-H.1 blood to 800 µL uninfected donor whole blood. Following this, Saponin solutions resulting in 1 %, 5 %, 10 % and 20 % Saponin concentrations were created as described in chapter 2 chapter 2 subsection 2.4.3 (*Saponin Lysis Method*). For each saponin concentration, 200 µL of *si*WB was aliquoted and a 10^{th} volume of saponin combined, resulting in a final saponin concentration of 0.1 %, 0.5 %, 1 % and 2 % respectively, in the *si*WB. Additionally, a control *si*WB sample was added where no manipulations were implemented. Following this, each sample was mixed by gentle inversion for 4 minutes at room temperature before being centrifuged for 5 minutes at 2500 rpm. The resulting supernatant was discarded and the pellet washed by resuspension in 1 mL PBS before centrifugation for 1 minute at 13 000 rpm. The supernatant was once again discarded and the wash steps repeated three more times with a final pellet suspension in 200 µL. Following this, DNA extraction was carried out as described in subsection 2.4.6 (*DNA Extraction*).

Eluted DNA concentration was quantified using both the Nanodrop 2000 and Qubit DNA quantification methods [see *chapter 2 subsection 2.4.7*]. Additionally, qPCR was carried

out using the preparation and methods described in chapter 2 subsection 2.4.2 however here, each sample was tested in triplicate. Analysis of the qPCR outputs were carried out using the calibration curve (see *chapter 2 subsection 2.4.2*); however no normalisation was necessary as the same starting volume was used for all five experimental conditions.

Investigating the effect of altering probes and primer concentration on DNA extraction

Unexpected inconsistencies were observed in preceding methods (see the inverse CD45 adaptation in *subsubsection 2.4.4, Saponin and CD45 DynaBeads method*), prompting a need to formulate new primers and probes from stock solution. Subsequently, evaluations were conducted to determine if the observed inconsistency was due to defective probes or accurate results. Firstly new *P. knowlesi* primers and probes were prepared from their 10 nM stock solution to make 250 nM working solutions. To test the efficacy of the new preparations, a cultured PkA1-H.1 sample was thawed and taken through DNA extraction directly [*chapter 2 subsection 2.4.6*] with the DNA eluted in 100 µL of nuclease-free water. Once eluted, the DNA concentration was measured using the Nanodrop 2000 [*chapter 2 subsection 2.4.7*] quantification method. Following this, the *P. knowlesi* and the hDNA standard solutions (described in *chapter 2 subsection 2.4.2*) were taken through a 1:5 dilution series, ready for qPCR. For this, two mastermix solutions were prepared, one using the previously formulated human and *P. knowlesi* primers and probes, and the other based on the newly generated primer and probe working solutions. Both mastermixes were prepared using volumes in Table 2.4. Each diluted DNA sample was run in duplicate on the Rotor-Gene Q platform.

2.5 Results

Generating a Polymerase Chain Reaction (PCR) calibration curve for human and *P. knowlesi* DNA

Although the lack of a *pk*DNA standard was not ideal, the use of the Cultured PkA1-H.1 strain allowed for an approximate in-house *P. knowlesi* standard. As such, the TaqMan™ Control Genomic DNA hDNA was serially diluted from 10 ng/µL to 0.016 ng/µL while the *P. knowlesi* Cultured PkA1-H.1 (*henceforth the P. knowlesi standard*) was diluted from 6.4 ng/µL to 0.01024 ng/µL. The concentration for the *P. knowlesi* standard was determined using the NanoDrop 2000.

The qPCR outputted calibration plots [Figure 2.3a,b] and distinct quantitation traces with thresholds of 0.021 and 0.0261 [Figure 2.3] for hDNA and *pk*DNA respectively. For the hDNA [Figure 2.3a,c], the threshold was manually adjusted from 0.021 to 0.008 after determining that the threshold calculated by the Rotor-Gene Q Series Software was too high. Manual adjustment was not done for the *pk*DNA. Using OriginPro [26], statistical analyses revealed the adjusted R squared values to be 0.9966 and 0.85657 for the hDNA and *pk*DNA respectively.

The resulting quantitation trace for both hDNA and *pk*DNA calibration plots [Figure 2.3c,d] allows visualisation of the spread of the dataset. From this, a clear separation of distinct concentrations is observed in the hDNA standard plots with each duplicate serial concentration step clustered together distinctly from other concentrations [Figure 2.3c]. However, [Figure 2.3d] reports a less clear delineation between concentrations. Additionally, one of the 6.4 ng/µL duplicate sample of *pk*DNA did not pass the set threshold. However, the calibration plot remained applicable for downstream experimental analyses and comparisons.

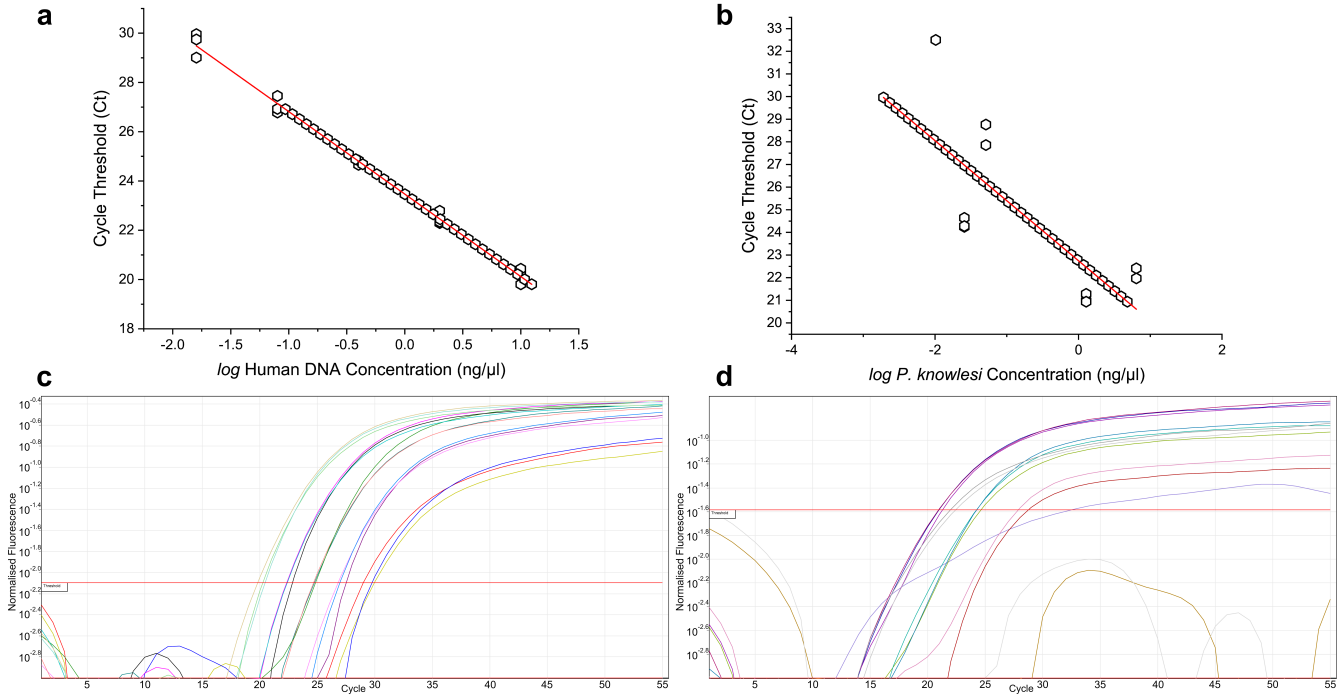


Figure 2.3: Calibration plots and quantitation traces generated from standard hDNA and *pk*DNA. (a) hDNA standard calibration plot generated to accurately estimate the concentration of human DNA within a mixed DNA sample. The standard plot was formulated using serial dilutions of TaqMan™ hDNA Control Genomic from 10 ng/µL to 0.016 ng/µL. **(b)** *pk*DNA standard calibration plot generated to accurately calculate the concentration of *P. knowlesi* DNA within a mixed DNA sample. The standard plot was formulated using serial dilutions of a Cultured PkA1-H.1 isolate from 6.4 ng/µL to 0.01024 ng/µL. Quantitation traces **(c,d)** were generated by the Rotor-Gene Q Series Software as part of the calibration plot. Each line refers to a serial diluted concentration with duplicate concentrations located in close proximity. Red horizontal line indicates the set thresholds of 0.008 for hDNA and 0.0261 for pkDNA. Using known DNA concentrations, the cycle threshold (Ct) was calculated as the cycle where the fluorescence exceeds the set threshold whereby the higher the DNA concentration present, the smaller the Ct value. For hDNA **(c)**, initial set threshold of 0.021 was manually lowered to 0.008 resulting in the equations: $conc.(ng/\mu L) = 10^{(-0.299*\mathrm{CT} + 7.011)}$ and $CT = -3.347 * log(conc.) + 23.469$ to calculate unknown DNA concentration and Ct, respectively. **(d)** The *pk*DNA threshold was set at 0.0261, resulting in a standard plot with the equations: $conc.(ng/\mu L) = 10^{(-0.377*\mathrm{CT} + 8.573)}$ and $CT = -2.653 * log(conc.) + 22.746$ to calculate unknown DNA concentration and Ct, respectively.

2.5.1 Preliminary Experiments

Leucocyte depletion of healthy donor blood using Saponin and CD45 DynaBeads

Saponin was first tested on healthy donor blood to determine its efficacy on whole blood alone and also in conjunction with the CD45 DynaBeads [Figure 2.4]. Experiment 1, carried out with the conditions '*Saponin_CD45_Donor*' [*Exp.1*,Appendix Table A.3] shows little change between the untreated Control and the Saponin-alone-treated sample [*Exp.1*,Table 2.6*i*], with a δ change of −0.38. On the other hand, leucocyte depletion using Saponin/CD45 DynaBeads [see *chapter 2 subsubsection 2.4.4*] showed a δ change of -10.06 between Control and Saponin/CD45-treated; indicating that there was a marked reduction in the human DNA present in the Saponin/CD45-treated sample [*Exp.1*,Table 2.6*j*]. The retained DynaBeads pellet [*Exp.1*,Table 2.6*o*] – *carrying the depleted leucocytes* – also suggest that a large proportion of the reduced hDNA content between the Control and Saponin/CD45-treated samples were present in the DynaBeads pellet. As the donor blood did not contain any parasites, no parasitic DNA, apart from the *pk*DNA control sample, was detected [*Exp.1*,Table 2.7].

Simulated infected whole blood leucocyte depletion

Experiment 2 ('*Saponin_CD45_Simulated*') and experiment 3 ('*Saponin_CD45_Simulated_2*') [Appendix Table A.3] assessed depletion using Saponin and CD45 DynaBeads on simulated infected whole blood [*chapter 2 subsubsection 2.4.4*, Figure 2.4]. Both experiments reported a positive δ change in the hDNA sample, between the simulated whole blood Control and the Saponin-alone treatment (*Exp. 2*: δ=3.58; *Exp. 3*: δ=3.82); indicating an increase in the hDNA present in the Saponin-alone-treated sample than the Control sample [*Exp.2*, *Exp.3*,Table 2.6*i*]. On the other hand, a negative δ change value for hDNA is observed between the Control and the Saponin/CD45-treated sample [*Exp.2*: δ= -0.17; *Exp.3*: δ= -0.98] [Table 2.6*j*].

For *pk*DNA, experiments 2 & 3 report a positive δ change in the Saponin-alone sample, indicating retention of *pk*DNA; suggesting the Saponin-alone treatment (δ=0.68; δ=4.69) [*Exp.2*, *Exp.3*, Table 2.7*i*] performed greater parasite retention/enrichment

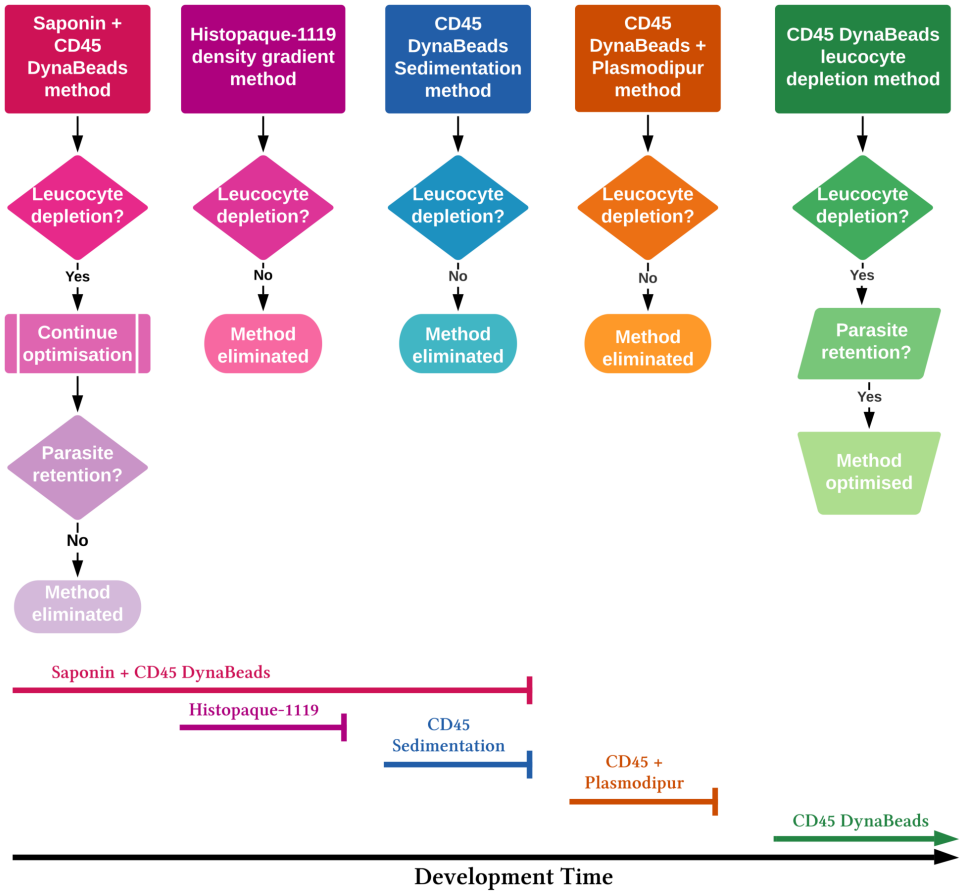


Figure 2.4: Timeline of the development of methods for leucocyte depletion. After a proof of concept for CD45 DynaBeads, multiple methods for leucocyte depletion in simulated infected patient blood were developed. Where leucocyte depletion was not achieved, the method is eliminated. Where depletion is successful, optimisation occurs to determine adequate parasite retention. Where this is not achieved, the method is also discarded. The directional timeline is shown below in a black arrow, with the timeline for each method displayed in corresponding colours.

than the combined Saponin/CD45 treatment (*Exp.2*: δ=-8.94; *Exp.3*: δ=not calculated) [Table 2.7*j*]. In a bid to increase the hDNA depletion, the method was further developed through the addition of a Histopaque step [*chapter 2 subsubsection 2.4.4*, Figure 2.4]. The mass separation using Histopaque-1119 did not improve hDNA depletion as well as the Saponin/CD45 DynaBeads method [*Exp.4*,Table 2.6*j,k,l*; Appendix Table A.3]. Indeed, the Histopaque-1119 densities resulted in δ changes of -1.39 (HPQ_200) and -4.08 (HPQ_1000), in comparison to a δ change of -9.30 from the Saponin/CD45 treatment [*Exp.4*,Table 2.6*j,k,l*]. Both Histopaque-1119 densities used in Exp 4, did not result in better *pk*DNA retention in comparison to the Saponin/CD45 treatment, rather, both Histopaque densities culminate in higher Cts and thus poorer *pk*DNA retention [*Exp.4*, Table 2.7*j,k,l*].

Exp.4 showed satisfactory hDNA depletion however, *pk*DNA required further developments. Thus, further attempts to increase *pk*DNA retention with maximum hDNA depletion using the simulated whole blood led to the sedimentation of the DynaBeads in order to avoid placing the microfuge tube in a magnetic field [*chapter 2 subsubsection 2.4.4*, Figure 2.4, Appendix Table A.3]. Leucocyte depletion was first observed in experiment 5, with a high hDNA δ change (δ= -21.33) and relatively low *pk*DNA δ change (δ= -4.96) [*Exp.5*,Table 2.6*j*; 2.7*j*]. However, this was not consistent and repeat iterations (*Exp.6*) indicated lower leucocyte depletion and an inability to detect *pk*DNA after treatment [*Exp.6*,Table 2.7*c,j*; Exp.6,Appendix Table A.6].

The effect of Saponin concentration on *P. knowlesi* DNA yield was determined by varying the concentration of Saponin used for RBC lysis [*chapter 2 subsection 2.4.3*, *subsubsection 2.4.10*]. DNA concentration results from NanoDrop 2000 and Qubit [Appendix Table A.8] showed considerable reduction at each Saponin concentration. However, the 5 % Saponin showed the highest overall DNA concentration (5.2 ng/µL) and the closest average 260/280 (1.67) to the ideal ratio of ~1.8. In contrast, the Qubit showed the 1 % Saponin sample retained the highest overall DNA concentration (5.04 ng/µL) but no 260/280 reading was available for the Qubit platform [Appendix Table A.8b].

qPCR outputs for Saponin-lysed simulated whole blood revealed similar Ct values between 1 % and 5 % Saponin [Appendix Table A.9*a*]. However, the greatest hDNA

Table 2.6: Average cycle threshold (Ct) of normalised human DNA from preliminary leucocyte depletion methods

Exp.	Sample	Starting vol. (µL)		Adj. fac	Normalised Average Ct values								Delta Change						
		Control	*Treated*		*Control (a)*	*Sap (b)*	*Sap + CD45 (c)*	*HPQ 200 (d)*	*HPQ 1000 (e)*	*PLDR (f)*	*CD45 (g)*	*Pellet (h)*	*Sap (i)*	*Sap + CD45 (j)*	*HPQ 200 (k)*	*HPQ 1000 (l)*	*PLDR (m)*	*CD45 (n)*	*Pellet (o)*
1	Healthy	200	200	0	21.27	21.64	31.32	N.A	N.A	N.A	N.A	28.52	-0.38	-10.06	N.A	N.A	N.A	N.A	-7.25
2	Simulated	200	200	0	30.12	26.54	30.29	N.A	N.A	N.A	N.A	30.62	3.58	-0.17	N.A	N.A	N.A	N.A	-0.51
3	Simulated	200	200	0	30.81	26.99	31.79	N.A	N.A	N.A	N.A	N.C	3.82	-0.98	N.A	N.A	N.A	N.A	N.C
4	Simulated	200	200	0	19.66	20.26	28.95	21.04	23.73	N.A	N.A	19.69	-0.61	-9.30	-1.39	-4.08	N.A	N.A	-0.03
5	Simulated	200	200	0	21.05	23.41	42.38	N.A	N.A	N.A	N.A	21.01	-2.36	-21.33	N.A	N.A	N.A	N.A	0.04
6	Simulated	200	200	0	20.27	19.46	20.35	N.A	N.A	N.A	N.A	22.61	0.82	-0.07	N.A	N.A	N.A	N.A	-2.34
7	Patient - sks265	50	200	2.00	18.40	N.A	19.71	N.A	N.A	N.A	N.A	21.17	N.A	-1.31	N.A	N.A	N.A	N.A	-2.77
8	Patient - sks078	200	250	0.32	17.34	N.A	20.52	N.A	N.A	N.A	N.A	18.42	N.A	-3.18	N.A	N.A	N.A	N.A	-1.08
9	Patient - sks367	50	210	2.07	18.58	N.A	19.76	N.A	N.A	N.A	N.A	20.42	N.A	-1.17	N.A	N.A	N.A	N.A	-1.84
10	Patient - sks134	50	600	3.58	17.36	N.A	19.02	N.A	N.A	N.A	N.A	N.A	N.A	-1.66	N.A	N.A	N.A	N.A	N.A
11	Patient - sks269	50	550	3.46	15.66	N.A	N.A	N.A	N.A	24.77	N.A	18.21	N.A	N.A	N.A	N.A	-9.10	N.A	-2.54
12	Patient - sks074	50	450	3.17	16.81	N.A	N.A	N.A	N.A	N.A	22.7	17.54	N.A	N.A	N.A	N.A	N.A	-5.89	-0.73

Cycle threshold (Ct) values of the hDNA channel from qPCR fluorescence capture over the course of several experiments. The adjustment factor was calculated using the starting volumes of the control and treated samples. Experimental conditions can be found in Appendix Table A.3. Adjustment factor was employed on raw Ct values of the control sample [Appendix Tables A.4; A.5] for normalisation. Delta change is the amount of change between the normalised control Ct and the treated sample Ct. Where a treatment was not implemented in a particular experiment, it is represented as N.A.

Control - Untreated blood acting as isolate and experiment control
Treated - Whole blood to be leucocyte depleted depending on the conditions of the experiment.
Adj. fac - Adjustment factor calculated based on the proportion of the 'Control' starting volume to the 'Treated' starting volume
Sap - Saponin leucocyte depletion method
Sap + CD45 - Saponin lysis method followed by CD45 DynaBeads leucocyte depletion method (and the inverse of this)
HPQ 200 - Histopaque leucocyte depletion method using 200 µL Histopaque-1119
HPQ 1000 - Histopaque leucocyte depletion method using 1000 µL Histopaque-1119
PLDR - Leucocyte depletion first using the CD45 DynaBeads and then the Plasmodipur filtration method
CD45 - Leucocyte depletion method using only the CD45 DynaBeads
Pellet - DNA extracted from the separated DynaBeads pellet from the associated experiment
N.A - Not Applicable. This refers to conditions not applicable to the corresponding experiment because the conditions were not carried out within that experiment
N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated

Table 2.7: Average cycle threshold (Ct) of normalised *pk*DNA from preliminary leucocyte depletion methods

Exp.	Sample	Starting vol. (µL)		Adj. fac	Normalised Average Ct values								Delta Change						
		Control	*Treated*		*Control Sap (b) (a)*		*Sap + CD45 (c)*	*HPQ 200 (d)*	*HPQ 1000 (e)*	*PLDR (f)*	*CD45 (g)*	*Pellet (h)*	*Sap (i)*	*Sap + CD45 (j)*	*HPQ 200 (k)*	*HPQ 1000 (l)*	*PLDR (m)*	*CD45 (n)*	*Pellet (o)*
1	Healthy	200	200	0	23.23	N.C	N.C	N.A	N.A	N.A	N.A	N.C	N.C	N.C	N.A	N.A	N.A	N.A	N.C
2	Simulated	200	200	0	21.84	21.17	30.78	N.A	N.A	N.A	N.A	N.C	0.68	-8.94	N.A	N.A	N.A	N.A	N.C
3	Simulated	200	200	0	26.51	21.82	N.C	N.A	N.A	N.A	N.A	N.C	4.69	N.C	N.A	N.A	N.A	N.A	N.C
4	Simulated	200	200	0	17.34	18.36	25.61	24.00	26.23	N.A	N.A	20.21	-1.02	-8.27	-6.66	-8.89	N.A	N.A	-2.87
5	Simulated	200	200	0	17.96	20.06	22.92	N.A	N.A	N.A	N.A	19.51	-2.10	-4.96	N.A	N.A	N.A	N.A	-1.55
6	Simulated	200	200	0	19.91	20.79	N.C	N.A	N.A	N.A	N.A	N.C	-0.88	N.C	N.A	N.A	N.A	N.A	N.C
7	Patient - sks265	50	200	2.00	N.C	N.A	N.C	N.A	N.A	N.A	N.A	N.C	N.A	N.C	N.A	N.A	N.A	N.A	N.C
8	Patient - sks078	200	250	0.32	15.25	N.A	18.78	N.A	N.A	N.A	N.A	21.15	N.A	-3.53	N.A	N.A	N.A	N.A	-5.89
9	Patient - sks367	50	210	2.07	N.C	N.A	N.C	N.A	N.A	N.A	N.A	N.C	N.A	N.C	N.A	N.A	N.A	N.A	N.C
10	Patient - sks134	50	600	3.58	4.09	N.A	N.C	N.A	N.A	N.A	N.A	N.A	N.A	N.C	N.A	N.A	N.A	N.A	N.A
11	Patient - sks269	50	550	3.46	N.C	N.A	N.A	N.A	N.A	N.C	N.A	N.C	N.A	N.A	N.A	N.A	N.C	N.A	N.C
12	Patient - sks074	50	450	3.17	17.41	N.A	N.A	N.A	N.A	N.A	21.15	23.15	N.A	N.A	N.A	N.A	N.A	-3.74	-5.74

Cycle threshold (Ct) values of the *pk*DNA channel from qPCR fluorescence capture over the course of several experiments. The adjustment factor was calculated using the starting volumes of the control and treated samples. Experimental conditions can be found in Appendix Table A.3. Adjustment factor was employed on raw Ct values of the control sample [Appendix Tables A.6; A.7] for normalisation. Delta change is the amount of change between the normalised control Ct and the treated sample Ct. Where a treatment was not implemented in a particular experiment, it is represented as N.A. Where a treatment was employed but no Ct reading was reported, it is represented as N.C.

Control - Untreated blood acting as isolate and experiment control
Treated - Whole blood to be leucocyte depleted depending on the conditions of the experiment.
Adj. fac - Adjustment factor calculated based on the proportion of the 'Control' starting volume to the 'Treated' starting volume
Sap - Saponin leucocyte depletion method
Sap + CD45 - Saponin lysis method followed by CD45 DynaBeads leucocyte depletion method (and the inverse of this)
HPQ 200 - Histopaque leucocyte depletion method using 200 µL Histopaque-1119
HPQ 1000 - Histopaque leucocyte depletion method using 1000 µL Histopaque-1119
PLDR - Leucocyte depletion first using the CD45 DynaBeads and then the Plasmodipur filtration method
CD45 - Leucocyte depletion method using only the CD45 DynaBeads
Pellet - DNA extracted from the separated DynaBeads pellet from the associated experiment
N.A - Not Applicable. This refers to conditions not applicable to the corresponding experiment because the conditions were not carried out within that experiment
N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated

Ct reduction is seen in the 10 % Saponin solution (Ct=22.88); a contrast to the control simulated whole blood Ct (Ct=19.38) [Appendix Table A.9]. On the other hand, *pk*DNA retention was less defined with at least one entry of each Saponin concentration triplicate failing to be detected or surpass the set threshold [Appendix Table A.9*b*]. In this instance, higher retention is desired, hence, the saponin concentration with the lowest numerical Ct value is the most appropriate. As such, 1 % Saponin reported the highest *pk*DNA content due to having the lowest Ct value (26.06); although, this remains a marked reduction from the average control Ct value (19.50).

Preliminary Leucocyte depletion using clinical samples

Table 2.8: Clinical patient *P. knowlesi*-infected Whole Blood used in preliminary experiments

Study Reference	Parasitaemia (parasites/ul)	% Life Stage (Rings : Trophs : Schizonts)
sks265	13752	90 : 6 : 0
sks078	31000	90 : 3 : 3
sks367	35550	100 : 0 : 0
sks134	38024	89 : 11 : 0
sks269	47850	44 : 24 : 10
sks074	164280	N/A

Patient isolates used in preliminary experiments for leucocyte depletion using adaptations of the inverse Saponin/CD45 DynaBeads method [*chapter 2 subsubsection 2.4.4*] and the CD45 depletion method [*chapter 2 subsection 2.4.5*]. Parasitaemia was recorded at the time of extracting whole blood at the hospital or local health centre. Parasite life stage was determined at the Universiti Malaysia Sarawak (UNIMAS) laboratories.

Trophs - Trophozoites

Upon using patient isolates, the order of the Saponin/CD45 DynaBeads leucocyte depletion method was inverted [Figure 2.4; *chapter 2 subsubsection 2.4.4*]. From this, patient isolate sks265 had 1.4 ng/µL and sks078, 0.8 ng/µL. Ct results show the presence of hDNA in the sample with the DynaBeads pellet [*Exp.7*,Table 2.6*o*]. No numerical Ct value returned for sks265, suggesting no *pk*DNA was recovered [Table 2.7, Appendix Table A.6, A.7]. Thus, new *P. knowlesi* qPCR primers and probes were prepared to test

the assay [Appendix Figure A.2 and Appendix Table A.10]. However, no improvement was observed between original and new formulations of the qPCR primers and probes [Appendix Table A.10], thereby supporting the observations made in experiment 7 with sks265.

A return to the inverse method of the Saponin and CD45 DynaBeads method [*chapter 2 subsubsection 2.4.4*], using sks367 and sks134 resulted in similar hDNA depletion with deltas of -1.17 and -1.66 respectively [*Exp.9*; *Exp.10*, Table 2.6*j*]. Once again, no *pk*DNA Ct values were reported in the qPCR results [*Exp.9*; *Exp.10*, Table 2.7, Appendix Table A.6, Table A.7].

With this, Plasmodipur was substituted for Saponin, to determine its effect in conjunction with the CD45 DynaBeads [*chapter 2 subsubsection 2.4.4*]. With hopes of increasing parasite retention, the CD45 Plasmodipur method [*chapter 2 subsubsection 2.4.4*] was assessed using patient isolate sks269. After treatment using Plasmodipur and CD45, sks269 reported a hDNA depletion δ change of -9.10 [*Exp.11*, Table 2.6*m*]. In contrast, the associated DynaBeads pellet reported δ of -2.54; indicating the DynaBeads were associated with a majority of the depleted leucocytes [*Exp.11*, Table 2.6*o*]. However, parasite *pk*DNA retention was poor, with no detectable *pk*DNA by qPCR [*Exp.11*, Table 2.7*m,o*]. The matched sks269 control sample also failed to detect *pk*DNA indicating that perhaps the parasitaemia was too low or the sample had degraded.

Human DNA (hDNA) depletion in the Plasmodipur/CD45 method [*chapter 2 subsubsection 2.4.4*] prompted the use of a CD45 DynaBeads only depletion method [*chapter 2 subsection 2.4.5*, Figure 2.4] using patient isolate sks074. Results show the hDNA δ change to be -5.89 for the CD45 leucocyte depleted sample [*Exp.12*, Table 2.6*n*]. While this is not as large a reduction in hDNA as observed in previous experimental conditions, the majority of the input hDNA are associated with the Bead pellet [*Exp.12*, Table 2.7*o*]. A *pk*DNA δ of -3.74 was reported, while the associated DynaBead pellet (*bound to depleted leucocytes*) had a δ of -5.74 [*Exp.12*, Table 2.7*n,o*]. Thus, although *pk*DNA is being lost, the action of the CD45 DynaBeads may not be the driver for the loss; resulting in the method being continued for subsequent patient isolates.

2.5.2 Leucocyte Depletion of *P. knowlesi*-infected patient Whole blood (ipWB) using CD45 DynaBeads

Method optimisation and development indicated a need to dilute the thawed patient whole blood in order to facilitate leucocyte capture and binding by the CD45 DynaBeads. This was demonstrated by the successful depletion of leucocytes and retention of *pk*DNA in the clinical sample sks074 above [*chapter 2 subsection 2.4.5*].

Table 2.9: DNA concentration measured in a subset of infected patient isolates after leucocyte depletion

Isolates	NanoDrop (ng/µL)			Qubit (ng/µL)			% DNA Depleted	
	Control	*CD45 Treated*	*Pellet*	*Control*	*CD45 Treated*	*Pellet*	*NanoDrop*	*Qubit*
PkA1H1(h)	-	-	-	44.85	2.48	27.2	-	94.47
PkA1H1(k)	-	-	-	31.668	23.6	25.4	-	25.48
PkA1H1(n)	-	-	-	109.98	10	29.6	-	90.91
sks047(b)	-	10.80	-	-	47.6	-	-	-
sks048(b)	-	-	-	100.1	9.12	14.6	-	90.89
sks050	220.80	9.65	12.30	-	93.4	-	95.63	-
sks058(a)	59.85	2.00	27.10	-	9.9	-	96.66	-
sks070(b)	-	-	-	212.4	13.7	36.8	-	82.67
sks234	18.20	0.80	24.85	-	-	-	95.60	-
sks339(a)	-	11.25	-	-	116.00	-	-	-

Full DNA concentration quantifications for all isolates processed with the CD45 DynaBeads leucocyte depletion method is presented in Appendix Table B.3 for Qubit and Appendix Table B.4 for the NanoDrop 2000. Where both the control and the CD45 treated were quantified, total DNA percentage reduction was calculated and presented for both the NanoDrop and the Qubit platforms. Percentage reduction acts as a quantifiable measure of the CD45 DynaBeads leucocyte depletion method's effect on the normalised control DNA measured.

- indicates where a sample was not quantified, due to a lack of the sample due to being discarded or more often, occurs when the sample is quantified prior to DNA sequencing (chapter 4 subsection 4.3.1)

To generate sufficient parasite DNA from the thawed patient samples necessary for genome sequencing, clinical samples with >80,000 parasites/µL were first processed

before reduction to as low as 18,000 parasites/μL [Appendix Table B.1]. Twenty unique patient isolates and twelve unique *P. knowlesi* clones PkA1-H.1 (5 - 7 % parasitaemia) were processed [*chapter 2 subsection 2.4.5*]. From this, a representative subset of the total DNA concentration before and after leucocyte depletion is given in Table 2.9 and all results in Appendix Tables B.3 and B.4. NanoDrop and Qubit DNA quantification were not available for all isolates processed, with NanoDrop used directly after extraction and Qubit prior to sequencing.

Once normalised by volume and adjustment factor, the matched un-depleted control samples show that hDNA depletion was successful. Across all isolates processed, average total DNA reduction reported was 95.88 % and 84.57 % for NanoDrop and Qubit respectively [Appendix Tables B.3 and B.4]. The qPCR outputs provide quantification for the hDNA and *pk*DNA content in the samples recorded as cycle threshold (Ct). Using the same subset of isolates selected above [Table 2.9], the recorded Cts for the hDNA and *pk*DNA are presented in Table 2.10 with a full list of processed isolates in Figure 2.5 and in the Appendix [Appendix Tables B.5, B.6, B.7 and B.8]. While the use of Ct allows for consistency, often, the qPCR outputs were unable to resolve the Ct, either due to not surpassing the set threshold of 10 % (*seen as NTC*) or not providing any fluorescence value, returning a blank reading (*seen as N.V*) [Appendix Tables B.5 and B.6]. Some isolates returned invalid Ct readings and are labelled as M.Ct. No average Ct values could be calculated in such cases and this represented as N.C [Appendix Tables B.7 and B.8].

After normalisation, average hDNA Ct values indicate that the clinical isolates possess higher hDNA than the cultured samples, as expected; however, similar average Cts are reported for the *pk*DNA of the cultured and clinical isolates [Appendix Tables B.7, B.8]. The effect of the CD45 DynaBeads treatment is displayed in Figure 2.5, where the CD45-treated (leucocyte depleted) and the pellet (*DynaBeads pellet of the depleted leucocytes*) samples for both hDNA and *pk*DNA show higher Cts and thus less target DNA than the matched control [Appendix Tables B.7 and B.8].

By quantifying the amount of hDNA and *pk*DNA associated with the CD45 DynaBeads magnetic pellet, the efficiency of the antibody-mediated isolation process can be assessed. As such, if 100 % efficient, all hDNA depleted in the CD45-treated sample will be

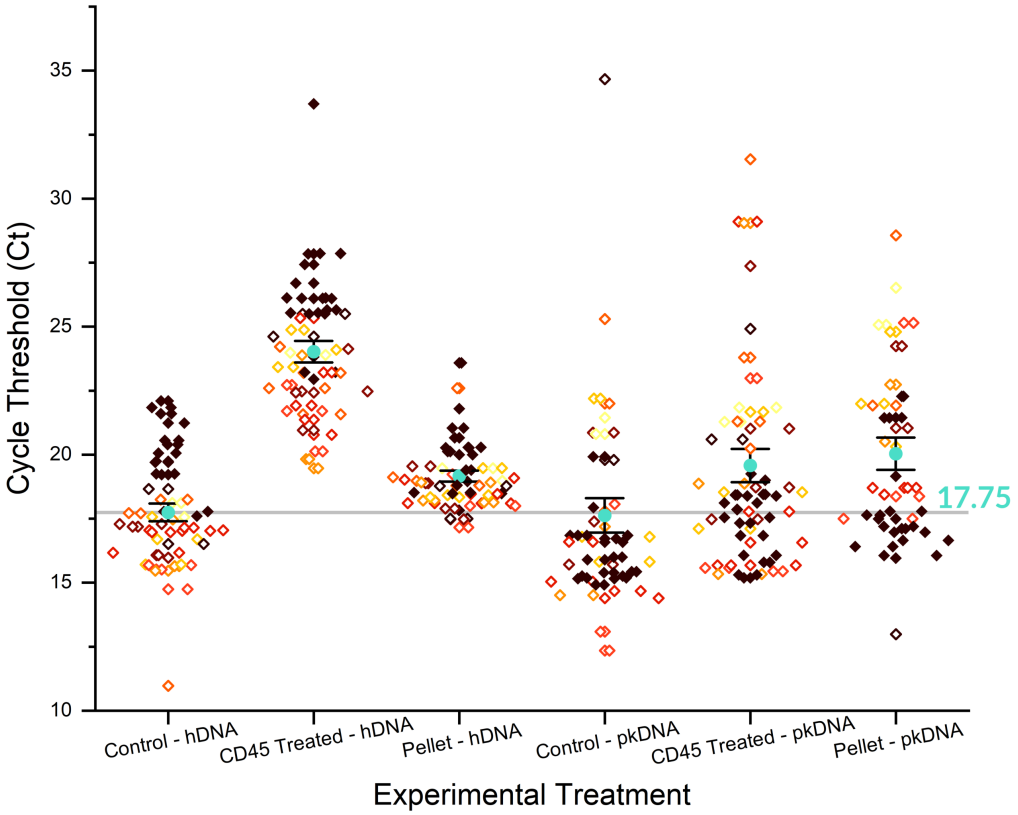


Figure 2.5: Ct Values of all isolates after leucocyte depletion using CD45 DynaBeads. hDNA and *pk*DNA refer to the Human DNA and *P. knowlesi* DNA content within each sample. Whiskers (black bars) show the standard error ± 1 and the grey horizontal line denotes the mean Ct value for the Human Control dataset at 17.75. No outliers were removed from the datasets. Cultured and clinical patient isolates are colour coded. Cultured PkA1H1 strains are filled diamonds while patient isolates have white centres. Isolates with more than one sample such as PkA1H1 (a - c) and sks047(a - b) possess the same colour. The mean for each dataset is denoted by a cyan circle.

associated with the corresponding pellet, and no *pk*DNA would be present within the pellet samples. Indeed, using the percentage input method for normalisation, it is possible to determine the theoretical percentage of hDNA and *pk*DNA associated with the pellet. Although, this is not perfect, as evidenced in the 136.09 % hDNA calculated to be associated with the pellet sample for PkA1H1(h) in contrast to the normalised control Ct [Table 2.10a]. A similar phenomenon can be seen in the percentage input for sks133

Table 2.10: Normalised Ct value of a subset of isolates which have been leucocyte depleted using CD45 DynaBeads

a *hDNA Ct Values*

Isolates	Adj. factor	Normalised Average Ct values			Delta Change		% inputs		% hDNA Reduction
		Control	CD45 Treated	Pellet	CD45 Treated	Pellet	CD45 Treated	Pellet	CD45 Treated
PkA1H1(h)	5.29	19.25	25.50	18.81	-6.25	0.44	1.31	136.09	98.69
PkA1H1(k)	5.29	21.84	25.66	19.40	-3.82	2.45	7.10	546.26	92.90
PkA1H1(n)	5.29	17.60	23.82	18.52	-6.22	-0.92	1.34	53.02	98.66
sks047(b)	6.49	16.08	20.96	17.48	-4.88	-1.39	3.40	38.11	96.60
sks048(b)	5.02	16.97	23.21	19.09	-6.23	-2.11	1.33	23.13	98.67
sks050	5.58	15.52	22.72	19.24	-7.20	-3.72	0.68	7.56	99.32
sks058	3.17	17.71	24.22	18.79	-6.51	-1.08	1.10	47.31	98.90
sks070(b)	4.21	15.47	19.83	18.21	-4.36	-2.74	4.87	14.92	95.13
sks234	4.70	17.03	21.37	18.47	-4.34	-1.44	4.95	36.97	95.05
sks339(a)	4.25	17.05	21.92	18.10	-4.86	-1.05	3.44	48.37	96.56

b *pkDNA Ct Values*

Isolates	Adj. factor	Normalised Average Ct values			Delta Change		% inputs		% *pk*DNA Loss
		Control	CD45 Treated	Pellet	CD45 Treated	Pellet	CD45 Treated	Pellet	CD45 Treated
PkA1H1(h)	5.29	16.58	19.25	17.64	-2.67	-1.06	15.76	47.95	84.24
PkA1H1(k)	5.29	15.26	15.19	15.96	0.07	-0.69	105.31	61.97	-5.31
PkA1H1(n)	5.29	15.43	16.07	21.44	-0.64	-6.01	64.15	1.56	35.85
sks047(b)	6.49	17.39	17.48	24.24	-0.09	-6.85	93.83	0.87	6.17
sks048(b)	5.02	14.68	15.68	18.44	-0.99	-3.75	50.27	7.42	49.73
sks050	5.58	12.35	15.45	17.50	-3.10	-5.15	11.66	2.82	88.34
sks058	3.17	17.80	20.24	21.92	-2.44	-4.12	18.37	5.75	81.63
sks070(b)	4.21	14.51	15.34	22.74	-0.83	-8.23	56.27	0.33	43.73
sks234	4.70	N.C	29.10	N.C	N.C	N.C	N.C	N.C	N.C
sks339(a)	4.25	15.04	16.57	18.71	-1.53	-3.67	34.56	7.87	65.44

Control Ct values, Delta change and percentage input values were normalised as described in *chapter 2 subsection 2.4.9*. Percentage hDNA reduction and percentage *pk*DNA loss were calculated by subtracting the percentage input from 100. Raw data are in Appendix Table B.5 and Appendix Table B.6 while the complete normalised data and subsequent calculations are in Appendix Table B.7 and Appendix Table B.8.

N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated

reporting the CD45-treated sample having 86 593 % *pk*DNA input in contrast to the control Ct value [Appendix Table B.8].

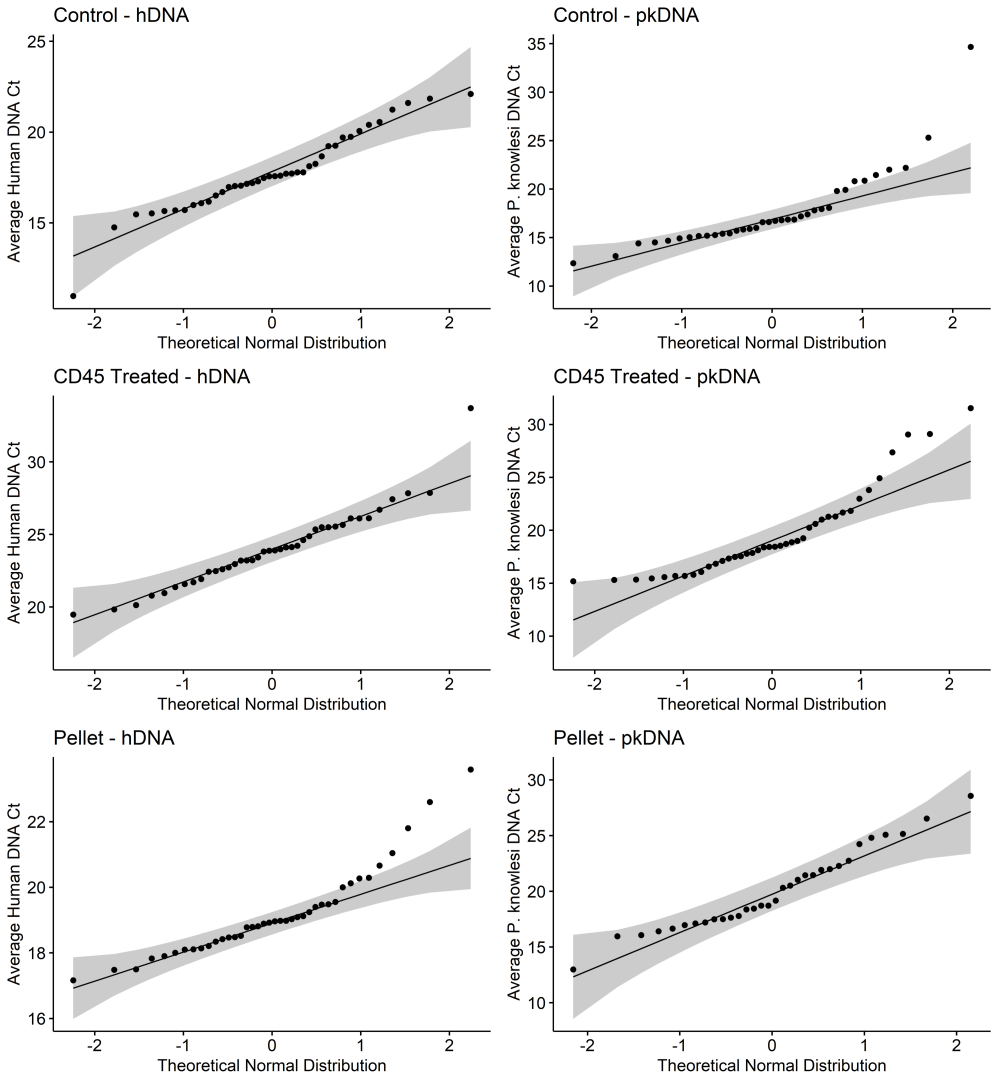


Figure 2.6: QQ plots of Ct results for human and *P. knowlesi* DNA. A visual representation for the normality test for each DNA subset of each treatment. All Ct values for the hDNA and *pk*DNA control, CD45 treated and pellet samples were plotted against the theoretical normal distribution to determine their normality. Grey bands represent error ranges within acceptable ranges of the normal distribution

However, this is likely due to the logarithmic nature of the Ct value, thus indicating the *pk*DNA in sks133 control sample was very low, and after CD45 treatment, the *pk*DNA had become enriched in the treated sample. Carrying out an inverse of the percentage input reveals the percentage DNA depleted in the CD45 treated sample and the percentage

DNA lost, which would be theoretically associated with the pellet sample. Doing this, Appendix Table B.7 reveals an overall average hDNA reduction of 97.83 % (median = 98.68 %) after CD45 treatment. Conversely, the pellet samples are reported to have only lost an average of 36.06 % (median = 60.12 %) of the total hDNA that was depleted from the CD45 treatment. Thus the DynaBeads pellets were associated with ~ 65 % of the leucocytes removed from the whole blood sample.

While the reduction of hDNA and retention of *pk*DNA can be observed in Figure 2.5 and Appendix B.2 - B.4, it was prudent to determine the statistical significance of the effect of the CD45 treatment on the Ct output of the qPCR. First, a Shapiro-Wilk test for normality was carried out on each treatment sample dataset (i.e. all hDNA and *pk*DNA Ct for the control and CD45-treated samples) and visually represented in Figure 2.6. From this the hDNA control subset (p = 0.08795) and the *pk*DNA pellet subset (p =0.2732) were normally distributed while the hDNA CD45-treated subset (p =0.02299), hDNA pellet subset (p =0.0009134), *pk*DNA control subset (p =7.38E-06) and *pk*DNA CD45-treated subset (p =0.0001358) were not normally distributed [Appendix Table B.9]. This is supported in Figure 2.6 where plots containing hits outside of the grey boundaries differ from the theoretical normal distribution. Due to the non-parametric outcome of most of the subsets, a Spearman's correlation test was employed to determine the relationship between each treatment. From this, a significant positive correlation is reported between all tested conditions with varying degrees of strength [Table 2.11].

Table 2.11: Spearman's correlation test to determine correlation between treatments and their Cts

Spearman's Correlation Test				
Condition 1	**Condition 2**	***p-value***	***Rho***	**Conclusion**
Control - hDNA	CD45 Treated - hDNA	1.48E-09	0.788911	Sig. strong positive
CD45 Treated - hDNA	Pellet - hDNA	7.60E-06	0.643212	Sig. moderate positive
Control - *pk*DNA	CD45 Treated - *pk*DNA	5.33E-10	0.82649	Sig. strong positive
CD45 Treated - *pk*DNA	Pellet - *pk*DNA	0.004651	0.487534	Sig. moderate positive

Non-parametric test to determine the relationship between the control and CD45 treated Ct to examine if the CD45 treated Ct outputs are affected by the Ct of the associated control sample Ct. The relationship between the CD45 treated and the corresponding pellet sample Ct are also examined for the same purpose. Conclusion on strength of correlation are determined using the standard scaling.

Table 2.12: Wilcoxon rank-sum test to determine effect of CD45 treatment on processed samples

Wilcoxon Rank-Sum Test				
Condition 1	**Condition 2**	***p-value***	**Pseudo-median**	**Conclusion**
Control - hDNA	CD45 Treated - hDNA	3.71E-08	-6.16	control is 6 Ct <CD45 treated
CD45 Treated - hDNA	Pellet - hDNA	5.46E-08	4.98	pellet is 5 Ct >CD45 treated
Control - *pk*DNA	CD45 Treated - *pk*DNA	7.96E-04	-1.15	control is 1 Ct <CD45 treated
CD45 Treated - *pk*DNA	Pellet - *pk*DNA	3.68E-04	-2.06	CD45 treated is 2 Ct <pellet

The medians of each compared condition differs significantly indicating that there is a statistical effect of the treatment on the Cts reported. Pesudo-median was calculated due to the presence of ties and empty values in the ranking within the equation hence no true p-value and median can be reported however, these are reported based on the data available.

Indeed, given that the subsets are paired, correlation would be expected however, a strong positive correlation bolsters the effect of the CD45 treatment on the *ip*WB processed. To determine the significance of the CD45 DynaBead-treatment's effect, a Wilcoxon-rank sum test was employed [Table 2.12]. From this, it can be concluded that due to the effect of the CD45 DynaBeads method, the CD45-treated sample reports significantly lower hDNA (p =3.71E-08) and *pk*DNA (p =7.96E-04) content in the processed samples [Table 2.12]. Importantly, while there is *pk*DNA loss, it is significantly less than the hDNA reduction, with an average *pk*DNA loss of 16 % in comparison to the un-depleted control sample.

2.6 Discussion

A method to deplete human DNA (hDNA), while retaining *P. knowlesi* DNA from thawed *P. knowlesi*-infected patient isolates through the use of anti-human CD45 DynaBeads was successfully developed. This involved assessing multiple methods of depletion and separation, ranging from a Saponin lysis method to a Plasmodipur filtration method. From these investigations, it is clear that depleting hDNA is rather trivial; via the use of almost any of the described methods. However, the difficulty lies in the retention of parasites to facilitate *pk*DNA extraction. Successive experiments greatly benefited

from previous findings, with improvements in one method informing procedures in the subsequent method.

While samples treated with only the Saponin lysis method showed increased *pk*DNA retention in comparison to the matched undepleted control sample, hDNA depletion was not satisfactory. Improvements on the technique were attempted throughout experiments 2 - 6 in a bid to improve hDNA depletion; however, these were unsuccessful, prompting the elimination of the Saponin lysis method. It is likely that while the Saponin was able to lyse the RBCs, the accompanying leucocytes were not affected and thus contributed to the increased hDNA observed downstream. Accordingly, the CD45 DynaBeads acted as a supplement to the action of the Saponin, removing the surviving leucocytes via antigen-antibody interactions. While we saw hDNA depletion as a consequence of the Saponin/CD45 method, *pk*DNA retention was lack-lustre in all experiments, regardless of the whole blood sample used (Simulated or patient). The loss of the precious parasite DNA could be due to the wash steps implemented in the Saponin lysis method after the rupturing of the infected-RBCs containing parasites.

As previously stated, Histopaque-1119 is usually combined with Histopaque-1077 to implement separation of red cells from granulocytic and mononuclear cells of the whole blood. As the aim was simply the separation of infected-RBCs from leucocytes that would be present in the plasma, Histopaque-1077 was omitted in the method described. Hence, it is unlikely that the absence of Histopaque-1077 affected the technique. Currently, we hypothesise that the loss of *pk*DNA in the Histopaque method might be due to free parasites present in the simulated-infected whole blood (*si*WB). Additionally, it is likely that the Histopaque method is not robust enough to be used on precious patient samples. In our experience, the delineation between the separated Histopaque/Whole blood mixture layers was challenging to discern visually, which may have led to both parasite loss and leucocyte retention.

We had hypothesised that the iron present in the parasites in the form of haemozoin could become magnetised and thus contribute to parasite loss when using the magnetic stand [27, 28]. However, this was not the case as haemozoin has been shown to increase over the course of the parasite's development, with more mature parasites like trophozoites containing more haemozoin than younger parasites like merozoites [29]. In our samples,

the largest proportion of parasites were in the ring stage, while trophozoites made up only ~20 % of parasites. Hence, it is unknown if this proportion would result in the considerable parasite loss observed in all methods involving the magnetic stand. Notwithstanding, the sedimentation method proved to be inconsistent, primarily due to the inconsistencies introduced by removing the supernatant from the sedimented DynaBeads. Additionally, other considerations which may affect the method, such as the speed at which the DynaBeads sediment/settle to the bottom of the microfuge tube, was unknown and may not be uniform for all experiments.

On the other hand, much like the Saponin lysis method, the Plasmodipur method likely lost *pk*DNA due to the number of washes imposed on the protocol. Alternatively, the filters utilised could have impeded the progress of larger or agglutinated parasites to pass through the filter pores. However, importantly, the method indicated that the leucocyte depletion was primarily done prior to the Plasmodipur filtration step. Indeed the depletion seen from the total DNA concentration measured shows evidence of the impact of the CD45 DynaBeads depletion. However, evidence also indicates that the DynaBeads do not account for all DNA depleted between the matched un-depleted control and the depleted CD45-treated samples. The additional lost DNA is likely due to free soluble DNA and free, circulating parasites within the infected patient whole blood.

Using our thawed patient samples, for the first time, we show evidence that the action of CD45 DynaBeads is statistically significant for hDNA depletion. Pinheiro et al. [12] determined that samples with <80 % hDNA contamination would be suitable for whole-genome sequencing. While we have no such proportional description, with average hDNA reductions of >97 %, it is likely that the resulting enriched DNA recovered would adhere to this threshold. To ensure this, infected patient isolates with sufficient starting parasitaemia would greatly benefit the yield of the CD45 depletion method described. However, the method is not perfect in its action as *pk*DNA loss is still present, though, from previous experiments, evidence suggests *pk*DNA loss is unlikely due to the action of the DynaBeads, but rather, the wash steps and dilutions prior to the addition of the DynaBeads.

It is this unknown loss of *pk*DNA which acts as a limiting factor to the CD45 depletion method. While the technique currently produces DNA of sufficient concentration

and quality from thawed infected whole blood, these are isolates with relatively high parasitaemia counts. To expand the optimal working parasitaemia range for the method to be below ~18 000 µL, further optimisation would be required; perhaps through the use of higher centrifugation speeds after the 1:50 dilution steps, ensuring lighter components of the whole blood are pelleted. Other improvements could include the introduction of a double pass of Buffer AL for the DNA elution after DNA extraction. As such, rather than eluting in 150 µL Buffer AL, a first pass of 75 µL Buffer AL is carried out, followed by the centrifugation as described. After this, a second pass of 75 µL Buffer AL is employed, with subsequent steps as described. There is some preliminary evidence that such a method might improve the DNA recovery yield due to the initial elution reducing any 'clogging', which may have occurred on the filtration column's surface.

Nevertheless, the aims set out were achieved with the development and optimisation of the CD45 depletion method. With sufficient, repeatable and robust depletion of hDNA from thawed infected patient whole blood, the recovered DNA can be taken forward for whole genome sequencing using the Oxford Nanopore sequencing platform.

2.7 References

[1] **S. Auburn**, **S. Campino**, **T. G. Clark**, **A. A. Djimde**, **I. Zongo**, **R. Pinches**, **M. Manske**, **V. Mangano**, **D. Alcock**, **E. Anastasi**, **G. Maslen**, **B. MacInnis**, **K. Rockett**, **D. Modiano**, **C. I. Newbold**, **O. K. Doumbo**, **J. B. Ouédraogo**, and **D. P. Kwiatkowski**. "An Effective Method to Purify Plasmodium Falciparum DNA Directly from Clinical Blood Samples for Whole Genome High-Throughput Sequencing". In: *PLOS ONE* **6**:7 (July 2011), e22213. DOI: 10.1371/journal.pone.0022213 (see pp. 47–50)

[2] **J. M. Carlton**, **J. H. Adams**, **J. C. Silva**, **S. L. Bidwell**, **H. Lorenzi**, **E. Caler**, **J. Crabtree**, **S. V. Angiuoli**, **E. F. Merino**, **P. Amedeo**, **Q. Cheng**, **R. M. R. Coulson**, **B. S. Crabb**, **H. A. del Portillo**, **K. Essien**, **T. V. Feldblyum**, **C. Fernandez-Becerra**, **P. R. Gilson**, **A. H. Gueye**, **X. Guo**, **S. Kang'a**, **T. W. A. Kooij**, **M. Korsinczky**,

E. V.-S. MEYER, **V. NENE**, **I. PAULSEN**, **O. WHITE**, **S. A. RALPH**, **Q. REN**, **T. J. SARGEANT**, **S. L. SALZBERG**, **C. J. STOECKERT**, **S. A. SULLIVAN**, **M. M. YAMAMOTO**, **S. L. HOFFMAN**, **J. R. WORTMAN**, **M. J. GARDNER**, **M. R. GALINSKI**, **J. W. BARNWELL**, and **C. M. FRASER-LIGGETT**. "Comparative Genomics of the Neglected Human Malaria Parasite Plasmodium Vivax". In: *Nature* **455**:7214 (Oct. 2008), 757–763. DOI: 10.1038/nature07327 (see p. 48)

[3] **M. J. GARDNER**, **N. HALL**, **E. FUNG**, **O. WHITE**, **M. BERRIMAN**, **R. W. HYMAN**, **J. M. CARLTON**, **A. PAIN**, **K. E. NELSON**, **S. BOWMAN**, **I. T. PAULSEN**, **K. JAMES**, **J. A. EISEN**, **K. RUTHERFORD**, **S. L. SALZBERG**, **A. CRAIG**, **S. KYES**, **M.-S. CHAN**, **V. NENE**, **S. J. SHALLOM**, **B. SUH**, **J. PETERSON**, **S. ANGIUOLI**, **M. PERTEA**, **J. ALLEN**, **J. SELENGUT**, **D. HAFT**, **M. W. MATHER**, **A. B. VAIDYA**, **D. M. A. MARTIN**, **A. H. FAIRLAMB**, **M. J. FRAUNHOLZ**, **D. S. ROOS**, **S. A. RALPH**, **G. I. MCFADDEN**, **L. M. CUMMINGS**, **G. M. SUBRAMANIAN**, **C. MUNGALL**, **J. C. VENTER**, **D. J. CARUCCI**, **S. L. HOFFMAN**, **C. NEWBOLD**, **R. W. DAVIS**, **C. M. FRASER**, and **B. BARRELL**. "Genome Sequence of the Human Malaria Parasite Plasmodium Falciparum". In: *Nature* **419**:6906 (Oct. 2002). DOI: 10.1038/nature01097 (see p. 48)

[4] **G. G. RUTLEDGE**, **U. BÖHME**, **M. SANDERS**, **A. J. REID**, **J. A. COTTON**, **O. MAIGA-ASCOFARE**, **A. A. DJIMDÉ**, **T. O. APINJOH**, **L. AMENGA-ETEGO**, **M. MANSKE**, **J. W. BARNWELL**, **F. RENAUD**, **B. OLLOMO**, **F. PRUGNOLLE**, **N. M. ANSTEY**, **S. AUBURN**, **R. N. PRICE**, **J. S. MCCARTHY**, **D. P. KWIATKOWSKI**, **C. I. NEWBOLD**, **M. BERRIMAN**, and **T. D. OTTO**. "Plasmodium Malariae and P. Ovale Genomes Provide Insights into Malaria Parasite Evolution". In: *Nature* **542**:7639 (Feb. 2017), 101–104. DOI: 10.1038/nature21038 (see p. 48)

[5] **A. PAIN**, **U. BÖHME**, **A. E. BERRY**, **K. MUNGALL**, **R. D. FINN**, **A. P. JACKSON**, **T. MOURIER**, **J. MISTRY**, **E. M. PASINI**, **M. A. ASLETT**, **S. BALASUBRAMMANIAM**, **K. BORGWARDT**, **K. BROOKS**, **C. CARRET**, **T. J. CARVER**, **I. CHEREVACH**, **T. CHILLINGWORTH**, **T. G. CLARK**, **M. R. GALINSKI**, **N. HALL**, **D. HARPER**, **D. HARRIS**, **H. HAUSER**, **A. IVENS**,

C. S. JANSSEN, **T. KEANE**, **N. LARKE**, **S. LAPP**, **M. MARTI**, **S. MOULE**, **I. M. MEYER**, **D. ORMOND**, **N. PETERS**, **M. SANDERS**, **S. SANDERS**, **T. J. SARGEANT**, **M. SIMMONDS**, **F. SMITH**, **R. SQUARES**, **S. THURSTON**, **A. R. TIVEY**, **D. WALKER**, **B. WHITE**, **E. ZUIDERWIJK**, **C. CHURCHER**, **M. A. QUAIL**, **A. F. COWMAN**, **C. M. R. TURNER**, **M. A. RAJANDREAM**, **C. H. M. KOCKEN**, **A. W. THOMAS**, **C. I. NEWBOLD**, **B. G. BARRELL**, and **M. BERRIMAN**. "The Genome of the Simian and Human Malaria Parasite Plasmodium Knowlesi". In: *Nature* **455**:7214 (Oct. 2008), 799–803. DOI: 10.1038/nature07306 (see pp. 48, 49)

[6] **THE 1000 GENOMES PROJECT CONSORTIUM**. "A Global Reference for Human Genetic Variation". In: *Nature* **526**:7571 (Oct. 2015), 68–74. DOI: 10.1038/nature15393 (see p. 48)

[7] **T. D. OTTO**, **J. C. RAYNER**, **U. BÖHME**, **A. PAIN**, **N. SPOTTISWOODE**, **M. SANDERS**, **M. QUAIL**, **B. OLLOMO**, **F. RENAUD**, **A. W. THOMAS**, **F. PRUGNOLLE**, **D. J. CONWAY**, **C. NEWBOLD**, and **M. BERRIMAN**. "Genome Sequencing of Chimpanzee Malaria Parasites Reveals Possible Pathways of Adaptation to Human Hosts". In: *Nature Communications* **5**:1 (Dec. 2014), 4754. DOI: 10.1038/ncomms5754 (see p. 49)

[8] **K. SRIPRAWAT**, **S. KAEWPONGSRI**, **R. SUWANARUSK**, **M. L. LEIMANIS**, **U. LEK-UTHAI**, **A. P. PHYO**, **G. SNOUNOU**, **B. RUSSELL**, **L. RENIA**, and **F. NOSTEN**. "Effective and Cheap Removal of Leukocytes and Platelets from Plasmodium Vivax Infected Blood". In: *Malaria Journal* **8**: (June 2009), 115. DOI: 10.1186/1475-2875-8-115 (see p. 49)

[9] **S. CHEVALLEY**, **A. COSTE**, **A. LOPEZ**, **B. PIPY**, and **A. VALENTIN**. "Flow Cytometry for the Evaluation of Anti-Plasmodial Activity of Drugs on Plasmodium Falciparum Gametocytes". In: *Malaria Journal* **9**:1 (Feb. 2010), 49. DOI: 10.1186/1475-2875-9-49 (see pp. 49, 50)

[10] **D. T. X. TRANG**, **N. T. HUY**, **T. KARIU**, **K. TAJIMA**, and **K. KAMEI**. "One-Step Concentration of Malarial Parasite-Infected Red Blood Cells and Removal of Contaminating White Blood Cells". In: *Malaria Journal* (2004), 7 (see pp. 49, 50)

[11] **A. BOISSIÈRE**, **C. ARNATHAU**, **C. DUPERRAY**, **L. BERRY**, **L. LACHAUD**, **F. RENAUD**, **P. DURAND**, and **F. PRUGNOLLE**. "Isolation of Plasmodium Falciparum by Flow-Cytometry: Implications for Single-Trophozoite Genotyping and Parasite DNA Purification for Whole-Genome High-Throughput Sequencing of Archival Samples". In: *Malaria Journal* **11**: (May 2012), 163. DOI: 10.1186/1475-2875-11-163 (see p. 49)

[12] **M. M. PINHEIRO**, **M. A. AHMED**, **S. B. MILLAR**, **T. SANDERSON**, **T. D. OTTO**, **W. C. LU**, **S. KRISHNA**, **J. C. RAYNER**, and **J. COX-SINGH**. "Plasmodium Knowlesi Genome Sequences from Clinical Isolates Reveal Extensive Genomic Dimorphism". In: *PLOS ONE* **10**:4 (Apr. 2015). Ed. by **O. KANEKO**, e0121303. DOI: 10.1371/journal.pone.0121303 (see pp. 49, 51, 83)

[13] **S. B. MILLAR**. "Gene Knock-in as a Tool to Phenotype Clinically Relevant Variant Alleles for Studies on Malaria Pathobiology: Proof of Concept Using the Plasmodium Knowlesi Normocyte Binding Protein Xa Gene". Thesis. St Andrews, Scotland: University of St Andrews, 2017 (see pp. 49–52)

[14] **D. R. ORESEGUN**, **C. DANESHVAR**, and **J. COX-SINGH**. "Plasmodium Knowlesi – Clinical Isolate Genome Sequencing to Inform Translational Same-Species Model System for Severe Malaria". In: *Frontiers in Cellular and Infection Microbiology* **11**: (2021). DOI: 10.3389/fcimb.2021.607686 (see p. 49)

[15] **T. J. EGAN**. "Haemozoin Formation". In: *Molecular and Biochemical Parasitology* **157**:2 (Feb. 2008), 127–136. DOI: 10.1016/j.molbiopara.2007.11.005 (see p. 50)

[16] **J. G. ALTIN** and **E. K. SLOAN**. "The Role of CD45 and CD45-Associated Molecules in T Cell Activation". In: *Immunology & Cell Biology* **75**:5 (1997), 430–445. DOI: 10.1038/icb.1997.68 (see p. 50)

[17] **M. L. HERMISTON**, **Z. XU**, and **A. WEISS**. "CD45: A Critical Regulator of Signaling Thresholds in Immune Cells". In: *Annual Review of Immunology* **21**: (Nov. 2003), 107–137. DOI: 10.1146/annurev.immunol.21.120601.140946 (see p. 50)

[18] **N. HOLMES**. "CD45: All Is Not yet Crystal Clear". In: *Immunology* **117**:2 (Feb. 2006), 145–155. DOI: 10.1111/j.1365-2567.2005.02265.x (see p. 50)

[19] **D. DESAI**, **J. SAP**, **O. SILVENNOINEN**, **J. SCHLESSINGER**, and **A. WEISS**. "The Catalytic Activity of the CD45 Membrane-Proximal Phosphatase Domain Is Required for TCR Signaling and Regulation." In: *The EMBO Journal* **13**:17 (Sept. 1994), 4002–4010. DOI: 10.1002/j.1460-2075.1994.tb06716.x (see p. 50)

[20] **A. M. AHMED**, **M. M. PINHEIRO**, **P. C. DIVIS**, **A. SINER**, **R. ZAINUDIN**, **I. T. WONG**, **C. W. LU**, **S. K. SINGH-KHAIRA**, **S. B. MILLAR**, **S. LYNCH**, **M. WILLMANN**, **B. SINGH**, **S. KRISHNA**, and **J. COX-SINGH**. "Disease Progression in Plasmodium Knowlesi Malaria Is Linked to Variation in Invasion Gene Family Members". In: *PLoS Neglected Tropical Diseases* **8**:8 (Aug. 2014). Ed. by **K. HIRAYAMA**, e3086. DOI: 10.1371/journal.pntd.0003086 (see p. 51)

[21] **C. H. KLAASSEN**, **M. A. JEUNINK**, **C. F. PRINSEN**, **T. J. RUERS**, **A. C. TAN**, **L. J. STROBBE**, and **F. B. THUNNISSEN**. "Quantification of Human DNA in Feces as a Diagnostic Test for the Presence of Colorectal Cancer." In: *Clinical Chemistry* **49**:7 (2003), 1185–1187 (see pp. 53, 54)

[22] **M. ROUGEMONT**, **M. VAN SAANEN**, **R. SAHLI**, **H. P. HINRIKSON**, **J. BILLE**, and **K. JATON**. "Detection of Four Plasmodium Species in Blood from Humans by 18S rRNA Gene Subunit-Based and Species-Specific Real-Time PCR Assays". In: *Journal of Clinical Microbiology* **42**:12 (Dec. 2004), 5636–5643. DOI: 10.1128/JCM.42.12.5636-5643.2004 (see pp. 53, 54)

[23] **P. C. DIVIS**, **S. E. SHOKOPLES**, **B. SINGH**, and **S. K. YANOW**. "A TaqMan Real-Time PCR Assay for the Detection and Quantitation of Plasmodium Knowlesi". In: *Malaria Journal* **9**: (Nov. 2010), 344. DOI: 10.1186/1475-2875-9-344 (see pp. 53, 54)

[24] **J. E. HYDE** and **M. READ**. *The Extraction and Purification of DNA and RNA from In Vitro Cultures of the Malaria Parasite Plasmodium Falciparum*. Methods in Molecular Biology™. Humana Press, 1993. DOI: 10.1385/0-89603-239-6:133 (see p. 54)

[25] **M. SLIFKIN** and **R. CUMBIE**. "Comparison of the Histopaque-1119 Method with the Plasmagel Method for Separation of Blood Leukocytes for Cy-

tomegalovirus Isolation." In: *Journal of Clinical Microbiology* **30**:10 (Oct. 1992), 2722–2724 (see p. 56)

[26] **ORIGINLAB CORPORATION**. *OriginPro*. Originlab Corporation. Northampton, MA, USA, 2021 (see p. 66)

[27] **M. INYUSHIN**, **Y. KUCHERYAVIH**, **L. KUCHERYAVIH**, **L. ROJAS**, **I. KHMELINSKII**, and **V. MAKAROV**. "Superparamagnetic Properties of Hemozoin". In: *Scientific Reports* **6**: (May 2016), 26212. DOI: 10.1038/srep26212 (see p. 82)

[28] **D. M. NEWMAN**, **J. HEPTINSTALL**, **R. J. MATELON**, **L. SAVAGE**, **M. L. WEARS**, **J. BEDDOW**, **M. COX**, **H. D. F. H. SCHALLIG**, and **P. F. MENS**. "A Magneto-Optic Route toward the In Vivo Diagnosis of Malaria: Preliminary Results and Preclinical Trial Data". In: *Biophysical Journal* **95**:2 (July 2008), 994–1000. DOI: 10.1529/biophysj.107.128140 (see p. 82)

[29] **S. E. FRANCIS**, **D. J. SULLIVAN**, and **A. D. E. GOLDBERG**. "Hemoglobin Metabolism In The Malaria Parasite Plasmodium Falciparum". In: *Annual Review of Microbiology* **51**:1 (Oct. 1997), 97–123. DOI: 10.1146/annurev.micro.51.1.97 (see p. 82)

CHAPTER THREE

PIPELINE DEVELOPMENT FOR *Plasmodium knowlesi* SEQUENCE DATA PROCESSING

Ebè kan ṣoṣo àkùrọ́ kúrò ní "Mo fẹ́rẹ̀ẹ́ ṣíwọ́ – **A single heap on the farm does not warrant saying "I am just about done"**

Yorùbá adage

Fundamentally, bioinformatics heavily relies upon various software tools and packages developed by independent researchers and companies with a vested interest in the sector. Such tools are necessary to carry out sequence data analyses and manipulations, resulting in an unwritten tenet of bioinformatics whereby the vast majority of bioinformatics tools are open-source. In cases where payment is required, a 'freemium' model is implemented. However, the rate of innovation in the field is seemingly increasing, resulting in further demand and usage. As such, tools are often updated or entirely superseded in an astonishingly short period, with pipelines– *a series of software, tools and scripts where one output acts as the input to the next to achieve an analytical purpose*— quickly becoming outdated. Thus, projects involving studies and analyses over an extended period must continuously test, develop, assess, and monitor the tools in their pipeline to ensure the most updated and appropriate tools are being implemented for their specific needs. The propagation and maintenance of these synergistic tools is pipeline development.

3.1 Genome Sequencing Technologies

3.1.1 History and Development

Genome sequencing has seen incredible democratisation in the last two decades with an impressive reduction in costs, while increasing access and facilitating innovative applications to answer long-held questions. Though genome sequencing appears to be a recent technological development, it has been a natural progression from the initial proposal of the double-helix deoxyribonucleic acid (DNA) structure by Watson and Crick, based on Franklin and Wilkins' work [1–3]. Initially, sequencing took the form of extensive analytical chemistry processes that were only able to determine nucleotide composition of ribonucleic acid (RNA) genomes due to their relatively short length, lack of complexity and increased availability in laboratories [2].

Indeed, even with a relatively rudimentary start, knowledge and associated technologies improved, resulting in further development of the fledgeling discipline. Once RNA sequencing was somewhat better understood and developed, the methods were adapted to DNA sequences. Using a combination of the 2-D fractionation method developed by Sanger et al [4], known oligonucleotides and a DNA polymerase, short stretches of DNA could be sequenced [2]. However, true sequencing did not begin until the adaptation of these methods to forgo 2-D fractionation for polyacrylamide gel electrophoresis to achieve nucleotide size separation. The resulting 'plus and minus' protocol developed by Sanger and Coulson [5] and the chemical cleavage method by Maxam and Gilbert [6], offered two differing means of forming the very first DNA sequences. However, the 'plus and minus' technique was time-consuming, could only produce ~50 base pair (bp) DNA sequences and was unable to resolve homopolymer regions [7]. On the other hand, Maxam and Gilbert's method provided better control and accuracy in its methodology. Thus it was the first widely accepted method, ushering in the first generation of sequencing.

3.1.2 First Generation Sequencing

Maxam and Gilbert's chemical cleavage method involved the use of ^{32}P radiolabelled DNA fragments that are exposed to specific chemicals; resulting in base-specific chemical reactions [2, 7, 8]. This meant that depending on which chemicals the DNA fragment was exposed to; breaks would occur at specific bases. For example, treating the radiolabelled DNA fragment with dimethyl sulfate would cause a break at the purine nucleotides (*adenine* and *guanine*) [2, 7]. To further delineate purine nucleotide, adenines can be isolated from guanines by treating the sequence with a weak acid like formic acid. This weakens the glycosidic bonds holding adenines more than it affects guanines, thus acting as a means of isolating adenine breakages [6]. With this, Maxam and Gilbert achieved that which the 'plus and minus' method could not. The chemical cleavage technique was able to resolve homopolymer regions as it could terminate at predictable nucleotide positions [7]. However, this specificity in its implementation also proved to be its downfall. The use of toxic chemicals to break the nucleotide bonds and the relatively complex method resulted in the abandonment of this method, favouring its successor.

Though Maxam and Gilbert's chemical cleavage technique was the first to be accepted, it was quickly superseded by the 'dideoxy' technique; commonly termed as the 'chain-termination' technique or even more simply as 'Sanger sequencing'. Sanger et al [9], with the 'dideoxy' technique solved the shortcomings of both the 'plus and minus' technique and the chemical cleavage method. Initially, Sanger sequencing involved the use of di-deoxyribonucleotides (ddNTPs) – chemical analogues to deoxyribonucleotides (dNTPs); the monomeric unit of DNA [2, 8, 9]. These ddNTPs lack a $3'$ hydroxyl group necessary to form the phosphodiester bond to other dNTPs; thus the sequence is terminated at that ddNTP.

By introducing radiolabelled ddNTPs into the elongation step of a PCR at a fraction of the concentration of non-labelled dNTP, allows the generation of terminated DNA fragments of varying lengths [2, 8]. Polyacrylamide gels can then be used to visualise the DNA fragments for each individual ddNTP. However, this iteration of the dideoxy technique required four different chemical reactions to represent each of the four nucleotide bases. Rather than four separate reactions, the use of fluorophore-labelled ddNTPs in later improvements increased the efficiency of the protocol by running all four ddNTP

individuals in the same reaction [2]. The miniaturisation of the process allowed for multichannel capillary-based electrophoresis that culminated in the development of various automated commercial platforms, though the Sanger method is currently mainly supported by Applied BioSystems (ABI) [8]. While ABI platforms can only produce 600 - 1000 bp reads, they facilitated the start of the Human Genome Project. Thus, even with more modern forms of sequencing available, the Sanger method is still in use for applications that do not require high throughput sequencing [8].

3.1.3 Second Generation Sequencing

'Next Generation Sequencing (NGS)' or 'short-read sequencing' is a progression of sequencing technology that took a drastically different route in comparison to Sanger et al [9]. Although NGS works on the same basis of sequencing by synthesis (SBS) like Sanger sequencing, where a polymerase is required to carry out observable synthesis of DNA, NGS does not utilise radioactively or fluorophore-labelled ddNTPs. Rather, NGS utilises fluorometric reactions of the conversion of pyrophosphate to ATP. The generated ATP fuels a secondary reaction of luciferase resulting in light being emitted that is proportional to the concentration of pyrophosphate converted [2, 8]. Initially, a variety of commercial platforms were released for this generation; the first of which was the '454 machine'[454 Life Sciences;Roche][2, 8, 10]. However, by far the most prominent NGS sequencing technology currently is the Illumina sequencing platform.

Illumina, with its variety of sequencing machines, achieves sequencing using highly modified fluorescent ddNTPs. The $3'$ hydroxyl group of these ddNTPs are temporarily blocked by a fluorophore, thus preventing the elongation of the DNA strand [2]. Upon excitement from a laser, the fluorophore simultaneously emits light of a known frequency. The fluorophore is cleaved from the ddNTP, allowing for the binding of another ddNTP adding to the growing chain [2, 8, 11]. From this, Illumina can produce millions of paired-end short reads with a maximum length of 2 x 300 bp reads. Essentially, Illumina carries out sequencing of both DNA strands at a high quality due to the mechanism of action of the reversible ddNTPs; as the synthesised DNA strand is only elongated after reading the joined nucleotide base [2, 8, 11]. However, it is difficult for Illumina sequencing to resolve regions of high variability or low complexity like homopolymer

regions that would emit the same consecutive wavelength of light [2]. Where there is an assemblage of projects requiring whole genome, targeted, deep, or RNA sequencing with a high-throughput, Illumina sequencing is often found – due to its high accuracy.

3.1.4 Third Generation sequencing

NGS was hindered by its length and inability to resolve or span regions of low complexity and high variability. This spurred the development of the third and most recent sequencing technology. Largely, this generation is characterised for its read lengths, where reads are consistently reaching up to and exceeding 10 000 bp or ten Kilobases (Kb). Other classifiers of this generation include the single-molecule sequencing technique it is based on and the real-time sequencing capabilities available [2, 8, 12]. The first commercially released application of this technology was the ‘ Single-molecule real time (SMRT) ’ sequencing platform by Pacific Biosciences (PacBio).

Through the use of tiny chambers called Zero-mode wave guides (ZMW), which have diameters small enough to prevent the entire wavelength of light to pass through, DNA synthesis via fluorescently-labelled dNTPs can be performed on non-amplified DNA [2, 8, 12]. The structure of the ZMW means that light (*in the form of an excitement laser*) is only able to illuminate the very bottom of the well [2, 12]. Here, a DNA polymerase that is attached to the DNA fragment to be sequenced is in close proximity to the labelled deoxyribonucleotides (dNTPs), which allows elongation of the chain. As each dNTP joins the strand being synthesised, the fluorophore is released and in the presence of the laser, fluoresces, and in turn, is read and analysed [2, 8, 12].

This appears similar in principle to previous generations; however, the applications of the SMRT technology has broadened the scope of sequencing. Here, PacBio has been able to implement the production of highly accurate, relatively long reads in the form of ‘Hi-Fi’ reads using the ‘Circular Consensus’ method. While, to generate extremely long reads up to and exceeding 50 Kb, the ‘Continuous long read’ sequencing method is used [2, 8, 12]. In either case, the products of PacBio sequencing are very long and relatively accurate reads. However, the inherent problem with all single-molecule sequencing technologies arises in the form of a high error rate. As the length of a particular PacBio

sequenced DNA fragment increases, so does the probability of any random individual nucleotide to be erroneous. While this can often be minimised with deep coverage of each nucleotide position, this results in a high input DNA concentration requirement and subsequently higher overall costs of sequencing [2, 8]. Rivalling this approach is the most recent entrant into the major commercial sequencing field. Through the use of biological nanopores, Oxford Nanopore Technologies (ONT) aim to reduce the cost of sequencing and the concentration of DNA required without the need for PCR amplification, while producing comparable quality sequence reads.

3.1.5 Nanopore Sequencing

Although seeming like a recent development, nanopore sequencing is a relatively old concept. Initially described by Deamer, Branton and Church [13, 14], it is a form of sequencing which utilises a 1 nanometer (nm) diameter nanopore [Figure 3.1]. A nanopore is a hole through a membrane that can be formed naturally by forming a transmembrane complex or synthetically made as solid-state nanopores. At present, all commercial forms of nanopore sequencing involve biological nanopores, and while solid-state nanopores are in development, they are currently not as precise [15]. However, if successful, solid-state nanopores would be more robust, adaptable, durable, chemically and thermally stable than natural biological nanopores [14]. In its infancy, nanopore sequencing had poor single-nucleotide resolution, which hindered several companies that sprang forth to capitalise on the new technology, all to no avail. However, from 2014, with the implementation of their 'early access' program, ONT began providing a third option for high throughput sequencing with impressively long output lengths [13, 14]; though, as with all things, this platform was not without issues.

Firstly, ONT miniaturised the technology to have the sequencing carried out on a flowcell [Figure 3.2, top] comprised of 2048 biological nanopores arranged into four groups. Here, ONT utilise the α-hemolysin biological nanopore [Figure 3.1] which is suspended in a synthetic electrically resistant polymer membrane [13, 16]. During library preparation, the DNA or RNA strand to be sequenced is attached to an adapter sequence, which in turn allows for the binding of a processive motor protein at the $5'$ end of the strand; forming a complex [13, 16]. The complex is then bound by a fixed tether on the flowcell to bring it

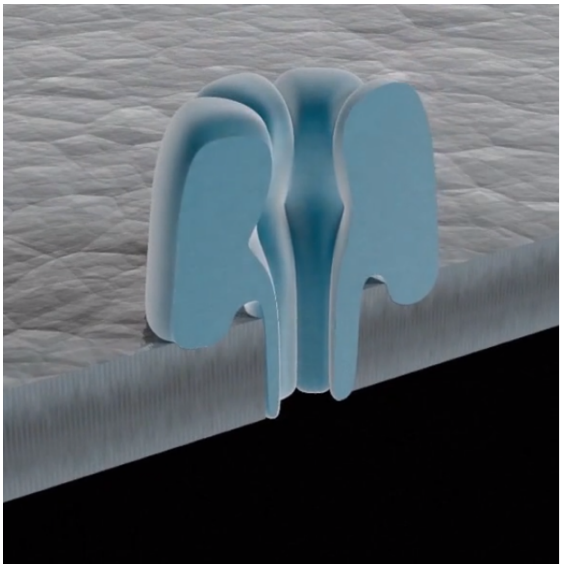


Figure 3.1: The α-hemolysin nanopore. A naturally occurring nanopore which facilitates sequencing in modern nanopore sequencing in particular Oxford Nanopore Technologies' long read sequencing platform [16].

closer to a nearby nanopore [16, 17] thereby initiating sequencing. A processive enzyme acts as a ratcheting device, allowing only one nucleotide to pass through the pore at a time in a unidirectional manner at a controlled speed. As the nucleotide passes through the pore, the ionic current running across the pore is affected, creating a membrane potential which is detected and recorded by the sequencing platform e.g. the MinION [Figure 3.2, bottom]. Every 3 - 6 nucleotides are seen as unique events, whose duration across the pore and the membrane potential amplitude are computationally translated into known nucleotide bases through the use of 'basecalling' algorithmic models [13, 16].

As such, ONT sequencing does not require amplification nor fragmentation of the genomic information to be sequenced. However, these processes can also be carried out to achieve various purposes, such as the sequences of large genomes being cleaved into relatively smaller fragments. Similar to the benefits of PacBio SMRT sequencing, single base modifications such as methylation sites can be analysed due to the lack of PCR amplification that could be applied. By far the largest advantage to using ONT sequencing is the portability of its small form sequencers like the MinION and Flongle and upcoming

Figure 3.2: The MinION sequencing platform. A small-form sequencer released by Oxford Nanopore Technologies (ONT), showing the flowcell [*top*] where the DNA/RNA library is loaded. The flowcell contains 2048 individual nanopores through which the DNA strand passes to allow sequencing. The flow cell is inserted into the MinION [*bottom*] which provides power to maintain a set temperature and voltage via USB connection to a computer [16].

SmidgION [18]. At ~90 %, the MinION is capable of performing nanopore sequencing in the field, producing comparable quality sequence data despite its diminutive size [Figure 3.2] [13]. This versatility and adaptability has facilitated the use of the MinION in increasingly inventive locales in the field; including the International Space Station [19]. Further applications include real-time, on-site monitoring for surveillance and tracking of both the Ebola epidemic and the SARS-COV2 Coronavirus pandemic [20, 21].

Larger desktop applications of the same technology can output considerably higher volumes of sequence data with the largest–*the PromethION*– producing 10-fold the sequence data possible on the MinION [18]. Finally, the starting and running costs for carrying out sequencing using ONT are also relatively small, with sequencing starting at £800 inclusive of all necessary components to carry out a sequencing experiment [18]. This is in stark comparison to the >$19,000 required to begin Illumina short-read sequencing. However, even with such advantages, the MinION, and nanopore sequencing in general, fall prey to the issues seen by other long-read sequencing technologies. The most prominent of this being high error rates, where ONT sequencing was reported to

have 66 % nucleotide base accuracy when it first launched before rapid improvement to 92 % a year later and settling to 95 % as of 2019 [13, 22]. However, with their recent release, ONT have reportedly increased their nucleotide base accuracy to ~ 96 %; though this would be under ideal conditions [23]. Further advancements have resulted in a reportedly 32 % reduction in read error rate in basecalling accuracy [24]. Though such an increase in sequence quality is impressive, it is still lower than PacBio SMRT sequencing at >99 % accuracy and the industry leader Illumina with >99.9 % accuracy. The far more impactful sector where ONT itself and the technology's adopters have exceeded is in the breadth and specificity of the software and *-omics* tools available to manipulate and analyse ONT sequence data.

3.2 Bioinformatics Analysis

As with NGS, ONT relies heavily on the downstream analyses performed on the sequenced reads to provide *in silico* context to a biological question. As such, multiple tools, software, pipelines and modes of analysis have been developed specifically for ONT sequence data. However, much like the progress of the sequencing quality, the progression of the tools available for ONT sequence data are ever-increasing and ever being improved. Described below are tools that were either used, tested or considered for various points of this project. Some tools represent direct successors of other tools; these will be stated where appropriate.

3.2.1 Basecalling and Demultiplexing

The first step with any ONT sequence data involves basecalling. Basecalling is the computational interpretation of the raw signals captured by the MinION during sequencing back into human-readable nucleotide bases. This is often done via the use of models that were tested and developed by ONT or by dedicated researchers with specific use-cases. At first, basecalling was carried out in real-time using ONTs's cloud service 'Metrichor' (*now 'EPI2ME'*) and while it provided an important service, it was slow and required an Internet connection; limiting its use in the field or in time-sensitive

applications [25].

Albacore

To provide local real-time basecalling, Albacore was released to carry out raw interpretations of the DNA sequence being read through the nanopore. Albacore uses the inherent features of Recurrent Neural Networks (RNN) to carry out basecalling. Here, using a model conceived and trained by ONT, Albacore begins to carry out the processive conversion of raw signals to nucleotide bases; however, as this continues, new raw signal inputs are influenced by previously basecalled signals [26]. Using this principle, the nucleotide with the highest probability is based on the raw signal currently being read and the neighbouring signals along the strand [27]. Albacore is a cross-platform command-line tool allowing for basecalling, calibration strand detection to assess individual run quality, demultiplexing of barcoded libraries and alignment of reads [28]. Inputs into Albacore are in the form of **FAST5** files holding the raw signals before outputting the inferred nucleotides in both **FAST5** and **FASTQ** files. At first, each FAST5 file only contained one read, however as the software matured, the models implemented were updated and improved, including expanding its capabilities to accept 'multi-read' FAST5 files as input, using the 'seamlessF5' extension [29]. However, Albacore was soon superseded by the release of Guppy.

Guppy

In many ways, Guppy was developed to expand on features of Albacore through the use of a Graphical Processing Unit (GPU) to significantly increase processing speeds by more than an order of magnitude [30]. Upon initial release, Guppy provided an updated basecalling algorithm that would provide higher accuracy basecalled reads. As with Albacore, Guppy is a RNN, allowing it the same advantage of using previous outputs to influence the current input [31]. Guppy is also a cross-platform command-line tool. After improvements, Guppy currently possesses modules to achieve basecalling, calibration strand detection, adapter trimming, demultiplexing, alignment and basecalling of modified bases [31]. To reduce memory and storage usage, Guppy was implemented

to take in and output 'multi-read FAST5' files [30].

Qcat

ONT is known to develop tools in parallel to their publicly available and supported tools. Thus, although Guppy possessed a demultiplexing module, Qcat was developed specifically for demultiplexing [32]. At release, Guppy's demultiplexing module was found unable to correctly classify ~29 % of reads into their distinct barcodes [33]. Using the same dataset, Qcat could classify 9 % more of the reads, indicating Qcat was better at detecting false positives and false negatives [33]. Although, this may be due to less stringent requirements within Qcat as it only requires 60 % identity while Guppy requires 70 % identity with a barcode to select it [34]. Qcat utilises a basecalling algorithm similar to that of the cloud-service basecaller EPI2ME rather than one similar to Guppy [32]. At the time of writing, Qcat is unsupported, and Guppy's demultiplexing algorithms are thought to have surpassed Qcat.

3.2.2 Data Parsing, Manipulation, Quality Assessment

Poretools

Poretools is a suite of tools developed to work with the native FAST5 file format providing explorative, quality control and downstream analysis [22]. To achieve this, poretools can convert the FAST5 file format to the more usable FASTQ format, thus allowing downstream analyses and data manipulation. Other capabilities of the suite include outputting and visualising the statistics of the run and measuring the nucleotide length and quality score [22, 35].

Porechop

Porechop was developed to allow the removal of adapters added during library preparation [36]. Porechop removes adapters from the start and end of the reads by matching the adapters within a subset of ~10000 reads from a list of known ONT adapters [36].

Adapter matching is calculated over the length of the adapter. To be accepted, the adapter must be present with a 90 % identity of its full length or 100 % identity on 90 % of the full length; otherwise, the adapter is rejected [36]. Porechop is also able to demultiplex barcoded sequence data where, rather than matching with adapters, the query sequence is matched with a list of known barcodes [36]. Although, this module was quickly superseded by ONT's in-house efforts.

NanoQC

Genomics studies using Illumina short read data and PacBio long reads have a wealth of tools and suites to aid in analysis and data manipulation. This was not the case for nanopore long reads due to the relative infancy of the technology. To cater for the lack of quality visualisation plots like FastQC, NanoQC was developed [37–39]. With this, NanoQC produces three plots which provide the sequence length distribution [Figure 3.3A], 'Per base sequence content and frequency' [Figure 3.3B] and the 'mean sequence quality' [Figure 3.3] [37].

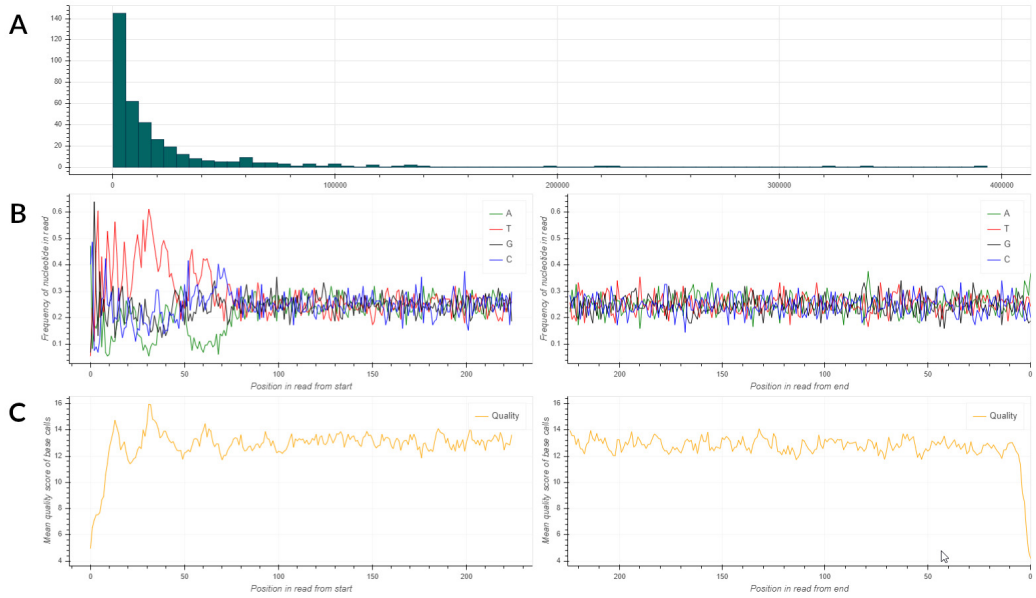


Figure 3.3: Example of NanoQC outputs. Plots produced by NanoQC visualising the quality of the input data at the nucleotide level. A sharp decrease in quality was seen at both ends of the read corresponding to the adapter sequences. Sequence length distribution shows a left-skewed distribution with most reads at 0 - 10 Kb.

NanoPlot

Sequence data is often huge, with no clear way to understand the entirety of the data. To allow for this, statistical analyses packages were developed. NanoPlot provides a means of visualising the statistical analyses of ONT sequence data without the need for the raw FAST5 outputs from the MinION. The tool produces plots that summarise and compare the key features of a dataset to facilitate comparisons based on average Phred quality score, alignment match identity and overall yield [38, 40].

Assembly-stats

Assembly-stats is a C++ package designed to carry out simple statistical analyses of the sequence data file provided to it in the form of FASTA and FASTQ files. Here, assembly-stats outputs simple descriptive metrics such as the total bp length of the input file, the number of reads, average read length and also the length of the longest read. It is intended to be used for assemblies, whereby the N50 metric can also be used. However, it is still helpful to quantify basic numerical metrics of the input sequence data before the assembly step.

BEDTools

BEDTools is a suite of packages built around the comparison, manipulation and analysis of the Browser Extensible Data (BED) file format. Due to the potentially large scale of data generated through various *omics* work, it becomes computationally difficult to analyse specific sections of the dataset effectively. However due to the inherent nature of BED files, BEDTools is able to carry out efficient comparisons between high-throughput datasets [41]. It achieves this by splitting the input BED file into separate bins/chunks. Then through the use of coordinates (like the start and stop positions) stored within it, BEDTools can carry out rapid processing, comparisons and manipulations of specific bins based on the coordinates indicated by the user [41].

Samtools

Samtools is a suite of different scripts and tools designed to produce, manipulate, visualise and assess the generic Sequence Alignment Mapping (SAM) file format for both short and long-read sequence data [42, 43]. Since its release in 2009, Samtools has seen an expansion in features and capabilities, with current builds capable of carrying out varying amounts of file manipulation such as sorting, merging and splitting, indexing of sequence data for faster parsing, detection of duplicated reads; rudimentary visualisation and different forms of alignment statistics recording and assessment of SAM formatted files [43].

3.2.3 Alignment

Nucmer

Nucmer is the nucleotide aligner module of the MUMmer package [44] that was initially created to align short bacterial genomes with relatively low error-rates [44]. However, this has been significantly expanded upon to allow for the alignment of longer, high error-rate sequence data, evolving the aligner to be compatible with much of the sequence data commonly used. With this expansion, nucmer's speed and efficiency have increased to become similar to other widely used genomic aligners, though this comes at the cost of potentially high memory usage [44]. In addition, Nucmer tends to be used for specific forms of alignments, commonly during structural variation analyses, where the module's '*all-against-all*' alignment algorithm is used.

Burrows-Wheeler Aligner (BWA)

The Burrows-Wheeler Aligner (BWA) is a sequence data aligner developed shortly after the acceptance and usage of third-generation sequencing techniques. The advent of long-read sequence data meant that previous aligners geared towards short reads of $<$70 bp were ill-equipped to sufficiently allow for longer reads while accounting for the increased error. Short read aligners were built using a philosophy of '*end-to-end*'

alignment where each nucleotide base had to align to the reference to be successful [45]. Such a method becomes complicated in long reads, which can potentially possess large gaps where nucleotide bases do not align due to structural variations covered by the read or due to admittedly high error rates [45]. BWA utilises the '*seed-and-extend*' principle whereby exact matches are passed over twice for each alignment (seeds) [45]. Co-linear seeds are called together as a 'chain', and chains are then filtered for length and quality before each seed in the chain is ranked. If a seed was repeated in a previous and higher ranked chain, the duplicated seed is dropped [45]. This implementation culminated in the rapid computing speed of BWA while maintaining or exceeding the accuracy of previously published short and long read aligners. However, BWA for long reads was expanded to create the successor to BWA in the form of minimap2.

Minimap2

Although many aligners utilise different algorithms, the same principle of pattern matching between the query sequence and the reference is still required. Minimap2 was developed as an extension of BWA [45] to provide a tool for processing and aligning large scale data of up to ~100 Kb [45, 46]. Minimap2 uses the standard '*seed-chain-align*' process where small *k*-mers of the reference sequence are indexed into a hash table [46]. *k*-mers of the reference genome are compared to *k*-mers or the '*seed*' of the query sequence for alignment. A seed match results in minimap2 attempting further matching of colinear seeds which map to the hash table in close proximity to the initial *k*-mer match, thus, creating a '*chain*' [46]. Extension of the chain results in longer read alignment outputs, meaning that minimap2 is versatile enough to cater to very long reads while maintaining fast processing speeds compared to other long-read alignment tools [46].

3.2.4 *De novo* Sequencing Assemblers

As in basecalling, the algorithm implemented by *de novo* assemblers influences the genome assembled downstream. Generally, there are two major classes of *de novo* assembly algorithms; the de Bruijn Graph (DBG) approach and the Overlap Layout

Consensus (OLC) approach [47]. DBG is mainly used by short-read assemblers as it uses *k*-mer graphs to determine the longest sequence possible [47]. Due to the length of short reads, they can effectively function as *k*-mers individually with little or no splitting. Conversely, OLC was developed for Sanger sequenced data; however, it has since been adapted for long-read technology; in particular ONT sequenced data. OLC overlaps all reads within the sequence data to determine a layout graph used to generate a consensus assembly sequence [47, 48].

There are multiple variants of these classes and other lesser-used classes of algorithms, all implemented with the rationale of resolving sequence data of ever-increasing complexity into a consensus genome assembly for further study. Thus, with increasing demand, there is an abundance of *de novo* assemblers available for researchers to utilise, depending upon their needs, data, and computing capabilities.

Canu

At a time, Canu was regarded as the 'gold-standard' *de novo* ONT sequence data assembler. Canu was built by Koren et al. [49] as a replacement and improvement on a previously released long-read assembler– Celera [50, 51]. Canu is built based on the OLC approach where it can assemble long, noisy error-prone sequence data that is ubiquitous of long-read sequencing technologies like ONTs nanopore sequencing. Canu is fully supported with ongoing updates and improvements to already robust features such as supporting high and low coverage data. The minimum and maximum thresholds are not stated, and to conserve memory usage when using high coverage data, only the longest 40x sequence data is used in the initial *de novo* assembly process [52]. Although Canu possesses a highly versatile range of features, it is also relatively slow in its assembly process, especially when assembling eukaryotic sequence data. However, it has been successfully used to resolve high repeat regions in *Escherichia coli* sequence data [49]. Nevertheless, further polishing steps using short reads are still necessary to produce a final *de novo* assembly of high quality and accuracy [49].

Redbean

Formerly known as 'wtdbg2', Redbean is a *de novo* genome assembler specifically developed to account for the increasingly slow processing time needed to carry out a *de novo* assembly on large mammalian genomes [53]. Algorithmically, it uses a basis of the OLC approach however rather than using the 'layout' step, a step similar to the DBG approach is substituted. Here, Redbean loads all input reads into the memory before separating the reads into 1024 bp segments (similar to very long *k*-mers) [53]. This contrasts with other OLC-based assemblers which load reads into memory in small batches, which, while reducing memory usage, increases processing duration [53]. The 1024 bp segments are used to construct a fuzzy DBG which is further cleaned to generate consensus contig sequences [Figure 3.4]. The assembled genome can then be polished downstream with Redbean's polishing module or another dedicated polishing tool like nanopolish [54]; thereby increasing the quality and accuracy of the final assembly [53]. However, increasing the speed of Redbean has been reported to negatively impact the accuracy of draft assemblies generated [55].

Flye

Much like Redbean, Flye attempts to carry out *de novo* genome assembly of ONT long-read sequence data using the basis of the DBG approach to generate accurate genomes. However, for this, rather than using exact *k*-mer matches to assemble a genome like traditional DBG assemblers, Flye implements repeat graphs that require approximate sequence matches; thereby allowing the processing of high-noise sequence data prevalent in long reads [56, 57]. Repeat graphs [Figure 3.5f] are a type of assembly graph which allow a visual representation of the repeats of a genome; facilitating further understanding of scaffolding and haplotyping due to alternative haplotypes being visible in the form of 'bubbles' in the graphs [*not shown*][56, 57].

Through the use of repeat graphs, Flye produces error-prone intermediate sequences called 'disjointigs', which are concatenated to form longer but also error-prone sequences [Figure 3.5]. These longer sequences are, in turn, used to generate a repeat graph representing the 'longest path' a genome fragment can take to form a single continuous

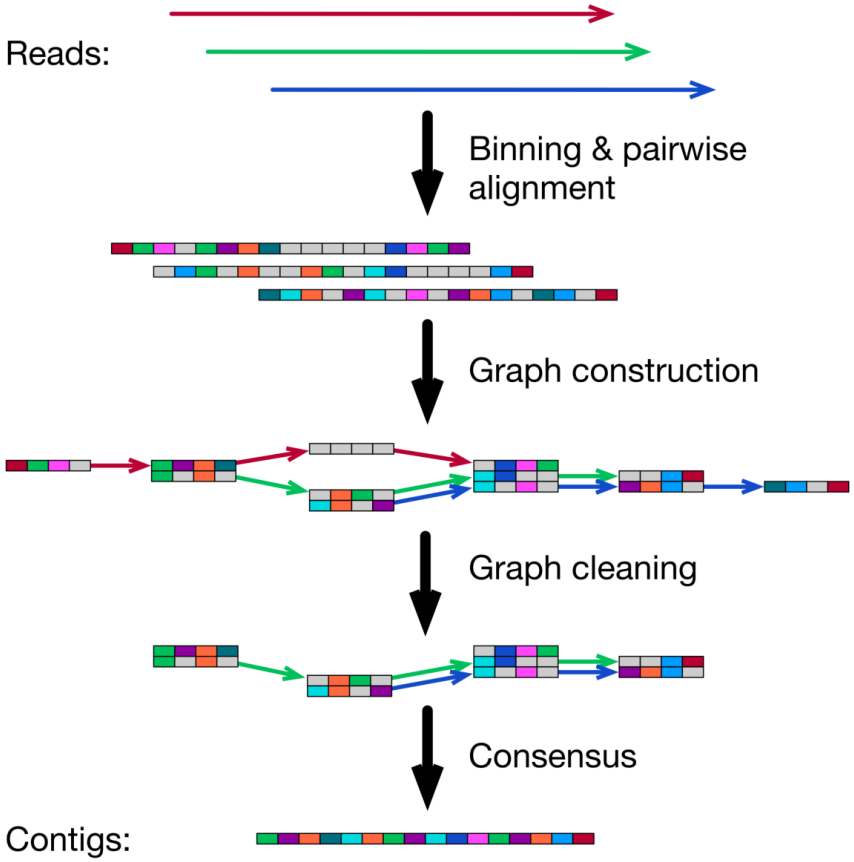


Figure 3.4: The workflow implemented in Redbean. Input reads in the form of raw, erroneous reads or error-corrected reads are separated into 1024 bp segments (coloured boxes). Segments of the same colour share similarly aligned regions, while grey boxes share no regions. An all-vs-all alignment is carried out, and the resulting alignments are used to construct a fuzzy DBG. The graph is cleaned by trimming the tips of the aligned segments and removing unsuitable intermediate pseudoassemblies. The alignment graph is used to generate a consensus contig sequence. Source: Ruan and Li [53].

string of nucleotide bases [56]. Once formed, the long sequence string is aligned against the initial input reads to improve regions of low resolution between the different disjointigs comprised within the long string [Figure 3.5h]. Repeats of the same region can also be further used to improve accuracy to form a consensus; thus, sufficient depth is necessary [Figure 3.5i][56]. While an improvement is made in this case to generate a final consensus draft *de novo* assembly, this is not error correction, but rather a crude form of consensus amalgamation. Once consensus is reached, the final assembly of

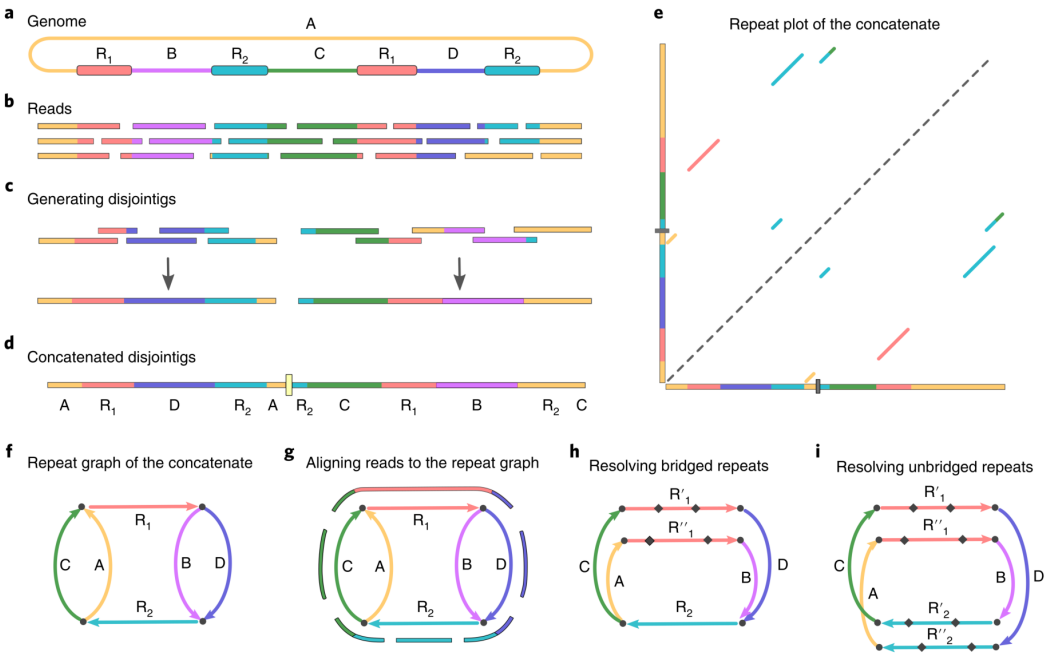


Figure 3.5: Flye *de novo* assembly workflow.(a) The Flye pipeline showing a genome with two regions that have identical repeats (R_1 and R_2) and non-repetitive regions A, B,C and D. **(b)** Reads are generated via sequencing and used to generate error-prone and misassembled disjointigs **(c)**, which are randomly concatenated to form a single string of sequence data (**d**). (**e**) A dot plot of the sequences is generated to determine the co-ordinates of the concatenated repeats in the sequence. (**f**) The co-ordinates are used to determine and join together points of connection between each of the disjointigs that make up the string to form the repeat graph. The input reads are aligned against the repeat graph to get overlaps (**g**) and resolve bridged repeats connected to the same disjointig but are slightly different (**h**). (**i**) A third-party module Trestle is used to extend and resolve the unbridged repeats using heterogenous regions between the repeat copies. Source: Kolmogorov et al. [56]

contigs is outputted.

3.2.5 Assembly Quality Assessment

BlobTools

Retrospective analyses of sequences generated in the early days of NGS have revealed the presence of either multi-organism or non-target genome sequences [58]. Here, an

extracted sample that unknowingly contained the genomes of multiple organisms or other genetic targets distinct from the target of interest was sequenced and analysed under an assumption of purity. The most affected class of sequences was eukaryotic

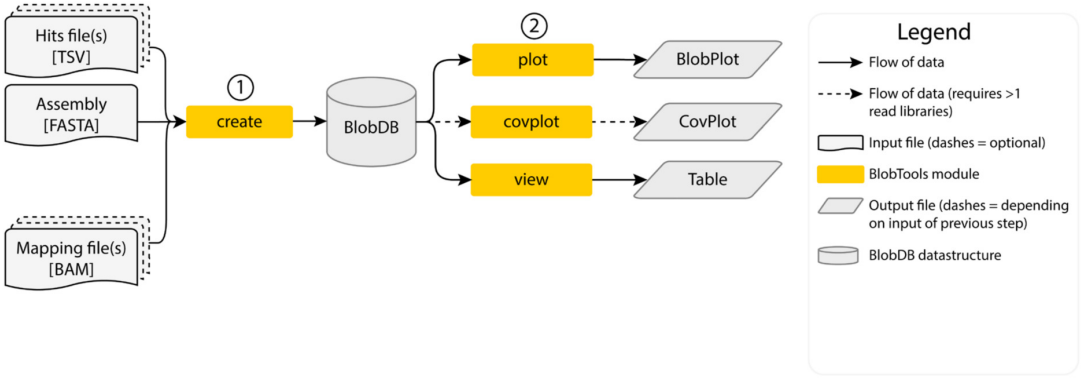


Figure 3.6: The Blobtools workflow. Blobtools carries out assessment, contamination detection and taxonomic interrogation of *de novo* genome assemblies. **1:** A blobDB data structure stores the information held within the three input files. **2:** The blobDB structure is used in plotting, and visualisation of the contaminants based on the best hits from the file and the taxonomic partitions determined. Adapted from: Kolmogorov et al. [56].

genomes, chiefly due to contamination by microbial sequences. As such, spurious errors were present within publicly available genome databases, which could potentially be recycled in perpetuity. More effort is now taken to ensure that contamination is reduced or eliminated in new genomes deposited on public databases. To detect and interrogate such occurrences, BlobTools was developed [58]. BlobTools works by carrying out a taxonomic assignment of input sequences by measuring and comparing the GC content of said sequences against a provided database [58]. The database could be a publicly available database or a curated one, resulting in sequences labelled by taxonomic ranks ranging from the species-level to superkingdom [58].

However, unlike similar tools, BlobTools achieves taxonomy ranking by accepting the highest sum of scores of similarity between multiple hits of a single sequence against the input database [58]. Thus, rather than simply accepting the ‘best’ hit (usually from Basic Local Alignment Search Tool (BLAST) outputs) and assigning a taxonomic rank based on said ‘best’ hit, BlobTools sums up all hits of a particular taxonomic rank and assigns the sequence to the taxonomic rank with the highest total score. This forms the

'blobDB' data structure, which is then used to generate visual plots [Figure 3.6]. The plots visualise the coverage of each assigned taxonomic rank and relative proportions of assigned taxonomies [58]. With this, users can graphically determine the proportion of their input data that would be classed as 'contaminants'.

Quast

Quast was initially developed to allow the evaluation of short-read genome assemblies [59] before being expanded to allow for long reads and large eukaryotic genomes [60]. In its first form, Quast could carry out analyses of a provided genome in the context of a reference genome [59]. From this, analyses of misassemblies, structural variation, N50 and other statistical metric data are generated and presented in easily understood reports and graphs. The expansion, Quast-LG, allowed the program to accept large genomes and long-read data as inputs for reference-free evaluation, where additional estimation of transposable elements (TEs) are carried out [60]. It achieves this by estimating the theoretical 'upper bound' of the input assembly's completeness [60]. However, it is still recommended to carry out TE prediction with dedicated tools [60]. With this, Quast-LG provides a significantly different approach to evaluating long-read genomes, while employing contemporary alignment methods suitable for large genomes.

BUSCO

The Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment tool was developed as a means of providing simple, quantitative quality control for genome assemblies. BUSCO analysis is done by quantifying their 'completeness' in comparison to expected gene-sets known to be ubiquitous within the selected lineage and evolving with single-copy control [61, 62]. Thus, these gene-sets are curated to contain genes that should be within a genome of a particular taxonomic rank in single copies [61]. With this, the number of genes available for a gene-set reduces as the taxonomic rank broadens with the Eukaryota gene-set containing a few hundred ubiquitous genes. In contrast, a gene-set for a specific species may contain thousands of genes. However, due to the extensive testing required to determine such ubiquity, the gene-sets available are restricted to many

model organisms and the larger taxonomic ranks. Where present, genes are marked as 'complete' where a 1:1 match occurs or 'fragmented' if the input genome contains fragmented forms of genes within the curated gene sets[61, 62]. In rare instances, duplications may occur, though such assertions may also be due to technical limitations of the tool where heterozygous regions were not sufficiently collapsed [61]. Further functionalities present in BUSCO include the capability of assessing transcriptomes, generating training sets for gene prediction and also identifying potential markers for metagenomic studies [61].

3.2.6 Polishing and Correction

Racon

Built to carry out rapid consensus, Racon is a stand-alone assembly consensus and correction tool developed to be highly modular and adaptable to any sequencing technology and assembly generating tool [63]. Genome assemblers often tend to involve time and processor-intensive procedures; however, with the release of miniasm, Li [64] produced an assembly tool capable of generating assemblies from erroneous reads without the need for prior correction. Nevertheless, such assemblies would still contain errors and require further improvement, hence the need for Racon. Racon is both sequencer and assembler independent, although it is optimised to work with the output of miniasm [63]. Additionally, Racon is scalable and adaptable to large genomes with complicated and highly variable sequences [63]. Thus Racon acts as a broad spectrum consensus sequence caller capable of iteratively correcting input sequences from a wide range of sources and organisms.

Medaka

Unlike Racon, which was designed to be a general tool for consensus and correction, Medaka was generated specifically for ONT long-read sequences. Designed in-house by ONT, Medaka is able to carry out relatively fast and accurate consensus generation and variant calling using long error-prone ONT reads [65]. However, Medaka is tuned to

correct sequences that have already been previously corrected at least once using racon [65]. Currently, Medaka is the recommended consensus sequence generating tool for draft assemblies of ONT origin.

Pilon

Pilon is a genome assembly improvement and variant calling tool that can detect differences between an input draft genome and a reference genome [66]. It is built to take in Illumina short-reads, and using large insert libraries of these highly accurate reads, it can correct misassemblies and fill gaps in the draft assembly to improve it [66]. The use of the large insert libraries also allows for the detection of large insertions and deletions (INDELs) resulting in the resolution of regions of high variability such as genes involved with pathogen-host interactions [66]. In addition, Pilon results in draft assemblies that are more contiguous with genes that would be usually difficult to resolve.

3.2.7 RepeatMasking

RepeatMasker

Building on the Smith-Waterman algorithm, RepeatMasker can screen and search draft assemblies for repetitive sequences like interspersed repeats and low complexity sequences [67]. It achieves this by aligning the input sequence against known repeats previously described in a database like RepBase [67]. Where no previous library exists for an organism, custom repeat libraries can be generated through a related tool — RepeatModeler, which can carry out *de novo* repeat finding [67].

One Code to Find them all

RepeatMasker is able to search and identify transposable elements (TEs) in draft assemblies although, it remains beholden to a reference sequence – even when using a custom *de novo* repeat library [68]. This could potentially hinder RepeatMasker's ability to discover new TEs in both model and non-model organisms. A larger consequence of

RepeatMasker lies in its inability to properly identify unique hits of TEs with multiple copies within the input genome [68]. One code to find them all (OCTFTA) attempts to solve these gaps in the repeat masking process left by RepeatMasker. OCTFTA is specifically designed to parse the output of RepeatMasker to identify the coordinates of TE copies and the distinct TE families they belong to [68]. Further information of each TE is also retrieved by OCTFTA resulting in the two packaged Perl scripts acting as an extension to RepeatMasker. Where copies of the same TE are found to meet the requirements set by OCTFTA i.e. the copies are within the same coordinates or close by; the copies are merged to represent a single copy at that coordinate [68].

TransposonPSI

Due to the diversity present in TEs, there exist a large variety of tools developed to search, identify and annotate them within genomes. One such tool is transposonPSI, which employs the use of PSI-BLAST to carry out repetitive element searches on the input genome [69]. TransposonPSI accomplishes this by BLAST aligning the input nucleic or protein sequence against a library of protein sequences corresponding to open reading frames for known families of transposons [69]. As such, transposonPSI can potentially find repetitive elements which are fragmented or divergent at the nucleic level that other TE-finding tools like RepeatMasker are unable to [69].

RaGOO/RagTag

To complete a draft genome, the contigs must be combined to form scaffolds and eventually chromosomes after protein prediction and annotation. While there are many ways employed to achieve this scaffolding step, they are often either time or memory processing-intensive [70]. However, to mitigate both of these problems, RaGOO was developed to utilise the rapid and accurate genome aligner Minimap2 [70]. By using Minimap2, RaGOO utilises a reference-based approach to chromosome ordering and structuring, where a closely related reference genome is required to align the draft assembly against [70]. There are consequences to this, namely, alignment bias introduced from the reference, though this issue is present in all reference-based chromosome

structuring and scaffolding tools. RaGOO also possesses functionalities to detect and break both chimeric contigs and misassemblies within the input contigs during the ordering process, though this is dependent on the user [70].

RagTag is the successor of RaGOO and expands its functionalities. To provide more utility in scaffolding, RagTag incorporates the capability to 'patch' and 'merge' input genomes. The new modules in RagTag add to the 'correct' and 'scaffold' modules first introduced in RaGOO [70]. The 'correct' is able to detect potential misassemblies between the input genome and the reference genome while the 'scaffold' acts to orient the contigs into pseudochromosomes based on the homology alignment with the reference sequence [71]. The 'patch' and 'merge' modules work in a similar manner however, 'patch' uses one genome to fill in the gaps of another while 'merge' uses multiple forms of the same scaffolded assembly to form a 'consensus' scaffolded assembly [71].

3.2.8 Gene Prediction and Genome annotation

Companion

In many ways, Companion represents the democratisation of genomics, which has been occurring in the past decade, albeit specifically for the annotation of parasite genomes. Companion provides a single platform that can carry out gene prediction, functional annotation and some preliminary analyses for parasite genomes [72]. Companion operates as both a web server and local software that the user can adapt further to fit a custom purpose. As it is designed with parasite genomes in mind, it provides the reference genomes of a set of parasites which are in turn used in the annotation pipeline [72]. Firstly, using ABACAS2, input contigs are ordered and oriented using a reference-based approach — similar to the scaffolding step of Ragtag, to produce pseudochromosomes [72]. After this, RATT is used to transfer annotated conserved gene models of the associated reference genome with little modifications to the input draft assembly; hence, the requirement for the reference to be a closely related organism [72]. To ensure new genes can be discovered, an *ab initio* approach using AUGUSTUS also takes place with *de novo* gene models generated in the presence of extrinsic information such as RNA-seq data [72]. Functional annotation information is added to the draft assembly

using OrthoMCL and additional small noncoding sequences annotated using ARAGORN and INFERNAL [72]. Predicted genes and associated annotations are merged, and a non-redundant consensus annotated draft genome is produced. Companion provides further results in the form of preliminary comparative genomics where information such as orthologous gene clusters as well as syntenic positions to the reference genomes are provided in easy, interactive plots and images [72]. With this, Companion can provide an optimised, automated pipeline encompassing multiple tools, which would require more time to learn and adequately implement individually.

3.2.9 Visualisation

Visualisation remains an essential step in multiple forms of *in silico* analyses, acting as a means of carrying out a form of analyses and a means of presenting analysed data in a succinct and easy to understand manner. As such, several tools are developed to display various forms of data to aid in biological research.

Artemis

Initially, Artemis was a simple stand-alone visualisation and annotation tool; however, it was soon expanded to provide a wide range of features to allow for genome visualisation, analysis and manipulation of high throughput data [73]. Other tools like 'BamView' further expanded this to allow for the visualisation of the Binary Alignment Mapping (BAM) file format [73, 74]. Functionally, Artemis can compare genomes, alignments and specified fragments based on the user inputs. A particularly unusual feature is its ability to display, analyse and manipulate variant calling files, allowing filtering and annotation of the called variants [73]. While this feature is not unique to Artemis, this, in conjunction with other capabilities, make Artemis one of the most robust and widely used visualisation and annotation tools available.

Integrative Genome Viewer

Integrative Genome Viewer (IGV) is a visualisation tool initially built to display large genomic datasets interactively, as part of a large project on cancer genomes [75]. Over time, IGV has been expanded to display genome information from NGS and third-generation sequencing technologies [75, 76]. IGV carries out visualisation of variant genome data with specific modes to allow easy understanding for the quality, proportion and coverage of the variant at a particular locus [75, 76]. Coupled with features that allow direct manipulation and incorporation of a multitude of data sources simultaneously, as well as other niche and bespoke implementations, IGV remains a highly robust and relevant visualisation tool that is widely used with bioinformatics.

3.3 *De novo* Genome Assembly

Genome assembly is the sequence of processes involving the reconstruction of a whole genome from disparate sequenced reads. As such, genome assembly acts a 'jigsaw puzzle' which aims to reform the 'picture' that is the whole genome. However, due to the inherent nature of genome sequencing and subsequent genome assembly tools, the generated assembly is expected to be shorter in length and fragmented into longer stretches of sequence data referred to as contigs [77]. Genome assembly can be carried out in the presence of a previously defined reference genome, or it can be carried out independently without using a reference.

The construction of a genome in such a manner, without a reference genome and using only information given within the sequence reads and computational inference is *de novo* genome assembly. Using only the sequence reads ensures no bias from a reference-guided assembly is introduced into the new assembly while also providing a means to detect novel regions of the genome [78]. This does not remove bias from the generated genome as any bias associated with the sequence reads are carried over. An example of this being sequence reads generated using sequence amplification would possess some of the amplification bias, which would impact any genome assembly carried out.

To achieve this, *de novo* genome assemblers utilise one of two assembly models: the

OLC model or the DBG model [77, 78], as described in chapter 3 subsection 3.2.4. DBG assemblers are suited and optimised for NGS short-read sequences, which are more sensitive to errors and repeats, resulting in very accurate but short contigs [78]. Alternatively, OLC assemblers are for longer error-prone sequence reads of third-generation sequencing technologies. As such, OLC assemblers generate assemblies of fewer but longer contigs that are prone to errors due to the inherent sequencing error present in the input reads. Indeed the approach taken is often dictated by the sequencing technology used, and thus the subsequent factors which may affect the resulting *de novo* genome.

One such factor is the coverage or read depth of the sequence reads to generate the genome assembly. With sufficient whole-genome coverage, the assembler can output an assembly representing a high proportion of the actual genome with minimal estimated errors. Hence, where there is insufficient coverage or poor-quality sequence reads, the subsequent accuracy and quality of the genome assembly will be called into question. In addition, other factors such as trimming and filtering choices and the choice of genome assemblers can significantly impact the genome assembly constructed.

3.4 Chapter 3: Aim and Rationale

Due to the protracted nature of the project, both in the length of time and the novel nature of the data, it was necessary to ensure that the tools and approaches implemented were appropriate and maintained. To achieve this, it is essential to evaluate current practices and software available for different analyses to develop a singular pipeline of tools that acts as the backbone for the larger project aim. Here, the aim was to assess basecalling, demultiplexing, the effects of reference-guided assembly approaches and also *de novo* assembly approaches available at different intervals during the project.

3.5 Methods and Results

3.5.1 Assessing basecallers and demultiplexers for Nanopore long reads

Assessing the performance of the Guppy basecaller

Upon release in late 2018, Guppy was recommended to supersede Albacore in carrying out basecalling. To determine the impact of Guppy on the sequence data, two previously sequenced multiplexed *P. knowlesi* samples were used [Table 3.1]. The data was sequenced using the SQK-RBK004 library preparation protocol with the R9.4.1 FLO-MIN106D flowcell on the MinION Mk1b platform [79] in March 2019, generating 220 multi-read FAST5 files. Albacore basecalling was carried out with the `sf5_read_fast5_basecaller` module of seamlessF5 and Albacore (*v2.3.4*) (Code. C.1), while Guppy basecalling was achieved with the `guppy_basecaller` module (*v2.3.7*) using default settings. seamlessF5 was used to allow multi-read FAST5 files as input for Albacore while Guppy natively accepts this file format [29].

Table 3.1: Representative clinical patient isolates

Isolate ID	Barcode	Parasitaemia (parasites/µL)	Input DNA concentration (ng/µL)
sks070a	BC01 ; BCO2	610402	18
sks339	BC03 ; BC04	321750	23

Sequenced clinical patient isolates used for basecaller comparison between Albacore (*v2.3.4*) and Guppy (*v2.3.7*). Each isolate was multiplexed with distinct barcodes to be appropriately sorted and removed by demultiplexing. Patient parasitaemia at whole blood extraction and subsequent enriched parasite DNA concentration are presented.

Basecalled reads from the `guppy_basecaller` were demultiplexed with `guppy_barcoder` in Guppy and the independent demultiplexing tool; Qcat (*v1.0.1*) [Figure 3.7, Appendix Code. C.1]. Demultiplexed reads were quality assessed with FastQC (*v0.11.8*) and reports for each barcode and associated basecaller were pooled and displayed with MultiQC

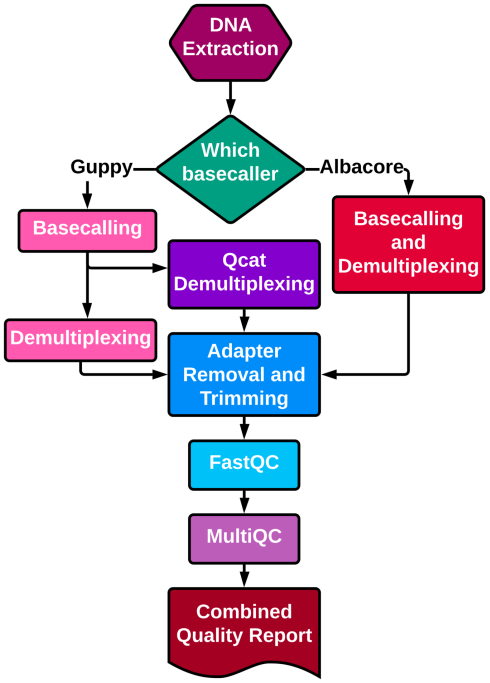


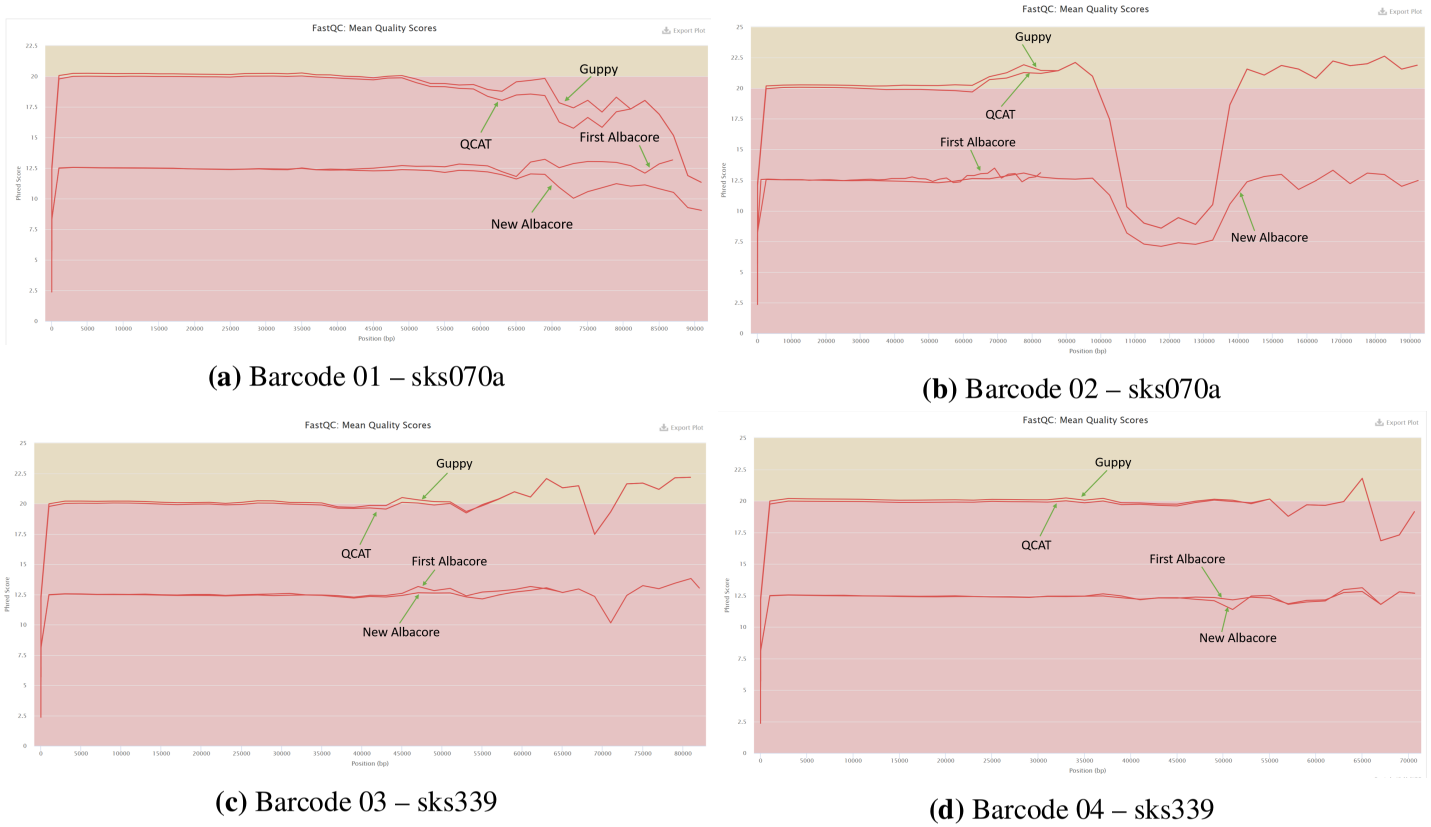
Figure 3.7: Pipeline to assess basecallers and demultiplexers. A pipeline taking sequence data through basecalling by either Albacore or Guppy. Albacore basecalls and demultiplexes, while Guppy has separate basecalling and demultiplexing modules. Guppy basecalled data are also demultiplexed using Qcat. Demultiplexed data are trimmed for adapters using Porechop before quality checks with FastQC. FastQC reports are combined using MultiQC and reported as a single report for each sequenced barcode/isolate. Code is in Appendix Code C.1

(*v1.4*) [Figure 3.7]. Statistical metrics were calculated using assembly-stats (*v1.0*) [80] on reads which surpassed the set quality threshold.

Guppy surpasses Albacore in sequencing quality and accuracy: Due to an unknown error, Albacore initially failed and had to be re-run. The incomplete run is still presented and denoted as 'First Albacore' while the successful run is 'New Albacore'. The 'First Albacore' is henceforth disregarded. Average sequence quality plots generated in MultiQC show the 'New Albacore' dataset for all barcodes possess the lowest quality scores at an average of ~Q12 and both Guppy and Qcat at ~Q20 [Figure 3.8].

Guppy appears to have slightly higher quality scores across the four barcodes, including longer reads with reduced sequencing quality. A notable drop in read quality can be observed in reads spanning 10.5 - 13.5 Kb across all basecallers, with a pronounced

decrease in the Guppy and Qcat basecalled read data [Figure 3.8b]. In assembly-stats outputs, across all barcodes, Qcat generated basecalled reads of longer lengths whilst having the lowest 'unclassified' total base-pair length [Figure 3.9]. Here 'unclassified' refers to reads that surpass the quality threshold but cannot be associated with a barcode. Similarly, Qcat reports more reads per barcode than both Guppy and Albacore, with Albacore generating the lowest read yield [Figure 3.9(b)]. Here, Qcat was unable to classify ~9 % of input reads while Guppy and Albacore were unable to resolve ~20 % of input reads into the distinct barcode bins [Figure 3.9(b)]. As such, a combination of Guppy basecalling and Qcat demultiplexing was be taken forward for basecalling.



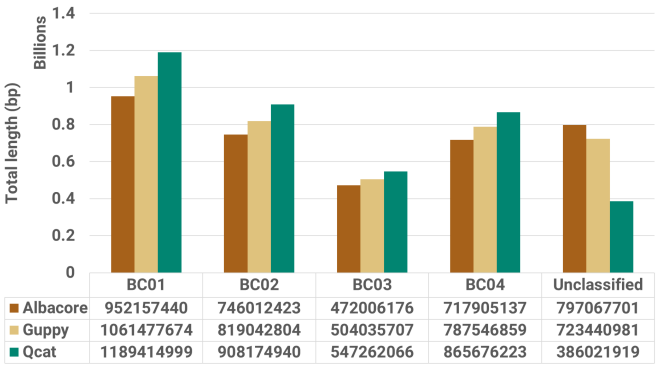
(a) Barcode 01 – sks070a

(b) Barcode 02 – sks070a

(c) Barcode 03 – sks339

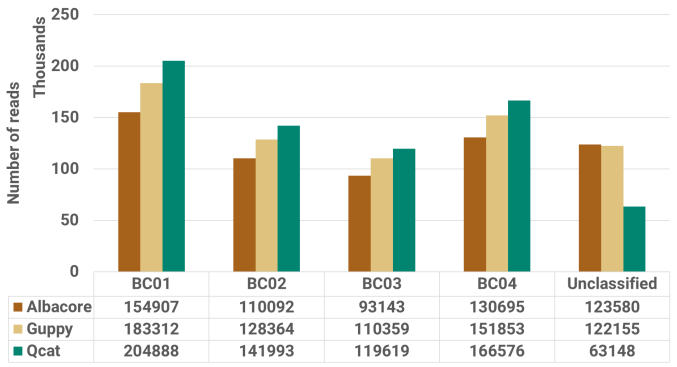
(d) Barcode 04 – sks339

Figure 3.8: FastQC plots aggregated in MultiQC after quality checks. Plots display mean quality score comparison of each basecaller/demultiplexer across the four barcodes described in Table 3.1. Quality scores use the Phred standard scaling [81] and the trace for each basecaller is annotated. All barcodes show an increase in quality score between Albacore (*New Albacore*) and Guppy. All plots show a decrease in quality in reads >40 000 bp with (**a**) greater deviation between Guppy and Albacore as reads become longer. A sharp drop in read quality is observed between 98 000 bp and 130 000 bp before returning to Q21 for reads >140 000 bp. However, an increase in quality is shown in (c) reads >70 000 bp for all basecallers, while (d) the inverse is apparent, with a rapid decline in quality after 70 000 bp



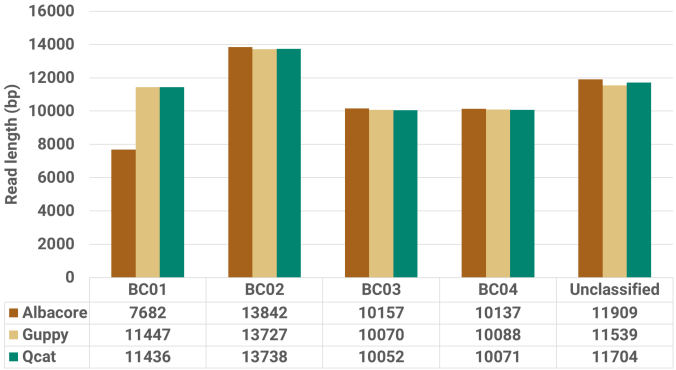
	BC01	BC02	BC03	BC04	Unclassified
Albacore	952157440	746012423	472006176	717905137	797067701
Guppy	1061477674	819042804	504035707	787546859	723440981
Qcat	1189414999	908174940	547262066	865676223	386021919

(a) Total base pair sequenced



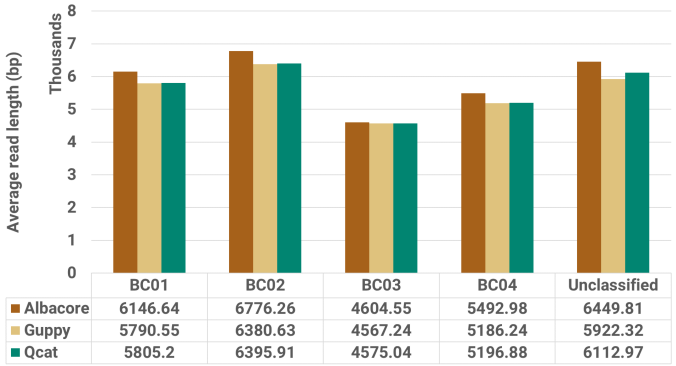
	BC01	BC02	BC03	BC04	Unclassified
Albacore	154907	110092	93143	130695	123580
Guppy	183312	128364	110359	151853	122155
Qcat	204888	141993	119619	166576	63148

(b) Total read count



	BC01	BC02	BC03	BC04	Unclassified
Albacore	7682	13842	10157	10137	11909
Guppy	11447	13727	10070	10088	11539
Qcat	11436	13738	10052	10071	11704

(c) N50



	BC01	BC02	BC03	BC04	Unclassified
Albacore	6146.64	6776.26	4604.55	5492.98	6449.81
Guppy	5790.55	6380.63	4567.24	5186.24	5922.32
Qcat	5805.2	6395.91	4575.04	5196.88	6112.97

(d) Average read length

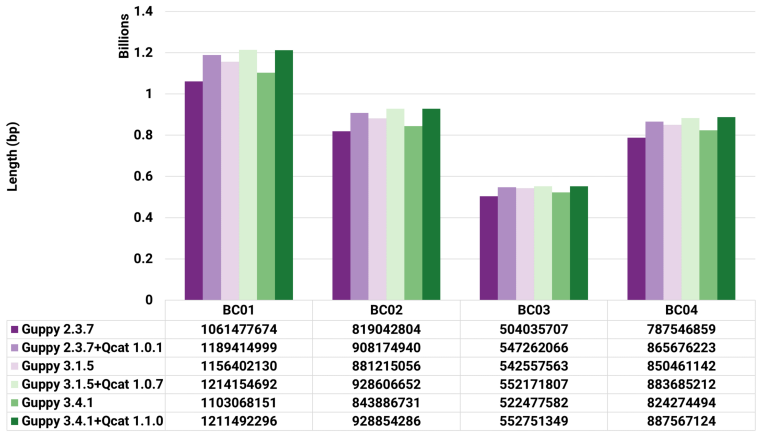
Figure 3.9: Comparison of the yield for parasite sequence data after basecalling and demultiplexing. Basecalling and demultiplexing was carried out for sks070a (BC01,BC02) and sks339 (BC03,BC04) using Albacore (v2.3.4), Guppy (v2.3.7) and Qcat (v1.0.1) to carry out basecalling and demultiplexing. The numeric metrics for each barcode is given in the data table below each graph. The unclassified dataset are reads which could not be resolved into the four barcode folders.

Assessing the performance of Guppy and Qcat for demultiplexing

Once Guppy was determined to be the most appropriate basecalling tool for the sequence data within this project, an appropriate demultiplexing tool was also necessary. As noted previously, Guppy contained the `guppy_barcoder` module for demultiplexing while ONT also provided an independent demultiplexer in the form of Qcat. Furthermore, ONT subsequently released the 'High-accuracy (HAC)' basecalling algorithm model which improved sequence accuracy by at least 5 %. The HAC algorithm is only available for use in conjunction with a Graphical Processing Unit (GPU) which was initially unavailable for use in this project. Upon gaining access to a GPU, it became necessary to assess Guppy_HAC's effect on a known dataset with known quality. With this, an assessment of the demultiplexing capabilities of Guppy was carried out for Guppy *v.2.3.7*, *v3.1.5* and *v3.4.1*.

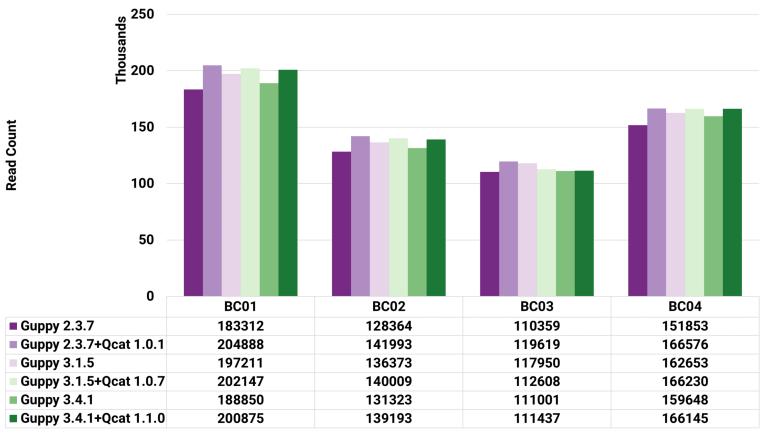
Prior to demultiplexing, basecalling was carried out on the same dataset sequenced in March 2019 [Table 3.1] using an expanded form of the `guppy_basecaller` command [Appendix Code. C.1] which included the '`-x auto`' parameter to activate the GPU mode in Guppy. Courtesy of Dr. Jon Thomson, the project gained access to an NVIDIA GTX GeForce™1060 GPU with 6GB of virtual memory to accelerate the Guppy basecalling process. Subsequently, Guppy_HAC basecalled sequenced data was demultiplexed using the `guppy_barcoder` module of Guppy (*v3.1.5*) and Guppy (*v3.4.1*), as well as Qcat (*v1.0.7*) and Qcat (*v1.1.0*). The outputs of all conditions were quality assessed using FastQC (*v0.11.8*) and numerical statistical metrics calculated with assembly-stats (*v1.0*). FastQC outputs were aggregated and presented on MultiQC (*v1.4*) [Figure 3.7].

Qcat generates reads of similar length and quality to Guppy: All Guppy and Qcat demultiplexing processes were completed with little difficulty. The addition of Guppy (*v2.3.7*) allowed for an assessment of the tool before the release of the HAC algorithm. A noticeable improvement can be seen between Guppy (*v2.3.7*) and Guppy (*v3.1.5*) with an increase in the total base pair (bp) basecalled and demultiplexed, as well as increases in the number of reads resolved for each barcode [Figure 3.10]. However, Guppy (*v3.4.1*) brought about a decrease in all four metrics (Assembly length, contigs, N50 and average contig length) while also appearing to resolve reads in proportionally lower quality.



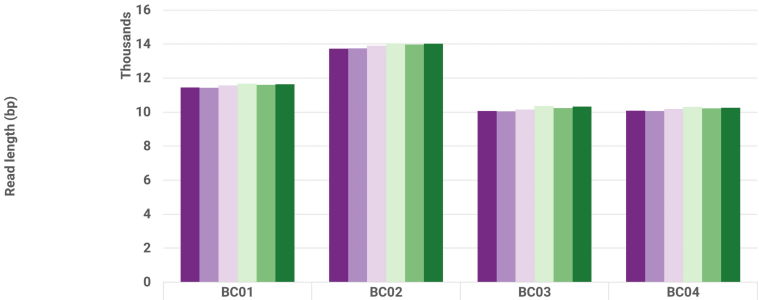
	BC01	BC02	BC03	BC04
Guppy 2.3.7	1061477674	819042804	504035707	787546859
Guppy 2.3.7+Qcat 1.0.1	1189414999	908174940	547262066	865676223
Guppy 3.1.5	1156402130	881215056	542557563	850461142
Guppy 3.1.5+Qcat 1.0.7	1214154692	928606652	552171807	883685212
Guppy 3.4.1	1103068151	843886731	522477582	824274494
Guppy 3.4.1+Qcat 1.1.0	1211492296	928854286	552751349	887567124

(a) Total base pair sequenced



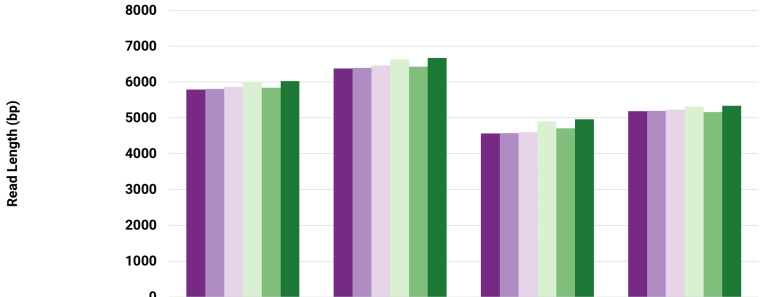
	BC01	BC02	BC03	BC04
Guppy 2.3.7	183312	128364	110359	151853
Guppy 2.3.7+Qcat 1.0.1	204888	141993	119619	166576
Guppy 3.1.5	197211	136373	117950	162653
Guppy 3.1.5+Qcat 1.0.7	202147	140009	112608	166230
Guppy 3.4.1	188850	131323	111001	159648
Guppy 3.4.1+Qcat 1.1.0	200875	139193	111437	166145

(b) Total read count



	BC01	BC02	BC03	BC04
Guppy 2.3.7	11447	13727	10070	10088
Guppy 2.3.7+Qcat 1.0.1	11436	13738	10052	10071
Guppy 3.1.5	11567	13899	10144	10181
Guppy 3.1.5+Qcat 1.0.7	11678	14044	10355	10299
Guppy 3.4.1	11595	13976	10239	10227
Guppy 3.4.1+Qcat 1.1.0	11640	14023	10316	10260

(c) N50



	BC01	BC02	BC03	BC04
Guppy 2.3.7	5790.55	6380.63	4567.24	5186.24
Guppy 2.3.7+Qcat 1.0.1	5805.2	6395.91	4575.04	5196.88
Guppy 3.1.5	5863.78	6461.8	4599.89	5228.68
Guppy 3.1.5+Qcat 1.0.7	6006.3	6632.48	4903.49	5316.04
Guppy 3.4.1	5840.98	6426.04	4706.96	5163.07
Guppy 3.4.1+Qcat 1.1.0	6031.08	6673.14	4960.21	5342.12

(d) Average read length

Figure 3.10: Metrics of parasite sequence data after demultiplexing with Guppy and Qcat. Comparison of the effects of different versions of Guppy and Qcat demultiplexers on the yield of sks070a (BC01,BC02) and sks339 (BC03,BC04). The same sequence data was processed using different versions of both software as they were released over a period of a year.

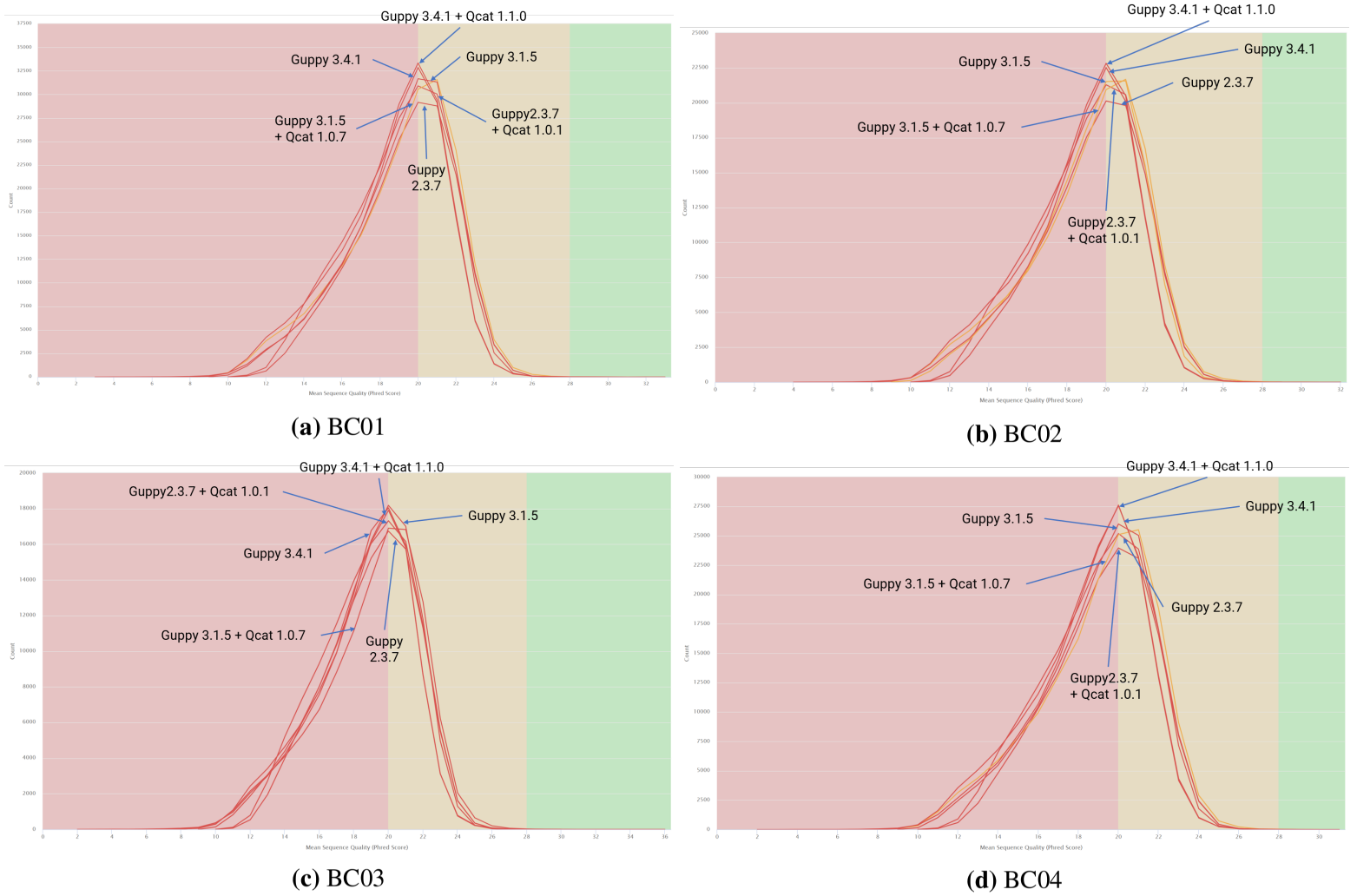


Figure 3.11: Aggregated quality scores of barcoded samples after demultiplexing with Guppy and Qcat. Average quality scores for sks070a (BC01,BC02) and sks339 (BC03,BC04) after demultiplexing with Guppy(*v2.3.7,v3.1.5,v3.4.1*) and Qcat(*v1.0.1,v1.0.7,1.1.0*). Each barcode shows a peak average quality at ~Q20 for each version of both demultiplexers. However, the yield of reads and the integrity of the reads for each demultiplexer version differs.

The lower read quality is indicated by a decrease in the proportion of reads, with relatively higher quality in *v3.4.1* in comparison to *v3.1.5* [Figure 3.11(a),3.11(b),3.11(d)]. Across the conditions set and for each barcode, Qcat-demultiplexed sequence yield consistently showed higher total length, read count and average read length than Guppy yields [Figure 3.10]. This is also maintained in the average quality metrics, where Qcat shows similar or better average read quality to its associated Guppy counterpart [Figure 3.11]. However, little difference is observed in the difference between Qcat (*v1.0.7*) and Qcat (*v1.1.0*).

3.5.2 Confirmation of contaminants in the *Plasmodium knowlesi* PKNH reference genome

Generating a *de novo* genome assembly from a mixed sample requires checking for and removing any undesired contaminant sequences within the generated assembly. In such circumstances, contamination may arise due to multiple strains of the same causative organism or, indeed, other infectious organisms. As such, detecting contaminants before sequencing can be complex, with most contamination detected being in the form of proteins, polyphenols and polysaccharides, which would affect the DNA sequencing process [82]. However, determining the contamination of an unwanted organism is only possible after sequencing, with further analyses to expose key indicators of contamination such as GC content, uneven coverage and random copy numbers. Under ideal conditions, no contamination would be present in an assembly, especially in published and available sequence data.

An assessment was carried out to determine the viability of carrying out a reference-guided genome assembly approach, utilising both the *P. knowlesi* PKNH [83] and PKNOH [84] reference genomes. Briefly, thawed *P. knowlesi*-infected patient whole blood was leucocyte depleted using the CD45 DynaBeads method [see *chapter 2 subsection 2.4.5*] [85]. Additionally, an isolate of the cultured *P. knowlesi* PkA1H1 experimental line was also leucocyte depleted and DNA extracted. Parasite-enriched DNA was subsequently sequenced using the ONT MinION sequencing platform before basecalling with Guppy (*v3.4.1*) and demultiplexing with Qcat (*v1.1.0*). Adapters were

removed with porechop (*v0.2.3*) and to retain sequence reads which only correspond to *P. knowlesi* DNA, adapter-removed reads were aligned against the PKNH reference genome [83]. Sequence reads which successfully aligned against the reference genome were extracted for genome assembly using Flye, and the generated draft assemblies were assessed using the BlobTools workflow [58]. These assemblies will be referred to as 'PKNH-guided' assemblies. Here, the PKNH-guided assemblies were aligned against their input reads (*reads which aligned to PKNH*) with minimap2 [46]; and the resulting alignment files sorted and compressed with SAMtools [42]. In BlobTools, the assemblies were processed through an *all-vs-all* BLAST alignment search against the National Center for Biotechnology Information (NCBI) nucleotide (nt) database [86]. The resulting plots were manually inspected and subsequently cleaned using a custom script to remove all contigs not labelled as "*Apicomplexa*", "*no-hit*" and "*undef*". Unexpectedly, human contaminant sequences were identified within all the assemblies, prompting an investigation into the presence of human genomic contamination with the PKNH reference genome.

The genome assembly process was repeated once human contamination was detected. However, here, the adapter-free reads were aligned against the PKNOH reference genome [84]. The PKNH-guided assemblies were compared against the PKNOH-guided assemblies using BlobTools and assembly-stats. Contigs identified by BlobTools to be "*Chordata*" in the PKNH-guided assemblies were extracted using seqtk(*v1.3r106*) [87] and aligned against the nt database, using the high similarity 'megablast' mode of the online BLAST aligner [88]. The coordinates of the top hits within the nt database corresponding to a Chordata genome were stored in BED files and subsequently used to extract the identified regions from the human GRCh38p.12 reference genome [89]. The extracted regions of the GRCh38p.12 genome were aligned against the PKNH reference genome with minimap2 to confirm their presence within the PKNH reference genome.

BlobTools confirms the presence of human contamination in PKNH-guided assemblies: To confirm that the PKNH reference genome introduces contamination, a merged reference genome was artificially created by combining the PKNH and PKNOH reference genomes. A third genome assembly process was carried out using the merged reference genome to guide the assembly as previously described. For this, sequence reads from the St. Andrews Cultured PKA1H1 isolate was used to generate three

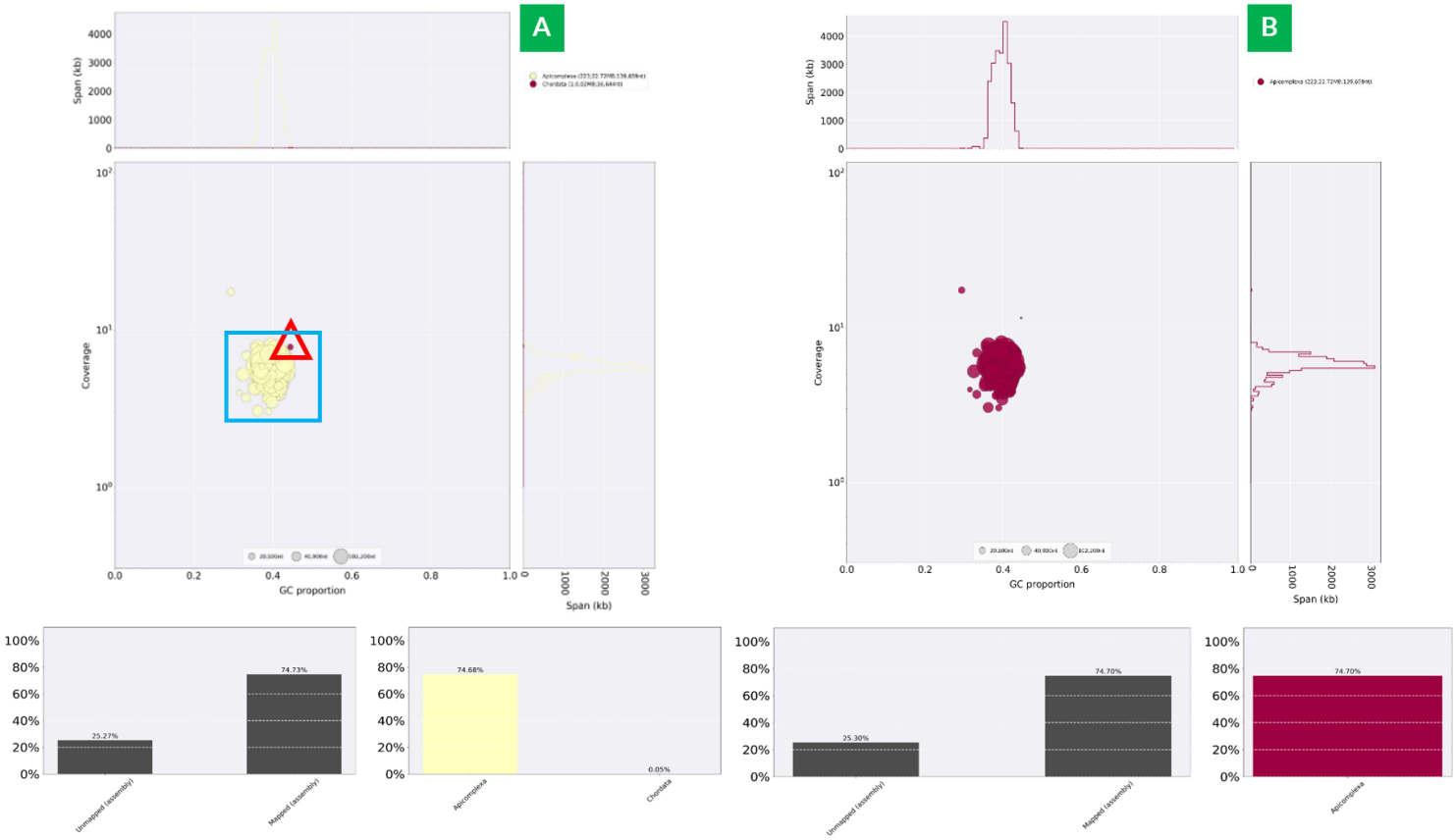


Figure 3.12: Blobplots for the Cultured *Plasmodium knowlesi* PkA1H1 draft assembly generated with a reference-guided approach using the PKNH reference genome [83]. Cultured PkA1H1 sequenced reads which mapped to the PKNH reference genome were used to generate the PkA1H1 genome assembly. **(a)** In the raw assembly, contamination in the form of Chordata (red triangle) is observed separated from the Apicomplexa (blue square) cluster. After cleaning the assembly, the Chordata contamination is removed, leaving just Apicomplexa sequences. Input reads used to generate PkA1H1 genome assembly were mapped back to the generated assembly **(A,B: lower left)**. The proportion of the assembly corresponding to identified taxonomic ranks **(A,B: lower right)** shows the presence of Chordata sequences in the raw assembly **(a)** and the absence of the Chordata sequences after cleaning **(b)**

reference-guided isolate genomes to act as an example. Assemblies of the three subsets – *PKNH-guided subset, PKNOH-guided subset and Merged-guided subset*– were compared for contaminants, and where found, identified contigs were extracted for further study.

Overall, BlobTools reported Apicomplexan (presumably *P. knowlesi*) GC content to be ~39 % (blue square) while non-Apicomplexan (Chordata) content had ~42 % (red triangle) [Figure 3.12a,top]. Some outlier Apicomplexan sequences can be observed with ~30 % GC content. No contigs were designated as 'no-hit' or 'undef', suggesting that all contigs were able to be classified into known taxonomic ranks. Coverage plots revealed only 74 % of input reads mapped to the generated Cultured PkA1H1 assembly [Figure 3.12a,bottom].
Of these, a tiny proportion (0.05 %) was identified as Chordata in the raw Cultured PkA1H1. After the raw assembly was cleaned, the contamination was not reported, resulting in an expected increase in the proportion of Apicomplexan sequences reported, and the overall proportion of the assembly, which mapped back to the input PkA1H1 reads [Figure 3.12b]. All other patient isolates (*not shown*) except for one, reported assemblies with lower Apicomplexan proportions, higher Chordata contamination and more sequences which were identified to be 'no-hit' or 'undef' (undefined taxonomic parent). The PKNOH-guided assembly dataset also clusters at ~39 % GC content with few observable outliers within them. However, no contaminants were reported in the assemblies before or after the cleaning phase [Figure 3.13]. Contigs identified as 'Chordata' were extracted, including in sks333 which possessed multiple contigs and a scaffold detected as a contaminant [Table 3.2, Appendix Table D.1].

Contaminated contigs are fragments of the human mitochondrion: BLAST alignment was done using the contaminated contig extracted from the Cultured PkA1H1. Possessing 16 644 bp, contig_168 [Table 3.2, Appendix Table D.1] was BLAST aligned against the whole PKNH reference genome of length (24 395 979 bp) using the high similarity 'megablast' algorithm. Here, a weak hit of low query identity was reported however, within the region of similarity, the hit possessed a high percentage identity corresponding to chromosome 00 of the PKNH reference genome [Figure 3.14].

With the megablast alignment showing only ~4000 bp mapped, the less stringent 'blastn' algorithm provided more hits, however, the hits reported were of shorter length, with

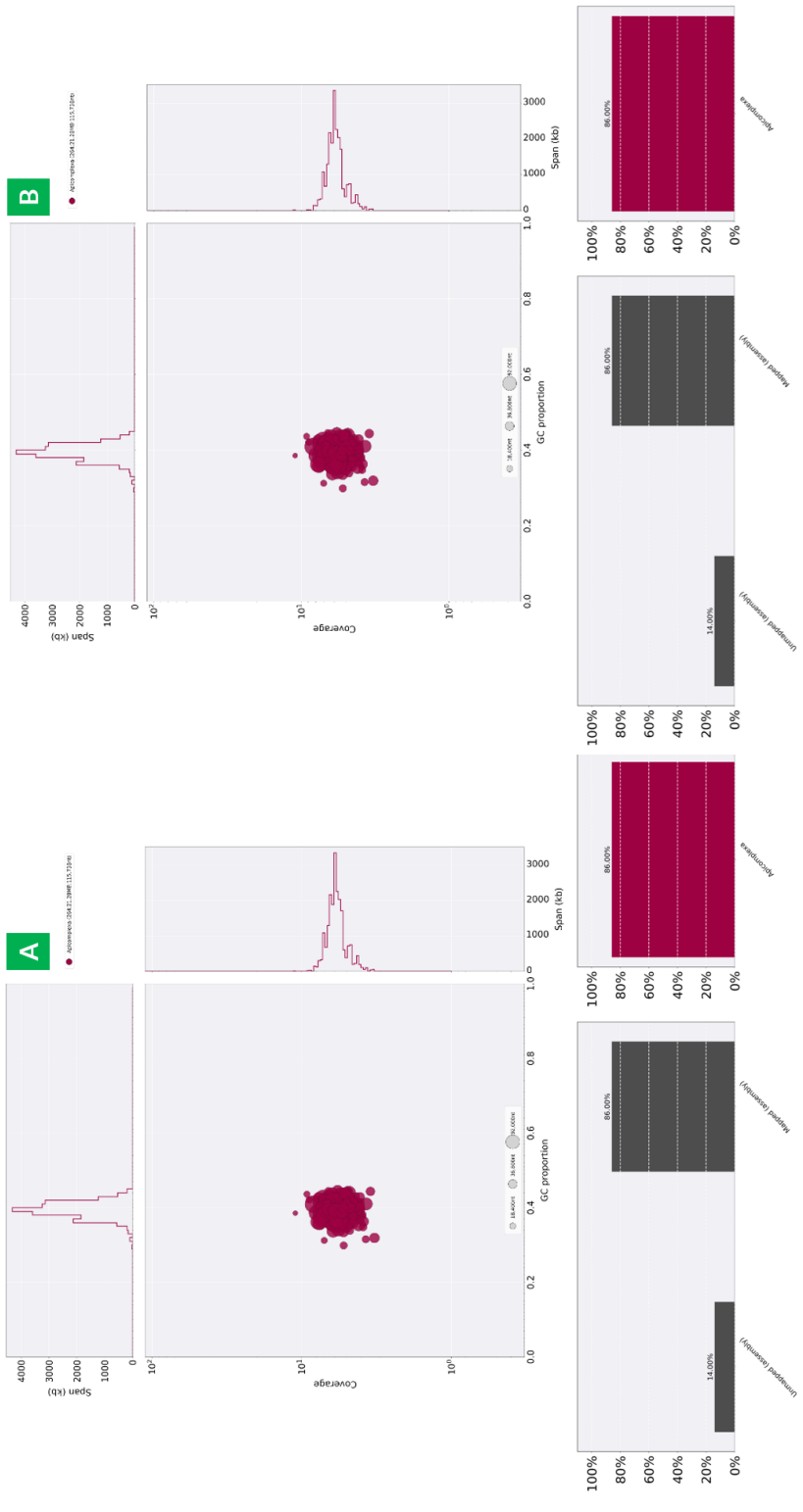


Figure 3.13: Blobplots for the Cultured *Plasmodium knowlesi* PkA1H1 draft genome assembly generated with a reference-guided approach using the PKNOH reference genome [84]. Cultured PkA1H1 sequenced reads which mapped to the PKNOH reference genome were used to generate the PkA1H1 genome assembly. In the raw assembly **(a)**, no contamination is observed before or after cleaning the assembly **(b)**

Table 3.2: Representive list of identified contaminated contigs

Assembly Isolate ID	Contaminated contig
Cultured PKA1H1	*contig_168*
sks047	*contig_24, contig_28, contig_29, contig_39, contig_48, contig_51*
sks048	*none*
sks050	*none*
sks058	*contig_124, contig_174, contig_182*
sks070	*contig_25, contig_28, contig_52*
sks074	*contig_102, contig_104, contig_138, contig_143, contig_159, contig_177, contig_40, contig_99*
sks078	*contig_166, contig_178, contig_166, contig_166, contig_166, contig_299, contig_31, contig_312, contig_39, contig_4, contig_40, contig_43, contig_44*
sks125	*contig_214, contig_233, contig_60*
sks325	*contig_4*
sks331	*contig_184, contig_190, contig_197, contig_227*
sks333	*contig_118, contig_119, contig_144, contig_154, contig_202, contig_215, contig_216, contig_238, contig_253, contig_282, contig_294, contig_315, contig_325, contig_348, contig_351, contig_354, contig_356, scaffold_279*
sks339	*contig_25, contig_32*
sks344	*none*

Representative list of contigs identified as Chordata by BlobTools within genome assemblies generated using the PKNH reference genome to guide the assembly process. The contigs were extracted from the genome assemblies using seqtk [87]. Complete list of contigs are presented in Appendix Table D.1.

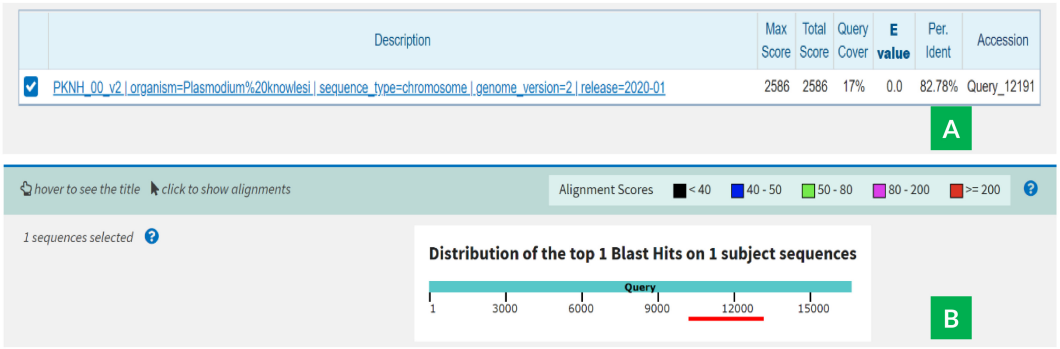


Figure 3.14: MegaBlast alignment of contaminated contigs. **(a)** Alignment of the Cultured PkA1H1 draft assembly contaminated contig_168 identified to be Chordata against the whole PKNH reference genome. **(b)** The graphical representation of the mapped region of similarity between the contaminated contig and the reference genome's chromsome 00.

Figure 3.15: MegaBlast alignment of contaminated contigs against the NCBI *nt* database. Alignment of the Cultured PkA1H1 draft assembly contaminated contig_168 identified to be Chordata against the NCBI nucleotide (nt) database showing **(a)** high similarity and high quality hits across the contig and **(b)** the graphical representation of the mapped region of similarity between the contaminated contig and the genes of similarity corresponding to the human mitochondrion complete genome.

lower quality and identity; often with 0 % coverage (*not shown*). However, when contig_168 was aligned against the entire NCBI *nt* database, results show that the entirety of the contig aligns with high similarity and identity to the human mitochondrion [Figure 3.15]. Analyses carried out for the other isolates (*sks047-sks344*) report similar outcomes with all hits being against the human mitochondirion genome; though, a few contigs also show top hits for the 'Pan troglodytes BAC clone' and 'Eukaryotic synthesis construct chromosome' [Appendix Table D.1].

Table 3.3: Representative subset of contaminated contigs and their top hits

Assembly Isolate ID	Contaminated contigs	Nucleotide database top hit(s)
Cultured PkA1H1	contig_168	Fragment of Human mitochondrion
sks048	none	-
sks058	contig_124	*Homo sapiens* clone BAC JH4
	contig_174	Human mitochondrion (full length)
	contig_182	Fragment of *Homo sapiens* 3 BAC RP11-512E23
sks070	contig_25	Fragment of *Homo sapiens* clone BAC JH4
	contig_28	Fragment of *Homo sapiens* BAC clone RP11-1396O13
	contig_52	*Homo sapiens* isolate KK23 mitochondrion (2x full-length copies)
sks125	contig_214	Fragment of *Homo sapiens* clone BAC JH4
	contig_233	Human mitochondrion (full length)
	contig_60	Fragment of *Homo sapiens* BAC clone RP11-1396O13
sks325	contig_4	*Homo sapiens* isolate BAL38 mitochondrion (full length)
sks339	contig_25	Fragment of *Homo sapiens* clone BAC JH4 and JH11
	contig_32	*Homo sapiens* mitochondrion (full length)

Contaminated contigs identified in selected isolates of the PKNH-guided assembly dataset by Blobtools. Upon extraction the contigs were aligned against the NCBI *nt* database. The top hits of each contig's alignment are presented here. Complete list of contigs and their top hits are found in Appendix Table D.1.

While the full length of the contig was aligned, alignment statistics showed the contig

Table 3.4: BLASTn alignment of the contaminated contig_168 against the human mitochondrion

Query	Subject	% identity	Length	M.match	Gaps	Q.start	Q.stop	S.start	S.stop	eval.	bitscore
J01415.2:6000-16500	PKNH_13	75.71	70	17	0	2152	2221	1293301	1293370	2.30E-04	50.9
J01415.2:6000-16500	PKNH_13	100	19	0	0	1945	1963	721417	721435	5.1	35.6
J01415.2:6000-16500	PKNH_13	100	19	0	0	5066	5084	1507804	1507786	5.1	35.6
J01415.2:6000-16500	PKNH_06	86.49	37	4	1	271	307	959183	959148	0.1	41.9
J01415.2:6000-16500	PKNH_06	87.10	31	3	1	8127	8156	578421	578391	5.1	35.6
J01415.2:6000-16500	PKNH_12	95.83	24	1	0	1726	1749	2163963	2163986	0.4	40.1
J01415.2:6000-16500	PKNH_12	91.67	24	2	0	4105	4128	2811438	2811415	5.1	35.6
J01415.2:6000-16500	PKNH_14	100	21	0	0	8019	8039	2566421	2566401	5.1	39.2
J01415.2:6000-16500	PKNH_14	81.08	37	7	0	8699	8735	53595	53631	5.1	36.5
J01415.2:6000-16500	PKNH_14	91.67	24	2	0	4540	4563	200293	200316	5.1	35.6
J01415.2:6000-16500	PKNH_09	88.89	27	3	0	3648	3674	545855	545829	5.1	36.5
J01415.2:6000-16500	PKNH_09	89.66	29	2	1	5202	5230	1373883	1373910	5.1	36.5
J01415.2:6000-16500	PKNH_09	86.21	29	4	0	10361	10389	1761346	1761374	5.1	35.6
J01415.2:6000-16500	PKNH_11	91.67	24	2	0	4935	4958	976761	976784	5.1	35.6
J01415.2:6000-16500	PKNH_08	87.10	31	3	1	8127	8156	1698256	1698226	5.1	35.6
J01415.2:6000-16500	PKNH_08	91.67	24	2	0	9439	9462	1883803	1883780	5.1	35.6
J01415.2:6000-16500	PKNH_07	91.67	24	2	0	6252	6275	829069	829046	5.1	35.6
J01415.2:6000-16500	PKNH_04	91.67	24	2	0	9439	9462	704016	703993	5.1	35.6

BLASTn alignment of the 6000 - 16 500 bp region of the J01415 human mitochondrion reference genome against the whole *P. knowlesi* reference genome. This region was identified in contig_168 of the Cultured PkA1H1 draft genome assembly to contain human contamination

M.match - Mismatch ; *Q.start* - Query start ; *Q.stop* - Query stop ; *S.start* - Subject start ; *S.stop* - Subject stop ; *eval* - E-value

only mapped to ~66 % of the human mitochondrial genome. The region mapped by contig_168 corresponds to 6000 - 16 500 bp of the human mitochondrion in the GRCh38p.12 reference genome with the accession code: J01415 [89]. Extraction of this region from J01415 and subsequent blastn alignment against the entire PKNH reference genome to determine the location of the contamination within the *P. knowlesi* PKNH reference genome showed little to no similarity between the two sequences [Table 3.4].

No clear evidence of contamination is seen in the PKNH reference genome, however, the Merged-guided dataset and subsequent BlobTools reported the presence of Chordata contamination with ~42 % GC content sequences [Figure 3.16a, top]; as seen in the PKNH-guided genome assemblies [Figure 3.12a].

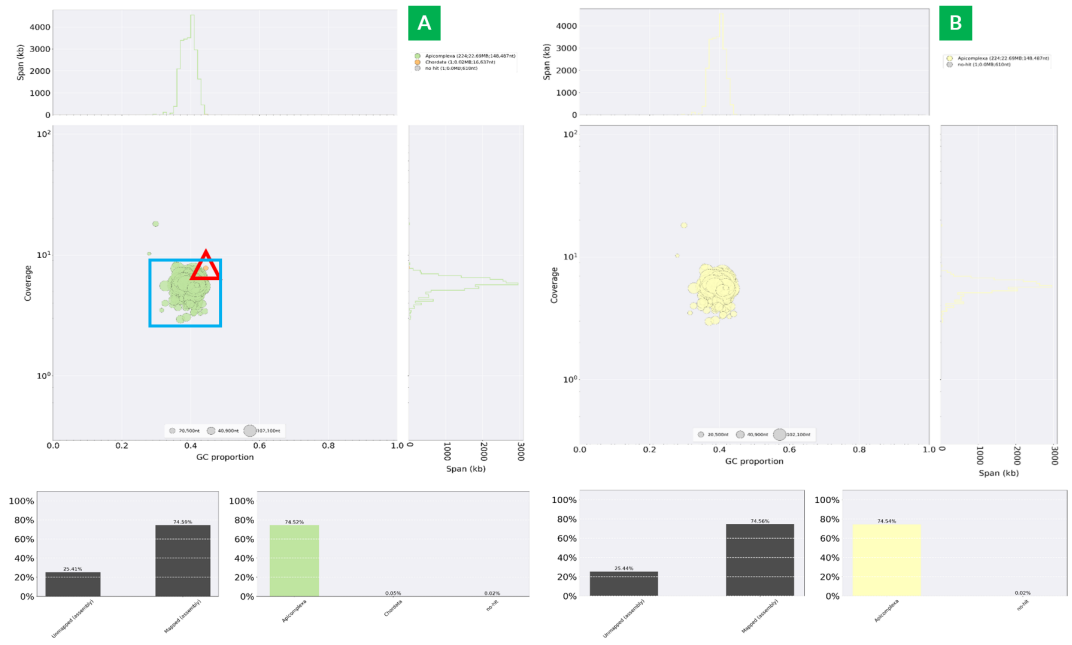


Figure 3.16: Blobplots for the Cultured *Plasmodium knowlesi* PkA1H1 genome assembly generated with a Merged reference-guided approach. A simulated reference genome manually produced from a combination of the PKNH and PKNOH reference genomes was used as a guide by the assembler [83, 84]. Cultured PkA1H1 sequenced reads which mapped to the Merged reference genome were used to generate the PkA1H1 genome assembly. In the raw assembly **(a)**, contamination is observed as Chordata, and after cleaning **(b)**, the contamination is not reported.

3.5.3 Comparison of *de novo* assemblers

To further expand on determining the most appropriate demultiplexing approach, an assessment of the demultiplexer's effect on the resulting data was implemented. Here, demultiplexed reads from Guppy (*v3.1.5*) and Qcat (*v1.0.7*) had adapters removed by Porechop (*v0.2.3*) [36] before being aligned against the human reference genome version 38; patch 12 (GRCh38p.12) [90, 91] using minimap2 (*v.2.11-r283-dirty*). Reads which did not align were separated into FASTQ files with samtools (*v1.6*) and bedtools (*v2.27.0*). The resulting FASTQ files were used as input for Canu (*v1.8*), Flye (*v2.5*) and Redbean (*v2.3*).

Canu was first implemented with default settings using a genome size of 24 Mb, the '*nanopore-raw*' preset, and only the longest 40-fold coverage of the input reads. To use the entirety of the input reads, Canu was repeated to include the '*corOutCoverage*' parameter. For Redbean, the recommended preset was first implemented using error-corrected reads generated by Canu as input. However, due to the increased error rate and uneven coverage of the input data, a parameter scan/search was also implemented using iterations of different parameters. These parameters accommodated for lower coverage isolates and determine the best *k*-mer size to use for the layout module of the assembly algorithm.

Table 3.5: Statistical metrics for PKNH and PKNOH reference genomes

Metrics	PKNH	PKNOH
Total Length (bp)	24395979	24771595
Contigs/Chromosomes	19	28
N50 (bp)	2162603	1832627
BUSCO Completedness (%)	68.9	68.9

Metrics for the *P. knowlesi* PKNH [83] and PKNOH [84] reference genomes calculated using BUSCO (*v3*) [62] and assembly-stats[80].

The resulting Redbean-generated assemblies were quantitatively assessed with the BUSCO (*v3*) tool [61, 62] and assembly-stats (*v1.0.0*) before being manually collated. Furthermore, the *P.knowlesi* PKNH [83] and PKNOH [84] reference genomes were

also assessed with BUSCO (*v3*) to provide contextual comparison to the *Plasmodium falciparum* BUSCO model used, as no BUSCO model is available specifically for *P. knowlesi* and *P. falciparum* is the closest protist relative [Table 3.5]. For each isolate and both Guppy and Qcat datasets, the best parameter was chosen based upon their BUSCO score, assembly length, the number of contigs resolved and N50 in descending order of priority. The Guppy and Qcat results were compared to determine which protocol gave the most accurate assemblies from each subset. Upon arriving at the most suitable protocol for each isolate, the parameters were averaged to result in a single set of parameters that would be suitable for the entire dataset.

Canu performs better with more coverage: Using the default Canu preset where only the longest 40x is used, only six out of eighteen sequenced isolates possessed sufficient coverage to be successfully assembled [Appendix Table D.2]. The 12 isolates which

Table 3.6: Numerical statistics of the assemblies generated using Canu with maximum coverage depth (corOutCoverage)

Archived Sample ID	Guppy				Qcat			
	Coverage^x	Length	Contigs	N50	Coverage^x	Length	Contigs	N50
PKNH*	*Unknown*	24395979	19	2162603	*Unknown*	24395979	19	2162603
PKNOH**	*Unknown*	24771595	28	1832627	*Unknown*	24771595	28	1832627
sks047	38.65X	22240229	158	415287	32.85X	21900894	180	309965
sks048	36.47X	22661103	140	386982	40.71X	23032387	132	464190
sks058	19.55X	20606041	229	227410	21.16X	20504144	282	199147
sks070	47.29X	23204224	85	736048	52.56X	23500482	83	785941
sks125	11.89X	18938141	469	87185	12.89X	19155919	462	95523
sks339	131.83X	23854232	56	1153367	141.11X	23973011	60	1152924

The same sequence data used in Appendix Table D.2 is used here, hence the same coverage depth is applicable, thus resulting in only six assemblies successfully assembling. Omitted isolates which failed de novo assembly are: *sks071*, *sks074*, *sks078*, *sks125*, *sks231*, *sks254*, *sks280*, *sks330*, *sks331*, *sks333*, *sks343*, *sks344*. All failed due to lack of coverage depth

* - The *P. knowlesi* PKNH reference genome generated using Sanger sequencing [83]

** - The *P. knowlesi* PKNOH genome generated using PacBio sequencing and HiC reads [84]

x - Total input read coverage based on the 24megabase estimate genome length. However, Canu only uses the longest 40x of reads

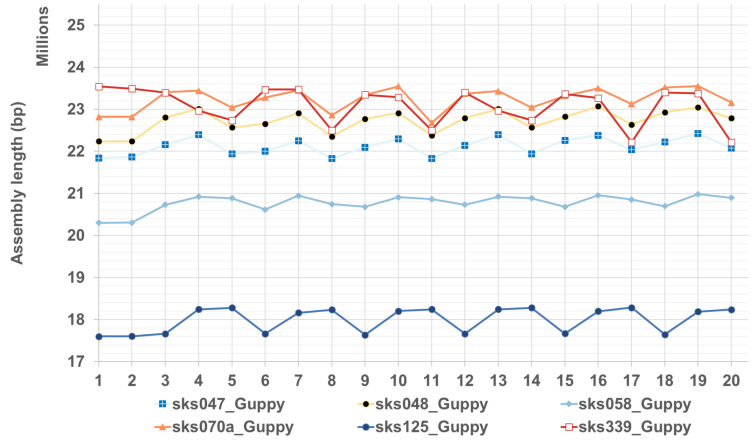
failed *de novo* assembly had less than the required 10-fold coverage depth, which Canu required to proceed with the assembly phase. Apart from sks048 in the Guppy basecalled and demultiplexed data, the coverage depth is directly proportional to the length of the

assembly generated. The proportional pattern is also evident in the N50 values, while the inverse is observed in the number of contigs the assemblies resolve.

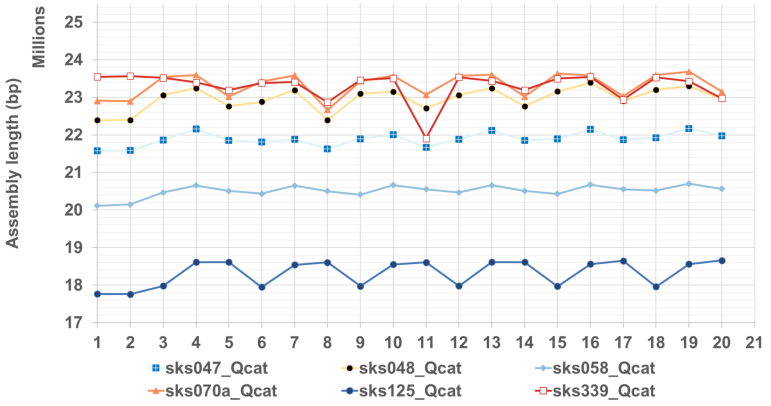
However, the use of the '*corOutCoverage*'parameter allows for an improvement in the successful assemblies (*henceforth maxCanu*); though no previously failed isolates successfully assembled here [Table 3.6]. Apart from sks125 of the Guppy dataset and sks047 of the Qcat dataset, improvements are observed on the total assembly lengths after using the entire reads sequence input, with the largest improvement seen in sks070a and sks339 for both Guppy and Qcat [Table 3.6]. Unlike the remaining isolates, sks058 and sks339 report decreased contigs while increasing their total length, suggesting that the contigs of the maxCanu assemblies were longer than the default Canu assemblies.

The parameters chosen for Redbean affect Redbean's ability to generate a complete assembly: For Redbean, assemblies were generated for all parameters described in Appendix Table D.3 apart from assemblies with the *k*-mer size of 24, which failed with unsolvable errors. Where successful, the length of assembled genomes for both Guppy and Qcat follow a similar, consistent pattern of direct proportionality between the coverage value and final assembled length [Figure 3.17(a), 3.17(b)]. It is evident that changing the parameters affect the outcome of the genome assembly process. As such, iterations 5, 8, 11, 14, 17 and 20, which all involve changing the parameters for subsampling (-AS) and minimum read depth (-e), produce the shortest assemblies using Redbean [Figure 3.17(a)]. Additionally, isolates with input coverage $<$30x (sks058 and sks125), showed a pattern where changing the *k*-mer parameter results in assemblies of shorter lengths [Figure 3.17(a), 3.17(b)].

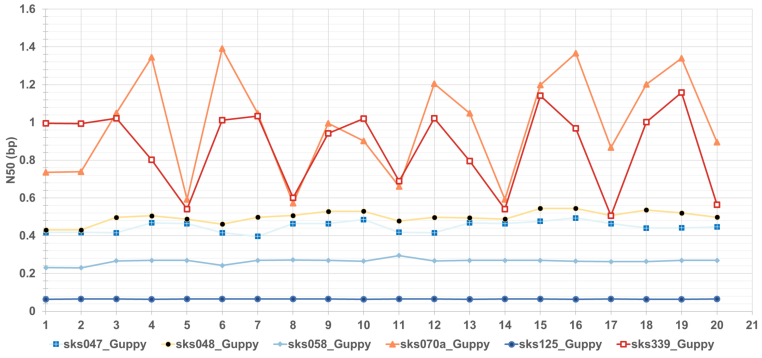
Both patterns were observed in the N50 for Guppy and Qcat; indicating that the contigs assembled for sks058 and sks125 in iterations 6, 9, 12, 15 and 18 showed comparably shorter sequences than the assemblies generated for the same isolates of different parameter iterations [Figure 3.17(c), 3.17(d)]. This was further observed in Figure 3.18(a) and 3.18(b) which show that sks058 and sks125 had the highest number of contigs in the generated assemblies. BUSCO comparisons showed very low completeness score percentages compared to both PKNH and PKNOH reference genome [Table 3.5, Figure 3.18(c),3.18(d)]. Between the two demultiplexing approaches, Guppy appeared to result in better results for sks047 and sks048. However, Qcat performed better for the



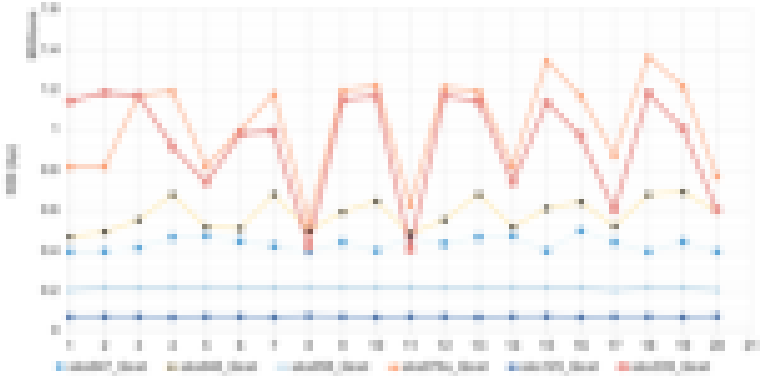
(a) Total Assembly Length – Guppy



(b) Total Assembly Length – Qcat

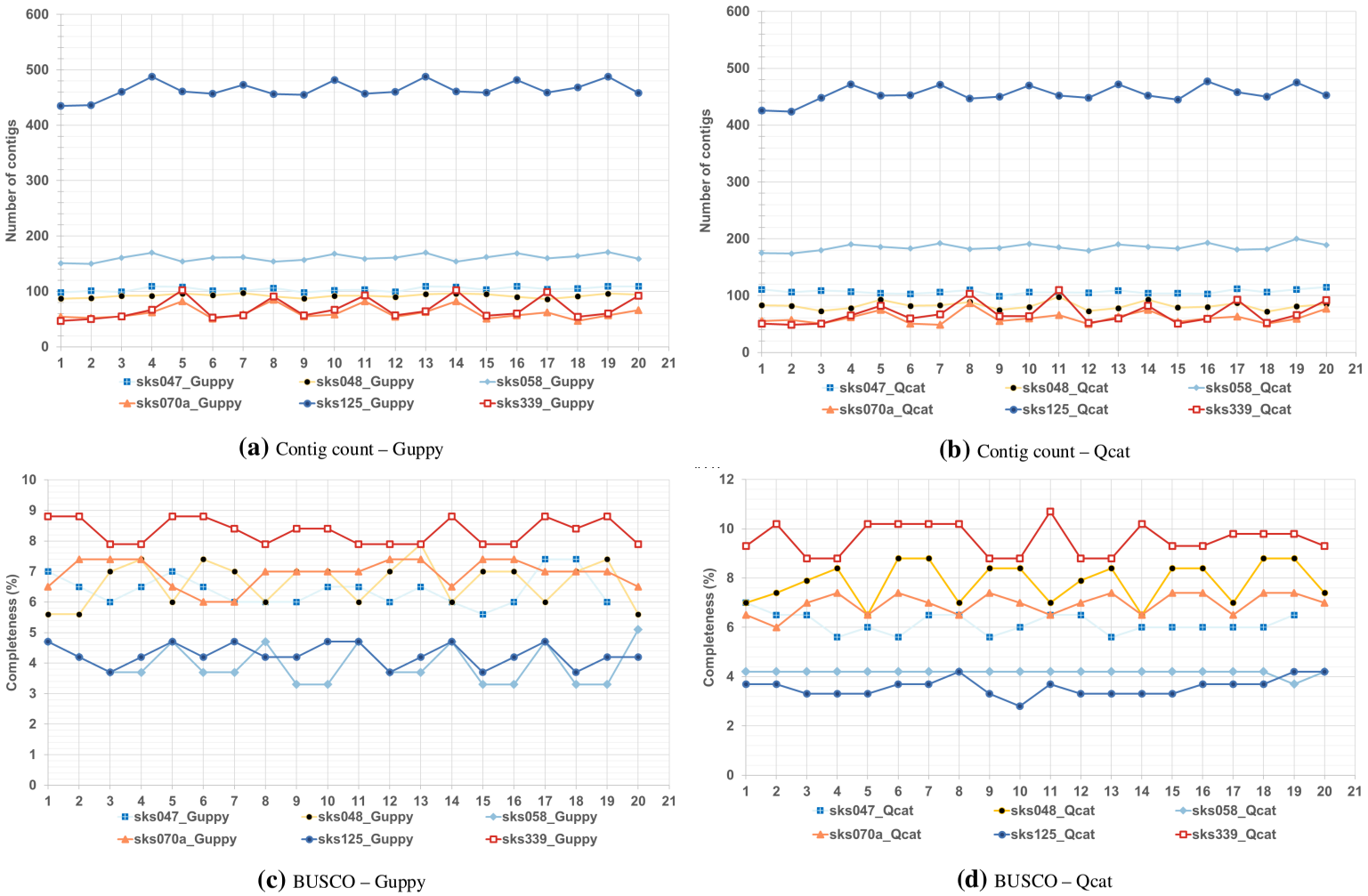


(c) N50 – Guppy



(d) N50 – Qcat

Figure 3.17: Assembly metrics for the assemblies generated for the parameters search for Redbean as described in Appendix Table D.3. Total assembly length and N50 values in base pair for each method used to prepare the reads for *de novo* assembly. *X*-axis represents the iteration numbers detailed in Appendix Table D.3. An estimated genome size of 25 Mb was used and each isolate uses the same colour in each plot.



(a) Contig count – Guppy

(b) Contig count – Qcat

(c) BUSCO – Guppy

(d) BUSCO – Qcat

Figure 3.18: Number of contigs and BUSCO scores for the assemblies generated for the Redbean parameter search as described in Appendix Table D.3. *X*-axis represents the iteration numbers detailed in Appendix Table D.3. An estimated genome size of 25 Mb was used and each isolate uses the same colour in Figure 3.17.

Table 3.7: Parameters concluded to produce the most suitable *de novo* genome assembly for the input isolates using Redbean

Isolate	Itr. Num.	Iteration condition name	Parameters
sks047	13	MIN2000.ctg.pdefault_e2_L2000	-p 21 -e 2 -l 2000
sks048	16	MIN2000.ctg.p22_e2_L2000	-p 22 -e 2 -l 2000
sks058	16	MIN2000.ctg.p22_e2_L2000	-p 22 -e 2 -l 2000
sks070a	15	MIN2000.ctg.p22_L2000	-p 22 -l 2000
sks125	20	MIN2000.ctg.p23_e2_AS2_L2000	-p 23 -e 2 -AS2 -l 2000
sks339	2	MIN2000.ctg.preset_L2000	-p 21 -k 0 -AS 4 -K 0.05 -s 0.5

For each isolate, generated assemblies are ranked in a descending hierarchy of BUSCO completeness, Total assembly length, Number of contigs and N50.
Itr. Num. - Iteration number

other isolates as assemblies generated from Qcat demultiplexed reads were longer than those generated using Guppy [Figure 3.17] while possessing similar or better BUSCO scores [Figure 3.18]. Using the stated criteria [Appendix Table D.3], parameters which produced the best assembly for each isolate showed an ideal *k*-mer size between 21 and 23 [Table 3.7]. Although each isolate had specific parameters which resulted in what was deemed as its 'best' assembly, a normalisation of the parameters of all isolates concluded in Redbean being run with the parameters "`-p 22 -l 2000`", and all other parameters left at their default values.

Flye assembles more isolates than Guppy and Redbean: Based on outputs from the assessments of demultiplexers and the Redbean parameter scan, only the Qcat demultiplexed reads were utilised for Flye *de novo* genome assembly. From this, Flye was able to assemble nine of the 18 isolates within the dataset [Table 3.8]. Once again, isolates that were unable to successfully assemble possessed very little input coverage to construct a draft genome assembly. Additionally, as previously observed, the more input coverage an isolate possessed, the longer the assembly. However, there appeared to be a threshold for how much coverage aided the assembly length. This was primarily evident in the assembly for sks048, which had the longest assembly length even though sks339 had $>$3-fold the input read coverage in comparison to sks048 [Table 3.8]. However, it is also observed that Flye was able to assemble three isolates that Canu and Redbean were unable to [Table 3.9].

Table 3.8: Descriptive statistics for *de novo* assemblies generated using Flye

Isolates	Coverage[x]	Total Length (bp)	Contigs	Max Contig (bp)	N50	Average Contig Length (bp)
sks047	32.85x	23119985	98	1852436	436014	235918.21
sks048	40.71x	24550273	81	1630708	556822	303089.79
sks058	21.16x	21959632	149	915685	273545	147380.08
sks070	52.56x	24415980	49	2631089	1241069	498285.31
sks074	9.66x	21924599	160	781734	245815	137028.74
sks078	7.48x	17540511	333	285633	80474	45499.54
sks125	12.89x	20252385	285	371997	111045	71061
sks333	9.09x	19058403	247	400799	110841	90842.68
sks339	141.11x	24117193	52	1855904	732744	463792.17

Nine of 18 isolates successfully assembled and subsequent assembly metrics were calculated using assembly-stats. Omitted isolates which failed de novo assembly are: *sks071*, *sks231*, *sks254*, *sks280*, *sks330*, *sks331*, *sks343*, *sks344*. All failed due to lack of coverage depth

x - Coverage calculated based on the 24 Mb estimate genome length used for Canu.

Table 3.9: Numerical comparisons of the *de novo* assemblers generated by Canu(maxCanu), Redbean and Flye genome assemblers

Isolates	Length			Contigs			Max			N50		
	Canu	Redbean	Flye	Canu	Redbean	Flye	Canu	Redbean	Flye	Canu	Redbean	Flye
sks047	21900894	21893857	23119985	180	104	98	1155389	1605769	1852436	309965	398205	436014
sks048	23032387	23155127	24550273	132	79	81	1658947	2376869	1630708	464190	606316	556822
sks058	20504144	20427528	21959632	282	183	149	710032	714009	915685	199147	214822	273545
sks070	23500482	23633198	24415980	83	54	49	1787230	2650389	2631089	785941	1343808	1241069
sks074	N.A	N.A	21924599	N.A	N.A	160	N.A	N.A	781734	N.A	N.A	245815
sks078	N.A	N.A	17540511	N.A	N.A	333	N.A	N.A	285633	N.A	N.A	80474
sks125	19155919	17969104	20252385	462	445	285	381369	323174	371997	95523	70100	111045
sks333	N.A	N.A	19058403	N.A	N.A	247	N.A	N.A	400799	N.A	N.A	110841
sks339	23973011	23500876	24117193	60	51	52	2602625	3130135	1855904	1152924	1133927	732744

Assembly metrics were calculated using assembly-stats.

N.A refers to isolates which failed de novo assembly in certain assemblers while being successfully assembled in other assemblers

Coverage has the greatest impact on the output of *de novo* assemblers: Using a Shapiro-Wilk test for normality, the coverage was not normally distributed for the successful isolates across the three assemblers [Table 3.10]. On the other hand, the

Table 3.10: Shapiro-Wilk normality test outputs for the input coverage, assembly length and number of contigs of all draft assemblies produced by Canu, Flye and Redbean

***de novo* Assembler**	**Shapiro-Wilk Test for normality (*p*-value)**			
	Assembly Length	***Cov.***	***Contigs***	***Distribution***
Canu	0.576	0.035	0.333	Assembly lengths and contigs normal. Coverage is not normal
Flye	0.373	0.002	0.302	Assembly lengths and contigs normal. Coverage is not normal
Redbean	0.205	0.035	0.088	Assembly lengths and contigs normal. Coverage is not normal

The test was carried out using Canu (*n=6*), Flye (*n=9*) and Redbean (*n=6*) successful assemblies at a 95 % confidence interval using the '*shapiro.test*' function in R. Coverage values are described in Table 3.8 while assembly length and contigs are presented in Table 3.9. *Cov.* - Coverage

assembly length and the number of contigs resolved by each assembly were normally distributed, with the p-values$>$0.05 for all three assemblers [Table 3.10]. Due to the coverage being determined by the extracted DNA and subsequent sequencing, the non-parametric Spearman's rank correlation assessed the relationship between the coverage and eventual draft assembly length and number of contigs. From this, all three assemblers showed a significant, strong correlation between the coverage and the lengths of generated draft assemblies [Table 3.11]. Conversely, a significant, negative correlation was observed between the coverage and the number of contigs each draft assembly resolves [Table 3.11].

Canu reported a smaller and presumably more precise range for the assemblies generated while Flye showed the largest range [Figure 3.19]. The interquartile ranges for all three assemblers overlap, with their medians seemingly close, indicating that the true median of the dataset does not differ and is only affected by the action of the assemblers, though not significantly [Figure 3.19]. This was supported using a Kruskal-Wallis One-way

Table 3.11: Spearman's rank correlation test to determine relationship between the input coverage, assembly length and the number of contigs resolved for Canu, Flye and Redbean

Conditions	Spearman's rank correlation						
	Canu		Flye		Redbean		Correlation
	p-value	*rho*	*p-value*	*rho*	*p-value*	*rho*	
Cov. vs Assembly	0.003	1.00	0.001	0.917	0.017	0.943	Significant positive
Cov. vs Contigs	0.003	-1.00	0.001	-0.933	0.017	-0.943	Significant negative

A correlation test of the assembly length and the number of contigs resolved for each successful draft *de novo* assembly for Canu (*n=6*), Flye (*n=9*) and Redbean (*n=6*) against the input read coverage for the corresponding isolates. Coverage was calculated as the total bp length of the reads against the total bp length of the PKNH reference genome (24 395 979 bp).
Cov. - Coverage

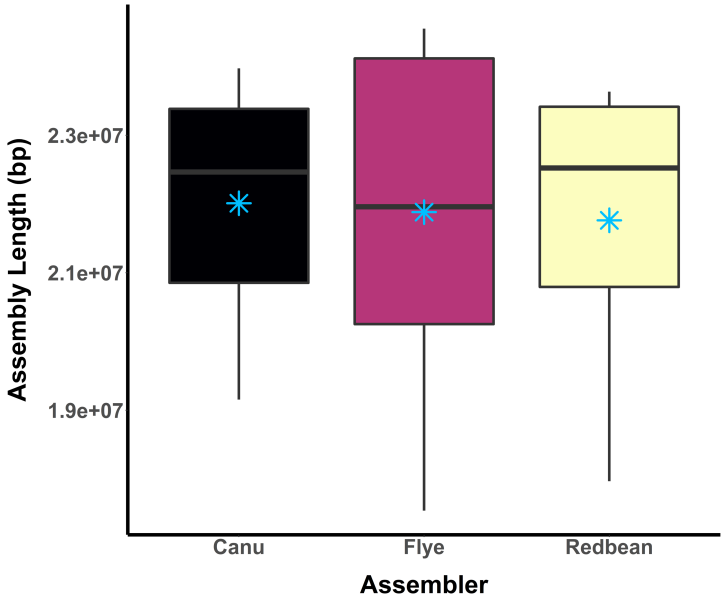


Figure 3.19: Median assembly length for all isolates generated using the three *de novo* assembly tools, Canu, Flye and Redbean. The blue star represents the mean value for each dataset. The sample sizes used for each dataset can be found in the Table 3.9.

Analysis of Variance (ANOVA) where the assemblers were reported to have no significant effect on the assembly length ($p=0.9563$). Furthermore, the assemblers had no significant effect on the number of contigs resolved ($p=0.4957$). However, in successful assemblies, Flye produced longer assemblies with fewer contigs while Redbean had assemblies with larger N50s, although this was still similar to Flye N50 values [Table 3.9].

3.6 Discussion

The impact of replacing the Albacore basecaller with the updated Guppy basecaller was assessed using an infected patient sequence dataset (March 2019 dataset). This assessment was expanded to include an evaluation of the demultiplexing capabilities of Albacore, Guppy and a dedicated demultiplexer – Qcat. Furthermore, the 'High-accuracy (HAC)' basecalling and demultiplexing algorithms of Guppy were also assessed. Guppy was shown to produce basecalled reads of higher quality, yield, and accuracy than Albacore. Qcat resulted in the highest demultiplexing yields whilst maintaining similar read quality to the Guppy demultiplexing module.

The consistency present in the four barcodes across the plots [Figure 3.8] indicates that the differences observed are due to the algorithms employed by the basecalling and demultiplexing tools. This is most evident in the total base pair yield length between Albacore and Guppy and subsequent improvements on the Guppy algorithm. Indeed, Albacore yields less total base-pair length and fewer reads than Guppy. The lower read count and total base-pair length reported by Guppy in comparison to Qcat suggest that the *guppy_barcoder* module is more conservative in its demultiplexing threshold than Qcat. This higher threshold in Guppy would also contribute to the slightly higher average quality score for Guppy observed in Figure 3.8(c). However, the differences between Guppy and Qcat appears to have been reduced in Guppy (*v3.4.1*) and Qcat (*v1.1.0*) outputs.

While Qcat reports more reads resolved into the distinct barcodes (BC01, BC02, BC03, BC04), the quality of the reads persist. This suggests that although appearing to be less stringent, Qcat can demultiplex more reads of comparable quality to Guppy. It is within this metric that the shortcomings of Albacore are most evident, as the increase in mean

quality score between Albacore and Guppy – and by extension, Qcat; shows a notable increase in the accuracy of RNN models designed by ONT. Albacore's average quality scores were ~Q12; representing ~92 % basecall accuracy while the ~Q20 in Guppy and Qcat increases basecall accuracy to 99 %. This shows simply changing the algorithm increased the basecall accuracy by an order of magnitude for the same dataset; though, the ideal quality score of Q40 or 99.99 % basecall accuracy is not possible with the algorithm models used here.

Thus, with incremental improvements on the basecalling algorithms utilised to translate raw signals captured by the MinION, basecall sequence quality can be improved. This means, where raw FAST5 reads are kept, they can be re-basecalled and re-investigated with the potential of increasing the quality of the basecalls. Such an improvement would increase confidence in downstream analyses and outputs observed. The removal of Albacore appears to be warranted and with benefit to the results downstream. While Qcat appears to allow for higher yields with comparable quality, it must be noted that Qcat's decreased threshold of 60 % match identity to the barcode library is a likely influence for this increased yield. As of writing, ONT have discontinued Qcat with a recommendation to use the *guppy_barcoder* module of Guppy.

The PKNH reference genome [83] was assumed to have no human contamination present due to no previous evidence indicating otherwise. However, the presence of contamination as indicated by BlobTools using parasite-aligned reads prompted further investigation. Confirmation of contamination in the PKNH reference genome is further complicated by recent realisations that the *P. knowlesi* H strain used to generate the PKNH reference genome had been inadvertently mislabelled [92]. Indeed, Assefa et al. [92] show evidence that the mislabelled H strain is instead the Malayan strain which had been extracted from an *Anopheles hackeri* mosquito in 1960 [84, 92]. On the other hand, Butcher and Mitchell [93] via personal communications with Collins et al. [94] now recognise that the Pk1(A+) A5 clone utilised for the PKNH reference genome [83] came from a Malaysian isolate of *P. knowlesi*. It is unknown if this Malaysian isolate is also the Malayan strain described by Lapp et al. [84] and Assefa et al. [92]. Nevertheless, the true H strain/experimental line, the Malaysian strain/experimental line and the Malayan strain/experimental line have been propagated in laboratory culture lines for over four decades. Hence it is unlikely for any human DNA to be present in these cultures prior to

DNA extraction and sequencing.

The alignment of suspected contigs against the larger and more comprehensive NCBI nucleotide *nt* database removed any bias which may be introduced by BlobTools' '*sum of best hits*' algorithm. No evidence of bias by BlobTools was observed in the outputs, and upon removal of these contigs, the assemblies did not possess the previously detected 'Chordata' contamination. However, with the contaminated contigs only having small yet strong alignments to the archived PKNH reference genome in the nt database, speculation arises as to the legitimacy of the contamination within the reference genome. It is currently unknown and admittedly difficult to prove if these regions within the PKNH reference genome show evidence of true *P. knowlesi* sequence or indeed spurious erroneous data that may be potentially transferred to other *P. knowlesi* genomes generated using the PKNH reference genome as a basis.

This is especially the case due to no hits of significant length or similarity being found by aligning the PKNH reference genome to the Human GRCh38.p12 reference genome. The inconsistency between the BlobTools results and the alignment results of the PKNH to the human reference could be due to an inappropriate method of testing or perhaps, due to the small proportion of the contamination, it is 'masked' within the larger genome. The extent of human contamination within the PKNH reference genome could not be quantified, though it could be speculated as a small proportion of the overall genome length. However, it is likely localised to the human mitochondrion, evidenced by the generated assemblies described here, possessing fragments corresponding to the human mitochondrion. Currently, the lack of the reads used to generate the PKNH reference genome limits this quantification. However, if the reads become available, a similar BlobTools assessment should elucidate this further.

The pipeline described above for the quantification and subsequent removal of contamination is effective. However, determining the presence and source of human contamination was important for the development of the study. With the PKNH reference genome used as the 'gold-standard' *P. knowlesi* reference genome, the presence of any contamination could be significant downstream. It must be noted that upon using the PKNOH reference genome as a basis, human contamination was not present. However, the human contamination was reintroduced within assemblies generated from the Merged

dataset of PKNH and PKNOH reference genomes. With this, it is clear that using a reference-guided assembly, though providing some accuracy and precision, may result in further propagation of unknown errors and sequencing mistakes from the reference source. Additionally, both the PKNH and PKNOH reference genomes were generated from experimental lines, which have been removed from natural selective pressure, thus may have introduced or indeed removed portions of the genome over time which the parasite organism does not require. Thus, it would be prudent for further genome assembly to be carried out using a *de novo* approach, hence reducing the introduction of previous biases and errors.

The output of the *de novo* assemblies suggest that Canu should be run using the '*corOutCoverage*'. While large input coverage –typically >100x— can be detrimental, this does not seem to be evident in this dataset. This can be observed in sks339, which has >120x input coverage. Upon assembly, both Guppy and Qcat report sks339 being resolved into the smallest number of contigs. The reduction in contig numbers as coverage increases is likely due to the assembler having access to relatively shorter reads that fill the gaps that would otherwise be left unfilled in assemblies generated using default Canu parameters.

On the other hand, Redbean's speedy approach to *de novo* assembly is achieved both by the algorithm used and by further parameter changes. In this study, the use of the 'subsampling' parameter resulted in $1/(subsamplingvalue)$ of reads being indexed. While this decreases the processing time for Redbean, it also reduces the match length of the contig being formed, thus reducing the length of the assembly generated. It is unknown why assemblies of *k*-mer size 24 did not successfully assemble, and due to this failure, no *k*-mers above 24 were attempted. Thus changing the *k*-mer size to >25 may have yielded more successful assemblies.

Once Qcat had been determined to be the demultiplexer to be carried forward, Flye was only run using the Qcat-demultiplexed reads. Impressively, Flye was able to assemble isolate sequences of meagre coverage of <10-fold, which is even less than those reported by Wick & Holt [55]. With this, isolates that would usually be unable to be assembled due to small coverage were successfully assembled, potentially facilitating the investigation of targeted regions of the genome that would be successfully assembled using Flye.

Although Wick & Holt [55] also report Redbean being able to assemble input coverage sequence data of 10-fold, this was not evident in this study. It is unknown if this is due to an unforeseen incompatibility with the input data or an upgrade to the features and capabilities of Redbean from *v2.3* to *v2.5* after this assessment was carried out. On the other hand, Flye produced longer assemblies with fewer contigs, thus reaching the set criteria for further downstream analyses. The effect of each assembler on the same data can be observed, evidenced in the interquartile range of the data overlapping in the boxplot [Figure 3.19]. As such, the actual mean and median of the data are likely within the overlapped range of the assemblies. However, Redbean appears to perform poorly due to the variability present in the assembly lengths produced; whilst possessing the same number of isolates successfully assembled with Canu [Table 3.9, Figure 3.19]. With Redbean having a more extensive range, particularly a larger bottom whisker than Canu, Redbean can be interpreted to be less precise than Canu with the same dataset.

Conversely, Flye assembled more isolates, including three with lower input coverage and thus lower assembly length due to the correlative and statistically significant relationship between coverage and assembly length. Indeed, this feature within Flye is responsible for the considerable variance in the assembly lengths reported in the Flye dataset. Thus, although Flye shows such variance, with its advantage of lower input coverage assembly, it is preferred to Canu. Furthermore, while Canu is likely to generate assemblies of consistent length, it is also prohibitively slow and processor intensive. On the other hand, Redbean is considerably faster in processing; however, it cannot assemble low-input coverage.

As such, no assembler showed significant variance on these output metrics to produce a statistical basis to proceed with a particular tool. Hence Flye is the most appropriate *de novo* assembly choice for this study. Flye strikes a balance between speed, accuracy, and usability whilst generating assemblies of longer contigs. While Flye's raw assembly may not be as accurate as Canu's, as with all forms of *de novo* long read assemblies, the raw assembly accuracy and quality can be increased with appropriate polishing, correcting and subsequent consensus.

3.7 References

[1] **J. D. WATSON** and **F. H. C. CRICK**. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* **171**:4356 (Apr. 1953), 737–738. DOI: 10.1038/171737a0 (see p. 92)

[2] **J. M. HEATHER** and **B. CHAIN**. "The Sequence of Sequencers: The History of Sequencing DNA". In: *Genomics* **107**:1 (Jan. 2016), 1–8. DOI: 10.1016/j.ygeno.2015.11.003 (see pp. 92–96)

[3] **D. T. ZALLEN**. "Despite Franklin's Work, Wilkins Earned His Nobel". In: *Nature* **425**:6953 (Sept. 2003), 15–15. DOI: 10.1038/425015b (see p. 92)

[4] **F. SANGER**, **G. G. BROWNLEE**, and **B. G. BARRELL**. "A Two-Dimensional Fractionation Procedure for Radioactive Nucleotides". In: *Journal of Molecular Biology* **13**:2 (Sept. 1965), 373–IN4. DOI: 10.1016/S0022-2836(65)80104-8 (see p. 92)

[5] **F. SANGER** and **A. R. COULSON**. "A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase". In: *Journal of Molecular Biology* **94**:3 (May 1975), 441–448. DOI: 10.1016/0022-2836(75)90213-2 (see p. 92)

[6] **A. M. MAXAM** and **W. GILBERT**. "A New Method for Sequencing DNA." In: *Proceedings of the National Academy of Sciences of the United States of America* **74**:2 (Feb. 1977), 560–564 (see pp. 92, 93)

[7] **C. A. HUTCHISON III**. "DNA Sequencing: Bench to Bedside and beyond †". In: *Nucleic Acids Research* **35**:18 (Sept. 2007), 6227–6237. DOI: 10.1093/nar/gkm688 (see pp. 92, 93)

[8] **B. E. SLATKO**, **A. F. GARDNER**, and **F. M. AUSUBEL**. "Overview of Next Generation Sequencing Technologies". In: *Current protocols in molecular biology* **122**:1 (Apr. 2018), e59. DOI: 10.1002/cpmb.59 (see pp. 93–96)

[9] **F. SANGER**, **G. M. AIR**, **B. G. BARRELL**, **N. L. BROWN**, **A. R. COULSON**, **J. C. FIDDES**, **C. A. HUTCHISON**, **P. M. SLOCOMBE**, and **M. SMITH**.

"Nucleotide Sequence of Bacteriophage φX174 DNA". In: *Nature* **265**:5596 (Feb. 1977), 687–695. DOI: 10.1038/265687a0 (see pp. 93, 94)

[10] **M. L. METZKER**. "Sequencing Technologies — the next Generation". In: *Nature Reviews Genetics* **11**:1 (Jan. 2010), 31–46. DOI: 10.1038/nrg2626 (see p. 94)

[11] **ILLUMINA**. *Illumina Sequencing by Synthesis.* https://www.youtube.com/watch?v=fCd6B5HRaZ Informational. Oct. 2016 (see p. 94)

[12] **PACBIO**. *Introduction to PacBio Highly Accurate Long-Read Sequencing.* https://www.youtube.com/watch?v=lT3NqhKD840. Informational. Feb. 2020 (see p. 95)

[13] **M. JAIN**, **H. E. OLSEN**, **B. PATEN**, and **M. AKESON**. "The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community". In: *Genome Biology* **17**: (Nov. 2016), 239. DOI: 10.1186/s13059-016-1103-0 (see pp. 96–99)

[14] **S. AGAH**, **M. ZHENG**, **M. PASQUALI**, and **A. B. KOLOMEISKY**. "DNA Sequencing by Nanopores: Advances and Challenges". In: *Journal of Physics D: Applied Physics* **49**:41 (Sept. 2016), 413001. DOI: 10.1088/0022-3727/49/41/413001 (see p. 96)

[15] **OXFORD NANOPORE TECHNOLOGIES**. *Types of Nanopores*. https://nanoporetech.com/how-it-works/types-of-nanopores. 2017 (see p. 96)

[16] **OXFORD NANOPORE TECHNOLOGIES**. *How It Works*. https://nanoporetech.com/how-it-works. 2017 (see pp. 96–98)

[17] **M. JAIN**, **I. FIDDES**, **K. H. MIGA**, **H. E. OLSEN**, **B. PATEN**, and **M. AKESON**. "Improved Data Analysis for the MinION Nanopore Sequencer". In: *Nature methods* **12**:4 (Apr. 2015), 351–356. DOI: 10.1038/nmeth.3290 (see p. 97)

[18] **OXFORD NANOPORE TECHNOLOGIES**. *Product Comparison*. http://nanoporetech.com/produc Mar. 2021 (see p. 98)

[19] **S. Stahl-Rommel**, **M. Jain**, **H. N. Nguyen**, **R. R. Arnold**, **S. M. Aunon-Chancellor**, **G. M. Sharp**, **C. L. Castro**, **K. K. John**, **S. Juul**, **D. J. Turner**, **D. Stoddart**, **B. Paten**, **M. Akeson**, **A. S. Burton**, and **S. L. Castro-Wallace**. "Real-Time Culture-Independent Microbial Profiling Onboard the International Space Station Using Nanopore Sequencing". In: *Genes* **12**:1 (Jan. 2021), 106. DOI: 10.3390/genes12010106 (see p. 98)

[20] **Department of Health and Social Care**. *Rapid Evaluation of Oxford Nanopore Technologies' LamPORE Assay*. Research and Analysis. Whitehall: Department of Health and Social Care, Jan. 2021 (see p. 98)

[21] **J. Quick**, **N. J. Loman**, **S. Duraffour**, **J. T. Simpson**, **E. Severi**, **L. Cowley**, **J. A. Bore**, **R. Koundouno**, **G. Dudas**, **A. Mikhail**, **N. Ouédraogo**, **B. Afrough**, **A. Bah**, **J. H. J. Baum**, **B. Becker-Ziaja**, **J. P. Boettcher**, **M. Cabeza-Cabrerizo**, **A. Camino-Sánchez**, **L. L. Carter**, **J. Doerrbecker**, **T. Enkirch**, **I. G. Dorival**, **N. Hetzelt**, **J. Hinzmann**, **T. Holm**, **L. E. Kafetzopoulou**, **M. Koropogui**, **A. Kosgey**, **E. Kuisma**, **C. H. Logue**, **A. Mazzarelli**, **S. Meisel**, **M. Mertens**, **J. Michel**, **D. Ngabo**, **K. Nitzsche**, **E. Pallasch**, **L. V. Patrono**, **J. Portmann**, **J. G. Repits**, **N. Y. Rickett**, **A. Sachse**, **K. Singethan**, **I. Vitoriano**, **R. L. Yemanaberhan**, **E. G. Zekeng**, **T. Racine**, **A. Bello**, **A. A. Sall**, **O. Faye**, **O. Faye**, **N. Magassouba**, **C. V. Williams**, **V. Amburgey**, **L. Winona**, **E. Davis**, **J. Gerlach**, **F. Washington**, **V. Monteil**, **M. Jourdain**, **M. Bererd**, **A. Camara**, **H. Somlare**, **A. Camara**, **M. Gerard**, **G. Bado**, **B. Baillet**, **D. Delaune**, **K. Y. Nebie**, **A. Diarra**, **Y. Savane**, **R. B. Pallawo**, **G. J. Gutierrez**, **N. Milhano**, **I. Roger**, **C. J. Williams**, **F. Yattara**, **K. Lewandowski**, **J. Taylor**, **P. Rachwal**, **D. J. Turner**, **G. Pollakis**, **J. A. Hiscox**, **D. A. Matthews**, **M. K. O. Shea**, **A. M. Johnston**, **D. Wilson**, **E. Hutley**, **E. Smit**, **A. D. Caro**, **R. Wölfel**, **K. Stoecker**, **E. Fleischmann**, **M. Gabriel**, **S. A. Weller**, **L. Koivogui**, **B. Diallo**, **S. Keïta**, **A. Rambaut**, **P. Formenty**, **S. Günther**, and **M. W.**

CARROLL. "Real-Time, Portable Genome Sequencing for Ebola Surveillance". In: *Nature* **530**:7589 (Feb. 2016), 228. DOI: 10.1038/nature16996 (see p. 98)

[22] **N. J. LOMAN** and **A. R. QUINLAN**. "Poretools: A Toolkit for Analyzing Nanopore Sequence Data". In: *Bioinformatics* **30**:23 (Dec. 2014), 3399–3401. DOI: 10.1093/bioinformatics/btu555 (see pp. 99, 101)

[23] **O. N. TECHNOLOGIES**. *R10.3: The Newest Nanopore for High Accuracy Nanopore Sequencing*. http://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store. Feb. 2021 (see p. 99)

[24] **C. WRIGHT**. *Rebasecalling of SRE and ULK GM24385 Dataset*. https://labs.epi2me.io/gm24385 May 2021 (see p. 99)

[25] **S. H. NGUYEN**, **T. D. S. DUARTE**, **L. J. M. COIN**, and **M. D. CAO**. "Real-Time Demultiplexing Nanopore Barcoded Sequencing Data With npBarcode". In: *bioRxiv* (May 2017), -. DOI: 10.1101/134155 (see p. 100)

[26] **ROBERT DIPIETRO** and **GREGORY D. HAGER**. "Chapter 21 - Deep Learning: RNNs and LSTM". In: *Handbook of Medical Image Computing and Computer Assisted Intervention*. Ed. by **S. KEVIN ZHOU**, **DANIEL RUECKERT**, and **GABOR FICHTINGER**. The Elsevier and MICCAI Society Book Series. Academic Press, 2020, 503–519. DOI: 10.1016/B978-0-12-816176-0.00026-0 (see p. 100)

[27] **Q. LIU**, **D. C. GEORGIEVA**, **D. EGLI**, and **K. WANG**. "NanoMod: A Computational Tool to Detect DNA Modifications Using Nanopore Long-Read Sequencing Data". In: *BMC Genomics* **20**:1 (Feb. 2019), 78. DOI: 10.1186/s12864-018-5372-8 (see p. 100)

[28] **OXFORD NANOPORE TECHNOLOGIES**. *Albacore Basecalling Software*. 2017 (see p. 100)

[29] **H. CHANG**. *Seamlessf5: Helpers for Smoother Transitioning to Multi-Read FAST5 Files*. 2019 (see pp. 100, 119)

[30] **R. DOKOS**, **J. PUGH**, **O. KUZNETSOVA**, and **R. RONAN**. *Pre-Release of Stand Alone Guppy - Guppy v2.1.3*. https://community.nanoporetech.com/posts/pre-release-of-stand-alone. 2018 (see pp. 100, 101)

[31] **OXFORD NANOPORE TECHNOLOGIES**. *Guppy Package v4.0.15*. Oxford Nanopore Technologies. 2020 (see p. 100)

[32] **OXFORD NANOPORE TECHNOLOGIES**. *QCAT: A Python Command-Line Tool for Demultiplexing Oxford Nanopore Reads from FASTQ Files*. Oxford Nanopore Technologies. May 2019 (see p. 101)

[33] **O. SILANDER**. *RBK004 with Guppy_barcoder 2.3.7 vs. Qcat*. Original Post. Mar. 2019 (see p. 101)

[34] **R. ALCANTARA**. *Demultiplexer Comparison*. Report. Mar. 2019 (see p. 101)

[35] **A. QUINLAN**. *Poretools: A Toolkit for Working with Oxford Nanopore Data*. Jan. 2018 (see p. 101)

[36] **R. WICK**. *Porechop: Adapter Trimmer for Oxford Nanopore Reads*. Jan. 2018 (see pp. 101, 102, 137)

[37] **DECOSTER**. *Per Base Sequence Content and Quality (End of Reads)*. June 2017 (see p. 102)

[38] **W. DE COSTER**, **S. D'HERT**, **D. T. SCHULTZ**, **M. CRUTS**, and **C. VAN BROECKHOVEN**. "NanoPack: Visualizing and Processing Long-Read Sequencing Data". In: *Bioinformatics* (2018). DOI: 10.1093/bioinformatics/bty149 (see pp. 102, 103)

[39] **W. DECOSTER**. *nanoQC: Quality Control Tools for Nanopore Sequencing Data*. Jan. 2018 (see p. 102)

[40] **W. DECOSTER**. *NanoPlot: Plotting Scripts for Long Read Sequencing Data*. June 2018 (see p. 103)

[41] **A. R. QUINLAN** and **I. M. HALL**. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features". In: *Bioinformatics* **26**:6 (Mar. 2010), 841–842. DOI: 10.1093/bioinformatics/btq033 (see p. 103)

[42] **H. LI**, **B. HANDSAKER**, **A. WYSOKER**, **T. FENNELL**, **J. RUAN**, **N. HOMER**, **G. MARTH**, **G. ABECASIS**, and **R. DURBIN**. "The Sequence Alignment/Map Format and SAMtools". In: *Bioinformatics* **25**:16 (Aug. 2009), 2078–2079. DOI: 10.1093/bioinformatics/btp352 (see pp. 104, 128)

[43] **P. DANECEK**, **J. K. BONFIELD**, **J. LIDDLE**, **J. MARSHALL**, **V. OHAN**, **M. O. POLLARD**, **A. WHITWHAM**, **T. KEANE**, **S. A. MCCARTHY**, **R. M. DAVIES**, and **H. LI**. "Twelve Years of SAMtools and BCFtools". In: *GigaScience* **10**:giab008 (Feb. 2021). DOI: 10.1093/gigascience/giab008 (see p. 104)

[44] **G. MARÇAIS**, **A. L. DELCHER**, **A. M. PHILLIPPY**, **R. COSTON**, **S. L. SALZBERG**, and **A. ZIMIN**. "MUMmer4: A Fast and Versatile Genome Alignment System". In: *PLOS Computational Biology* **14**:1 (Jan. 2018), e1005944. DOI: 10.1371/journal.pcbi.1005944 (see p. 104)

[45] **H. LI**. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM". In: *arXiv:1303.3997 [q-bio]* (May 2013). arXiv: 1303.3997 [q-bio] (see p. 105)

[46] **H. LI**. "Minimap2: Pairwise Alignment for Nucleotide Sequences". In: *Bioinformatics* **34**:18 (Sept. 2018), 3094–3100. DOI: 10.1093/bioinformatics/bty191 (see pp. 105, 128)

[47] **Z. LI**, **Y. CHEN**, **D. MU**, **J. YUAN**, **Y. SHI**, **H. ZHANG**, **J. GAN**, **N. LI**, **X. HU**, **B. LIU**, **B. YANG**, and **W. FAN**. "Comparison of the Two Major Classes of Assembly Algorithms: Overlap–Layout–Consensus and de-Bruijn-Graph". In: *Briefings in Functional Genomics* **11**:1 (Jan. 2012), 25–37. DOI: 10.1093/bfgp/elr035 (see p. 106)

[48] **A. R. KHAN**, **M. T. PERVEZ**, **M. E. BABAR**, **N. NAVEED**, and **M. SHOAIB**. "A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective". In: *Evolutionary Bioinformatics Online* **14**: (Feb. 2018). DOI: 10.1177/1176934318758650 (see p. 106)

[49] **S. KOREN**, **B. P. WALENZ**, **K. BERLIN**, **J. R. MILLER**, **N. H. BERGMAN**, and **A. M. PHILLIPPY**. "Canu: Scalable and Accurate Long-read Assembly via Adaptive K-mer Weighting and Repeat Separation". In: *Genome Research* (Mar. 2017). DOI: 10.1101/gr.215087.116 (see p. 106)

[50] **J. R. MILLER**, **A. L. DELCHER**, **S. KOREN**, **E. VENTER**, **B. P. WALENZ**, **A. BROWNLEY**, **J. JOHNSON**, **K. LI**, **C. MOBARRY**, and **G. SUTTON**. "Aggressive Assembly of Pyrosequencing Reads with Mates". In: *Bioinformatics (Oxford,*

England) **24**:24 (Dec. 2008), 2818–2824. DOI: 10.1093/bioinformatics/btn548 (see p. 106)

[51] **E. W. MYERS**, **G. G. SUTTON**, **A. L. DELCHER**, **I. M. DEW**, **D. P. FASULO**, **M. J. FLANIGAN**, **S. A. KRAVITZ**, **C. M. MOBARRY**, **K. H. J. REINERT**, **K. A. REMINGTON**, **E. L. ANSON**, **R. A. BOLANOS**, **H.-H. CHOU**, **C. M. JORDAN**, **A. L. HALPERN**, **S. LONARDI**, **E. M. BEASLEY**, **R. C. BRANDON**, **L. CHEN**, **P. J. DUNN**, **Z. LAI**, **Y. LIANG**, **D. R. NUSSKERN**, **M. ZHAN**, **Q. ZHANG**, **X. ZHENG**, **G. M. RUBIN**, **M. D. ADAMS**, and **J. C. VENTER**. "A Whole-Genome Assembly of Drosophila". In: *Science* **287**:5461 (Mar. 2000), 2196–2204. DOI: 10.1126/science.287.5461.2196 (see p. 106)

[52] **A. PHILLIPPY**, **S. KOREN**, and **B. WALENZ**. *Canu Documentation*. Tech. rep. Maryland Bioinformatics Labs, Dec. 2018, 53 (see p. 106)

[53] **J. RUAN** and **H. LI**. *Fast and Accurate Long-Read Assembly with Wtdbg2*. Preprint. Bioinformatics, Jan. 2019. DOI: 10.1101/530972 (see pp. 107, 108)

[54] **J. SIMPSON**. *Nanopolish: Signal-Level Algorithms for MinION Data.* June 2019 (see p. 107)

[55] **R. R. WICK** and **K. E. HOLT**. "Benchmarking of Long-Read Assemblers for Prokaryote Whole Genome Sequencing". In: *F1000Research* **8**: (Feb. 2021), 2138. DOI: 10.12688/f1000research.21782.4 (see pp. 107, 149, 150)

[56] **M. KOLMOGOROV**, **J. YUAN**, **Y. LIN**, and **P. A. PEVZNER**. "Assembly of Long, Error-Prone Reads Using Repeat Graphs". In: *Nature Biotechnology* **37**:5 (May 2019), 540–546. DOI: 10.1038/s41587-019-0072-8 (see pp. 107–110)

[57] **M. KOLMOGOROV** and **J. YUAN**. *Flye with Mikhail Kolmogorov and Jeffrey Yuan - Episode 4*. Apr. 2019 (see p. 107)

[58] **D. R. LAETSCH** and **M. L. BLAXTER**. "BlobTools: Interrogation of Genome Assemblies". In: *F1000Research* **6**: (July 2017), 1287. DOI: 10.12688/f1000research.12232.1 (see pp. 109–111, 128)

[59] **A. GUREVICH**, **V. SAVELIEV**, **N. VYAHHI**, and **G. TESLER**. "QUAST: Quality Assessment Tool for Genome Assemblies". In: *Bioinformatics* **29**:8 (Apr. 2013), 1072–1075. DOI: 10.1093/bioinformatics/btt086 (see p. 111)

[60] **A. MIKHEENKO**, **A. PRJIBELSKI**, **V. SAVELIEV**, **D. ANTIPOV**, and **A. GUREVICH**. "Versatile Genome Assembly Evaluation with QUAST-LG". In: *Bioinformatics* **34**:13 (July 2018), i142–i150. DOI: 10.1093/bioinformatics/bty266 (see p. 111)

[61] **R. M. WATERHOUSE**, **M. SEPPEY**, **F. A. SIMÃO**, **M. MANNI**, **P. IOANNIDIS**, **G. KLIOUTCHNIKOV**, **E. V. KRIVENTSEVA**, and **E. M. ZDOBNOV**. "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics". In: *Molecular Biology and Evolution* **35**:3 (Mar. 2018), 543–548. DOI: 10.1093/molbev/msx319 (see pp. 111, 112, 137)

[62] **F. A. SIMÃO**, **R. M. WATERHOUSE**, **P. IOANNIDIS**, **E. V. KRIVENTSEVA**, and **E. M. ZDOBNOV**. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs". In: *Bioinformatics* **31**:19 (Oct. 2015), 3210–3212. DOI: 10.1093/bioinformatics/btv351 (see pp. 111, 112, 137)

[63] **R. VASER**, **I. SOVIĆ**, **N. NAGARAJAN**, and **M. ŠIKIĆ**. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads". In: *Genome Research* **27**:5 (May 2017), 737–746. DOI: 10.1101/gr.214270.116 (see p. 112)

[64] **H. LI**. "Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences". In: *Bioinformatics* **32**:14 (July 2016), 2103–2110. DOI: 10.1093/bioinformatics/btw152 (see p. 112)

[65] **OXFORD NANOPORE TECHNOLOGIES**. *Medaka: Consensus Sequence Tool for Nanopore Sequences*. Oxford Nanopore Technologies. May 2019 (see pp. 112, 113)

[66] **B. J. WALKER**, **T. ABEEL**, **T. SHEA**, **M. PRIEST**, **A. ABOUELLIEL**, **S. SAKTHIKUMAR**, **C. A. CUOMO**, **Q. ZENG**, **J. WORTMAN**, **S. K. YOUNG**, and **A. M. EARL**. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement". In: *PLoS ONE* **9**:11 (Nov. 2014). Ed. by **J. WANG**, e112963. DOI: 10.1371/journal.pone.0112963 (see p. 113)

[67] **M. TARAILO-GRAOVAC** and **N. CHEN**. "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences". In: *Current Protocols in Bioinformatics*. Ed. by **A. D. BAXEVANIS**, **G. A. PETSKO**, **L. D. STEIN**, and **G. D. STORMO**. Hoboken, NJ, USA: John Wiley & Sons, Inc., Mar. 2009, bi0410s25. DOI: 10.1002/0471250953.bi0410s25 (see p. 113)

[68] **M. BAILLY-BECHET**, **A. HAUDRY**, and **E. LERAT**. ""One Code to Find Them All": A Perl Tool to Conveniently Parse RepeatMasker Output Files". In: *Mobile DNA* **5**:1 (May 2014), 13. DOI: 10.1186/1759-8753-5-13 (see pp. 113, 114)

[69] **B. J. HAAS**. *TransposonPSI: An Application of PSI-Blast to Mine (Retro-)Transposon ORF Homologies*. http://transposonpsi.sourceforge.net/. 2010 (see p. 114)

[70] **M. ALONGE**, **S. SOYK**, **S. RAMAKRISHNAN**, **X. WANG**, **S. GOODWIN**, **F. J. SEDLAZECK**, **Z. B. LIPPMAN**, and **M. C. SCHATZ**. "RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes". In: *Genome Biology* **20**:1 (Oct. 2019), 224. DOI: 10.1186/s13059-019-1829-6 (see pp. 114, 115)

[71] **M. ALONGE**. *RagTag*. May 2021 (see p. 115)

[72] **S. STEINBISS**, **F. SILVA-FRANCO**, **B. BRUNK**, **B. FOTH**, **C. HERTZ-FOWLER**, **M. BERRIMAN**, and **T. D. OTTO**. "Companion: A Web Server for Annotation and Analysis of Parasite Genomes". In: *Nucleic Acids Research* **44**:Web Server issue (July 2016), W29–W34. DOI: 10.1093/nar/gkw292 (see pp. 115, 116)

[73] **T. CARVER**, **S. R. HARRIS**, **M. BERRIMAN**, **J. PARKHILL**, and **J. A. MCQUILLAN**. "Artemis: An Integrated Platform for Visualization and Analysis of High-Throughput Sequence-Based Experimental Data". In: *Bioinformatics* **28**:4 (Feb. 2012), 464–469. DOI: 10.1093/bioinformatics/btr703 (see p. 116)

[74] **K. RUTHERFORD** and **M. BERRIMAN**. "Viewing and Annotating Seqeunce Data with Artemis". In: *Briefings in Bioinformatics* **4**:2 (2003), 124–132 (see p. 116)

[75] **H. THORVALDSDÓTTIR**, **J. T. ROBINSON**, and **J. P. MESIROV**. "Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration". In: *Briefings in Bioinformatics* **14**:2 (Mar. 2013), 178–192. DOI: 10.1093/bib/bbs017 (see p. 117)

[76] **J. T. ROBINSON**, **H. THORVALDSDÓTTIR**, **A. M. WENGER**, **A. ZEHIR**, and **J. P. MESIROV**. "Variant Review with the Integrative Genomics Viewer". In: *Cancer Research* **77**:21 (Nov. 2017), e31–e34. DOI: 10.1158/0008-5472.CAN-17-0337 (see p. 117)

[77] **R. CHIKHI**. *De Novo Assembly & k -Mers*. Training. 2018 (see pp. 117, 118)

[78] **T. SEEMANN**. *De Novo Genome Assembly of NGS Data*. Training. Monash University, 2011 (see pp. 117, 118)

[79] **OXFORD NANOPORE TECHNOLOGIES**. *Rapid Barcoding Sequencing (SQK-RBK004) Protocol*. Oct. 2020 (see p. 119)

[80] **WELLCOME SANGER INSTITUTE**. *Assembly-Stats: Get Assembly Statistics from FASTA and FASTQ Files*. Pathogen Informatics, Wellcome Sanger Institute. June 2019 (see pp. 120, 137)

[81] **ILLUMINA**. *Quality Scores for Next-Generation Sequencing*. 2011 (see p. 122)

[82] **V. DOMINGUEZ DEL ANGEL**, **E. HJERDE**, **L. STERCK**, **S. CAPELLA-GUTIERREZ**, **C. NOTREDAME**, **O. VINNERE PETTERSSON**, **J. AMSELEM**, **L. BOURI**, **S. BOCS**, **C. KLOPP**, **J.-F. GIBRAT**, **A. VLASOVA**, **B. L. LESKOSEK**, **L. SOLER**, **M. BINZER-PANCHAL**, and **H. LANTZ**. "Ten Steps to Get Started in Genome Assembly and Annotation". In: *F1000Research* **7**: (Feb. 2018), 148. DOI: 10.12688/f1000research.13598.1 (see p. 127)

[83] **A. PAIN**, **U. BÖHME**, **A. E. BERRY**, **K. MUNGALL**, **R. D. FINN**, **A. P. JACKSON**, **T. MOURIER**, **J. MISTRY**, **E. M. PASINI**, **M. A. ASLETT**, **S. BALASUBRAMMANIAM**, **K. BORGWARDT**, **K. BROOKS**, **C. CARRET**, **T. J. CARVER**, **I. CHEREVACH**, **T. CHILLINGWORTH**, **T. G. CLARK**, **M. R. GALINSKI**, **N. HALL**, **D. HARPER**, **D. HARRIS**, **H. HAUSER**, **A. IVENS**, **C. S. JANSSEN**, **T. KEANE**, **N. LARKE**, **S. LAPP**, **M. MARTI**, **S. MOULE**, **I. M. MEYER**, **D. ORMOND**, **N. PETERS**, **M. SANDERS**, **S. SANDERS**, **T. J. SARGEANT**, **M. SIMMONDS**, **F. SMITH**, **R. SQUARES**, **S. THURSTON**,

A. R. TIVEY, **D. WALKER**, **B. WHITE**, **E. ZUIDERWIJK**, **C. CHURCHER**, **M. A. QUAIL**, **A. F. COWMAN**, **C. M. R. TURNER**, **M. A. RAJANDREAM**, **C. H. M. KOCKEN**, **A. W. THOMAS**, **C. I. NEWBOLD**, **B. G. BARRELL**, and **M. BERRIMAN**. "The Genome of the Simian and Human Malaria Parasite Plasmodium Knowlesi". In: *Nature* **455**:7214 (Oct. 2008), 799–803. DOI: 10.1038/nature07306 (see pp. 127–129, 136–138, 147)

[84] **S. A. LAPP**, **J. A. GERALDO**, **J.-T. CHIEN**, **F. AY**, **S. B. PAKALA**, **G. BATUGEDARA**, **J. HUMPHREY**, **THE MAHPIC CONSORTIUM**, **J. D. DEBARRY**, **K. G. LE ROCH**, **M. R. GALINSKI**, and **J. C. KISSINGER**. "PacBio Assembly of a Plasmodium Knowlesi Genome Sequence with Hi-C Correction and Manual Annotation of the SICAvar Gene Family". In: *Parasitology* (July 2017), 1–14. DOI: 10.1017/S0031182017001329 (see pp. 127, 128, 131, 136–138, 147)

[85] **D. R. ORESEGUN**, **C. DANESHVAR**, and **J. COX-SINGH**. "Plasmodium Knowlesi – Clinical Isolate Genome Sequencing to Inform Translational Same-Species Model System for Severe Malaria". In: *Frontiers in Cellular and Infection Microbiology* **11**: (2021). DOI: 10.3389/fcimb.2021.607686 (see p. 127)

[86] **BETHESDA (MD): NATIONAL LIBRARY OF MEDICINE**. *Nucleotide*. https://www.ncbi.nlm.nih.gov/nuccore. 1988 (see p. 128)

[87] **H. LI**. *Seqtk - Toolkit for Processing Sequences in FASTA/Q Formats*. 2012 (see pp. 128, 132)

[88] **C. CAMACHO**, **G. COULOURIS**, **V. AVAGYAN**, **N. MA**, **J. PAPADOPOULOS**, **K. BEALER**, and **T. L. MADDEN**. "BLAST+: Architecture and Applications". In: *BMC Bioinformatics* **10**: (Dec. 2009), 421. DOI: 10.1186/1471-2105-10-421 (see p. 128)

[89] **INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM**, **E. S. LANDER**, **L. M. LINTON**, **B. BIRREN**, **C. NUSBAUM**, **M. C. ZODY**, **J. BALDWIN**, **K. DEVON**, **K. DEWAR**, **M. DOYLE**, **W. FITZHUGH**, **R. FUNKE**, **D. GAGE**, **K. HARRIS**, **A. HEAFORD**, **J. HOWLAND**, **L. KANN**, **J. LEHOCZKY**, **R. LEVINE**, **P. MCEWAN**, **K. MCKERNAN**, **J. MELDRIM**, **J. P. MESIROV**, **C. MIRANDA**, **W. MORRIS**, **J. NAYLOR**, **C. RAYMOND**, **M. ROSETTI**,

R. SANTOS, A. SHERIDAN, C. SOUGNEZ, N. STANGE-THOMANN, N. STOJANOVIC, A. SUBRAMANIAN, D. WYMAN, J. ROGERS, J. SULSTON, R. AINSCOUGH, S. BECK, D. BENTLEY, J. BURTON, C. CLEE, N. CARTER, A. COULSON, R. DEADMAN, P. DELOUKAS, A. DUNHAM, I. DUNHAM, R. DURBIN, L. FRENCH, D. GRAFHAM, S. GREGORY, T. HUBBARD, S. HUMPHRAY, A. HUNT, M. JONES, C. LLOYD, A. MCMURRAY, L. MATTHEWS, S. MERCER, S. MILNE, J. C. MULLIKIN, A. MUNGALL, R. PLUMB, M. ROSS, R. SHOWNKEEN, S. SIMS, R. H. WATERSTON, R. K. WILSON, L. W. HILLIER, J. D. MCPHERSON, M. A. MARRA, E. R. MARDIS, L. A. FULTON, A. T. CHINWALLA, K. H. PEPIN, W. R. GISH, S. L. CHISSOE, M. C. WENDL, K. D. DELEHAUNTY, T. L. MINER, A. DELEHAUNTY, J. B. KRAMER, L. L. COOK, R. S. FULTON, D. L. JOHNSON, P. J. MINX, S. W. CLIFTON, T. HAWKINS, E. BRANSCOMB, P. PREDKI, P. RICHARDSON, S. WENNING, T. SLEZAK, N. DOGGETT, J.-F. CHENG, A. OLSEN, S. LUCAS, C. ELKIN, E. UBERBACHER, M. FRAZIER, R. A. GIBBS, D. M. MUZNY, S. E. SCHERER, J. B. BOUCK, E. J. SODERGREN, K. C. WORLEY, C. M. RIVES, J. H. GORRELL, M. L. METZKER, S. L. NAYLOR, R. S. KUCHERLAPATI, D. L. NELSON, G. M. WEINSTOCK, Y. SAKAKI, A. FUJIYAMA, M. HATTORI, T. YADA, A. TOYODA, T. ITOH, C. KAWAGOE, H. WATANABE, Y. TOTOKI, T. TAYLOR, J. WEISSENBACH, R. HEILIG, W. SAURIN, F. ARTIGUENAVE, P. BROTTIER, T. BRULS, E. PELLETIER, C. ROBERT, P. WINCKER, A. ROSENTHAL, M. PLATZER, G. NYAKATURA, S. TAUDIEN, A. RUMP, D. R. SMITH, L. DOUCETTE-STAMM, M. RUBENFIELD, K. WEINSTOCK, H. M. LEE, J. DUBOIS, H. YANG, J. YU, J. WANG, G. HUANG, J. GU, L. HOOD, L. ROWEN, A. MADAN, S. QIN, R. W. DAVIS, N. A. FEDERSPIEL, A. P. ABOLA, M. J. PROCTOR, B. A. ROE, F. CHEN, H. PAN, J. RAMSER, H. LEHRACH, R. REINHARDT, W. R. MCCOMBIE, M. DE LA BASTIDE, N. DEDHIA, H. BLÖCKER, K. HORNISCHER, G. NORDSIEK, R. AGARWALA, L. ARAVIND, J. A. BAILEY, A. BATEMAN, S. BATZOGLOU, E. BIRNEY, P. BORK, D. G. BROWN, C. B. BURGE, L. CERUTTI, H.-C. CHEN, D. CHURCH, M. CLAMP, R. R. COPLEY, T. DOERKS, S. R. EDDY, E. E. EICHLER, T. S. FUREY, J. GALAGAN, J. G. R. GILBERT, C. HARMON,

Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M. J. Morgan, C. f. G. R. Whitehead Institute for Biomedical Research, The Sanger Centre: Washington University Genome Sequencing Center, US DOE Joint Genome Institute: Baylor College of Medicine Human Genome Sequencing Center: RIKEN Genomic Sciences Center: Genoscope and CNRS UMR-8030: I. o. M. B. Department of Genome Analysis, GTC Sequencing Center: Beijing Genomics Institute/Human Genome Center: T. I. f. S. B. Multimegabase Sequencing Center, Stanford Genome Technology Center: University of Oklahoma's Advanced Center for Genome Technology: Max Planck Institute for Molecular Genetics: L. A. H. G. C. Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology: a. includes individuals listed under other headings): *Genome Analysis Group (listed in alphabetical order, U. N. I. o. H. Scientific management: National Human Genome Research Institute, Stanford Human Genome Center: University of Washington Genome Center: K. U. S. o. M. Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas: U. D. o. E. Office of Science, and The Wellcome Trust: "Initial Sequencing and Analysis of the Human Genome".

In: *Nature* **409**:6822 (Feb. 2001), 860–921. DOI: 10.1038/35057062 (see pp. 128, 136)

[90] **V. A. SCHNEIDER**, **T. GRAVES-LINDSAY**, **K. HOWE**, **N. BOUK**, **H.-C. CHEN**, **P. A. KITTS**, **T. D. MURPHY**, **K. D. PRUITT**, **F. THIBAUD-NISSEN**, **D. ALBRACHT**, **R. S. FULTON**, **M. KREMITZKI**, **V. MAGRINI**, **C. MARKOVIC**, **S. MCGRATH**, **K. M. STEINBERG**, **K. AUGER**, **W. CHOW**, **J. COLLINS**, **G. HARDEN**, **T. HUBBARD**, **S. PELAN**, **J. T. SIMPSON**, **G. THREADGOLD**, **J. TORRANCE**, **J. M. WOOD**, **L. CLARKE**, **S. KOREN**, **M. BOITANO**, **P. PELUSO**, **H. LI**, **C.-S. CHIN**, **A. M. PHILLIPPY**, **R. DURBIN**, **R. K. WILSON**, **P. FLICEK**, **E. E. EICHLER**, and **D. M. CHURCH**. "Evaluation of GRCh38 and de Novo Haploid Genome Assemblies Demonstrates the Enduring Quality of the Reference Assembly". In: *Genome Research* **27**:5 (May 2017), 849–864. DOI: 10.1101/gr.213611.116 (see p. 137)

[91] **THE 1000 GENOMES PROJECT CONSORTIUM**. "A Global Reference for Human Genetic Variation". In: *Nature* **526**:7571 (Oct. 2015), 68–74. DOI: 10.1038/nature15393 (see p. 137)

[92] **S. ASSEFA**, **C. LIM**, **M. D. PRESTON**, **C. W. DUFFY**, **M. B. NAIR**, **S. A. ADROUB**, **K. A. KADIR**, **J. M. GOLDBERG**, **D. E. NEAFSEY**, **P. DIVIS**, **T. G. CLARK**, **M. T. DURAISINGH**, **D. J. CONWAY**, **A. PAIN**, and **B. SINGH**. "Population Genomic Structure and Adaptation in the Zoonotic Malaria Parasite *Plasmodium Knowlesi*". In: *Proceedings of the National Academy of Sciences* **112**:42 (Oct. 2015), 13027–13032. DOI: 10.1073/pnas.1509534112 (see p. 147)

[93] **G. A. BUTCHER** and **G. H. MITCHELL**. "The Role of *Plasmodium Knowlesi* in the History of Malaria Research". In: *Parasitology* **145**:1 (Jan. 2018), 6–17. DOI: 10.1017/S0031182016001888 (see p. 147)

[94] **W. E. COLLINS**, **P. G. CONTACOS**, **J. C. SKINNER**, **W. CHIN**, and **E. GUINN**. "Fluorescent-Antibody Studies on Simian Malaria: I. Development of Antibodies to Plasmodium Knowlesi". In: *The American Journal of Tropical Medicine and Hygiene* **16**:1 (Jan. 1967), 1–6. DOI: 10.4269/ajtmh.1967.16.1 (see p. 147)

CHAPTER FOUR

GENERATION OF *Plasmodium knowlesi de novo* REFERENCE GENOMES FROM CLINICAL ISOLATES

Òtítọ́ korò; bì omi tooro nirọ́ ri —
Truth is bitter; lies are sweet like meat stew

Yorùbá adage

The advent of genome assembly and the subsequent *de novo* assembly of both new and ancient genomes has allowed for the ever-increasing boom in the scientific and biologically relevant questions researchers can ask – and in turn answer. In many cases, the first set of genomes sequenced and assembled for a particular organism have been the experimental lines painstakingly developed and adapted for laboratory research purposes. These lines were removed from active natural selection pressure, resulting in genetic variation from the contemporary organisms found in nature. Due to this and advancements in sequencing technology, non-laboratory restricted strains/lines are continually being investigated to provide insights into wild-type genomes. Such understanding can improve currently available genome sequences and provide new targets for treatment and prevention if necessary.

4.1 Introduction

Plasmodium knowlesi occupies a unique niche within malaria research. The organism is zoonotic, meaning, it has successfully surpassed a species barrier for infection in humans, thus it is able to act as an experimental model for human malaria [1, 2]. *P. knowlesi* has historically been an animal and *in* vitro model organism for malaria, due to its ability to infect non-human primates (NHP) and also be maintained as different experimental lines [2–4]. Indeed the *P. knowlesi* malaria model has been the source of advancements in malaria research such as antigenic variation [5]. As previously mentioned, *P. knowlesi* transmission occurs in and around the jungle fringes of Southeast Asia, with the largest known incidence in Malaysian Borneo [6]. As such, *P. knowlesi* has the potential to act as both a human pathogen and animal model for malaria infection and pathophysiology with clearly defined regions of incidence [2, 7, 8].

Despite this, unlike *P. falciparum*, *P. knowlesi* remains unsupported for full-fledged extended study with adequate financial backing. The currently available *P. knowlesi* whole genome, initially released by Pain et al. [9] and subsequently improved upon by GeneDB (Wellcome Sanger Institute), remains the 'gold-standard' representation of the *P. knowlesi* genome. While the genome was painstakingly constructed and accurate, it was also unable to resolve variable multigene families that had been identified to be associated with important processes of the parasite's infection. In particular, the Schizont Infected Cell Agglutination variant antigen (SICAvar) genes (see *chapter 1 subsection 1.3.2*), which are a family of ~100 - 136 (*or more*) variable genes that encode proteins implicated in host immune system evasion [9, 10]. However, the *SICAvar* genes remain poorly understand, with no clear definitive biological function associated with the group apart from antigenic variation.

Pain et al. [9] successfully characterised the *SICAvar*s and estimated a total of 107 gene family members within the *P. knowlesi* genome. However, with improved technologies and targeted investigations, this estimated total has increased, with Lapp et al. [10] currently estimating at least 136 *SICAvar* family gene members. However, both of these genomes – the 'gold-standard' PKNH reference genome [9] and the published PKNOH genome [10] – suffer from the same shortcoming, namely, the experimental

line from which the respective genome was assembled. Both Pain et al's PKNH [9] and Lapp et al's PKNOH [10] genomes were derived from experimental lines, which have been maintained by artificial passage in laboratory settings since they were first isolated in the 1960s [9–12]. Through artificial passage (via a NHP) in this manner, the *P. knowlesi* cultures can be maintained and continuously propagated, effectively removing host selection pressure on these lines.

The removal of natural selective pressure on the *P. knowlesi* genome of experimental lines over the decades since initial isolation could potentially result in artificial genome signatures, including vital structural variances present in the contemporary isolates and absent in the chronically stagnated experimental lines. Alternatively, the experimental lines can also act as a time capsule into the *P. knowlesi* infective genome. A view to a time with less deforestation and environmental upheaval in the endemic regions frequented by the Anopheline vectors and the simian and inadvertent human hosts. Thus, the importance of experimental line-derived genomes remains paramount. However, updated genomes generated from contemporary wild type *P. knowlesi* parasites that have been under continuous selection pressure using modern methods and techniques would only benefit *P. knowlesi* research.

The PKNH reference genome was constructed using the Sanger sequencing method [chapter 3 subsection 3.1.2]; thus, the sequence data is accurate and relatively long enough to resolve a majority of discrepancies in the released genome. However, the combined use of high-throughput sequencing technologies in the form of Illumina short reads and Nanopore long reads can result in further resolution of the *P. knowlesi* genome. With Nanopore long reads able to consistently output reads $>$4500 bp (*and up to* 2 Mb), the reads produced are able to span regions of high variability. In contrast, incorporating highly accurate Illumina short reads can resolve most sequencing errors inherent to long-read sequencing. These new techniques and technologies (*explored in chapter 3*) tend to be facilitated by a myriad of post-processing and analysis in the form of a pipeline; ensuring reproducible results.

4.2 Chapter 4: Aim and Rationale

With successful extractions of parasite-enriched DNA material, it was necessary to successfully carry out whole genome sequencing of the extracted material on the Oxford Nanopore Technologies (ONT) MinION sequencing platform. We first aimed to optimise a sequencing protocol to maximise the sequencing yield from DNA available and generated from our Biobank. To achieve this, we carried out evaluations of available sequencing protocols to carry out PCR-free genome sequencing.

Through the use of the tools we previously validated for similar data, we aimed to manipulate the generated sequence data and publicly available genome sequence data to construct *de novo* whole genomes of *Plasmodium knowlesi* from clinical samples. The assembled genomes were to be subsequently analysed and improved to investigate the genome structure and organisation as well as the identification of genes of interest – particularly the Schizont Infected Cell Agglutination variant antigen (SICAvar) and *Plasmodium knowlesi* interspersed repeat (*kir*) multigene families.

4.3 Methods

Throughout this project, the sequencing protocols implemented for genome sequencing changed as they were continuously improved upon by Oxford Nanopore Technologies (ONT). In the first instance, the SQK-RAD002 sequencing protocol [13] was used to determine the viability of sequencing *Plasmodium knowlesi* DNA on the Oxford Nanopore MinION platform. Following this, the first multiplexed sequencing protocol developed by ONT –SQK-RBK001 [14]– was employed to sequence five previously extracted *P. knowlesi* DNA samples. Whole-genome sequencing of enriched *P. knowlesi* DNA extracted from thawed clinical isolates using the CD45 DynaBeads method [chapter 2 subsection 2.4.5, *CD45 DynaBeads leucocyte depletion method*] began in earnest with the SQK-RBK004 sequencing protocol [15]. Additionally the SQK-LWB001 (*now the SQK-PBK004*) [16] was utilised for low genomic input isolates. The majority of isolates were sequenced using the SQK-RBK004, and the protocol is given below in detail. Methods for the other protocols (SQK-RAD002, SQK-RBK001, SQKPBK004/LWB001)

are presented in Appendix section E.1. Although the SQK-RAD002, SQK-RBK001 and SQK-RBK004 are PCR-free, i.e. with no need for genomic amplification, the SQK-PBK004/LWB001 is a protocol that uses PCR amplification to fortify the low genomic input DNA to be sequenced.

4.3.1 Long read sequencing using Oxford Nanopore Sequencing

Upon completing DNA extraction [chapter 2 subsection 2.4.6,DNA Extraction], the eluted DNA was either stored at -30°C or directly taken forward for whole-genome sequencing on the MinION sequencing platform. In either case, sequencing was carried out using a modified SQK-RBK004 sequencing library [15] as described below. Sequencing using the SQK-RBK004 library requires 400ng of genomic DNA, regardless of the number of barcodes being used. In addition to the improved protocols, the flowcells upon which genome sequencing occurred were also improved towards the latter half of this study. For the majority of this project, whole-genome sequencing was carried out on ONT's FLO-MIN106D R9.4.1 flowcell. However, some experiments were also sequenced using the FLO-MIN110/FLO-MIN111 R10 flowcells at the latter end of this project. Information regarding the isolates, barcodes, flowcells and sequencing kits used for each experiment can be found in Appendix Table E.2.

AMPure XP beads wash step

The DNA solution was washed using the AMPure XP beads at a 1:1 ratio by volume to concentrate the DNA while removing shorter DNA fragments. Much like the CD45 DynaBeads, the AMPure beads are superparamagnetic beads that bind to DNA, isolating the bound fragments.

To begin, the DNA concentration of the sample was quantified using the Qubit fluorometer as described in chapter 2 subsection 2.4.7 [*DNA Quantification*]. After DNA quantitation, AMPure beads were added at a 1:1 ratio by volume to each isolate DNA sample before each LoBind microfuge tube was gently inverted for five minutes at room temperature. The microfuge tubes were briefly spun to recover any fluids and promptly placed in a

magnetic rack until the supernatant became clear. The supernatant was carefully removed, leaving the pellet in the rack. 200 µL of freshly made ethanol (70 %) was added to each microfuge tube by running the alcohol over the bead pellet, taking care not to disturb the pellet. The ethanol wash was then discarded, and the wash step was repeated. After the second alcohol removal, the microfuge was briefly spun and placed back in the magnetic field. Residual alcohol was then removed by pipetting or left open to evaporate. To reach the 400 µL target DNA (eluted in 7.5 µL nuclease water) required by SQK-RBK004, a DNA concentration of $\geq$54 ng/µL (nuclease-free water elution was often $\geq$7.5 µL) was needed.

The tubes were removed from the magnetic rack, and the beads were resuspended in the nuclease-free water to make up a concentration of $\geq$54 ng/µL. Where the concentration was insufficient, the target DNA was stored to either be used in a multiplexed sequencing experiment with isolates of significantly high concentration or sequenced using other library preparations [Appendix section E.1]. The resuspended beads were incubated at room temperature for two minutes and placed in a magnetic rack until the beads were coalesced by the magnetic field. 7.5 µL of the re-eluted DNA was aliquoted into 0.2 mL PCR tubes to be taken forward for the SQK-RBK004 library preparation. Where there was an excess in the volume of eluted DNA, multiple samples of the same isolate DNA were sequenced to be concatenated once basecalling and demultiplexing were completed [seen as isolate (a,b) in *Appendix Table E.2*]. 1 µL of eluted DNA was also taken through DNA quantification using the Qubit fluorimeter [see: chapter 2 subsubsection 2.4.7, Qubit Quantification].

SQK-RBK004 Library preparation

The 0.2 mL PCR tubes containing 7.5 µL of purified DNA was placed on ice before adding 2.5 µL of the barcode fragmentation mix. Each 0.2 mL tube, i.e. each DNA sample was given a unique barcode between 1 and 12. The tubes were mixed by flicking and then briefly spun down before being placed in a thermocycler using the settings: 30°C for 1 minute then 80°C for 1 minute. The barcoded DNA content in all tubes was then pooled into a single LoBind microfuge tube with the total volume recorded. Following this, the AMPure XP beads washing step was repeated as described above.

After the second ethanol wash, the AMPure XP bead pellet bound to the barcoded DNA was resuspended in 10 µL Tris-HCl (pH7.5 - 8.0) with 50 Millimolar (mM) NaCl. The mixture was incubated for two minutes at room temperature to elute DNA bound to the beads, before being placed on the magnetic rack to remove beads from the eluted DNA. The pooled DNA, now eluted in the buffer, was transferred into a new LoBind microfuge tube, to which 1 µL of the rapid adapter (RAP) was added before mixing by gentle flicking [15]. The microfuge tube was then incubated for 5 minutes at room temperature before being placed on ice, while the sequencing flowcell was prepared.

Sequencing on the MinION platform

The MinION flowcell was prepared for sequencing using a priming mix made by adding 30 µL of flush tether (FLT) to a tube of flush buffer (FB). Following this, the priming port of the flowcell was opened, and an air bubble was removed per the manufacturer's instructions. 800 µL of the priming mix was added into the flowcell via the priming port, ensuring no air bubbles were introduced. The flowcell was left to incubate at room temperature for 5 minutes, after which the SpotOn port was opened. 200 µL of the priming mix was then added into the priming port once again, with the SpotOn port open. Subsequently, 34 µL of sequencing buffer (SQB), 25.5 µL of loading beads (LB) and 4.5 µL of nuclease-free water were added to the pooled 11 µL DNA library. Using a P200 pipette, the final 75 µL library was loaded into the flowcell via the SpotOn port in a dropwise fashion. Sequencing was started using the MinKNOW software with recommended settings for each sequencing kit, including the default starting voltage (-180 millivolts). No real-time basecalling was carried out due to access to a high performance cluster (HPC) for accelerated basecalling and analysis.

Sequencing on the Illumina MiSeq platform

Isolates that possessed adequate concentration were kindly sequenced on the Illumina MiSeq platform by Dr Susana Campino of the London School of Hygiene and Tropical Medicine. A standard PCR amplification-based sequencing run was implemented, resulting in reads of 2 x 150 bp. Earlier work (Pinheiro et al. [17]) resulted in the

Table 4.1: Accession codes for previously sequenced matched Illumina sequences used in this project

Isolate	Accession Code
sks047	ERR366425
sks048	ERR274221
sks050	ERR274223
sks058	ERR274224

Isolates were first extracted and sequenced by Pinheiro et al. [17] for an unrelated study. The sequence reads generated were stored on a public repository and a copy of these sequence reads were downloaded for use in this project

production and publication of high-quality Illumina short-read sequences of isolates also available in this project. Hence, matched Illumina short-read sequences for isolates with duplicate archival whole blood were carried forward to be used in this study. The accession codes for the used isolates are shown [Table 4.1].

4.3.2 *De novo* genome assembly of extracted *Plasmodium knowlesi* DNA

Basecalling, Adapter Removal, Quality Assessment and Alignment

The sequencing outputs of the MinION are stored in custom HDF5 format files called FAST5 files. Upon sequencing completion, FAST5 files were basecalled on a Linux based system running Ubuntu *v*19.10, equipped with a NVIDIA GeForce™GTX1060 GPU. As previously described, basecalling was carried out using different versions of ONT's basecalling algorithms [see chapter 3 subsection 3.5.1, *Assessing basecallers and demultiplexers for Nanopore long reads*]. However, within the pipeline presented below, basecalling was achieved with Guppy v.4.0.15 (*released August 2020*) using the command shown in Appendix Code. C.2.

Following this, demultiplexing was achieved with qcat (*v1.1.0*) [Figure 4.1] with the commands shown in Appendix Code. C.2, generating sequence reads in the FASTQ format. Using default parameters, Porechop (*v0.2.4*) was subsequently used to remove

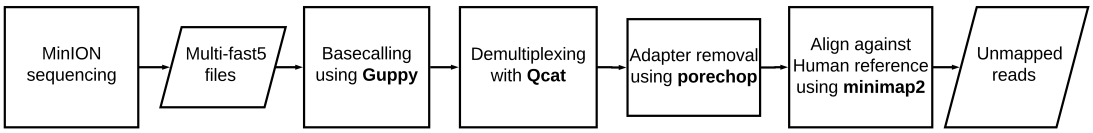


Figure 4.1: Pipeline to carry out basecalling, quality checks and alignments prior to *de novo* assembly. A sequence of steps to take raw sequenced reads through basecalling, demultiplexing, adapter removal and alignment as a means of generating high quality reads for genome assembly.

adapters and remnant barcodes attached to each read [Appendix Code. C.2]. Sequenced reads generated from human DNA (hDNA) which remained after the CD45 leucocyte depletion were removed by alignment against human reference genome. Briefly, using the script in Appendix Code. C.2, sequence reads were aligned against the human GRCh38.p13 reference genome (NCBI accession: GCF_000001405.39) [18] with minimap2 [19]. Subsequently, reads which did not align to the human reference genome were separated from the alignment file using samtools (*v1.10*, [20, 21], Figure 4.1). The extracted unmapped alignments were converted back to sequence reads using bedtools (*v2.29.2*, [22]).

De novo genome assembly and Decontamination

FASTQ files containing sequence reads which did not align to the human reference genome (*henceforth Pk-input reads*) were used as inputs for Flye (*v*2.8.1, [23, 24], Figure 4.2) using the script in Appendix Code. C.3. Following this, assemblies which were completed successfully, were checked for contamination using Blobtools (*v*1.0.1, [25]) as described in chapter 3 subsection 3.5.2 [Figure 4.2]. Briefly, a script containing two functions was written, where one function assessed contamination on raw assembly outputs and the second function assessed assemblies which had already been decontaminated, acting as a second checkpoint to confirm decontamination. Both sets of commands function in a similar fashion, hence only the raw assembly blobtools assessment function is provided in Appendix Code. C.3. Here, the generated assembly was aligned against its *Pk*-input sequence reads using minimap2 (*v*2.17, [19]), generating Binary Alignment Mapping (BAM) files. Subsequently, a local Basic Local Alignment Search Tool (BLAST) alignment of the assembly against the NCBI nt database was

carried out [Appendix Code. C.3]. The BAM and the BLAST output file were used as inputs for BlobTools to assess GC distribution within the assembly and its aligned reads, as well as coverage information. Contigs which were identified to be contaminated as non-*Plasmodium* spp. were manually removed and discarded from the assembly using the decontamination script in Appendix Code. C.3 and Appendix Code. C.4. The decontaminated assemblies will henceforth be referred to as 'cleaned draft assemblies'.

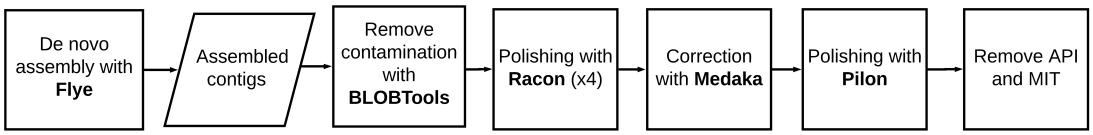


Figure 4.2: Pipeline detailing steps carried out for *de novo* genome assembly, decontamination, polishing and correction. The outputs from the alignment steps are assembled using Flye and successful assemblies are subsequently assessed for contamination, length and accuracy. Accuracy and consensus is improved using Pilon, Medaka and Racon.

4.3.3 Assembly polishing and Correction

After decontamination, the draft assemblies were polished with the long-read *Pk*-input reads (reads which did not align to the human GRCh38.p13 reference), using four rounds of racon (*v*1.4.13, [26]) with default settings [Figure 4.2]. To achieve this, the cleaned draft assemblies were first aligned against the *Pk*-input reads to produce Sequence Alignment Mapping (SAM) files. After this, the SAM files, alongside the *Pk*-input reads and the clean draft assemblies were used as inputs into racon [Appendix Code. C.5, Figure 4.2]. The assembly output of the first racon iteration was then aligned against *Pk*-input reads once more using minimap2. Then racon was repeated, using the output of iteration one as an input for iteration two. The process of alignment using minimap2 and subsequent racon polishing was repeated twice more. The output of the fourth racon iteration was then used as input for medaka (*v*1.0.3) to produce a consensus sequence [Appendix Code. C.5, Figure 4.2]. For isolates which possessed matched Illumina short sequence reads, Pilon (*v*1.23, [27]) was used for further correction with modified default parameters [Appendix Code. C.5, Figure 4.2]. Here, three rounds of Pilon was carried out, where the first round used the consensus output of medaka and the BAM file

containing alignments of the medaka output against the matched isolate Illumina short reads [Appendix Code. C.5]. The alignments of the short reads against the consensus draft assembly generated by medaka and each subsequent round of Pilon was carried out using bwa (*v*0.7.17).

4.3.4 Apicoplast and Mitochondrial Circularisation

The prokaryotic *P. knowlesi* apicoplast (API) and mitochondrial (MIT) sequences were extracted from the consensus draft assemblies of each isolate using a series of custom scripts. To achieve this, the polished and corrected draft *P. knowlesi* assemblies (after Medaka or Pilon) were aligned against the *P. knowlesi* reference API and MIT [9] using default parameters of the '*megablast*' function of BLAST (*v*2.9, [28]). Contigs which aligned to the Pain et al. [9] reference API or MIT were extracted from the draft assemblies using custom Python scripts [Appendix Codes C.6 and C.7]. The apicoplast/mitochondria free (API/MIT-free) genomes were taken forward for repeatmasking. The extracted prokaryotic sequences were processed separately from the remaining eukaryotic sequences to be circularised with Circlator [[29], Figure 4.3]. Here, Circlator (*v*1.5.5) was used with the command `circlator all --data_type nanopore-raw --bwa-opts \ -x ont2d" --merge_min_id 85 --merge_breaklen 1000` and the outputs analysed manually.

Concurrently, a prokaryotic *de novo* approach was implemented using the Canu assembler [30]. Briefly, the Pk-input reads (which did not align to the human GRCh38.p13 reference genome [chapter 4 subsubsection 4.3.2]) were aligned to reference apicoplast (API) and mitochondrial (MIT) sequences extracted from the PKNH reference genome [9] using minimap2 [19]. Following this, reads which aligned to the reference sequences were extracted as previously described [chapter 4 subsubsection 4.3.2]. Successfully extracted reads were used as input for the Canu assembler (*v*1.8, Appendix Code. C.8). From this, successfully assembled contigs were circularised using Circlator [29] with an expansion of the previously described command above [Appendix Code. C.8] before annotation of all successfully assembled and circularised assemblies was implemented using Prokka (*v*1.14.6) [31]. Manual curation was also implemented to split large contigs into constituent fragments, which span the length of the desired prokaryotic sequence.

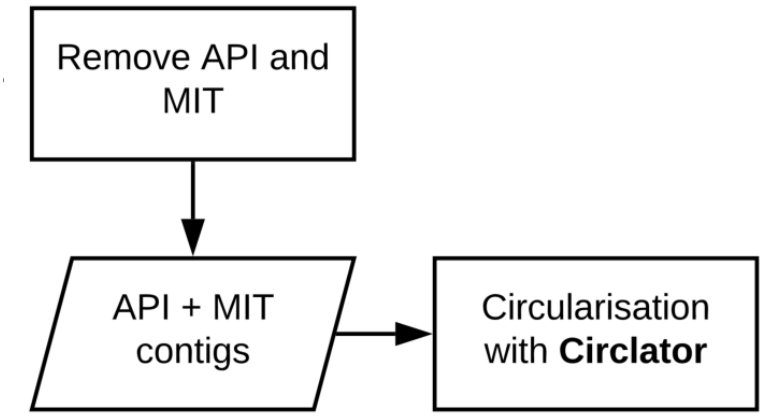


Figure 4.3: Flowchart for the processing of the apicoplast and mitochondral genomes for circularisation. The apicoplast and mitochondrial genomes from each successful draft assembly were separated for circularisation using Circlator.

4.3.5 Masking Repetitive elements

Subsequently the API/MIT-free draft assemblies were taken through a custom pipeline for masking repeat elements [Figure 4.4]. The pipeline was initially developed by Szitenberg [32] however, was extensively modified for this project. Briefly, the API/MIT-free draft assemblies were first processed using RepeatModeler (*v*1.0.10) and the *de novo* repeats outputted were used as inputs for Censor [33] (*www.girinst.org/censor/*), with the '*Eukaryota*' and '*Report simple repeats*' options chosen. The outputs of Censor were used to provide further classifications for the *de novo* repeat clusters from RepeatModeler, resulting in a repeat library for each isolate. Using CD-HIT (*v*4.8.1, [34, 35]), the individual isolate repeat libraries were combined and redundant repeats were removed to generate a singular 'master' library of non-redundant repeats represented in all the isolates sequenced [Figure 4.4, Appendix Code. C.9].

Following this, RepeatMasker (*v*4.0.7, [36]) was run using the master repeat library and the output of repeatmodeler for each isolate. The repeats identified in the output of RepeatMasker for each isolate API/MIT-free draft assembly was further parsed and classified using 'One code to find them all (OCTFTA)' [Appendix Code. C.9]. Furthermore, alternative approaches to investigate and classify specific repeat classes was carried out. In this capacity, the LTRHarvest module of GenomeTools (*v*1.6.1, [37]) was

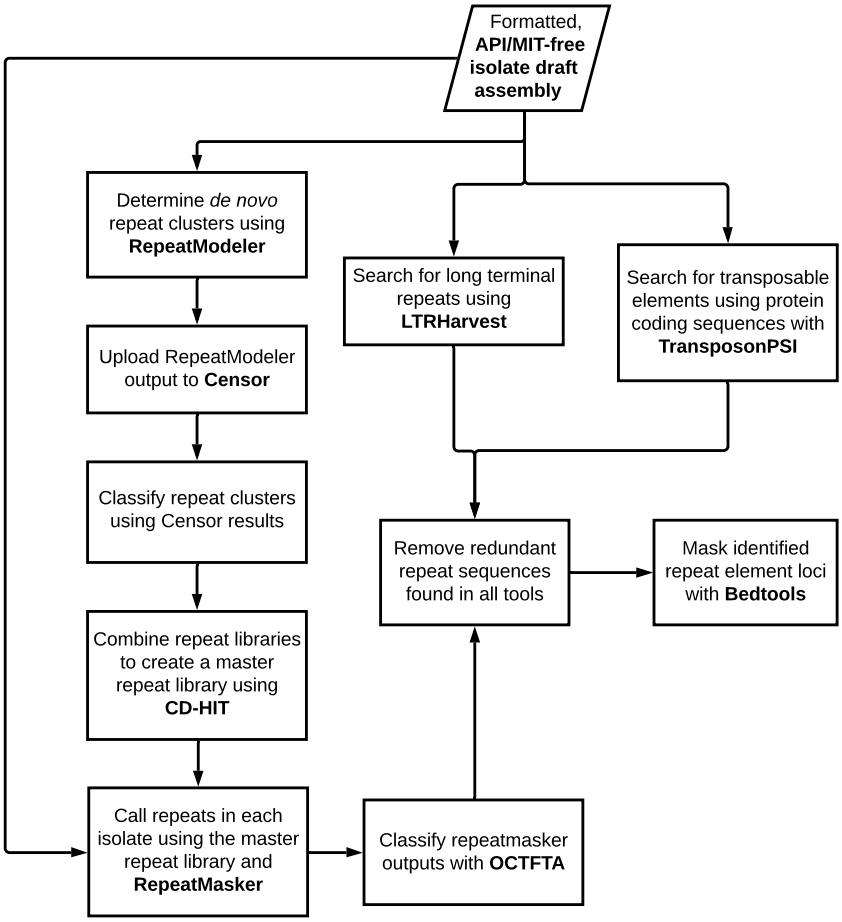


Figure 4.4: Masking of transposable elements in generated *Plasmodium knowlesi* genomes. The contigs of the input draft assembly were renamed and formatted to ensure no downstream errors. The assembly file was then used as inputs for repeatmodeler, LTRHarvest and TransposonPSI. RepeatModeler carried out *de novo* repeat cluster prediction, which was in turn classified by Censor. The resulting repeat libraries were combined with CD-HIT and used to find repeats within the isolate draft assemblies. These repeats were then further classified using One Code To Find Them All. LTRHarvest and TransposonPSI carried out repeat element discovery for long terminal repeats and protein-coding sequences, respectively. Results from all three branches informed the generation of a set of repeat loci soft-masked using Bedtools. Pipeline adapted from: Szitenberg [32]

used to investigate the secondary structures on long terminal repeats (LTR) [Figure 4.4], using the '*suffixerator*' and '*ltrharvest*' functions of genometools [Appendix Code. C.9]. At the same time, TransposonPSI (*v*1.0, [38]) was used on the API/MIT-free draft assemblies using default parameters, to find repeat elements using coding sequences.

The outputs of RepeatMasker, OCTFTA, LTRHarvest and TransposonPSI were combined and subsequently collapsed to remove redundancies between them, generating a General Feature Format (GFF) holding the loci of the repeat elements for each isolate [Figure 4.4]. The identified loci within each API/MIT-free draft assembly were soft masked using the default settings of the '*maskfasta*' function of bedtools (*v*2.27, [39], Appendix Code. C.9).

4.3.6 De-chimerisation, Prediction and Annotation

Following this, RagTag (*v*1.0.1,[40, 41]) was used to check the masked draft assemblies for chimeric contigs [Figure 4.5]. Both the '*correct*' and '*scaffold*' functions of the RagTag package were run using the options `--debug --aligner nucmer --nucmer-params='-maxmatch -l 100 -c 500'`.

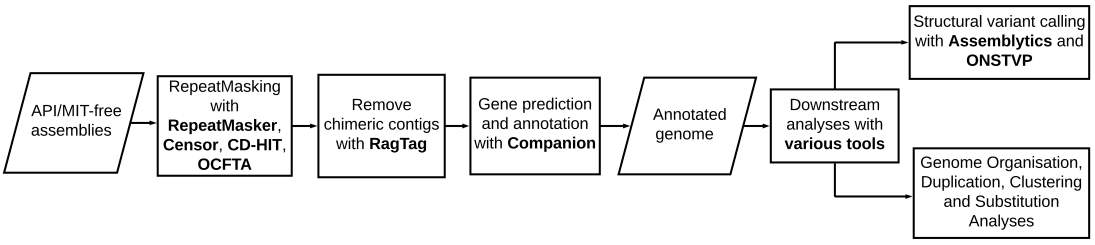


Figure 4.5: Process for carrying out final stages of the pipeline. The apicoplast and mitochondria-free draft assemblies are repeatmasked, then checked for chimeric contigs before being uploaded on Companion for protein prediction and genome annotation. Post-annotation analyses are subsequently carried out, including structural variant calling and duplication analyses of various genes of interest.

Gene prediction and annotation were implemented using the Companion webserver [42] (Sanger server; *decommissioned*). Here, each masked assembly was uploaded onto Companion with a unique sequence prefix, where the Cultured experimental line was denoted as 'PKA1H1_STAND' and the patient isolate assemblies were labelled as

'PKCLINC' e.g. PKCLINC047. Companion was subsequently run with no transcript evidence provided, minimum match length of 500 bp, 80 % match similarity for contig placement, AUGUSTUS score threshold of 0.8 and taxid of 5851. Furthermore, pseudochromosomes were contiguated, reference proteins were aligned to the target sequence, pseudogenes were detected, and RATT [42] used the reference genome's gene models for prediction.

4.3.7 Structural Variant Analyses

A preliminary investigation into the structural variation present in the patient isolates compared to the experimental line was carried out [Figure 4.5]. The two patient isolates (sks047 and sks048) were ideal representatives of the dataset as they represented the two dimorphic clusters of *P. knowlesi* – Cluster 1 and 2, respectively. Additionally, both isolates possessed high quality matched Illumina sequence reads [17], that were used to polish the *de novo* genomes assembled on Flye. Thus, both genomes would provide appropriate input read coverage, post-assembly polishing and processing. The StAPkA1H1 genome was used as the control in the methods described below to ensure parity across sequencing technologies. Apart from the availability of matched Illumina short reads, which were unavailable for StAPkA1H1, the three genomes were assembled and processed with similar commands and techniques.

Alignment-based structural variant calling

Alignment-based structural variants (SVs) calling was carried out using the Oxford Nanopore structural variation pipeline (ONTSV) (*v*2.0.2) [43–46]. The snakefile containing the parameters for ONTSV was modified and tuned for the input genomes. Briefly, ONTSV first parses the input reads for sks047 and sks048 (see: *chapter 4 subsubsection 4.3.2*) using catfishq [47] and seqtk [48]. The parsed reads are aligned to the StAPkA1H1 genome (used as the reference) using lra [44] using parameters '`-ONT -p s`'. The lra alignment outputs were sorted and indexed using samtools and read coverage was calculated using mosdepth [45] with parameters '`-x -n -b 1000000`'.

Following this, SV calling was performed with cuteSV [46] using the parameters '*--min-size 30 --max-size 1000000 --retain_work_dir --report_readid --min_support 2*'. Variants identified by cuteSV were then filtered by their length (30 bp), read depth (8 reads), alignment and read quality (>Q30), as well as the type of SV. SVtypes such as INDELs, inversions (INV), duplications (DUP) and translocations (TRA) were identified, sorted, filtered and subsequently indexed. SVtypes which failed the first pass filteration were further manually filtered based on length and quality alone, with no read depth limitation given, in order to identify the presence of high-quality, low-occurrence variants.

Assembly-based structural variant calling

Assembly-based SV calling involved the alignment of the complete and annotated genomes of sks047 and sks048 against the StAPkA1H1 genome. Alignments were carried out using nucmer [49] with the parameters '*--maxmatch -l 100 -c 500*' and the aligned outputs were uploaded onto Assemblytics for variant calling [50]. Assemblytics was run using default settings and a minimum SV length of 30 bp. Using SURVIVOR (*v*1.0.7, [51]), BEDfile outputs of Assemblytics were converted into Variant Call Format (VCF) files for further analyses and annotation.

Variant Annotation and Analyses

VCF files produced from the alignment-based and assembly-based approaches were further filtered for variants of lengths >49 bp using bcftools [Appendix Code. C.11]. However, no quality filter was applied to the Assemblytics-derived VCF file due to the lack of quality information in the original BEDfile output of Assemblytics.

Annotation of identified and filtered variants was achieved using vcfanno (*v*0.3.2, [52]) and the annotated output sorted and indexed [Appendix Code. C.11]. Using IGV [53], visual assessments of called variants was performed; including assessments of the *PknbpXa* gene locus, previously associated with genome-wide dimorphism in *P. knowlesi* [17]. Summary statistics of annotated variants was performed using SURVIVOR [51] with parameters '*-1 -1 -1*' and plots were generated using vcfstats [54]. Comparisons

of generated VCF files was done using the '`isec`' function of bcftools with default settings, allowing for quantification of variants present within the two representative isolates.

4.3.8 Comparative Genomics, Quality Assessment and Visualisations

As the pipeline progressed, numerical metrics for the related sequence data were performed using assembly-stats (*v*1.0.1, [55]) and pomoxis (*v*0.3.4, [56], Appendix Code. C.10). Accuracy and completeness of the draft genomes were assessed using BUSCO (*v*5.0, [57]) and QUAST (*v*5.0.2, [58]). BUSCO was run through a singularity container, while QUAST was run using the QUAST-LG mode for large eukaryotic genomes [Appendix Code. C.10]. Further metrics were provided by Qualimap (*v*2.2.2, [59]) by aligning the chromosomes of annotated genomes against those of the Pain et al. [9] PKNH reference genome [Appendix Code. C.10].

Subsequently, the Companion outputs were manually parsed and analysed for regions of synteny, orthologous clustering and genes of interest, focusing on variant multi-gene families known to span the core genome and telomeric regions. Visualisations of gene density, chromosome structure and gene loci were completed using karyoploteR [60]. Further visualisations of chromosomal comparisons were completed using Mauve (*progressiveMauve*, *v*2.4.0, [61]) with default settings. Additionally, to visually represent regions of differences between the annotated genomes, dotplots were generated using dotPlotly (parameters: '`pafCoordsDotPlotly.R -l -x -p 15 -m 10000`', [62]). dotPlotly was implemented using the PAF format generated by minimap2, whereby the entire genomes of the annotated isolates were aligned against the PKNH reference genome [9].

4.4 Results

The completed pipeline developed for this project is represented in Figure 4.6, and the scripts used to generate the results are hosted on Github at www://github.com/damioresegun/Pknowlesi_denovo_genome_assembly. The outputs from each step of the pipeline are presented below and in Appendix E.

4.4.1 Sequencing enriched parasite DNA on the MinION and Data processing

Sequencing on ONT's MinION sequencing platform was carried out for 25 unique isolates (24 patient, one cultured), resulting in varying sequence read yield [Figure 4.7, Appendix Table E.7]. The first sequencing experiments 'Nov_2017' and 'Dec_2017' showed the lowest sequence length yield with 0.0286 Gb and 0.0929 Gb, respectively [Figure 4.7*a*, Appendix Table E.7]. However, both have the expected average read length between 4000 - 5000 bp, with the Nov_2017 sequencing experiment reporting multiple reads $>$175 000 bp. The greatest sequence length and reads yield came from both June 2019 (a,b) experiments, with June_2019_b in particular resulting in $>$9 Gb sequence length yield [Figure 4.7]. The final sequencing experiments of the project, which primarily involved isolates of the cultured PkA1-H.1, showed poor sequence read and length yield. However, the average read length and maximum read length appear similar across experiments [Figure 4.7].

After carrying out demultiplexing and adapter removal, the simple numerical metrics calculated for each isolate are shown in Table 4.2. Here, the greatest yield comes from StAPkA1H1 (*St. Andrews cultured PkA1-H.1*), sks070 and sks339, all of which possessed multiple samples that were sequenced and concatenated together after demultiplexing [Table 4.2]. This is also apparent in other isolates with multiple sequenced samples having relatively larger yield lengths. However, sks074, sks276 and sks367 also have relatively large yields from a single sequenced sample. Although large starting concentrations are seen for some isolates like 30.60 ng/µL for sks254, there is no correlation between starting concentration and the yield of an isolate (*Spearman's rank correlation test* p-

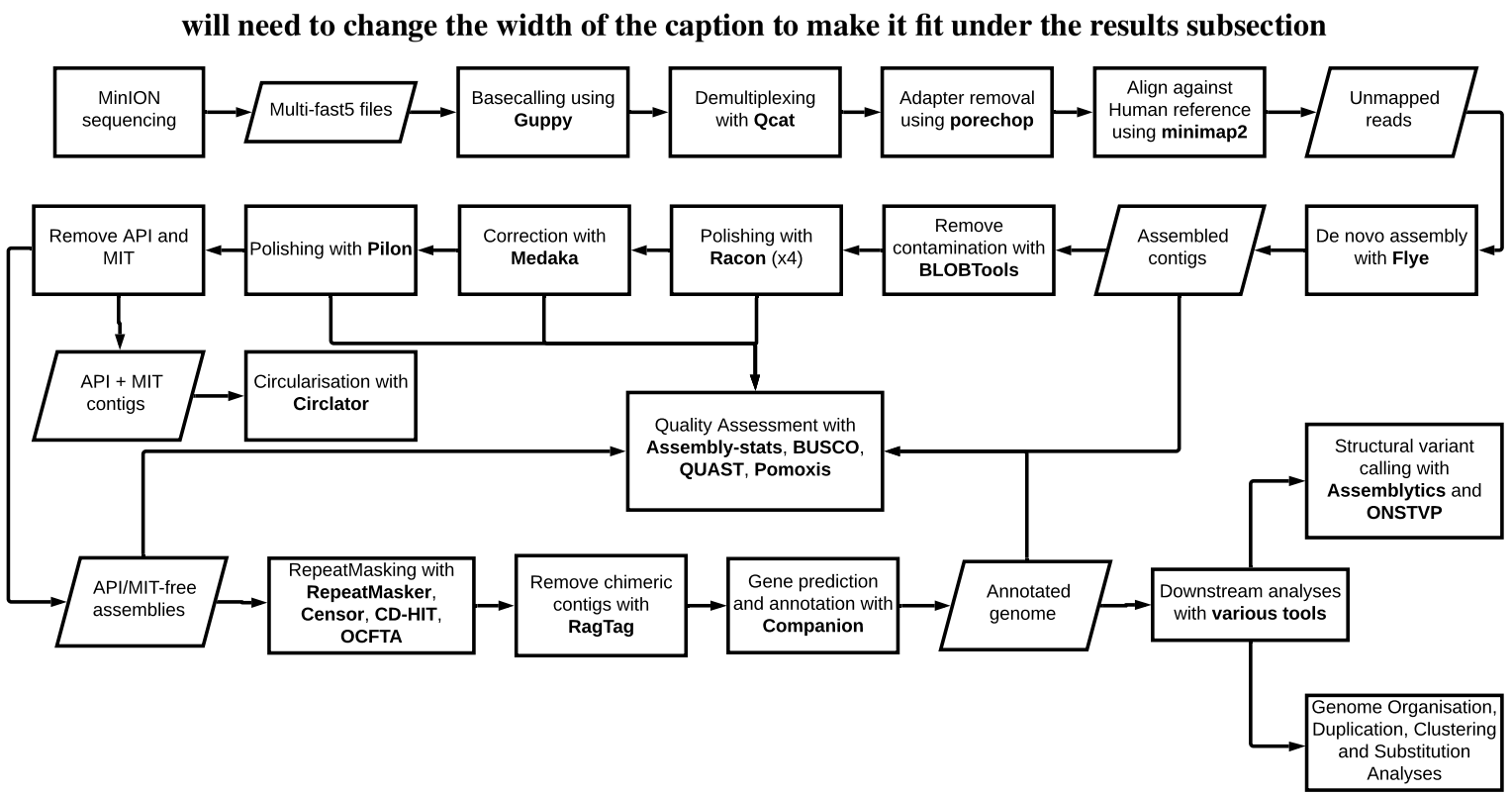


Figure 4.6: A pipeline to generate and analyse high quality *de novo Plasmodium knowlesi* genomes. Sequence data generated from enriched *P. knowlesi* extracted from thawed whole blood isolates collected from infected patients. After basecalling, remnant human DNA is removed before *de novo* genome assembly. The assembly is checked for contamination before it is polished and corrected. Prokaryotic extra-chromosomal sequences are separated for a prokaryotic-specific assembly pipeline. The assemblies are repeatmasked and checked for chimeric contigs before protein prediction. Further analyses can be done on the annotated outputs

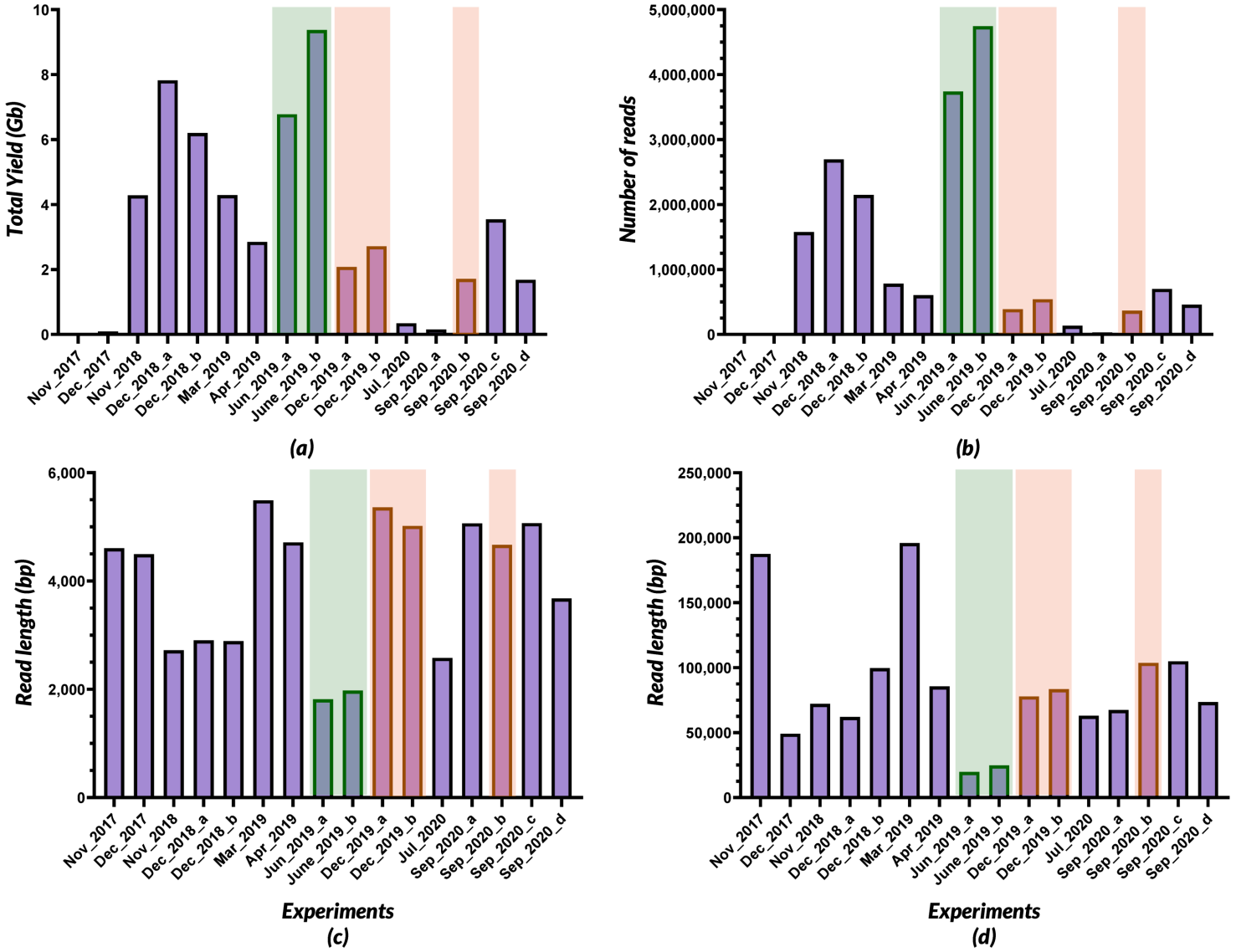


Figure 4.7: Descriptive metrics of each sequencing experiment carried out within this project. The total yield in base pair (bp) **(a)**, number of reads **(b)**, average read length in bp and **(c)** maximum read length **(d)** of each sequencing experiment carried out is presented. When multiple sequencing experiments occur in the same month, they are given an alphabetical denotation. Additionally, each experiment was multiplexed, where an average of 4 isolates was sequenced at a time. Experiments carried out with the SQK-PBK004/LWB001 are shown in green, and experiments carried out using the R10 flowcells are shown in orange. There was no statistical correlation between the input DNA concentration and the sequencing yield per experiment (*Spearman's rank correlation test, p=0.58247*). There are statistically significant correlations between the concentration and average read count (*p=5.56E-04*) and the yield and read count (*p=7.24E-12*).

value=0.58247; Appendix Table E.3.2). On the other hand, the yield and read count (p=7.24E-12) as well as the starting concentration and average read length (p=5.56E-04) are significantly positively correlated [Appendix Table E.3.2]. FastQC results of the adapter removed sequence data shows a median Phred quality score of 23 and a mean quality score of 21 [Table 4.2].

Table 4.2: Sequencing yield outputs for each sequence isolate after adapter removal

Isolates	Avg. Starting Parasite Conc. (ng/µL)	Yield (Gigabases (Gb))	Read count	Average read length (bp)	Quality Score
StAPkA1H1^{+}	7.92	6.17	1,340,355	4603.46	Q23
sks047^{+}	34.3	4.00	1,215,835	3286.25	Q24
sks048^{+}	33.26	1.81	352,122	5152.69	Q23
sks050	93.40	0.49	89,195	5474.97	Q19
sks058^{+}	21.75	1.72	674,945	2547.64	Q24
sks070^{+}	13.45	5.29	1,582,056	3341.8	Q23
sks071	-	0.01	3,432	3751.8	Q15
sks074	2.38*	1.03	247,131	4162.75	Q23
sks078	-	0.99	416,589	2377.07	Q23
sks125	6.52	0.79	366,335	2153.93	Q24
sks133	0.38	2.24	1,425,401	1568.62	Q23
sks134	0.76	2.55	1,705,491	1496.51	Q23
sks201	–	0.03	6,150	4587.04	Q17
sks231	–	0.02	3,523	4281.53	Q15
sks254	30.60	1.71	825,358	2072.55	Q21
sks276	3.38	4.25	2,392,770	1775.69	Q23
sks280	6.08	0.56	325,845	1732.28	Q24
sks325	12.50	0.13	29,545	4341.66	Q19
sks330	9.04	0.18	81,415	2236.52	Q24
sks331	6.78	0.56	248,840	2263.04	Q24
sks333	16.40	1.60	572,643	2791.62	Q24

Continued on next page

Table 4.2 – *Continued from previous page*

Isolates	Avg. Parasite Starting Conc. (ng/µL)	Yield (Gb)	Read count	Average read length (bp)	Quality Score
sks339+	63.70	5.21	1,355,203	3842.69	Q24
sks343	–	0.01	3,628	4108.75	Q16
sks344	71.20	0.46	85,326	5440.88	Q19
sks367	5.78	2.73	1,603,707	1701.29	Q23

Basecalled reads were demultiplexed, and adapters were removed. Isolates with multiple samples sequenced across the different sequencing experiments were combined to determine the yield. DNA concentrations were measured using the Qubit fluorometer and the average presented above, and the full concentration information is presented in Appendix Table B.3. Descriptive statistics were calculated using assembly-stats.

* - starting concentration were taken from the NanoDrop values

+ - Isolates with multiple samples. The multiple samples were averaged and the resulting concentration is presented above

Alignments to the human GRCh38p.13 reference genome resulted in an average mapping percentage of 55.35 % [Appendix Table E.8]. However, mapping ranged from as little as 6.61 % in sks231 to as large as 98.33 % in sks367, with the cultured PkA1.H-1 isolates (StAPkA1H1) still possessing 20.23 % of human mapped sequence reads [Appendix Table E.8]. A Spearman's rank correlation test shows the isolate's total base-pair length is significantly positively correlated (p=0.008) to the percentage of sequence mapped to the human reference genome. At the same time, the starting concentration is significantly negatively correlated to the percentage mapped (p=0.004, Appendix Table E.5). Unmapped sequence read length for each isolate is presented in Figure 4.8 showing the isolates which had multiple samples of the same patient/experimental isolate combined (*StAPkA1H1, sks047, sks048, sks058, sks070, sks339*) resulted in the largest total length *–and subsequently largest read count–* [Appendix Table E.4.2], that did not map to the human reference genome. By far the largest of these is the cultured StAPkA1H1 isolate sequence data with >5.5 billion base pairs with the smallest being sks071 with 11.6 million base pairs [Figure 4.8, Appendix Table E.4.2]. Quality assessments using FastQC showed very little change between the adapter-removed sequence reads, and the unmapped sequence reads with the only change in quality seen in

sks280, which reports a quality score of Q24; an increase of 1 from the adapter-removed sequence data shown in Table 4.2.

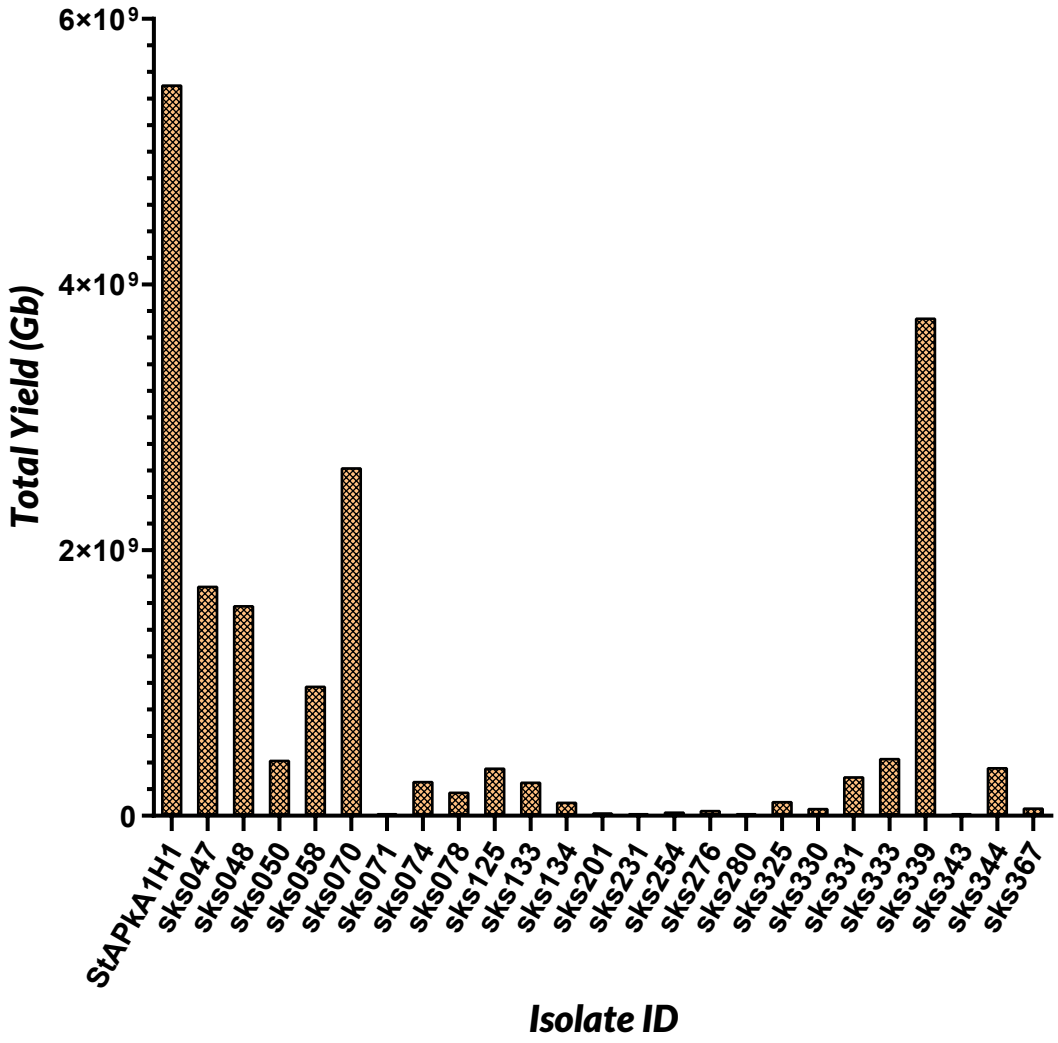


Figure 4.8: Base pair length for each isolate's sequence reads which did not map to the human GRCh38p.13 reference genome

4.4.2 Evaluating draft *de novo* assemblies

To determine a coverage threshold for our dataset, and represent the two dimorphic clusters, all twenty-five unique (24 patients, one cultured) isolate sequence data were put forward for *de novo* genome assembly. The overall sequence coverage used as input for Flye *de novo* genome assembly are presented in Appendix Table E.10. Isolate input sequence reads, which were the product of combining multiple samples of the same isolate, generated greater coverage based on the length of the reference genome [Table 4.2]. Of the 25 isolates, two isolates (sks071, sks280) failed to result in any form of genome assembly. A further eight isolates were eliminated from further study due to insufficient assembled genome length, where the largest genome assembled was sks133 with 1.9 Mb total sequence length [Appendix Table E.10].

The remaining 14 isolates (13 patient, 1 cultured) resolved into *de novo* draft assemblies ranging from 17.7 Mb to 24.4 Mb with a median assembly length of 23.1 Mb [Figure 4.9, Appendix Table E.10]. Figure 4.9 reveals that the longest draft assemblies also resolved in less contigs and larger N50 lengths with the longest draft assemblies resolving around $2.4x10^7$ or 24 Mb base pair lengths. Subsequent BlobTools analyses on the successful 14 draft genomes revealed no human sequence contamination within the contigs of all draft assemblies.

Overall, the improvements implemented by Racon, Medaka and Pilon are not very evident in the descriptive metrics of the assemblies [Figure 4.10]. Generally, Racon increased the sequence length of the assemblies while the number of contigs, N50 and average contig length remain similar to the raw assembly outputs [Figure 4.10, Appendix Table E.10]. On the other hand, Medaka increased the number of contigs and reduced the assembly length, average contig length and N50 of the Racon outputs. The most striking effect of Medaka can be seen in the Figure 4.10*d* with a marked reduction in the length of contigs resolved for the assembly dataset. While Pilon was used for a smaller set of isolates (*due to lack of matched Illumina short reads*), the median assembly length for the seven isolates [Appendix Table E.10] with matched Illumina short reads is higher than other improvements employed by Racon and Medaka; while also including recovery of the average contig length per assembly [Figure 4.10] that was reduced in the Medaka correction step.

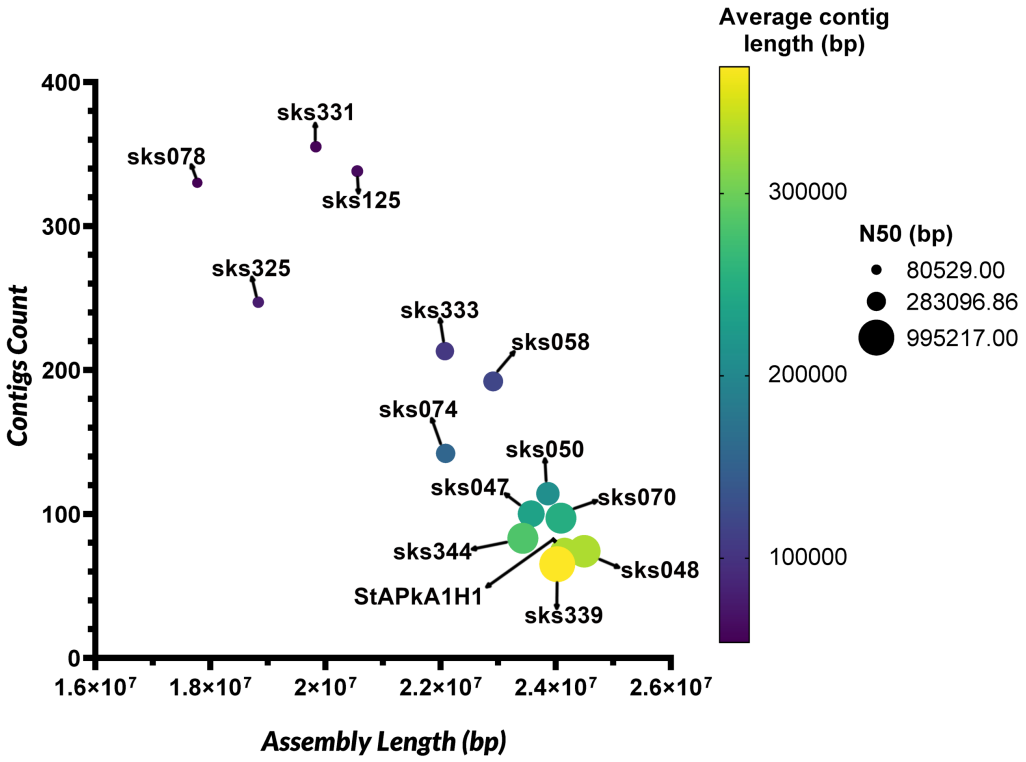


Figure 4.9: The relationship of the assembly length, contig count, average contig length and N50 of raw *de novo* draft assemblies on Flye. The outputs of Flye were assessed using assembly-stats. Each point describes a successful isolate draft assembly with the size of the circle indicating the N50 and the colour determining its average contig size.

The impact of the three *de novo* assembly improvement softwares are only apparent using quality assessment. Here, BUSCO scores reveal a consistent increase in genome completeness with median completeness scores of 66.80 %, 65.40 %, 71.80 %, 95.70 %, 86.55 % for the initial Blobtools-checked assembly, Racon, Medaka, Pilon and the annotated genomes, respectively [Figure 4.11]. BUSCO completeness quality was improved with the combined Racon and Medaka processes resulting in a 15.16 % average increase to the BUSCO completeness score (*not shown*, Appendix Table E.12). Furthermore, after carrying out all assembly quality improvements [Figure 4.2], BUSCO scores improved by an average of 27.66 % in comparison to the BlobTools-cleaned *de novo* assembly.

Pilon-corrected draft assemblies show the highest BUSCO scores with the smallest

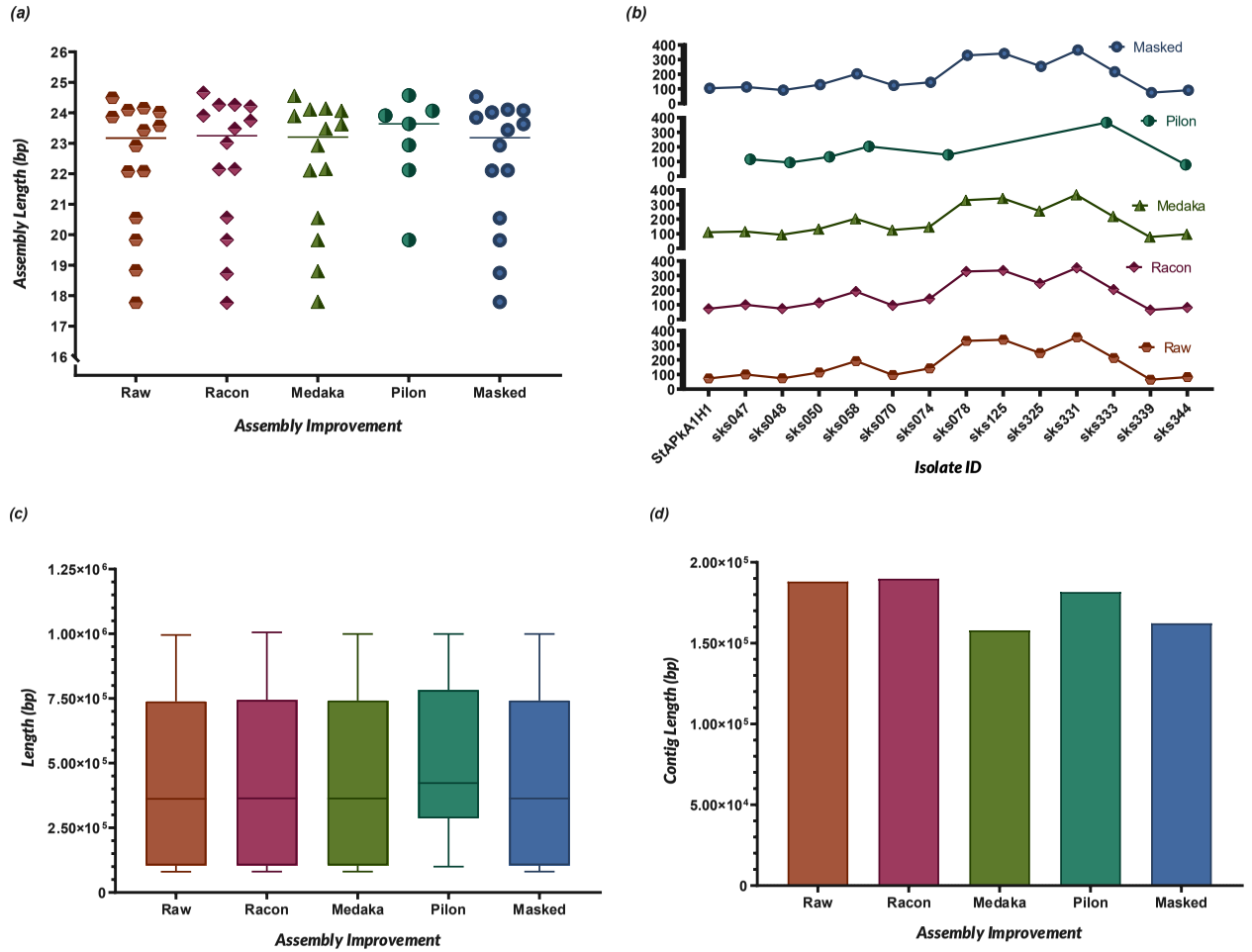


Figure 4.10: Step-wise descriptive statistical metrics of successful draft assemblies as they are improved. (a) The total sequence length and (b) the number of contigs of draft assemblies in each improvement dataset going from raw draft assembly outputs from Flye to the final masked assembly outputs of the repeatmasking pipeline. (c) shows the range of the N50 for each dataset with the medians identified across the box plot. (d) is the average contig length for each dataset. The full numeric values for each dataset are presented in Appendix Table E.10

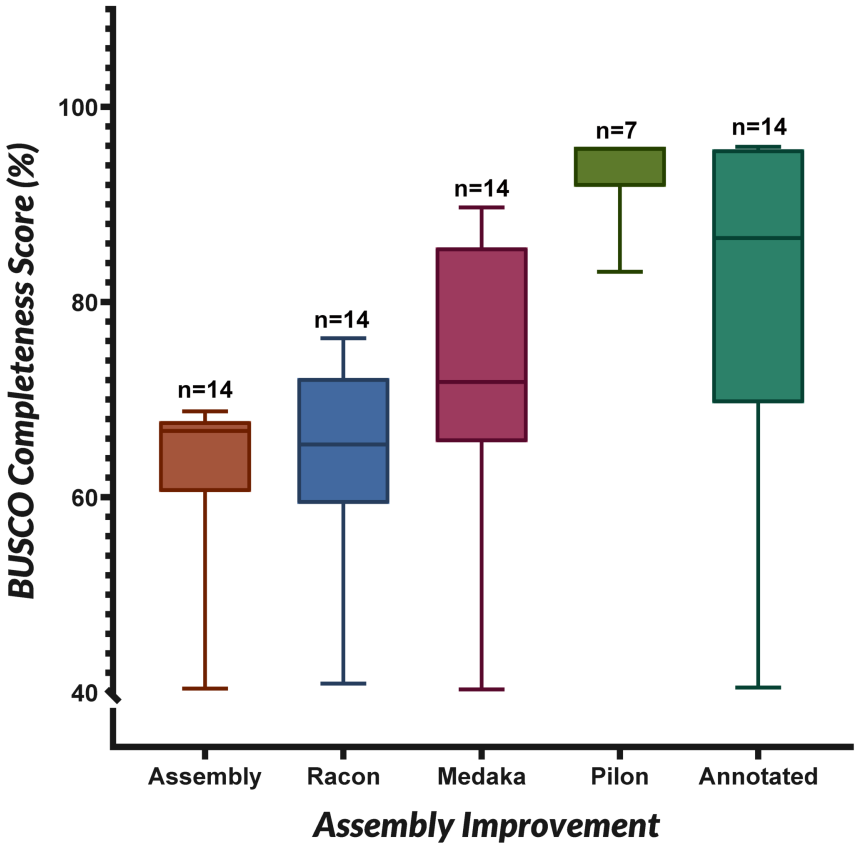


Figure 4.11: BUSCO scores for successful assemblies as they are improved in a step-wise manner. Quality assessment for completeness using BUSCO was done for each improvement step from the BlobTools cleaned assembly (Assembly) through the Racon, Medaka and Pilon polishing and correcting processes. After annotation on the Companion webserver, the final draft genomes were also BUSCO assessed.

variance. However, the small sample size within the Pilon dataset also aids in the precise spread of BUSCO scores. The annotated dataset is a combination of the Medaka-corrected and Pilon-corrected assemblies. Thus the presence of the Pilon-corrected dataset acts as a means of improving the precision for the entire dataset – in conjunction with Medaka-corrected assemblies [Figure 4.11]. Pilon-corrected assemblies show a significant positive correlation (*p=0.00119*) to the Annotated dataset [Appendix Table E.6]. Although the same significance is not present between the Medaka and Pilon dataset (*p=0.20833*), likely due to a difference in sample size. Overall, the post-assembly manipulations implemented (*Racon, Medaka, Pilon, Repeatmasking and Chimeric contigs removal*) resulted in a significant positive correlation for improved BUSCO scores (*p=0.00289*) [Appendix Table E.6].

4.4.3 Apicoplast and Mitochondrial sequences Circularisation

Apicoplast (API) and Mitochondrial (MIT) sequences extracted from the Flye assembled *de novo* genomes (after Medaka/Pilon) largely failed to circularise [Appendix Table E.14 - Appendix Table E.15]. Indeed no API sequences circularised while only four MIT isolates successfully circularised, with an average length of 6094.5 bp [Appendix Table E.14 - Appendix Table E.15]. The subsequent Canu *de novo* prokaryotic assembly resulted in greater success, however, like the Flye assembler, API sequences generated using Canu were not circularised [Appendix Table E.16].

On the other hand, the majority of Canu-assembled MIT sequences circularised successfully with lengths similar to that of the PKNH reference mitochondrial (MIT) [Appendix Table E.17]. No clear reason is apparent for the sks070 and sks339 failing to assemble a mitochondrial genome via Canu, with both isolates possessing $>$500-fold input read coverage [Appendix Table E.17]. Contrasting this, API assembly failures are based on the input read coverage, as reported by the assembler during its initialisation process.

Isolates with MIT sequences which were unable to be circularised (*sks078, sks125, sks344*) possessed greater than one contig in the initial Canu assembly. Each isolate assembly was manually curated to retrieve only the contig which aligned to the PKNH

reference MIT genome. A second attempt for circularisation resulted in none of the three isolate MIT sequences being successfully circularised. No such manual curation was implemented for API sequences.

Table 4.3: Annotated gene content of sequenced apicoplast (top) and mitochondrial sequences (bottom)

API Isolates	Length	# Genes	# rRNA	# tRNA
PKNH (Ref.)	30638	57	2	25
StAPkA1H1	29531	51	2	16
sks048	34377	72	4	29
sks339	34407	72	6	32

MIT Isolates	Length	# Coding Genes	# tRNA	# rRNA
PKNH (Ref.)	5957	3	–	34
StAPkA1H1	6082	4	2	–
sks047	5960	4	1	–
sks048	6083	4	1	–
sks050	6082	5	1	–
sks058	6081	4	1	–
sks074	6072	3	1	–
sks078	5931*	5	1	–
sks125	5993*	5	1	–
sks325	6034	4	1	–
sks331	6085	4	1	–
sks333	6082	4	1	–
sks344	5992*	4	1	–

Apicoplast sequences (top) failed to be circularised by Circlator [29], thus Canu outputs were annotated using Prokka [31]. Mitochondrial sequences (bottom) successfully circularised, with unsuccessful isolates possessing similar lengths to the reference genome. Genome content was assessed using AGAT [63].
– no results
* - Isolates which were manually curated to generate a single contig which aligned and covered the full span of the PKNH reference MIT sequence.

Annotation of the Circlator outputs (*and Canu outputs for apicoplast sequences*) resulted in Table 4.3. From this, the experimental StAPkA1H1 apicoplast sequence is more similar to the PKNH reference apicoplast sequence, with both apicoplast genomes sharing ribosomal RNA values, with a similar number of genes [Table 4.3, top]. The

patient isolates sks048 and sks339 are more similar to one another, with equal gene counts. General Feature Format (GFF) files generated by Prokka require further manual analyses for full protein identification and characterisation.

4.4.4 RepeatMasking

The masking of repetitive elements within the apicoplast/mitochondrion-free isolate draft assemblies resulted in Table 4.4. Here, the proportion of Class I retroelements masked remains similar between isolate genomes, with minor variances within each repetitive element type. However, the largest single interspersed repeat element subclass was DNA elements with a variance of 2.54 [Table 4.4]. Overall, an average of 14.3 % of the genome was masked in each isolate genome, with 18 % being masked in sks048 [Table 4.4].

Table 4.4: Percentage of each isolate draft assembly masked

Isolate	% masked
StAPkA1H1	17.2
sks047	15.68
sks048	18.29
sks050	16.11
sks058	14.38
sks070	16.79
sks074	13.02
sks078	9.11
sks125	11.8
sks325	10.2
sks331	12.57
sks333	13.81
sks339	16.73
sks344	15.04

Masking metrics were measured by RepeatMasker for Class I retroelements prior to concatenating the results of RepeatMasker, TransposonPSI and LTRHarvest. Final percentage masked was not calculated.

4.4.5 Scaffolding and De-chimerisation

Initial scaffolding of the masked draft assemblies (*assemblies which have been repeat-masked*) was directly implemented within the Companion workflow through the use of the included ABACAS2 module [42]. However, results from these preliminary runs (*not shown*) revealed translocations of large sections of chromosomes within other chromosomes. A particular example is chromosome two of sks047 and sks048 being scaffolded into chromosome 8 of their preliminary draft genomes. A BLAST alignment of the reference PKNH chromosome 2 against the sks047 and sks048 preliminary genomes showed an alignment similarity seen in Table 4.5.

Table 4.5: Selected chromosomes from preliminary draft genomes of sks047 and sks048 aligned against the reference PKNH chromosome 2

Isolate	PKNH chromosome*	Aligned isolate chromosome**	% similarity
sks047	2	2	56
sks047	2	8	33
sks047	2	14	22
sks048	2	2	38
sks048	2	8	58

Chromosome 2 of the reference PKNH genome was aligned against patient isolate sks047 and sks048 genomes and alignment similarity is shown, with alignments to different patient isolate chromosomes evident. The patient genomes used here were preliminary assemblies during early stages of the annotation pipeline development.

* – Pain et al. [9] *P. knowlesi* reference genome

** – The chromosome within the patient isolate genome which has aligned against the reference chromosome.

The translocation seen within the preliminary draft genomes showed evidence of chimeric contigs which could not be resolved using ABACAS within Companion. Hence, the introduction of Ragtag resulted in the removal of chimeric contigs, as evidenced in the lack of the translocations described in Table 4.5 being present downstream annotated genomes. Additionally, Ragtag provided early scaffolding of the assemblies prior to final scaffolding within Companion. Here, Ragtag joined contigs to form scaffolds of greater lengths with the introduction of 'N's' used to bridge gaps between contigs [Table 4.6].

Table 4.6: Descriptive statistical metrics of masked and de-chimered assemblies

Isolate	Assembly Length		Contigs		N50	
	Masked	*De-chimered*	*Masked*	*De-chimered*	*Masked*	*De-chimered*
StAPkA1H1	2.41E+07	2.44E+07	105	71	690944	1896268
sks047	2.36E+07	2.42E+07	113	69	556526	2085900
sks048	2.45E+07	2.48E+07	92	50	782439	2207148
sks050	2.38E+07	2.45E+07	129	67	423137	2133551
sks058	2.29E+07	2.41E+07	203	112	304054	2086040
sks070	2.41E+07	2.45E+07	124	87	741531	2176838
sks074	2.21E+07	2.34E+07	145	52	286541	2087157
sks078	1.78E+07	2.16E+07	329	48	80983	1968374
sks125	2.05E+07	2.30E+07	342	113	106823	2082141
sks325	1.87E+07	2.26E+07	254	56	98909	1846231
sks331	1.98E+07	2.29E+07	366	122	99889	2029450
sks333	2.21E+07	2.36E+07	217	105	259183	2119152
sks339	2.40E+07	2.42E+07	75	50	999134	2170181
sks344	2.34E+07	2.39E+07	91	41	743388	2131793

After the action of Ragtag to remove chimeric contigs, Ragtag also carried out scaffolding of the contigs using gap-inference. De-chimered assemblies resulted in longer assembly lengths due to the addition of N's between contigs within a scaffold. Construction of a chromosome/pseudo-chromosome was not possible here; however, the N50 for each assembly indicate scaffolds of large sizes dominate each assembly, with contigs of smaller lengths unable to be resolved.

4.4.6 Comparative Genomics

Genome Annotation and Gene Content

Annotation by Companion resulted in all 14 draft assemblies being resolved into annotated genomes with 15 chromosomes each [Table 4.7]. 14 chromosomes correspond

to the nuclear chromosomes of *P. knowlesi* while an additional 'bin' chromosome (chromosome 00) containing sequences which could not be placed into a chromosome by Companion was added. Genome length varied between 2.1 x 10^7 - 2.5 x 10^7 [Figure 4.12]; all fitting within previously described genome lengths for *P. knowlesi* [9, 64]. The N50 length – *an indicator of the length of the constituent chromosomes* – of all isolates are similar to the PKNH reference, with the greatest deviation seen in sks325 and sks078 [Table 4.7].

Table 4.7: Descriptive statistical metrics for the complete annotated experimental and patient isolates

Isolate	Assembly length*	Chr.	N50 (bp)	# Ns	Gaps	Genes^^	Pseudo genes	Gene Density
PKNH	24359384	15	2162603	11381	98	5327	12	22.05
StAPkA1H1	24391456	15	2132014	288598	127	5358	973	18.16
sks047	24176682	15	2085900	544896	109	5327	441	20.25
sks048	24815742	15	2207148	283076	84	5398	494	19.83
sks050	24491956	15	2154880	650306	128	5342	445	20.03
sks058	24119513	15	2114396	1188005	207	5240	423	19.97
sks070	24529770	15	2184201	452898	118	5404	1581	15.76
sks074	23396003	15	2104427	1283428	138	5102	854	18.20
sks078	21558481	15	1975334	3762686	317	4837	2211	12.08
sks125	23053158	15	2106383	2507703	331	5012	1989	13.03
sks325	22628919	15	1891393	3880205	244	4758	2462	10.12
sks331	22900901	15	2029450	3083330	362	4624	720	16.99
sks333	23616369	15	2119152	1510500	220	5333	2131	13.49
sks339	24247965	15	2170181	234672	74	5401	734	19.47
sks344	23873823	15	2131793	437495	86	5702	2591	13.08

Descriptive statistics generated using assembly-stats, pomoxis and AGAT [63]. Gene counts and number of pseudogenes were extracted from the GFF3 annotation file produced by Companion. The *P. knowlesi* PKNH reference genome is provided for comparison.

* - Assembly count excludes the separated apicoplast and mitochondrial sequences

^^ - Gene counts are a combination of the number of genes with a function and total pseudogenes. Non-coding genes are excluded from the count.

The number of N's added to bridge gaps in the isolate genomes are generally associated with the number of gaps present in each assembly, with a significant positive correlation

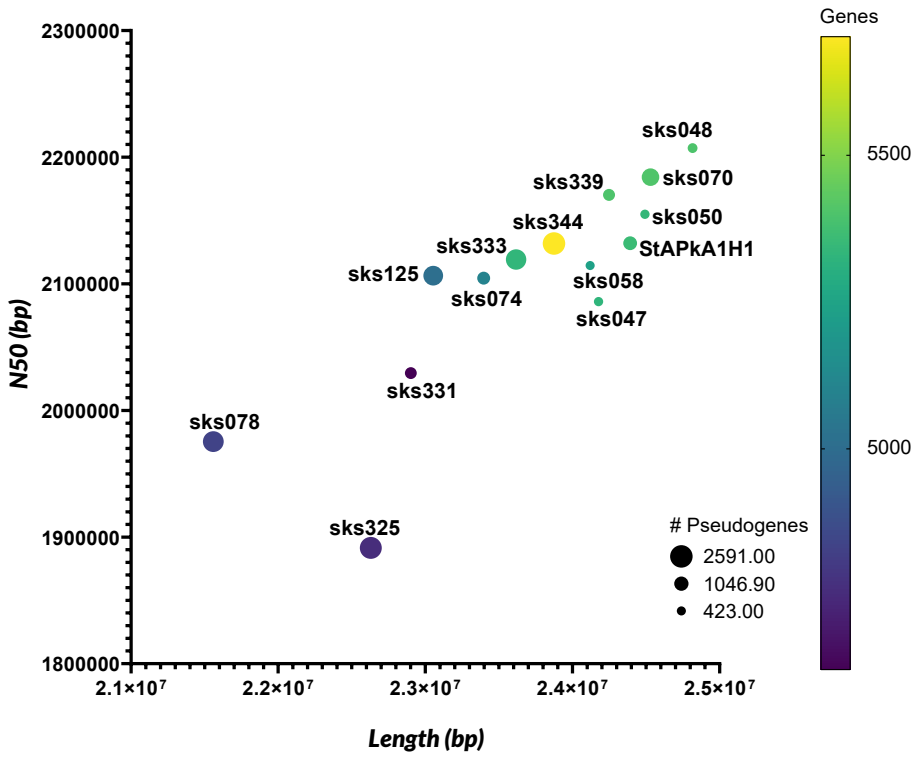


Figure 4.12: A representation of the genome length, N50, number of genes and pseudogenes present in the isolate genomes after prediction and annotation on Companion. A colour gradient representing the number of genes from purple (low) to yellow (high) as well as size gradient for pseudogene content is presented. Larger pseudogenic content are more likely in short length genomes while longer genomes possess more genes.

between gaps and N count (p=4.04 x 10^{-6}, Appendix Figure E.1). On the other hand, the assembly length, N50 and gene count all show a significant negative correlation to the N count and the number of gaps present in the assembly [Appendix Figure E.1]. Isolates with the smallest input read coverage [Appendix Table E.10] possess the most gaps and subsequently, more N's added [Table 4.7]. The starting coverage remains significantly positively correlated to other metric variables like the assembly length (p=0.0002), N50 (p=0.002) and gene count (p=0.003) whilst being significantly negatively correlated to the N count and number of gaps in the isolate genomes [Appendix Figure E.1].

The PKNH reference genome [9] possessed 5327 genes (*depending on the version used*), made up of complete protein-encoding genes as well as pseudogenes with a predicted

function. Such pseudogenes are often associated with missing start/stop codons or other insertions and deletions within the gene. The 14 isolate genomes (13 patient + 1 Cultured experimental) possess an average of 5203 genes made up of complete coding genes and the total pseudogenes detected within each assembly [Table 4.7]. Non-coding genes such as small nuclear RNA and transcription RNA were excluded from the total count for the generated isolate genomes; however, this information was unavailable for the PKNH reference genome. Pseudogenes are significantly positively correlated to the assembly length (p=0.037) with no significant correlation to other variables.

Genome Structure

Gaps present in the generated isolates exceed those reported in the PKNH reference genome, resulting in higher N counts within our generated *P. knowlesi* genomes. While Ragtag successfully carried out the gap inference as expected during its scaffolding process, some gaps in the assemblies are likely due to sequences that could not be placed in the appropriate loci. These are often highly variable genes or sequences that do not singularly align to an individual chromosome during contig – and subsequently, scaffold – construction. In such cases, bin chromosomes can reveal the presence of preferential scaffolding (*and lack thereof*) for certain chromosomes. Synteny plots of isolate chromosome 00 against the entire PKNH reference genome sequence reveal no such clustering. However, an exception is seen in Figure 4.13 where the StAPkA1H1 draft genome possesses sequences in its chromosome 00 that appear highly clustered to the PKNH reference chromosome 00. On the other hand, patient isolates such as sks047, sks048, sks125, sks325 and sks339 do not display any observed clustering, with sequences within the bin chromosome dispersed across the genome [Appendix Figures E.2,E.3]. This is further exhibited in the coverage values reported per chromosome, with chr00 of StAPkA1H1 showing coverage of 0.33 (33 %) of the PKNH reference chr00, while patient isolates (*sks****) show $<$1 % coverage to the PKNH reference chr00 [Appendix Table E.18]

Apart from the bin chromosomes (Chr 00), true *P. knowlesi* chromosomes (Chr 1 - 14) of the isolate draft genomes and the PKNH reference genome show high levels of synteny with average chromosome coverage $>$79 %, and six isolates (sks047, sks048, sks050,

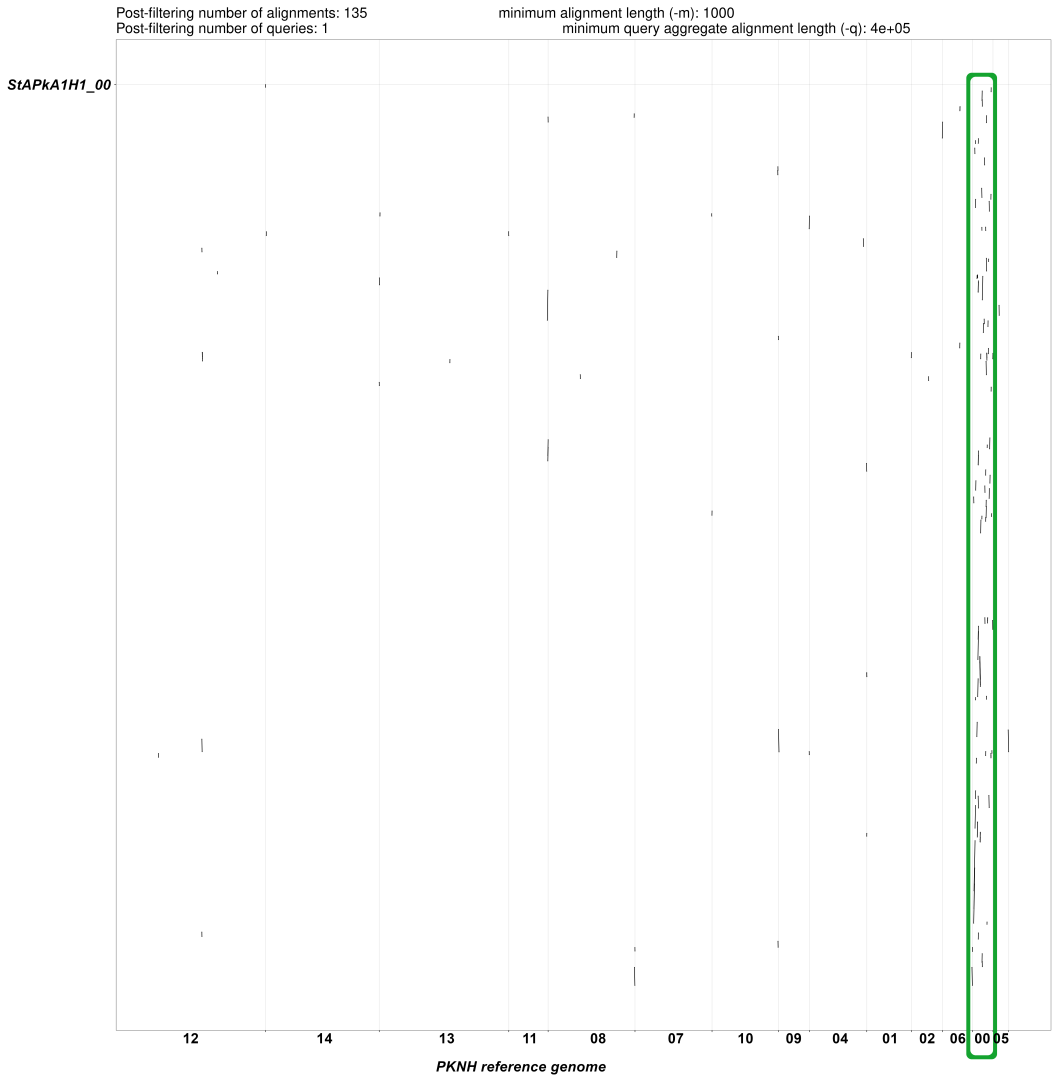


Figure 4.13: Alignments of chromosome 00 (bin) for StAPkA1H1 against the whole PKNH reference genome with a 1 Kb **alignment length filter.** The 'bin' chromosome contain sequence fragments that could not be confidently resolved into a particular chromosome during the scaffolding process. Here, StAPkA1H1 shows a concentration of sequences aligned to the PKNH 'bin' chromosome 00 (green box). In comparison, no clustering is evident in sks047() and sks048 ().

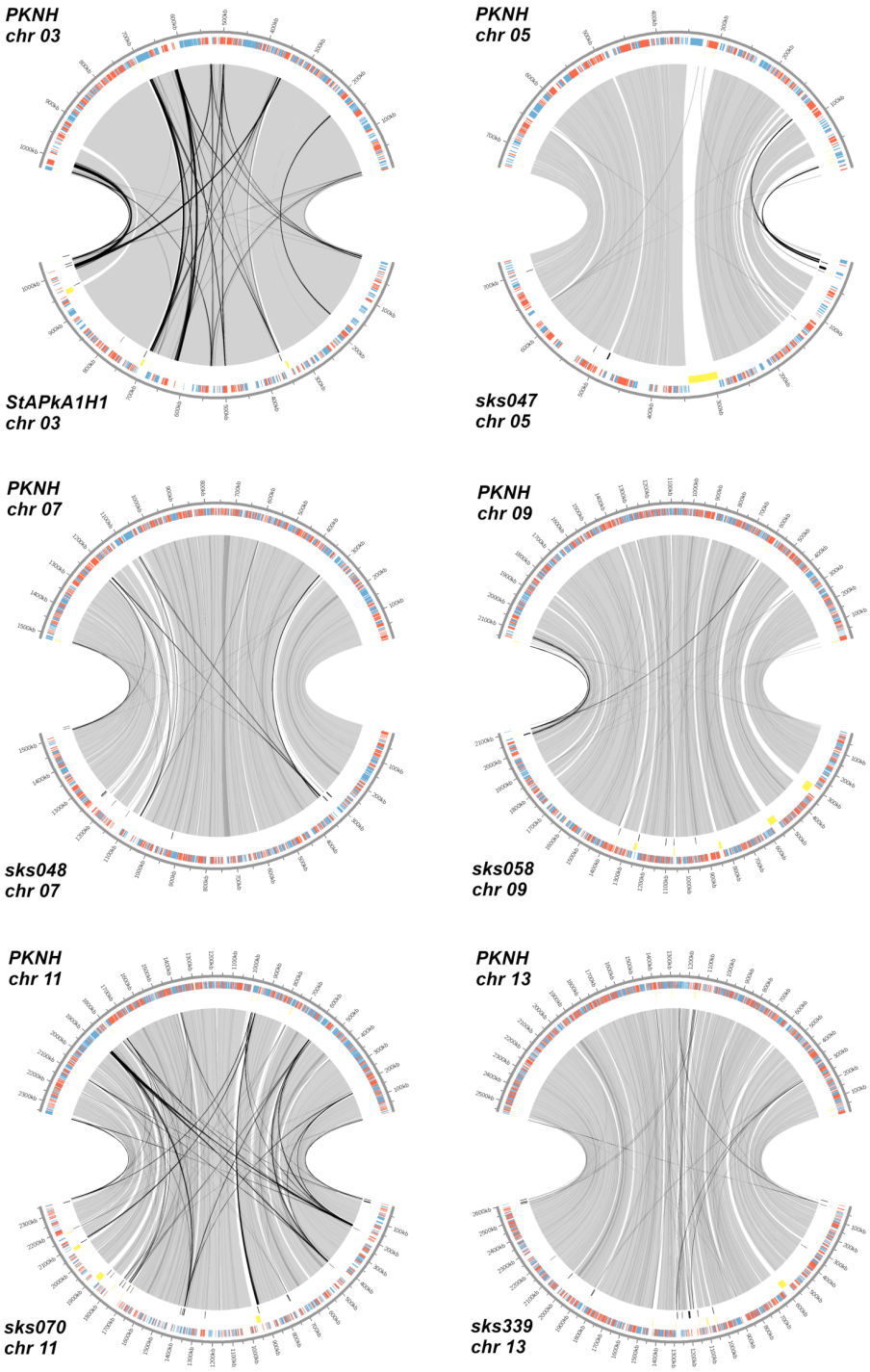


Figure 4.14: Circos plots of synteny showing chromosome alignments of representative isolates against matched PKNH reference chromosome. Plots were generated on Companion as part of the prediction, annotation and analysis process. The PKNH reference genome is shown at the top of the plot and the isolate genome at the bottom. The central grey ribbons represent regions of similarity. The blue and red track/bars represent the forward and reverse strands, respectively. The yellow track/bars represent gaps; singletons are the black track/bars and missing core genes as green track/bars.

sks058, sks070 and sks339) showing >85 % per-chromosome coverage [Figure 4.14, Appendix Table E.18]. The lack of clustering suggests that no single chromosome within the patient isolate were more challenging to scaffold after de-chimerisation. However, the high average coverage (and synteny) reported between StAPkA1H1 and the PKNH reference genome also suggests the structure of the experimental line was more similar to the reference genome [Figure 4.14].

Table 4.8: Orthologous clusters of genes identified in all isolate draft genomes

Isolate	Shared Orthologous clusters w. reference	Unique orthologous clusters	Singleton clusters
StAPkA1H1	4172	3	62
sks047	4666	9	82
sks048	4664	11	100
sks050	4669	16	108
sks058	4630	13	95
sks070	3530	13	150
sks074	4058	8	134
sks078	1846	2	699
sks125	2581	4	340
sks325	1375	4	895
sks331	3747	4	113
sks333	2711	8	354
sks339	4438	12	95
sks344	2355	8	609

Genes that are identified as orthologs and paralogs within the *P. knowlesi* species based on the PKNH reference genome are grouped together into distinctive clusters using OrthoMCL. Here, clustering occurs with at least two genes which share similarities. Shared clusters are clusters of genes present in both the isolate draft genome and the PKNH reference genome, while unique clusters are grouped genes which are sample/genome specific. Singletons are genes which are not similar to any other genes and thus occupy an individual ortho-group. Isolates with <3500 shared orthologous clusters (red) to the PKNH reference genomes are removed from further analyses.

The high synteny, coverage and gene counts reported for each isolate genome are further supported by the orthologous gene clustering from OrthoMCL via Companion. Here, the isolates share an average of >3500 orthologous clusters whilst maintaining ~8 unique

clusters [Table 4.8]. Four isolate genomes (sks047, sks048, sks050 and sks058) report >4600 shared orthologous clusters to the PKNH reference genome, which is larger than the 4100 clusters reported for StAPkA1H1 [Table 4.8]. Furthermore the number of singletons reported for each genome is significantly negatively correlated to the input coverage (Spearman's rank correlation test: p=0.0012, *rho=-0.759*) indicating that increasing input coverage decreases the number of non-clustered singleton genes. However, isolates that report <3500 shared orthologous genes clusters with the PKNH reference genome were eliminated from further analyses. Once eliminated, the remaining nine isolates (*henceforth the 'curated dataset'*) still show a significant negative correlation between the input coverage and the number of singleton gene clusters resolves (Spearman's rank correlation: p=0.029, *rho=0.626*; Wilcoxon signed-rank: p=0.0020).

Genome-wide alignments of the isolate draft genomes (curated dataset) against the PKNH reference genome reveals similar chromosomal structure and orientation to all generated genomes. However, large gaps within and between chromosomes of generated genomes persist, with dotplot visualisations indicating the presence of frameshifts and other large structural variations between the isolate draft genomes and the PKNH reference genome. This variation is predominantly present within the patient isolates, with observations of insertions, deletions, inversions, and duplications present within the patient isolates in comparison to the PKNH reference genome [Figure 4.15].

Such variations are widespread across the curated dataset genomes, including frameshifts which may affect the correct characterisation and annotation of genes. This has led to a mischaracterisation of complete genes as pseudogenes, contributing to the higher pseudogene count reported for the patient isolate genomes [Table 4.7]. As such, mean annotated gene density for the PKNH reference genome is 22.05 genes per 100 Kb, while the experimental line StAPkA1H1 reported 18.15 genes per 100 Kb [Table 4.7, Figure 4.16]. Draft genomes generated from patient isolates report lower gene density than the PKNH reference genome and the StAPkA1H1, with the patient isolates reporting an average of 16.33 genes/kb. However, visual representations of the gene density [Figure 4.16] display a similar pattern of genes within the chromosomes of the reference, experimental and patient genomes. The impact of the pseudogenic count is present in the gene density [Table 4.7], where pseudogenes show a strong and significant negative correlation between pseudogenes and gene density (Spearman's rho=-0.891;p=0.001;

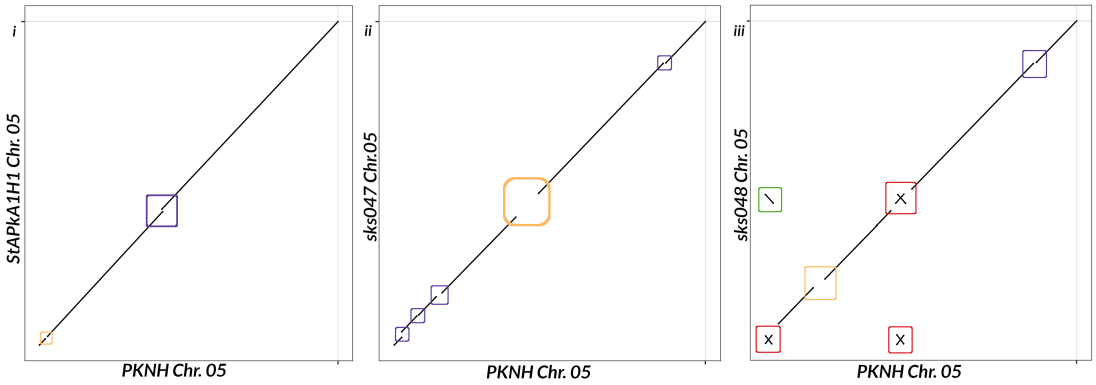


Figure 4.15: Dot plots showing draft genomes aligned against the PKNH reference genome with a minimum alignment of 10 Kb**.** Chromosome 5 is given for StAPkA1H1 (i), sks047 (ii) and sks048 (iii), as an example of variation observed within chromosomes of the generated isolate genomes. Here frameshifts are outlined in purple, gaps outlined in orange, inversions outlined in green and inverted repeats in red.

Wilcoxon signed-rank: p=0.0020), indicating that the low gene density reported is directly due to the pseudogenic count of each isolate. Taken together, this suggests gene density may increase as pseudogenes are decreased, potentially via manual pseudogene curation.

4.4.7 Multigene Families of *Plasmodium knowlesi*

The annotation of the genomes by Companion allowed a search for known multigene families of *P. knowlesi*. From this, biologically relevant genes and multigene families like the merozoite surface protein (MSP) and the CSP were confirmed to be present in the generated genomes. StAPkA1H1 revealed further similarities to the PKNH reference genome, with all but one of the representative multigene families having equal members to the reference [Table 4.9]. Here, the MSP gene family with 13 members in the PKNH reference genome, was reported to have $\leq$10 members within the generated experimental and patient genomes [Table 4.9]. Further deviation from the reference is observed in the resolution of the reticulocyte-binding protein (or the *P. knowlesi normocyte binding protein (pknbp)*), which, with two copies, were described to be associated with dimorphism in *P. knowlesi* [65]. However, of the 9 annotated genomes,

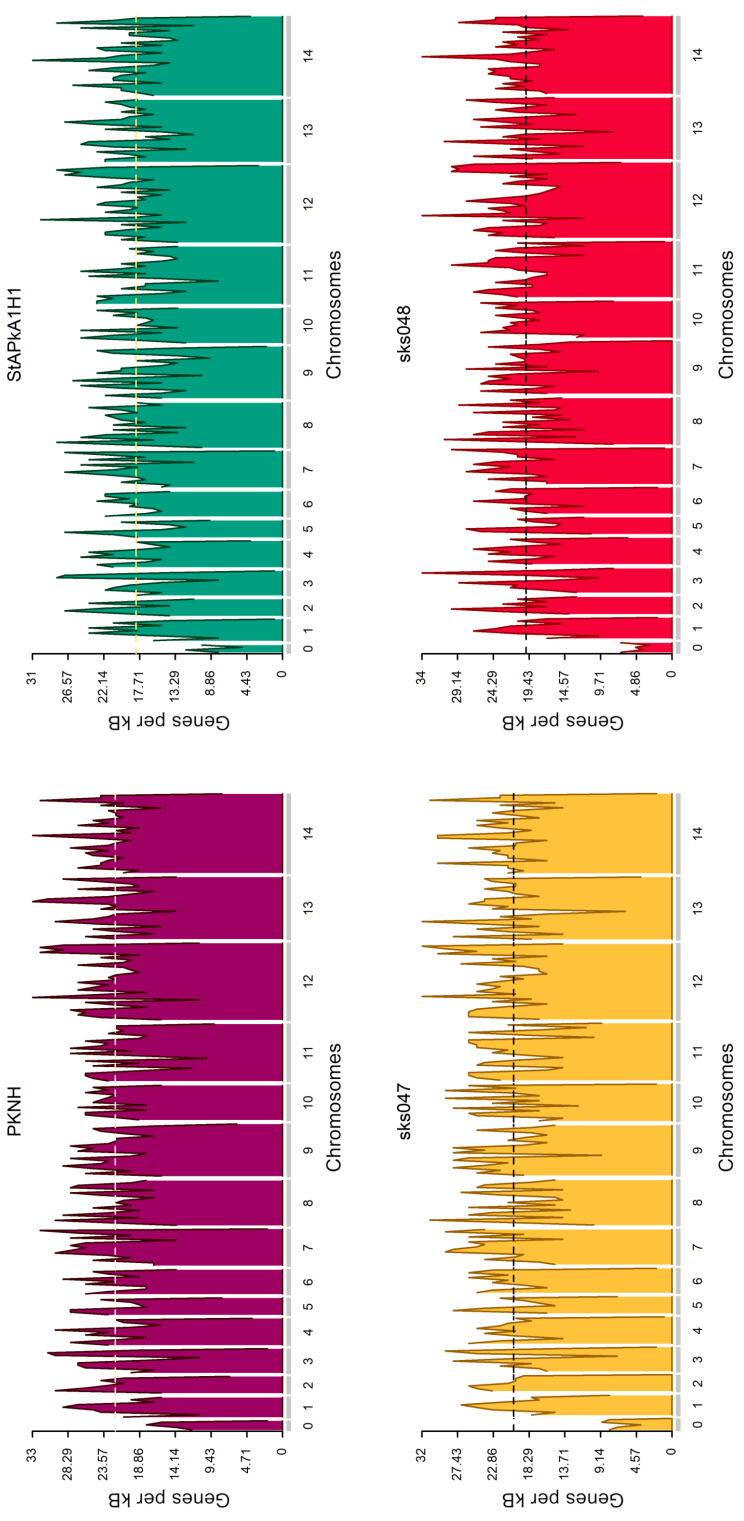


Figure 4.16: Gene density plots for the P. knowlesi PKNH reference genome, StPkA1H1, sks047 and sks048 draft genomes. Gene density was calculated based on the number of identified genes within a sliding window of 100 Kb. Mean density (dashed line) shows the PKNH reference genome to have 22.05 genes per 100 Kb, StAPkA1H1 to have 18.16, sks047 to have 20.25 and sks048 with 19.83 genes/100 Kb. Gene density for the remaining isolates of the curated dataset are presented in the Appendix [Appendix Figure E.4]. Plots were generated using karyoploteR [60].

Table 4.9: Multigene family members characterised in the generated experimental and patient genomes

Genes	PKNH	StAPkA1H1	sks047	sks048	sks050	sks058	sks070	sks074	sks331	sks339
CSP	2	2	2	2	2	2	2	2	2	2
CLAG	2	2	2	2	2	1	2	1	1	2
DBP	3	3	3	3	3	3	3	1	0	3
ETRAMP	9	9	9	9	9	9	9	7	5	9
TrpRA	29	29	30	29	32	29	30	16	1	33
MSP	13	10	10	10	10	10	10	10	9	10
PKNBP	2	2	2	2	2	2	2	0	0	2
SPIAP	2	2	2	2	2	2	2	1	1	2
ABCtrp	15	15	15	15	16	15	15	15	14	15

A representative selection of multigene families known to be present in *Plasmodium knowlesi* genomes and are deemed to be biologically significant in the function and activity of the parasite. Gene counts were manually determined from the annotated General Feature Format (GFF).

ˆ - CSP = *Circumsporozoite protein*; CLAG = *Cytoadherence linked asexual protein*; DBP = *Duffy binding protein*; ETRAMP = *Early transcribed membrane protein*; TrpRA = *Tryptophan-rich antigen*; MSP = *Merozoite surface protein*; PKNBP = P. knowlesi *Reticulocyte binding protein*; SPIAP = *Sporozoite invasion-associated protein*; ABCtrp = *ABC Transporter*

two isolates (sks074, and sks331) reported no *pknbp* gene [Table 4.9] present in their genomes. Both of these genomes with absent *pknbp* genes were generated from the lowest input read coverage (*Pk-input reads*), with sks074 and sks331 having had 10.77 and 12.23-fold input coverage prior to the *de novo* assembly [Appendix Table E.10].

Schizont-infected Cell Agglutination variant (SICAvar) genes

The investigation into multigene families of *P. knowlesi* allowed the search for prominent multigene families such as the *SICAvar* and the *kir* genes. Of the 9 annotated genomes of the curated dataset, six possessed greater *SICAvar*s than the PKNH reference genome, with these six isolates (StAPkA1H1, sks047, sks048, sks050, sks070 and sks339) possessing an average of 152 annotated *SICAvar* genes [Table 4.10]. As described by Pain et al. [9], *SICAvar* type I genes vastly exceed *SICAvar* type II genes, with the patient isolates consistently reporting less type II genes than the PKNH reference genome [Table 4.10].

Table 4.10: Member gene count of the *SICAvar* and *kir* multigene families

Isolate	Input Coverage	Interspersed Repeats		SICAvars	
		KIRs	***Total PIRs***	***Type I***	***Type II***
PKNH (Ref.)		56	61	89	20
StAPkA1H1	226.09	50	58	191	15
sks047	71.12	26	47	115	9
sks048	65.13	25	47	153	7
sks050	17.36	23	45	115	5
sks058	40.25	29	44	87	4
sks070	107.77	27	49	131	4
sks074	10.77	31	42	54	4
sks331	12.23	18	28	42	5
sks339	154.02	40	58	163	5

Complete *SICAvar* and *kir* genes present in the generated and annotated genomes of the curated dataset. *SICAvar* gene fragments and domains are not included. Multigene families are distributed across the genome with the *SICAvar*s and *kirs* associated with high variability. *SICAvar*s are consituted by a combination of three domains. *kirs* are better characterised and belong to a collection of genes – *Plasmodium interspersed repeats* known to be orthologous within human-adapted *Plasmodium* spp. The orthologous gene forms of the *kir* genes were also annotated separately by Companion. The *vir* or the *Plasmodium vivax*-interspersed repeat genes are associated with *Plasmodium vivax* while the *pir* genes are *Plasmodium* interspersed repeats which is the superfamily the *kir* and *vir* genes are classified within.

On the other hand, seven identified genomes (6 patient + 1 experimental) reported higher type I genes present than the reference genome [Table 4.10]. *SICAvar* fragments and domains were also found to be present within all genomes, including within the bin chromosomes of the reference genome, however, the fragments were excluded from the total gene count and thus, not presented in the genic loci plots [Figure 4.17, Appendix

Figure E.6 - Figure E.11]. Unlike the PKNH reference genome, the bin chromosomes of the annotated genomes, including the experimental line StAPkA1H1, contained complete *SICAvar* genes [Figure 4.17, Appendix Figure E.6 - Figure E.11]. *SICAvar*s are observed to be randomly distributed across all chromosomes of the genomes. Furthermore, gene locus visualisation provides further evidence of the *SICAvar*s present within both the telomeric and core regions of each chromosome, with no apparent bias for the location.

Knowlesi interspersed repeats (KIRs)

For the *Plasmodium knowlesi* interspersed repeat (*kir*), an issue arose while validating the predicted number of *kir* genes. As previously mentioned, *kir* are members of the larger *pir* superfamily of interspersed repeats that have been identified in *P. knowlesi*. As such, there is a high level of similarity between interspersed repeat genes from different *Plasmodium* spp., resulting in certain interspersed genes in one *Plasmodium* species being annotated as *pir* gene family members from another *Plasmodium* species. This can be seen in the PKNH reference genome which identifies 56 *kir* genes and 5 *pir* genes [Table 4.10]. Similarly, StAPkA1H1 reported 50 annotated *kir* genes with an additional 9 *pir* genes. While the patient isolates were also observed to contain genes annotated as *pirs* and *kirs*, they also possessed genes annotated as *virs*, the interspersed repeats for *Plasmodium vivax*. Due to this, the number of genes uniquely identified as *kir* and the total number of genes annotated as part of the *pir* superfamily (*including kir genes*) are present in Table 4.10.

From this, StAPkA1H1 showed similar annotated *pir* gene members (n = 58) to the PKNH reference genome (n = 61), while the patient isolates reported consistently fewer *KIR* and *PIR* gene family members, with the largest cohort seen in sks339 (n = 58) [Table 4.10]. Indeed, the similarly high copy number retention observed between the PKNH and the StAPkA1H1, coupled with the parity of processing and manipulation between the StAPkA1H1 and the patient isolates, suggests gene retention of the *kir* genes within the StAPkA1H1 and the PKNH reference genome through artificial passage. It is possible that genes that would have naturally decayed in the patient isolates due to natural selection and gene loss, have been retained in the experimental line samples. Positional analyses of the *kir* genes reveal a preference for the core-fringe regions of the chromosomes. From

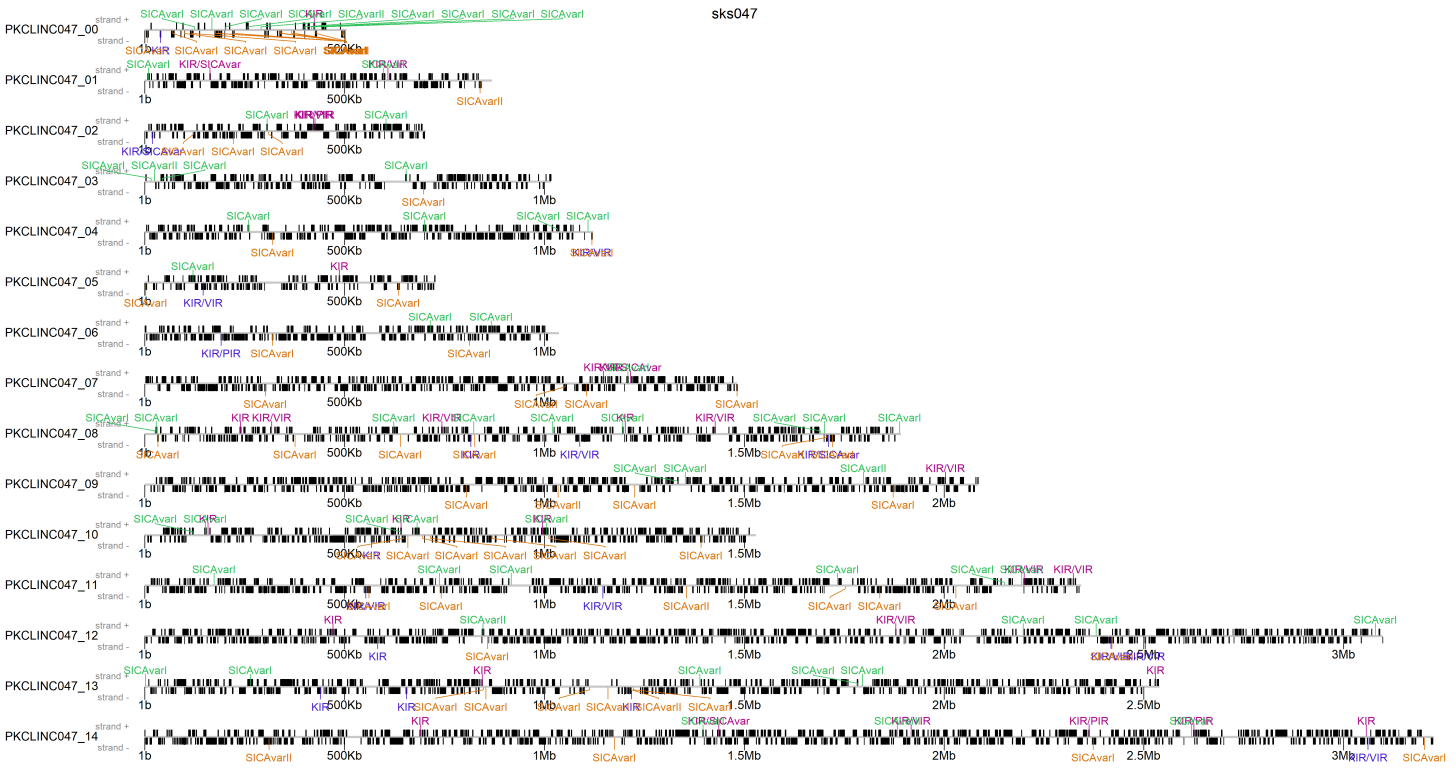


Figure 4.17: Genome loci of *SICAvar*, *kir*, *vir* and *pir* genes in sks047 to demonstrate gene spread and loci in respective chromosomes. Genome annotation was completed on Companion and the resulting General Feature Format (GFF) was parsed for annotated *SICAvar* and *kir* genes. *SICAvar* genes loci are represented in green and orange for the positive and negative strand, respectively. *Kir* genes are shown in plum and purple for the positive and negative strand respectively, per chromosome. Genome loci plots were generated using a custom R script and the karyoploteR package [60]. In addition to the sks047, similar plots of the *SICAvar*, *kir*, *vir* and *pir* genes for the PKNH reference genome, StAPkA1H1 experimental line and the other patient isolates are presented in the Appendix Figure E.5 - Appendix Figure E.11.

this dataset, only *KIR*s on the 3' ends of chromosomes 04, 09 and 14 are consistently in the telomeric regions while appearing to forgo the core regions of the chromosome [Figure 4.17]. However, to our knowledge, a strict definition of the intra-chromosomal region and their boundaries has not been performed for *Plasmodium knowlesi* and would be beneficial for further investigation. The diversity of annotation described here showed that for some isolates, >30 % of an isolate's annotated interspersed gene family members were alternatively labelled as *Plasmodium* spp. interspersed repeat (*pir*) or *Plasmodium vivax* interspersed repeat (*vir*). Without accounting for such alternatives, genes annotated uniquely as a *kir* gene family members can be up to 50 % reduced. sks339 reported the highest number of *kir* genes in the patient isolates whilst also showing 17 *vir* genes – similar to sks047 and sks048 [Table 4.10]. As described in chapter 4 subsection 4.4.7, sks074 and sks331 showed the lowest *kir* genes present in their genome. As such, sks074 and sks331 were eliminated from further analyses.

No clear evidence is given in the annotation files of the genomes to dictate why genes were characterised as either *vir* or *pir*. A neighbour-joining tree constructed using 1000 bootstraps and all genes identified as members of the *pir* superfamily (*including kirs*) from the PKNH, StAPkA1H1, sks047 and sks048, also showed no clear evidence for the diversity of annotations reported.

4.4.8 Structural Variation

The use of StAPkA1H1 allowed for parity between the patient isolates and the dataset's reference genome in the form of StAPkA1H1. With this, it was possible to determine that variants present were isolate-specific and not due to idiosyncrasies present in differing sequencing technologies. Raw variant calls by Assemblytics reported sks047 possessing a total of 864 structural variants (SVs) consisting mainly of deletions (INS/DEL ratio of 0.86) [Table 4.11]. After filtering for length (>49 bp), total variants reported was 856, although the INS/DEL ratio remained at 0.86 [Table 4.11]. While sks048 possesses similar total variants to sks047, the INS/DEL ratio is reported to be 1.33 and 1.34 for the raw and annotated variant subsets, respectively [Table 4.11], signifying a considerable bias for insertions in sks048. Assemblytics is not tuned to report other SVtypes like inversions (INV), translocations (TRA), duplications (DUP), thus these are reported as 0

(zero).

Table 4.11: Annotated structural variants identified using the assembly-based approach of Assemblytics

Variants	sks047						sks048					
	Total	***DEL***	***DUP***	***INS***	***INV***	***TRA***	***Total***	***DEL***	***DUP***	***INS***	***INV***	***TRA***
Raw	864	465	0	399	0	0	844	363	0	481	0	0
Annotated	856	460	0	396	0	0	839	359	0	480	0	0

Patient isolates sks047 and sks048 were aligned against a similarly generated experimental line (StAPkA1H1) using nucmer and variants were called using Assemblytics. Variants were filtered and annotated, however, further filtering could not be done due to no quality information being provided in Assemblytics outputs.

DEL - Deletions; INS - Insertions; DUP - Duplications; INV - Inversions; TRA - Translocations

The reads-based approach using the Oxford Nanopore structural variation pipeline (ONTSV) reported 1439 and 1593 total variant for sks047 and sks048, respectively [Table 4.12]. Unlike the assemblytics outputs, the ratio of insertion to deletion is variable with sks047 seeing an increase in INS:DEL ratio (from 0.73 to 0.75), indicating more deletions present in sks047 in comparison to the StAPkA1H1 reference used. On the other hand, raw sks048 reported an INS:DEL ratio of 0.83, however after filtration, annotation and further filtration, the INS:DEL ratio increased to 1.1, favouring insertions within sks048; supporting the same indication seen in the Assemblytics results.

Table 4.12: Annotated structural variants identified using the reads-based approach of the Oxford Nanopore structural variant pipeline

Variants	sks047						sks048					
	Total	***DEL***	***DUP***	***INS***	***INV***	***TRA***	***Total***	***DEL***	***DUP***	***INS***	***INV***	***TRA***
Raw	1439	834	0	605	0	0	1593	879	0	712	2	0
Annotated	1316	752	0	564	0	0	1398	667	0	731	0	0

Patient isolates sks047 and sks048 were aligned against a similarly generated experimental line (StAPkA1H1) and variants were called using ONTSV. Variants were filtered by length, quality and depth before annotation. Annotated variants were further filtered for variants which passed all filter parameters.

DEL - Deletions; INS - Insertions; DUP - Duplications; INV - Inversions; TRA - Translocations

The use of reads for variant calling allowed for granular investigations of specific SVtypes, as such after the initial cuteSV filtration (minimum size 30 bp and supported by two

reads), two inversions were reported in the raw sks048 cuteSV variant calls [Table 4.12]. However, further filtration using a read depth of 8 removed the inversions from the final annotated variants. Such complex SVtypes like INVs, DUPs and TRAs often failed due to lack of reads to support them. An unfiltered cuteSV variant file reported 5 INVs for sks047 and 6 INVs and 3 DUPs for sk048 (*not shown*). However, only the two previously mentioned inversions surpassed the set cuteSV thresholds.

To determine the presence of shared variant sites between sks047 and sks048 for both variant calling approaches applied, the annotated VCF file outputs were compared. From this, 101 shared variant sites were found between sks047 and sks048 in the assembly-based approach [Table 4.10]. On the other hand, only two similar sites were found between the two isolates using the reads-based approach (ONTSV). Of 101 similar variant sites found using the assembly-based approach, 68 variants were within annotated genes, including genes of interest like the *SICAvar* and *kir* multigene families [Table 4.10].

Table 4.13: Comparisons of structural variants between patient isolates and the experimental line

Approach	# Shared	# Unique to sks047	# Unique to sks048	# Shared within genes
Assembly-based	101	749	732	68
Reads-based	2	1381	1300	1

The count of shared, unique variants found between sks047 and sks048 based on using StAPkA1H1 as the reference genome. The number of shared variants within genes is also provided. Comparison was achieved after analysis using intersection (isec) function. Assembly-based SV calling approach utilised assemblytics to call variants between the isolate drat genomes (sks047, sks048) and the StAPkA1H1 draft reference genome. Reads-based SV calling approach used input reads of the isolate draft genomes against the StAPkA1H1draft reference genome to call variants with the Oxford Nanopore Structural Variation pipeline

4.5 Discussion

Through the use of parasite-enriched DNA extracted from thawed clinical *Plasmodium knowleis*-infected whole blood, Oxford Nanopore long-read whole-genome sequencing was carried out. Additionally, to act as a control for the sequencing, downstream data manipulation and eventual analyses, the PkA1H1 cultured experimental line was also sequenced [66]. From this, 16 sequencing experiments were carried out, consisting of 25 unique *P. knowlesi*-enriched isolates. The sequenced data were appropriately basecalled, demultiplexed, filtered, and human contamination was removed before *de novo* genome assembly.

Although human DNA content had been considerably reduced prior to DNA extraction [chapter 2 subsection 2.5.2], human DNA content was still present in the sequenced data. This included the PkA1H1 cultured isolate (*henceforth StAPkA1H1*), which had little to no human leucocytes in the culture medium. While it was known to be unlikely for the complete elimination of human content in the DNA and subsequent sequence reads, human aligned sequenced reads made up ~50 % of all reads generated per isolate. Although, even with such a significant presence of human contamination, isolates still retained enough *P. knowlesi* sequence data to carry out *de novo* whole-genome assembly. As such, given the archival nature of the whole blood used in this project, as well as the additional manipulation to deplete human leucocytic content, there remains sizeable *P. knowlesi* sequencing yields of adequate quality and length.

By far, the most significant determinant of assembly success was each isolate's input coverage, which was in turn influenced by the sequencing yield. While Flye can carry out genome assembly for low input coverage isolates (<10-fold), it became evident that the assembler flourished with isolates of higher input coverage. However, after removing human-aligned sequence data, no further human contamination was detected, following the preliminary pipeline development results. It remains essential for this contamination step to be present in the pipeline, acting as a 'sanity check' to determine the presence of any unforeseen contaminant organism from distinct phyla.

Indeed, the use of BlobTools in this manner would be unable to determine different species of Plasmodium or other members of the Apicomplexan phylum or strains of the

same species. An improvement to this step would implement capabilities to determine the classification level for comparison, i.e., species would be an option rather than phyla. Furthermore, the addition of an estimation for the multiplicity of infection would aid in determining the presence of different strains. However, estMOI, a tool to carry such a measure, was not suitable for this project due to being tuned for Illumina short reads. To our knowledge, no similar tool is available for long read parasite sequence data.

Interestingly, Flye successfully assembled genomes with considerably lower coverage than initially described, with genomes being generated with as little as 4.5-fold input coverage. Although, as expected, such genomes were of shorter overall length, lower quality and contained more constituent contigs. Shortcomings like these could not be sufficiently adjusted using the improvement techniques implemented. Racon, Medaka and Pilon's action was entirely evident in the quality of the resulting assemblies, seen in the BUSCO score improvements. However, the scale of quality improvements by Racon and Medaka appears to be considerably less than those by Pilon. However, Pilon indeed builds on Racon and Medaka's improvements, providing a noticeable increase in quality (as calculated by BUSCO). While nanopore long-read sequencing is a viable means of producing high-quality whole-genome assemblies, these genomes still possess errors even after consensus by Medaka. Furthermore, the positive action of Pilon was observed to occur whilst using impressively low Illumina read coverage, with as little as 5-fold Illumina short-read coverage providing a noticeable increase in quality.

With this, Pilon-polished genomes in this dataset further supports previously described work exhibiting the need for matched Illumina short reads to aid in completing long-read whole-genome assemblies. However, no improvements by Medaka or Pilon were able to retain and resolve the extra-chromosomal genomes present in *P. knowlesi*. Within this dataset, the resolution of the apicoplast (API) and mitochondrial (MIT) proved to be more challenging than would have been expected. Both Flye and Canu could not assemble and resolve the apicoplast in most isolates. Indeed, while many isolates possessed $<$10-fold apicoplast input coverage for Canu assembly – thus explaining the lack of resolution –it remains unclear if the lack of apicoplast sequences is due to an unforeseen aspect of the leucocyte depletion method for parasite enrichment. Alternatively, the sequencing platform or the alignment to the human reference genome could also be facilitating the apicoplast loss. Although given the evolutionary history of the apicoplast and its

similarity to plant plastids [67], it is unlikely an alignment to the human reference genome would occur.

The mitochondrial (MIT) sequence was largely successfully assembled using both Flye and Canu, suggesting it is less challenging to complete and circularise the MIT genome than its plastid counterpart. However, Canu consistently overassembled the MIT, with often >4-fold the length of the PKNH reference mitochondrial sequence. However, while Flye was able to curtail this, with MIT contig lengths closer to the PKKNH reference genome, Circlator was able to use Canu outputs to circularise the target sequence. This is likely because Circlator has a preferential input of corrected reads, which only Canu produces.

Further peculiarities of the dataset are present in the formation of chimeric contigs, which were detected and removed by Ragtag. Chimeric contigs can occur during genome assembly, although preliminary assemblies during the pipeline development stage did not produce chimeric contigs. The significant translocations seen in later assemblies was determined to be due to improper alignment heuristics by minimap2, likely introduced during one of the quality improvement steps by Racon or Medaka. However, Ragtag successfully split identified chimeric contigs to scaffold the assemblies prior to annotation on Companion. While this approach allowed the removal of chimeric contigs and tangentially resolved the contigs into viable scaffolds, this was achieved using a reference-aware method. As such, the generated scaffold may contain some bias for the structure and organisation of the PKNH reference genome, which was used as a reference for Ragtag. Although it must be noted that the ABACAS2 module of Companion also uses a similar approach to scaffolding; hence, it is unknown the impact Ragtag has on the scaffolding of this dataset.

Nevertheless, comparisons to the published PKNH genome shows the generated StAPkA1H1, and curated dataset genomes (sks047, sks048, sks050, sks058, sks070 and sks339) are of high quality and accuracy; particularly as first versions produced, with no manual curation. However, considerable differences were also reported in the structure and organisation of this dataset's newly generated experimental line and patient isolate genomes. The gaps, inferred sizes, and N's that occupied these gaps provided the stark difference in the descriptive statistics between this dataset and the PKNH reference

genome. Even in isolates such as sks047, sks048 and sks339, which possessed the highest input DNA concentrations, sequence yield, sequence reads coverage, and further improvements after *de novo* assembly, gaps were determined to be larger than those in the PKNH reference genome. Undoubtedly, some of the unplaced sequences within each isolate genome's bin chromosome 00 (chr00) would be able to occupy these gaps. However, it is unlikely that these unplaced sequences can fill all the gaps present in each genome. Other solutions could involve revisiting the raw sequence data and regenerating these genomes using basecalled data from an updated –and more accurate– version of the Guppy basecaller. With Oxford Nanopore Technologies (ONT) frequently updating and improving the basecalling algorithms used in Guppy, it is feasible to see an improvement in quality and likely quantity of basecalled reads, which may include further coverage for these gapped regions in the isolates of the curated dataset.

Further differences are observed in the number of pseudogenes present in these genomes. Generally, pseudogenes are thought to be a form of inactive coding genes which has been gradually lost due to the evolution of the organism [68]. However, such a large and disparate amount of genes being lost in a relatively small genome as *P. knowlesi* remains problematic. A more probable cause for such an increase in pseudogenic content would be a loss of a start/stop codon which would cause the designation of a pseudogene within the genome. As manual curation has not been performed on these genomes, it is expected that manual curation targetting the pseudogenes would considerably reduce these values. This is particularly the case for missing start/stop codons, frameshift and other minor errors. Pseudogenes resulting from single nucleotide polymorphisms (SNP), single nucleotide variation (SNV) or larger structural variants (SVs) would be more challenging to resolve.

Adding to the complexity of manual curation are the sequences present in the chr00 of each isolate. Broadly, these sequences are highly variable, with regions of low complexity repeats and stretches of homopolymers. These sequences contribute to the difficulty the Flye assembler and subsequent improvement and scaffolding tools faced during assembly and genome construction. An assessment of chromosome 00 of the PKNH, StAPkA1H1, sks047 and sks048 genomes shows that the two patient isolates (sks047 and sks048) have more extensive chr00 sequences which cover 2.09 % and 1.94 % of their respective genomes (and 2.07 % and 1.97 % respectively, of the proportional PKNH reference

genome length). This contrasts the 1.73 % and 1.59 % covered by the chr00 of the PKNH and StAPkA1H1 genomes, respectively. However, all three new genomes possessed fewer genes in their respective chr00, with 18, 35 and 25 genes in the StAPkA1H1, sks047 and sks048, respectively, compared to the 62 genes in the PKNH genome. As such, it appears that more genes are being successfully placed within the 14 *P. knowlesi* genomes, rather than in their respective chr00 chromosomes.

Curiously, alignments of the StAPkA1H1 chr00 to the PKNH reference genome reveal sequence clustering between the two chromosomes, whereby a majority of StAPkA1H1 chr00 aligned explicitly to the reference chr00. Such clustering would suggest some structural similarity between the two experimental lines, thus suggesting the likelihood of further unseen similarities. However, orthologous clustering of translated proteins does not support this. Indeed five out of six patient isolates of the curated dataset share more orthologous clusters with the PKNH reference genome than the StAPkA1H1, even though the nucleotide sequence and synteny plots show considerably more similarity between StAPkA1H1 and the PKNH genome. There is no apparent reason for this juxtaposition within this study. The historical mislabelling of the strains has confused the origin of the strains used in ongoing experimental lines. As presented by Butcher and Mitchell [4], it is currently understood that the PKNH reference genome was not sequenced from the H strain, but rather the 'Malaysian' strain, isolated from Peninsular Malaysia by Collins et al. [12]. This raises further confusion, as the PkA1H1 experimental line (source of StAPkA1H1), which may have come from the true H strain or a Malayan strain, was also isolated in Peninsular Malaysia. On the other hand, the patient isolates were all isolated in Malaysian Borneo, which has been shown to have similar, but distinct sub-populations of *P. knowlesi* in contrast to Peninsular *P. knowlesi* [69]. As such, the higher orthologous cluster similarity observed between the patient isolate and the PKNH reference is a cause for interest. Chromosomal-segment exchanges have been seen to occur between sub-populations within Malaysian Borneo and between Malaysian Borneo and Peninsular Malaysia [69]. The similarity seen between the patient isolate genomes, particularly between sks047, sks048 and the PKNH reference genome, could be indications of genetic exchanges occurring earlier than previously thought. However, this remains to be corroborated, and further work is required for the complete elucidation of this finding.

Retaining and resolving the *SICAvar*s and *kir* genes represents an impressive achievement for these patient samples. With the different manipulations carried out on the whole blood and subsequent loss of parasite DNA and loss of sequence reads over the course of the pipeline, the presence of this highly variable and notoriously difficult to resolve genes cannot be overstated. As described by Pain et al. [9], the *SICAvar*s are randomly distributed across the genome in all chromosomes, with *SICAvar* family members found in both the core genome and the telomeric regions. *SICAvar* type II genes remain rare within the patient isolate genomes, which show low complete type II genes in comparison to the StAPkA1H1 and the PKNH reference genome. Due to the challenge of resolving these genes, it is not surprising that this is the case, particularly as *SICAvar* type II genes are predicted to be comprised of 2 - 4 exons and thus would only possess 1 - 3 combined SICA protein domains. As such, the difficulty in completing these genes becomes clear. Importantly the SICA domains are abundant in all genomes sequenced and annotated in this study, including the PKNH reference genome, although in varying proportions. In the PKNH reference genome, 127 single domain SICA fragments were identified while StAPkA1H1, sks047 and sks048 possessed 88, 181 and 196 single SICA domain fragments, respectively. As such, some of the unresolved *SICAvar* type I and II genes may be represented in the single domain fragments. Manual curation of these domains to further resolve the *SICAvar*s would be particularly difficult due to the highly variable yet similar nature of the multigene family. Such an endeavour would undoubtedly be worthwhile to complete, increasing our understanding of not only the *P. knowlesi* genome but perhaps as *P. knowlesi* has done previously, increase our knowledge of other relevant human-adapted *Plasmodium* spp. However, there remains some potential to achieve this and restore these genes to their full length, as evidenced in the PKNH reference only possessing 29 full-length *SICAvar* genes at first release. This number has subsequently been gradually improved upon in the years since release, with the most recent version still containing ~109 full-length *SICAvar* genes (*VEuPathDB Bioinformatics Resource Center, PlasmoDB data release 55; November 2021*).

Unlike the *SICAvar*s, the *kir* genes in the patient isolates and StAPkA1H1 were reported with reduced abundance than the PKNH reference genome. However, the difference between the generated experimental line and patient isolate genomes, and the PKNH reference could be due to genes deemed to be pseudogenic due to a loss of the start/stop

codon or other minor errors, which can be resolved manually. Additionally, the number of *kir*s reported in each isolate of this dataset did not include genes labelled to be 'KIR-like', as to our knowledge, no literature has given a characterisation of this denotation. As such, it is unknown if these genes were labelled as KIR-like due to them being other forms of *Plasmodium* spp. interspersed repeat (*pir*) or perhaps an isoform of the *kir*s. Such ambiguity has been a subject of discussion with Fundel and Zimmer [70] determining a need to ensure adequate and accurate naming conventions for publicly available databases, with all alternatives of a gene given during the annotation process. However, this ambiguity has only increased with the democratisation of genome sequencing, with the ambiguity seen here with the *kir*s being anecdotal of the problem. Here, the annotation of genes as *pir* or *vir* in the automated annotations of the genomes produced in this study confuses the true identity of these genes. With extensive studies by Janssen et al. [71] and Harrison et al. [72] into the orthology of the *Plasmodium* spp. interspersed repeat (*pir*) superfamily of *Plasmodium* proteins, it is evident that genes labelled *pir* or *vir* can be thought to be the same as the *kir*s in *P. knowlesi*, with Harrison et al. [72] showing 97.7 % likelihood of shared structures between members of the *pir* superfamily. While the number of genes annotated as *vir* is likely due to the phylogenetic similarities between *P. vivax* and *P. knowlesi* [71], it is unclear why the StAPkA1H1 genome did not have any genes annotated as such. Interestingly, each patient isolate genome annotated also contained *pir* genes, which may be due to the *kir*s being more variable than most other *Plasmodium* spp. interspersed repeat superfamily members (apart from the *vir*) [71, 73].

Equal to the resolution of the multigene families, these newly generated patient isolates have allowed an examination of structural variation present in the *P. knowlesi* genome. Currently, the sks047 and sks048 isolate genomes have been assessed and analysed to represent reference *de novo* genome from clinical isolates. Through these, it is possible to evaluate the effect of restricting the StAPkA1H1 (and, by extension, the PKNH) in a laboratory setting, free from the pressures of natural selection. The use of StAPkA1H1 as a reference for structural variation allowed for parity with the patient isolates, with any structural variants (SVs) found being solely due to changes in the isolate genomes. StAPkA1H1 was sequenced and processed in precisely the same manner as sks047 and sks048, with the only deviation being sks047 and sks048 were polished with Illumina

short reads, which was unavailable to StAPkA1H1. Indeed, a proportion of SVs found between sks047 and sks048 would be due to the extra polishing carried out for sks047 and sks048. In such cases, through Pilon correction, sks047 and sks048 could likely resolve a gap that remained in the StAPkA1H1 genome. However, importantly some of the identified SVs in sks047 and sks048 are too large to be resolved with Illumina short reads and, as such, would not be due to polishing improvements in the patient isolates.

SV calling using long-read sequences remains in its infancy, especially in non-human, non-diploid sequences. The vast majority of tools and pipelines available have been designed for and tuned to large human datasets and thus require significant manipulation to use for other organisms. In this project, neither the Assemblytics nor the Oxford Nanopore structural variation pipeline (ONTSV) was ideal. Assemblytics was unsatisfactory in its identification and analysis and lacked options for adequate optimisation. Contrasting this, ONTSV while providing considerable granular control and settings, the set-up was challenging and impressively convoluted, with multiple issues encountered. Additionally and perhaps most important, ONTSV was developed for human samples and as such required further tuning for these *Plasmodium* spp. samples. Both approaches still provided insights into the large variations present in contemporary *P. knowlesi* genomes in comparison to an experimental line genome. However, the validity of these variant sites requires further investigation to ensure these are true variants and not due to gapped differences in the genomes. Likely, sequencing additional StAPkA1H1 isolates using Illumina short-read sequencing and subsequent polishing would be able to provide some validation for the SVs calling.

The predicted biological characteristics of the *SICAvar*s and *kir*s ensure that they are highly variable genes due to them interacting with the host cells, whether for infection or evasion. As such, these genes encode antigenically variable proteins and are thought to be part of the virulence cascade of the parasite, though this is yet to be proven.

The patient isolates sequenced, assembled and annotated in this project provide evidence for the utility of long-read nanopore sequencing for whole-genome sequencing of *Plasmodium knowlesi* clinical samples – particularly of adequate parasite DNA input. Robust outputs were generated for one experimental line isolate and six patient isolates, two of which were taken forward to represent the two dimorphic clusters identified in

our biobank. The two representative *de novo* genomes (sks047 and sks048) provide an important insight into the contemporary *P. knowlesi* genome as a result of being sequenced from enriched parasite DNA extracted from clinical whole blood. Both sks047 and sks048 (*and the remaining isolates of this project*) were extracted considerably more recently than experimental lines, which were first isolated over four decades ago. The patient isolates, and by extension, the method used to generate them, are further supported with the resolution of several notable multigene families as well as other important biologically relevant gene families of *P. knowlesi*. Sufficiently retaining and resolving these important functional gene families supports the leucocyte depletion method that extracted the parasite DNA and the sequencing pipeline. This indicates that the use of long-read sequencing is a viable and accurate means of generating *P. knowlesi* whole genomes from small volume whole blood inputs. As such, this allows contemporary investigations of genes like the Duffy-binding protein (DBP), merozoite surface protein (MSP) and Sporozoite invasion-associated protein (SPIAP) to build on previous findings observed in experimental lines like the PKNH and PKA1H1. Furthermore, this allows *P. knowlesi* to also be used as a viable malaria model as suggested in Oresegun, Daneshvar, and Cox-Singh [2].

Indeed, this pipeline and procedure were not without difficulties nor without need for improvement. However, the outputs generated, act as a proof of concept to be built and expanded upon, furthering the knowledge available for *P. knowlesi*.

4.6 References

[1] **WORLD HEALTH ORGANISATION**. *WHO — Malaria*. http://www.who.int/ith/diseases/malaria/en/. 2018 (see p. 166)

[2] **D. R. ORESEGUN**, **C. DANESHVAR**, and **J. COX-SINGH**. "Plasmodium Knowlesi – Clinical Isolate Genome Sequencing to Inform Translational Same-Species Model System for Severe Malaria". In: *Frontiers in Cellular and Infection Microbiology* **11**: (2021). DOI: 10.3389/fcimb.2021.607686 (see pp. 166, 221)

[3] **E. M. PASINI**, **A.-M. ZEEMAN**, **A. V.-V. D. WEL**, and **C. H. M. KOCKEN**. "Plasmodium Knowlesi: A Relevant, Versatile Experimental Malaria Model". In: *Parasitology* **145**:1 (Jan. 2018), 56–70. DOI: 10.1017/S0031182016002286 (see p. 166)

[4] **G. A. BUTCHER** and **G. H. MITCHELL**. "The Role of *Plasmodium Knowlesi* in the History of Malaria Research". In: *Parasitology* **145**:1 (Jan. 2018), 6–17. DOI: 10.1017/S0031182016001888 (see pp. 166, 217)

[5] **K. N. BROWN** and **I. N. BROWN**. "Immunity to Malaria: Antigenic Variation in Chronic Infections of Plasmodium Knowlesi". In: *Nature* **208**: (1965), 1286–1288 (see p. 166)

[6] **A. Z. CHIN**, **M. C. M. MALUDA**, **J. JELIP**, **M. S. B. JEFFREE**, **R. CULLETON**, and **K. AHMED**. "Malaria Elimination in Malaysia and the Rising Threat of Plasmodium Knowlesi". In: *Journal of Physiological Anthropology* **39**:1 (Nov. 2020), 36. DOI: 10.1186/s40101-020-00247-5 (see p. 166)

[7] **J. COX-SINGH** and **R. CULLETON**. "Plasmodium Knowlesi: From Severe Zoonosis to Animal Model". In: *Trends in Parasitology* **31**:6 (June 2015), 232–238. DOI: 10.1016/j.pt.2015.03.003 (see p. 166)

[8] **J. COX-SINGH**, **J. HIU**, **S. B. LUCAS**, **P. C. DIVIS**, **M. ZULKARNAEN**, **P. CHANDRAN**, **K. T. WONG**, **P. ADEM**, **S. R. ZAKI**, **B. SINGH**, et al. "Severe Malaria - A Case of Fatal Plasmodium Knowlesi Infection with Post-Mortem Findings: A Case Report". In: *Malaria journal* **9**:1 (2010), 10 (see p. 166)

[9] **A. PAIN**, **U. BÖHME**, **A. E. BERRY**, **K. MUNGALL**, **R. D. FINN**, **A. P. JACKSON**, **T. MOURIER**, **J. MISTRY**, **E. M. PASINI**, **M. A. ASLETT**, **S. BALASUBRAMMANIAM**, **K. BORGWARDT**, **K. BROOKS**, **C. CARRET**, **T. J. CARVER**, **I. CHEREVACH**, **T. CHILLINGWORTH**, **T. G. CLARK**, **M. R. GALINSKI**, **N. HALL**, **D. HARPER**, **D. HARRIS**, **H. HAUSER**, **A. IVENS**, **C. S. JANSSEN**, **T. KEANE**, **N. LARKE**, **S. LAPP**, **M. MARTI**, **S. MOULE**, **I. M. MEYER**, **D. ORMOND**, **N. PETERS**, **M. SANDERS**, **S. SANDERS**, **T. J. SARGEANT**, **M. SIMMONDS**, **F. SMITH**, **R. SQUARES**, **S. THURSTON**, **A. R. TIVEY**, **D. WALKER**, **B. WHITE**, **E. ZUIDERWIJK**, **C. CHURCHER**, **M. A. QUAIL**, **A. F. COWMAN**, **C. M. R. TURNER**, **M. A. RAJANDREAM**,

C. H. M. KOCKEN, **A. W. THOMAS**, **C. I. NEWBOLD**, **B. G. BARRELL**, and **M. BERRIMAN**. "The Genome of the Simian and Human Malaria Parasite Plasmodium Knowlesi". In: *Nature* **455**:7214 (Oct. 2008), 799–803. DOI: 10.1038/nature07306 (see pp. 166, 167, 175, 181, 195, 197, 198, 206, 218)

[10] **S. A. LAPP**, **J. A. GERALDO**, **J.-T. CHIEN**, **F. AY**, **S. B. PAKALA**, **G. BATUGEDARA**, **J. HUMPHREY**, **THE MAHPIC CONSORTIUM**, **J. D. DEBARRY**, **K. G. LE ROCH**, **M. R. GALINSKI**, and **J. C. KISSINGER**. "PacBio Assembly of a Plasmodium Knowlesi Genome Sequence with Hi-C Correction and Manual Annotation of the SICAvar Gene Family". In: *Parasitology* (July 2017), 1–14. DOI: 10.1017/S0031182017001329 (see pp. 166, 167)

[11] **W. CHIN**, **P. G. CONTACOS**, **G. R. COATNEY**, and **H. R. KIMBALL**. "A Naturally Acquired Quotidian-Type Malaria in Man Transferable to Monkeys". In: *Science* **149**:3686 (Aug. 1965), 865–865. DOI: 10.1126/science.149.3686.865 (see p. 167)

[12] **W. E. COLLINS**, **P. G. CONTACOS**, **J. C. SKINNER**, **W. CHIN**, and **E. GUINN**. "Fluorescent-Antibody Studies on Simian Malaria: I. Development of Antibodies to Plasmodium Knowlesi". In: *The American Journal of Tropical Medicine and Hygiene* **16**:1 (Jan. 1967), 1–6. DOI: 10.4269/ajtmh.1967.16.1 (see pp. 167, 217)

[13] **OXFORD NANOPORE TECHNOLOGIES**. *MinION Rapid Sequencing (SQK-RAD002) Protocol*. 2016 (see p. 168)

[14] **OXFORD NANOPORE TECHNOLOGIES**. *Rapid Barcoding Sequencing (SQK-RBK001) Protocol*. 2017 (see p. 168)

[15] **OXFORD NANOPORE TECHNOLOGIES**. *Rapid Barcoding Sequencing (SQK-RBK004) Protocol*. Oct. 2020 (see pp. 168, 169, 171)

[16] **OXFORD NANOPORE TECHNOLOGIES**. *PCR Barcoding Kit - Ligation (SQK-PBK004) Protocol*. May 2019 (see p. 168)

[17] **M. M. PINHEIRO**, **M. A. AHMED**, **S. B. MILLAR**, **T. SANDERSON**, **T. D. OTTO**, **W. C. LU**, **S. KRISHNA**, **J. C. RAYNER**, and **J. COX-SINGH**. "Plasmodium Knowlesi Genome Sequences from Clinical Isolates Reveal Extensive Genomic Dimorphism". In: *PLOS ONE* **10**:4 (Apr. 2015). Ed. by **O. KANEKO**, e0121303. DOI: 10.1371/journal.pone.0121303 (see pp. 171, 172, 179, 180)

[18] **INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM**, **E. S. LANDER**, **L. M. LINTON**, **B. BIRREN**, **C. NUSBAUM**, **M. C. ZODY**, **J. BALDWIN**, **K. DEVON**, **K. DEWAR**, **M. DOYLE**, **W. FITZHUGH**, **R. FUNKE**, **D. GAGE**, **K. HARRIS**, **A. HEAFORD**, **J. HOWLAND**, **L. KANN**, **J. LEHOCZKY**, **R. LEVINE**, **P. MCEWAN**, **K. MCKERNAN**, **J. MELDRIM**, **J. P. MESIROV**, **C. MIRANDA**, **W. MORRIS**, **J. NAYLOR**, **C. RAYMOND**, **M. ROSETTI**, **R. SANTOS**, **A. SHERIDAN**, **C. SOUGNEZ**, **N. STANGE-THOMANN**, **N. STOJANOVIC**, **A. SUBRAMANIAN**, **D. WYMAN**, **J. ROGERS**, **J. SULSTON**, **R. AINSCOUGH**, **S. BECK**, **D. BENTLEY**, **J. BURTON**, **C. CLEE**, **N. CARTER**, **A. COULSON**, **R. DEADMAN**, **P. DELOUKAS**, **A. DUNHAM**, **I. DUNHAM**, **R. DURBIN**, **L. FRENCH**, **D. GRAFHAM**, **S. GREGORY**, **T. HUBBARD**, **S. HUMPHRAY**, **A. HUNT**, **M. JONES**, **C. LLOYD**, **A. MCMURRAY**, **L. MATTHEWS**, **S. MERCER**, **S. MILNE**, **J. C. MULLIKIN**, **A. MUNGALL**, **R. PLUMB**, **M. ROSS**, **R. SHOWNKEEN**, **S. SIMS**, **R. H. WATERSTON**, **R. K. WILSON**, **L. W. HILLIER**, **J. D. MCPHERSON**, **M. A. MARRA**, **E. R. MARDIS**, **L. A. FULTON**, **A. T. CHINWALLA**, **K. H. PEPIN**, **W. R. GISH**, **S. L. CHISSOE**, **M. C. WENDL**, **K. D. DELEHAUNTY**, **T. L. MINER**, **A. DELEHAUNTY**, **J. B. KRAMER**, **L. L. COOK**, **R. S. FULTON**, **D. L. JOHNSON**, **P. J. MINX**, **S. W. CLIFTON**, **T. HAWKINS**, **E. BRANSCOMB**, **P. PREDKI**, **P. RICHARDSON**, **S. WENNING**, **T. SLEZAK**, **N. DOGGETT**, **J.-F. CHENG**, **A. OLSEN**, **S. LUCAS**, **C. ELKIN**, **E. UBERBACHER**, **M. FRAZIER**, **R. A. GIBBS**, **D. M. MUZNY**, **S. E. SCHERER**, **J. B. BOUCK**, **E. J. SODERGREN**, **K. C. WORLEY**, **C. M. RIVES**, **J. H. GORRELL**, **M. L. METZKER**, **S. L. NAYLOR**, **R. S. KUCHERLAPATI**, **D. L. NELSON**, **G. M. WEINSTOCK**, **Y. SAKAKI**, **A. FUJIYAMA**, **M. HATTORI**, **T. YADA**, **A. TOYODA**, **T. ITOH**, **C. KAWAGOE**, **H. WATANABE**, **Y. TOTOKI**, **T.**

TAYLOR, J. WEISSENBACH, R. HEILIG, W. SAURIN, F. ARTIGUENAVE, P. BROTTIER, T. BRULS, E. PELLETIER, C. ROBERT, P. WINCKER, A. ROSENTHAL, M. PLATZER, G. NYAKATURA, S. TAUDIEN, A. RUMP, D. R. SMITH, L. DOUCETTE-STAMM, M. RUBENFIELD, K. WEINSTOCK, H. M. LEE, J. DUBOIS, H. YANG, J. YU, J. WANG, G. HUANG, J. GU, L. HOOD, L. ROWEN, A. MADAN, S. QIN, R. W. DAVIS, N. A. FEDERSPIEL, A. P. ABOLA, M. J. PROCTOR, B. A. ROE, F. CHEN, H. PAN, J. RAMSER, H. LEHRACH, R. REINHARDT, W. R. MCCOMBIE, M. DE LA BASTIDE, N. DEDHIA, H. BLÖCKER, K. HORNISCHER, G. NORDSIEK, R. AGARWALA, L. ARAVIND, J. A. BAILEY, A. BATEMAN, S. BATZOGLOU, E. BIRNEY, P. BORK, D. G. BROWN, C. B. BURGE, L. CERUTTI, H.-C. CHEN, D. CHURCH, M. CLAMP, R. R. COPLEY, T. DOERKS, S. R. EDDY, E. E. EICHLER, T. S. FUREY, J. GALAGAN, J. G. R. GILBERT, C. HARMON, Y. HAYASHIZAKI, D. HAUSSLER, H. HERMJAKOB, K. HOKAMP, W. JANG, L. S. JOHNSON, T. A. JONES, S. KASIF, A. KASPRYZK, S. KENNEDY, W. J. KENT, P. KITTS, E. V. KOONIN, I. KORF, D. KULP, D. LANCET, T. M. LOWE, A. MCLYSAGHT, T. MIKKELSEN, J. V. MORAN, N. MULDER, V. J. POLLARA, C. P. PONTING, G. SCHULER, J. SCHULTZ, G. SLATER, A. F. A. SMIT, E. STUPKA, J. SZUSTAKOWKI, D. THIERRY-MIEG, J. THIERRY-MIEG, L. WAGNER, J. WALLIS, R. WHEELER, A. WILLIAMS, Y. I. WOLF, K. H. WOLFE, S.-P. YANG, R.-F. YEH, F. COLLINS, M. S. GUYER, J. PETERSON, A. FELSENFELD, K. A. WETTERSTRAND, R. M. MYERS, J. SCHMUTZ, M. DICKSON, J. GRIMWOOD, D. R. COX, M. V. OLSON, R. KAUL, C. RAYMOND, N. SHIMIZU, K. KAWASAKI, S. MINOSHIMA, G. A. EVANS, M. ATHANASIOU, R. SCHULTZ, A. PATRINOS, M. J. MORGAN, C. F. G. R. WHITEHEAD INSTITUTE FOR BIOMEDICAL RESEARCH, THE SANGER CENTRE: WASHINGTON UNIVERSITY GENOME SEQUENCING CENTER, US DOE JOINT GENOME INSTITUTE: BAYLOR COLLEGE OF MEDICINE HUMAN GENOME SEQUENCING CENTER: RIKEN GENOMIC SCIENCES CENTER: GENOSCOPE AND CNRS UMR-8030: I. O. M. B. DEPARTMENT OF GENOME ANALYSIS, GTC SEQUENCING CENTER: BEIJING GENOMICS INSTITUTE/HUMAN GENOME CENTER: T. I. F. S. B. MULTIMEGABASE SEQUENCING CENTER, STANFORD GENOME TECH-

NOLOGY CENTER: UNIVERSITY OF OKLAHOMA'S ADVANCED CENTER FOR GENOME TECHNOLOGY: MAX PLANCK INSTITUTE FOR MOLECULAR GENETICS: L. A. H. G. C. COLD SPRING HARBOR LABORATORY, GBF—GERMAN RESEARCH CENTRE FOR BIOTECHNOLOGY: A. INCLUDES INDIVIDUALS LISTED UNDER OTHER HEADINGS): *GENOME ANALYSIS GROUP (LISTED IN ALPHABETICAL ORDER, U. N. I. O. H. SCIENTIFIC MANAGEMENT: NATIONAL HUMAN GENOME RESEARCH INSTITUTE, STANFORD HUMAN GENOME CENTER: UNIVERSITY OF WASHINGTON GENOME CENTER: K. U. S. O. M. DEPARTMENT OF MOLECULAR BIOLOGY, UNIVERSITY OF TEXAS SOUTHWESTERN MEDICAL CENTER AT DALLAS: U. D. O. E. OFFICE OF SCIENCE, and THE WELLCOME TRUST: "Initial Sequencing and Analysis of the Human Genome". In: *Nature* **409**:6822 (Feb. 2001), 860–921. DOI: 10.1038/35057062 (see p. 173)

[19] **H. LI**. "Minimap2: Pairwise Alignment for Nucleotide Sequences". In: *Bioinformatics* **34**:18 (Sept. 2018), 3094–3100. DOI: 10.1093/bioinformatics/bty191 (see pp. 173, 175)

[20] **H. LI**, **B. HANDSAKER**, **A. WYSOKER**, **T. FENNELL**, **J. RUAN**, **N. HOMER**, **G. MARTH**, **G. ABECASIS**, and **R. DURBIN**. "The Sequence Alignment/Map Format and SAMtools". In: *Bioinformatics* **25**:16 (Aug. 2009), 2078–2079. DOI: 10.1093/bioinformatics/btp352 (see p. 173)

[21] **H. LI**. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data". In: *Bioinformatics (Oxford, England)* **27**:21 (Nov. 2011), 2987–2993. DOI: 10.1093/bioinformatics/btr509 (see p. 173)

[22] **A. QUINLAN**. *Bedtools2: A Powerful Toolset for Genome Arithmetic*. Jan. 2018 (see p. 173)

[23] **M. KOLMOGOROV**, **J. YUAN**, **Y. LIN**, and **P. A. PEVZNER**. "Assembly of Long, Error-Prone Reads Using Repeat Graphs". In: *Nature Biotechnology* **37**:5 (May 2019), 540–546. DOI: 10.1038/s41587-019-0072-8 (see p. 173)

[24] **M. KOLMOGOROV**. *Fast and Accurate de Novo Assembler for Single Molecule Sequencing Reads: Fenderglass/Flye*. July 2019 (see p. 173)

[25] **D. R. LAETSCH** and **M. L. BLAXTER**. "BlobTools: Interrogation of Genome Assemblies". In: *F1000Research* **6**: (July 2017), 1287. DOI: 10.12688/f1000research.12232.1 (see p. 173)

[26] **R. VASER**, **I. SOVIĆ**, **N. NAGARAJAN**, and **M. ŠIKIĆ**. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads". In: *Genome Research* **27**:5 (May 2017), 737–746. DOI: 10.1101/gr.214270.116 (see p. 174)

[27] **B. J. WALKER**, **T. ABEEL**, **T. SHEA**, **M. PRIEST**, **A. ABOUELLIEL**, **S. SAKTHIKUMAR**, **C. A. CUOMO**, **Q. ZENG**, **J. WORTMAN**, **S. K. YOUNG**, and **A. M. EARL**. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement". In: *PLoS ONE* **9**:11 (Nov. 2014). Ed. by **J. WANG**, e112963. DOI: 10.1371/journal.pone.0112963 (see p. 174)

[28] **A. MORGULIS**, **G. COULOURIS**, **Y. RAYTSELIS**, **T. L. MADDEN**, **R. AGARWALA**, and **A. A. SCHÄFFER**. "Database Indexing for Production MegaBLAST Searches". In: *Bioinformatics* **24**:16 (Aug. 2008), 1757–1764. DOI: 10.1093/bioinformatics/btn322 (see p. 175)

[29] **M. HUNT**, **N. D. SILVA**, **T. D. OTTO**, **J. PARKHILL**, **J. A. KEANE**, and **S. R. HARRIS**. "Circlator: Automated Circularization of Genome Assemblies Using Long Sequencing Reads". In: *Genome Biology* **16**: (Dec. 2015), 294. DOI: 10.1186/s13059-015-0849-0 (see pp. 175, 193)

[30] **S. KOREN**, **B. P. WALENZ**, **K. BERLIN**, **J. R. MILLER**, **N. H. BERGMAN**, and **A. M. PHILLIPPY**. "Canu: Scalable and Accurate Long-read Assembly via Adaptive K-mer Weighting and Repeat Separation". In: *Genome Research* (Mar. 2017). DOI: 10.1101/gr.215087.116 (see p. 175)

[31] **T. SEEMANN**. "Prokka: Rapid Prokaryotic Genome Annotation". In: *Bioinformatics* **30**:14 (July 2014), 2068–2069. DOI: 10.1093/bioinformatics/btu153 (see pp. 175, 193)

[32] **A. SZITENBERG**. *TE Discovery in a Genome Assembly*. HullUni-bioinformatics. July 2021 (see pp. 176, 177)

[33] **O. KOHANY**, **A. J. GENTLES**, **L. HANKUS**, and **J. JURKA**. "Annotation, Submission and Screening of Repetitive Elements in Repbase: RepbaseSubmitter and Censor". In: *BMC Bioinformatics* **7**:1 (Oct. 2006), 474. DOI: 10.1186/1471-2105-7-474 (see p. 176)

[34] **L. FU**, **B. NIU**, **Z. ZHU**, **S. WU**, and **W. LI**. "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data". In: *Bioinformatics (Oxford, England)* **28**:23 (Dec. 2012), 3150–3152. DOI: 10.1093/bioinformatics/bts565 (see p. 176)

[35] **W. LI** and **A. GODZIK**. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences". In: *Bioinformatics (Oxford, England)* **22**:13 (July 2006), 1658–1659. DOI: 10.1093/bioinformatics/btl158 (see p. 176)

[36] **J. M. FLYNN**, **R. HUBLEY**, **C. GOUBERT**, **J. ROSEN**, **A. G. CLARK**, **C. FESCHOTTE**, and **A. F. SMIT**. "RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families". In: *Proceedings of the National Academy of Sciences of the United States of America* **117**:17 (Apr. 2020), 9451–9457. DOI: 10.1073/pnas.1921046117 (see p. 176)

[37] **D. ELLINGHAUS**, **S. KURTZ**, and **U. WILLHOEFT**. "LTRharvest, an Efficient and Flexible Software for de Novo Detection of LTR Retrotransposons". In: *BMC bioinformatics* **9**: (Jan. 2008), 18. DOI: 10.1186/1471-2105-9-18 (see p. 176)

[38] **B. J. HAAS**. *TransposonPSI: An Application of PSI-Blast to Mine (Retro-)Transposon ORF Homologies*. http://transposonpsi.sourceforge.net/. 2010 (see p. 178)

[39] **A. R. QUINLAN** and **I. M. HALL**. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features". In: *Bioinformatics* **26**:6 (Mar. 2010), 841–842. DOI: 10.1093/bioinformatics/btq033 (see p. 178)

[40] **M. ALONGE**, **S. SOYK**, **S. RAMAKRISHNAN**, **X. WANG**, **S. GOODWIN**, **F. J. SEDLAZECK**, **Z. B. LIPPMAN**, and **M. C. SCHATZ**. "RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes". In: *Genome Biology* **20**:1 (Oct. 2019), 224. DOI: 10.1186/s13059-019-1829-6 (see p. 178)

[41] **M. ALONGE**. *RagTag*. May 2021 (see p. 178)

[42] **S. STEINBISS**, **F. SILVA-FRANCO**, **B. BRUNK**, **B. FOTH**, **C. HERTZ-FOWLER**, **M. BERRIMAN**, and **T. D. OTTO**. "Companion: A Web Server for Annotation and Analysis of Parasite Genomes". In: *Nucleic Acids Research* **44**:Web Server issue (July 2016), W29–W34. DOI: 10.1093/nar/gkw292 (see pp. 178, 179, 195)

[43] **OXFORD NANOPORE TECHNOLOGIES**, **P. RESCHENEDER**, and **STEPHEN RUDD**. *Oxford Nanopore Structural Variation Pipeline*. Oxford Nanopore Technologies. Nov. 2021 (see p. 179)

[44] **J. REN** and **M. J. CHAISSON**. *Lra: The Long Read Aligner for Sequences and Contigs*. Preprint. Bioinformatics, Nov. 2020. DOI: 10.1101/2020.11.15.383273 (see p. 179)

[45] **B. S. PEDERSEN** and **A. R. QUINLAN**. "Mosdepth: Quick Coverage Calculation for Genomes and Exomes". In: *Bioinformatics* **34**:5 (Mar. 2018), 867–868. DOI: 10.1093/bioinformatics/btx699 (see p. 179)

[46] **T. JIANG**, **Y. LIU**, **Y. JIANG**, **J. LI**, **Y. GAO**, **Z. CUI**, **Y. LIU**, **B. LIU**, and **Y. WANG**. "Long-Read-Based Human Genomic Structural Variation Detection with cuteSV". In: *Genome Biology* **21**:1 (Aug. 2020), 189. DOI: 10.1186/s13059-020-02107-y (see pp. 179, 180)

[47] **P. RESCHENEDER**. *Catfishq*. Oxford Nanopore Technologies. Nov. 2021 (see p. 179)

[48] **H. LI**. *Seqtk - Toolkit for Processing Sequences in FASTA/Q Formats*. 2012 (see p. 179)

[49] **G. MARÇAIS**, **A. L. DELCHER**, **A. M. PHILLIPPY**, **R. COSTON**, **S. L. SALZBERG**, and **A. ZIMIN**. "MUMmer4: A Fast and Versatile Genome Alignment System". In: *PLOS Computational Biology* **14**:1 (Jan. 2018), e1005944. DOI: 10.1371/journal.pcbi.1005944 (see p. 180)

[50] **M. NATTESTAD** and **M. C. SCHATZ**. "Assemblytics: A Web Analytics Tool for the Detection of Variants from an Assembly". In: *Bioinformatics* **32**:19 (Oct. 2016), 3021–3023. DOI: 10.1093/bioinformatics/btw369 (see p. 180)

[51] **D. C. JEFFARES**, **C. JOLLY**, **M. HOTI**, **D. SPEED**, **L. SHAW**, **C. RALLIS**, **F. BALLOUX**, **C. DESSIMOZ**, **J. BÄHLER**, and **F. J. SEDLAZECK**. "Transient Structural Variations Have Strong Effects on Quantitative Traits and Reproductive Isolation in Fission Yeast". In: *Nature Communications* **8**:1 (Jan. 2017), 14061. DOI: 10.1038/ncomms14061 (see p. 180)

[52] **B. S. PEDERSEN**, **R. M. LAYER**, and **A. R. QUINLAN**. "Vcfanno: Fast, Flexible Annotation of Genetic Variants". In: *Genome Biology* **17**:1 (June 2016), 118. DOI: 10.1186/s13059-016-0973-5 (see p. 180)

[53] **H. THORVALDSDÓTTIR**, **J. T. ROBINSON**, and **J. P. MESIROV**. "Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration". In: *Briefings in Bioinformatics* **14**:2 (Mar. 2013), 178–192. DOI: 10.1093/bib/bbs017 (see p. 180)

[54] **P. WANG**. *Vcfstats*. Oct. 2021 (see p. 180)

[55] **WELLCOME SANGER INSTITUTE**. *Assembly-Stats: Get Assembly Statistics from FASTA and FASTQ Files*. Pathogen Informatics, Wellcome Sanger Institute. June 2019 (see p. 181)

[56] **OXFORD NANOPORE TECHNOLOGIES**. *Pomoxis - Bioinformatics Tools for Nanopore Research*. Oxford Nanopore Technologies. May 2021 (see p. 181)

[57] **F. A. SIMÃO**, **R. M. WATERHOUSE**, **P. IOANNIDIS**, **E. V. KRIVENTSEVA**, and **E. M. ZDOBNOV**. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs". In: *Bioinformatics* **31**:19 (Oct. 2015), 3210–3212. DOI: 10.1093/bioinformatics/btv351 (see p. 181)

[58] **A. GUREVICH**, **V. SAVELIEV**, **N. VYAHHI**, and **G. TESLER**. "QUAST: Quality Assessment Tool for Genome Assemblies". In: *Bioinformatics* **29**:8 (Apr. 2013), 1072–1075. DOI: 10.1093/bioinformatics/btt086 (see p. 181)

[59] **K. OKONECHNIKOV**, **A. CONESA**, and **F. GARCÍA-ALCALDE**. "Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data". In: *Bioinformatics* **32**:2 (Jan. 2016), 292–294. DOI: 10.1093/bioinformatics/btv566 (see p. 181)

[60] **B. GEL** and **E. SERRA**. "karyoploteR: An R/Bioconductor Package to Plot Customizable Genomes Displaying Arbitrary Data". In: *Bioinformatics* **33**:19 (Oct. 2017), 3088–3090. DOI: 10.1093/bioinformatics/btx346 (see pp. 181, 205, 209)

[61] **A. E. DARLING**, **B. MAU**, and **N. T. PERNA**. "progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement". In: *PLOS ONE* **5**:6 (June 2010), e11147. DOI: 10.1371/journal.pone.0011147 (see p. 181)

[62] **T. POORTEN**. *dotPlotly*. July 2020 (see p. 181)

[63] **J. DAINAT**, **D. HEREÑÚ**, and **PASCAL-GIT**. *NBISweden/AGAT: AGAT-v0.8.0*. Zenodo. Aug. 2021. DOI: 10.5281/zenodo.5336786 (see pp. 193, 197)

[64] **E. D. BENAVENTE**, **D. R. ORESEGUN**, **P. F. DE SESSIONS**, **E. M. WALKER**, **C. ROPER**, **J. G. DOMBROWSKI**, **R. M. DE SOUZA**, **C. R. F. MARINHO**, **C. J. SUTHERLAND**, **M. L. HIBBERD**, **F. MOHAREB**, **D. A. BAKER**, **T. G. CLARK**, and **S. CAMPINO**. "Global Genetic Diversity of Var2csa in Plasmodium Falciparum with Implications for Malaria in Pregnancy and Vaccine Development". In: *Scientific Reports* **8**: (Oct. 2018). DOI: 10.1038/s41598-018-33767-3 (see p. 197)

[65] **A. M. AHMED**, **M. M. PINHEIRO**, **P. C. DIVIS**, **A. SINER**, **R. ZAINUDIN**, **I. T. WONG**, **C. W. LU**, **S. K. SINGH-KHAIRA**, **S. B. MILLAR**, **S. LYNCH**, **M. WILLMANN**, **B. SINGH**, **S. KRISHNA**, and **J. COX-SINGH**. "Disease Progression in Plasmodium Knowlesi Malaria Is Linked to Variation in Invasion Gene Family Members". In: *PLoS Neglected Tropical Diseases* **8**:8 (Aug. 2014). Ed. by **K. HIRAYAMA**, e3086. DOI: 10.1371/journal.pntd.0003086 (see p. 204)

[66] **R. W. MOON**, **J. HALL**, **F. RANGKUTI**, **Y. S. HO**, **N. ALMOND**, **G. H. MITCHELL**, **A. PAIN**, **A. A. HOLDER**, and **M. J. BLACKMAN**. "Adaptation of the Genetically Tractable Malaria Pathogen Plasmodium Knowlesi to Continuous Culture in Human Erythrocytes". In: *Proceedings of the National Academy of Sciences of the United States of America* **110**:2 (Jan. 2013), 531–536. DOI: 10.1073/pnas.1216457110 (see p. 213)

[67] **G. I. McFadden** and **E. Yeh**. "The Apicoplast: Now You See It, Now You Don't". In: *International journal for parasitology* **47**:2-3 (Feb. 2017), 137–144. DOI: 10.1016/j.ijpara.2016.08.005 (see p. 215)

[68] **R. Albalat** and **C. Cañestro**. "Evolution by Gene Loss". In: *Nature Reviews Genetics* **17**:7 (July 2016), 379–391. DOI: 10.1038/nrg.2016.39 (see p. 216)

[69] **E. D. Benavente**, **A. R. Gomes**, **J. R. De Silva**, **M. Grigg**, **H. Walker**, **B. E. Barber**, **T. William**, **T. W. Yeo**, **P. F. de Sessions**, **A. Ramaprasad**, **A. Ibrahim**, **J. Charleston**, **M. L. Hibberd**, **A. Pain**, **R. W. Moon**, **S. Auburn**, **L. Y. Ling**, **N. M. Anstey**, **T. G. Clark**, and **S. Campino**. "Whole Genome Sequencing of Amplified Plasmodium Knowlesi DNA from Unprocessed Blood Reveals Genetic Exchange Events between Malaysian Peninsular and Borneo Subpopulations". In: *Scientific Reports* **9**: (July 2019). DOI: 10.1038/s41598-019-46398-z (see p. 217)

[70] **K. Fundel** and **R. Zimmer**. "Gene and Protein Nomenclature in Public Databases". In: *BMC bioinformatics* **7**: (Aug. 2006), 372. DOI: 10.1186/1471-2105-7-372 (see p. 219)

[71] **C. S. Janssen**, **R. S. Phillips**, **C. M. R. Turner**, and **M. P. Barrett**. "Plasmodium Interspersed Repeats: The Major Multigene Superfamily of Malaria Parasites". In: *Nucleic Acids Research* **32**:19 (Oct. 2004), 5712–5720. DOI: 10.1093/nar/gkh907 (see p. 219)

[72] **T. E. Harrison**, **A. J. Reid**, **D. Cunningham**, **J. Langhorne**, and **M. K. Higgins**. "Structure of the Plasmodium-Interspersed Repeat Proteins of the Malaria Parasite". In: *Proceedings of the National Academy of Sciences* **117**:50 (Dec. 2020), 32098–32104. DOI: 10.1073/pnas.2016775117 (see p. 219)

[73] **E. F. Merino**, **C. Fernandez-Becerra**, **A. M. Durham**, **J. E. Ferreira**, **V. F. Tumilasci**, **J. d'Arc-Neves**, **M. da Silva-Nunes**, **M. U. Ferreira**, **T. Wickramarachchi**, **P. Udagama-Randeniya**, **S. M. Handunnetti**, and **H. A. del Portillo**. "Multi-Character Population Study of the Vir Subtelomeric Multigene Superfamily of Plasmodium Vivax, a Major Human Malaria Parasite". In: *Molecular and Biochemical Parasitology*

149:1 (Sept. 2006), 10–16. DOI: 10.1016/j.molbiopara.2006.04.002 (see p. 219)

CHAPTER FIVE

CONCLUSION AND AFTERWORD

Ajá t'ó bá re'lé Ẹkùn t'ó bá bọ̀ l'á yọ̀, ńṣe l'ó yẹ k'a kíi kú Orí Ire —
The dog that visits the tiger and returns unscathed should be congratulated

Yorùbá adage

5.1 Conclusion

Plasmodium knowlesi is a biologically and experimentally relevant organism that requires further understanding and research. As previously mentioned, *P. knowlesi* malaria is prevalent across Southeast Asia, with the largest known incidence being in Malaysia. However, given the large geographical range available to the parasite, its Anopheline vector and the non-human primates (NHP) host, it remains unknown if the large incidence in Malaysia is partly due to actively testing for the organism, human activity in the jungle or increased deforestation. While *P. knowlesi* currently appears to cause relatively few human malaria infections, further encroachment into forest spaces by humans in *P. knowlesi*-endemic regions is likely to increase the incidence. Large-scale environmental changes in the natural transmission sites may destabilise the current low incidence of *P. knowlesi* in humans, causing a rapid increase. For *P. knowlesi*, the NHP host have been seen to be able to migrate into urban spaces in search of food; however, the parasite has not been found in these urbanised NHP. However, this appears to be due to the Leucosphyrus group mosquito vectors of *P. knowlesi* being sparse in these

urbanised areas. A drastic change in the natural conditions of their jungle habitat may force adaptation in the mosquitoes leading to an Anopheline migration – although a more likely occurrence would be the rise of another *Plasmodium* spp. organism to occupy the niche currently held by *P. knowlesi*.

In any case, understanding the mechanisms of *P. knowlesi* as a parasite, and subsequently, its action in the human host is necessary. Not only would this increase the knowledge for zoonotic malaria infections, but due to its potential to act as a human malaria model, it would also provide a foundation for investigating fundamental malaria modes of action in other *Plasmodium* spp. organisms, chiefly *P. falciparum* and *P. vivax*. This project aimed to investigate the multigene families of *P. knowlesi* using robust whole-genome sequences derived from infected patient whole blood hitherto sparsely, if at all described in the literature.

Through the use of Oxford Nanopore MinION whole-genome sequencing, the Schizont Infected Cell Agglutination variant antigen (SICAvar) and *Plasmodium knowlesi* interspersed repeat (*kir*) multigene families of *Plasmodium knowlesi* have been described in clinical patient whole blood from infected humans. A highly effective, efficient and relatively inexpensive depletion method was developed for the depletion of human leucocytes in the thawed whole blood, resulting in *P. knowlesi*-enriched DNA for sequencing. *De novo* sequencing of the resulting sequence data proved to show differences between contemporary isolate genomes and the experimental line *P. knowlesi* genomes. These differences are mainly observed in the quantification of the multigene families present in the genomes; however, it is evident that these preliminary outputs give a basis to ask more questions of the *P. knowlesi* genome. Such questions have been described in this report, with a collaborative investigation into the organisation, duplication and selective pressure the *SICAvar* and *kir* genes of interest experience in the genome (See Afterword). This work provides evidence for the power of this sequencing technology and its utility in malaria research, providing a means of generating high-quality, contemporary genomes of organisms of importance and interest.

5.2 Afterword

5.2.1 Utility of newly generated *Plasmodium knowlesi* whole genome sequences from clinical isolates.

Towards the tail end of this project, during the preparation of a publication of generated data, the annotated patient whole genomes constructed in this project were further analysed collaboratively by this researcher and Dr Peter Thorpe. Due to the collaborative nature of the outputs, this was not included in the main body of this report; however, a summary will be presented below.

The high quality and stringency of the annotated genomes for patient isolate sks047 and sks048 allowed for an investigation into their genomic organisation and internal structure. As both genomes represent the two dimorphic clusters present in clinical *P. knowlesi* infections – sks047 representing the more diverse cluster and sks048, the less diverse –, further understanding of their internal organisation was of interest, particularly pertaining to the *SICAvar* and *kir* gene families. Outputs from the work presented, support the dispersal of the highly diverse *SICAvar* and *kir* multigene families across chromosomes and raise the question of the impact these highly variable genes would have in the core genome.

The presence of variable genes in telomeric and sub-telomeric loci is understood to play a role in the biological adaptability of the parasite due to the highly recombinatory nature and high rates of copy number variation in these regions. Particularly as the *var* genes in *P. falciparum* are situated in the telomeric and sub-telomeric regions of the *P. falciparum* genome. Additionally, the predicted biological function of the encoded proteins for the *SICAvar* gene family suggests that they are necessary for the parasite's survival and pathophysiology. As such, *Plasmodium* spp. convention would suggest localisation of both gene families in these variable telomeric regions to facilitate their translational variability, which would be essential to their evolution and multiplicity.

To investigate these questions, we first performed preliminary gene duplication analyses by detecting regions of synteny and collinearity in the genomes. We were able to analyse

chromosomal structural changes and intra-species gene expansion through this. After which, we carried out further orthologous clustering using all the amino acids of the patient isolates sks047 and sks048 as well as the generated StAPkA1H1 genome and the PKNH reference genome. By pooling all the amino acids available in these genomes and comparing them to stable BUSCO genes – that are known to only have one copy in the genome–, we were able to infer the ratio of non-synonymous to synonymous substitutions (dN/dS) present in the whole genome and the *SICAvar* and *kir* gene families of interest. The outputs of this prompted us to carry out additional investigations into the distances of the members of the *SICAvar* and *kir* multigene families to other genes within the *P. knowlesi* genome.

To determine the duplication profiles for the StAPkA1H1, sks047 and sks048, we defined duplication depending on the distance separating two individual duplicated genes. We were able to identify singleton, proximal, dispersed, tandem, segmental and whole-genome duplications from this. As expected, no significant duplication was found in the identified BUSCO genes in each isolate genome; however, this was not the case for either the *SICAvar* or *kir* genes. Members of both multigene families were highly duplicated and classified as dispersed duplications where $>$20 genes separate them. From this we found that the *SICAvar* and *kir* genes are significantly more duplicated than the BUSCO genes (Mann-Whitney U: $p<1.0e\text{-}9$). Determining the dN/dS allowed us to understand the selection pressure placed on the genes of interest.

From this, we found that while the BUSCO gene clusters only had a dN/dS of 0.353, the *SICAvar* and *kir* genes had dN/dS ratios $>$2.3; indicating they are under more selective pressure than the BUSCO genes. With the genes under such selective stress, their dispersed genomic organisation was counter-intuitive. Further analyses found that all *SICAvar* and *kir* genes of interest were statistically significantly farther away from their neighbouring genes than the BUSCO control genes. Both the *SICAvar* and *kir* genes occupy gene-sparse regions, with the stable BUSCO genes found to be present in high gene density regions. Outcomes such as this indicate the value of these newly generated genomes acting as a resource to provide other means of pursuing different avenues of *P. knowlesi* research.

5.2.2 Publications and Presentations

Over the course of this project, outputs generated have been successfully presented in multiple conferences and symposiums including the 'Tenovus Scotland Researchers' Networking Symposium' in 2019, 'Molecular Approaches to Malaria' 2020 meeting and the 2021 meeting of 'BioMalPar'.

Two journal articles have been published from outputs from the research presented in this report. The first has been published under the title: "*Plasmodium knowlesi* – clinical isolate genome sequencing to inform translational same-species model system for severe malaria"; DOI: 10.3389/fcimb.2021.607686. The article is provided below.

The second has been accepted and awaiting release with the title: "*De novo* assembly of *Plasmodium knowlesi* genomes from clinical samples explain the counter-intuitive intra-chromosomal organization of variant *SICAvar* and *KIR* multiple gene family members"; DOI: 10.3389/fgene.2022.855052. The abstract to the article is provided below in [Figure 5.1].

De novo assembly of Plasmodium knowlesi genomes from clinical samples explain the counterintuitive intrachromosomal organization of variant SICAvar and KIR multiple gene family members.

Damilola R. Oresegun[1], Peter Thorpe[1], Ernest D. Benavente[2], Susana Campino[2], Muh Fauzi[1], Robert W. Moon[2], Taane G. Clark[2, 3] and Janet Cox-Singh[4*]

[1]School of Medicine, University of St Andrews, United Kingdom
[2]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, United Kingdom
[3]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, United Kingdom
[4]School of Medicine, University of St Andrews, United Kingdom

Plasmodium knowlesi, a malaria parasite of old-world macaque monkeys, is used extensively to model Plasmodium biology. Recently P. knowlesi was found in the human population of Southeast Asia, particularly Malaysia. P. knowlesi causes uncomplicated to severe and fatal malaria in the human host with features in common with the more prevalent and virulent malaria caused by Plasmodium falciparum.

As such P. knowlesi presents a unique opportunity to develop experimental model systems for malaria pathophysiology informed by clinical data from same-species human infections.

Experimental lines of P. knowlesi represent well characterised genetically static parasites and to maximise their utility as a backdrop for understanding malaria pathophysiology, genetically diverse contemporary clinical isolates, essentially wild-type, require comparable characterization.

The Oxford Nanopore PCR-free long-read sequencing platform was used to sequence and de novo assemble P. knowlesi genomes from frozen clinical samples. The sequencing platform and assembly pipelines were designed to facilitate capturing data and describing, for the first time, P. knowlesi schizont-infected cell agglutination (SICA) var and Knowlesi-Interspersed Repeats (KIR) multiple gene families in parasites acquired from nature.

The SICAvar gene family members code for antigenically variant proteins analogous to the virulence-associated P. falciparum erythrocyte membrane protein (PfEMP1) multiple var gene family. Evidence presented here suggest that the SICAvar family members have arisen through a process of gene duplication, selection pressure and variation. Highly evolving genes including PfEMP1family members tend to be restricted to relatively unstable sub-telomeric regions that drive change with core genes protected in genetically stable intra-chromosomal locations. The comparable SICAvar and KIR gene family members are counter intuitively located across chromosomes. Here we demonstrate that, in contrast to conserved core genes, SICAvar and KIR genes occupy otherwise gene-sparse chromosomal locations that accommodate rapid evolution and change.

The novel methods presented here not only offer the malaria research community new tools to generate comprehensive genome sequence data from small clinical samples but new insight into the complexity of clinically important real-world parasites.

Figure 5.1: Abstract of the accepted manuscript for publication. Research outcomes from this project were summarised and presented as a manuscript for publication to Frontiers in Genetics. The manuscript has been accepted and is awaiting publication.

frontiers in Cellular and Infection Microbiology

PERSPECTIVE
published: 02 March 2021
doi: 10.3389/fcimb.2021.607686

Plasmodium knowlesi – Clinical Isolate Genome Sequencing to Inform Translational Same-Species Model System for Severe Malaria

Damilola R. Oresegun, Cyrus Daneshvar and Janet Cox-Singh *

Division of Infection, School of Medicine, University of St Andrews, St Andrews, United Kingdom

Malaria is responsible for unacceptably high morbidity and mortality, especially in Sub-Saharan African Nations. Malaria is caused by member species' of the genus *Plasmodium* and despite concerted and at times valiant efforts, the underlying pathophysiological processes leading to severe disease are poorly understood. Here we describe zoonotic malaria caused by *Plasmodium knowlesi* and the utility of this parasite as a model system for severe malaria. We present a method to generate long-read third-generation *Plasmodium* genome sequence data from archived clinical samples using the MinION platform. The method and technology are accessible, affordable and data is generated in real-time. We propose that by widely adopting this methodology important information on clinically relevant parasite diversity, including multiple gene family members, from geographically distinct study sites will emerge. Our goal, over time, is to exploit the duality of *P. knowlesi* as a well-used laboratory model and human pathogen to develop a representative translational model system for severe malaria that is informed by clinically relevant parasite diversity.

Keywords: *Plasmodium knowlesi*, MinION, parasite virulence, severe malaria, translational model system

OPEN ACCESS

Edited by:
Takeshi Annoura, National Institute of Infectious Diseases (NIID), Japan

Reviewed by:
Celio Geraldo Freire-de-Lima, Federal University of Rio de Janeiro, Brazil
Guan Zhu, Jilin University, China
Chaturong Putaporntip, Chulalongkorn University, Thailand

***Correspondence:**
Janet Cox-Singh
jcs26@st-andrews.ac.uk

Specialty section:
This article was submitted to Parasite and Host, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: *17 September 2020*
Accepted: *27 January 2021*
Published: *02 March 2021*

Citation:
Oresegun DR, Daneshvar C and Cox-Singh J (2021) Plasmodium knowlesi – Clinical Isolate Genome Sequencing to Inform Translational Same-Species Model System for Severe Malaria. Front. Cell. Infect. Microbiol. 11:607686. doi: 10.3389/fcimb.2021.607686

BACKGROUND

Malaria is a vector-borne disease that has impacted human health in tropical and sub-tropical regions since ancient time and continues to outwit human endeavors to control and eradicate. Malaria parasites, genus *Plasmodium*, have a highly complex lifecycle, intimately dependant on an invertebrate mosquito host for the diploid sexual stage of reproduction and equally dependant on specific vertebrate hosts for asexual replication and transmission. Lifecycle complexity, including adaptation to specific vertebrate hosts, invertebrate host restriction to particular Anopheline vector species with spatial and ecological niche requirement may augur unfavourably for *Plasmodium* spp. survival in a dynamic world. Yet, despite sustained efforts, human malaria persists to the extent that the altruistic World Health Organization (WHO) malaria eradication goal of the 1950s, was downgraded to country and at times species-specific elimination https://www.who.int/malaria/areas/elimination/en/. Even so, eradication is not a forgotten dream and may well be achievable within a new 30-year time-frame (Feachem et al., 2019).

The human host-adapted *Plasmodium* species; *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, and *Plasmodium ovale*, two sub-species (Sutherland et al., 2010) are responsible for most of the reported cases of malaria. *P. falciparum*, in particular, and *P. vivax* are responsible for the global health burden of disease. *P. falciparum* infections carry a high level of morbidity and mortality in adults and children. Severe malaria manifests variously, for example as severe malaria with coma, acute kidney injury and severe malarial anaemia (Plewes et al., 2018; White, 2018; World-Health-Organization, 2019). Understanding the underlying pathophysiology of severe malaria is thwarted by the absence of a translational model system. In practice, malaria elimination remains the most effective strategic method to reduce indigenous transmission of *P. falciparum* and/or *P. vivax* and consequently the impact of severe malaria. Malaria elimination is a long-term goal and in the meantime people will continue to be infected and succumb to severe malaria.

Malaria elimination status is awarded to each country by the WHO even though the country need not necessarily be malaria free. A case in point is Malaysia where indigenous human-host adapted *Plasmodium* species transmission is zero and malaria elimination status was expected to be awarded to Malaysia by the WHO in 2020 (Liew et al., 2018; Jiram et al., 2019; Noordin et al., 2020) https://www.who.int/malaria/areas/elimination/e2020/malaysia/en/. However, malaria - the disease, persists in Malaysia, particularly in the eastern states of Sabah and Sarawak where for the past 20 years *Plasmodium knowlesi*, a malaria parasite of macaque monkeys, has been regularly diagnosed in symptomatic patients in Sabah and Sarawak (Lee et al., 2009a; Barber et al., 2017; Cooper et al., 2020; Raja et al., 2020).

Plasmodium knowlesi Malaria

As one millennium closed and a new one began, a substantial number of cases of *P. knowlesi* were identified in the human population in the Kapit division of Sarawak Malaysia Borneo (Singh et al., 2004). The entry of *P. knowlesi* into the human population became apparent as the number of cases of *P. falciparum* and *P. vivax* declined in response to robust control programmes. Up to that point *P. knowlesi*, a parasite morphologically similar to both *P. malariae* and the early trophozoites of *P. falciparum*, was misdiagnosed by routine microscopy (Lee et al., 2009b). Misdiagnosis as *P. falciparum* had little clinical consequence as both infections require urgent treatment and management. Misdiagnosis as the more benign *P. malariae* resulted in delayed treatment and the development of severe disease and preventable fatality (Cox-Singh et al., 2008).

There is no indication that the cases of *P. knowlesi* malaria are decreasing, 69% of the 16,500 reported cases of malaria in Malaysia between 2013 and 2017 were caused by *P. knowlesi* (Raja et al., 2020) (Hussin et al., 2020). In 2018, more than 4,000 cases of malaria were reported in Malaysia and with *P. falciparum* and *P. vivax* close to elimination, *P. knowlesi* accounted for most of those (World-Health-Organization, 2019; Chin et al., 2020).

P. knowlesi malaria is also widespread across South East Asia where the natural habitat supports sylvan transmission – areas where the specific Anopheline vectors of *P. knowlesi*, the natural macaque hosts and the parasites co-exist and where humans enter these habitats (Singh and Daneshvar, 2013; Shearer et al., 2016). Zoonotic malaria is unlikely to fill the void left by the removal of *P. falciparum* and *P. vivax*, however, communities living close to and individuals who enter the jungle transmission areas for work or leisure activities are at risk of this newly emergent potentially life threatening disease.

P. knowlesi malaria is associated with severe disease in 10 – 12% of cases with death in vulnerable and untreated individuals (Cox-Singh et al., 2008; Daneshvar et al., 2009; William et al., 2011; Rajahram et al., 2012; Grigg et al., 2018; Hussin et al., 2020). Although *P. knowlesi* infections are associated with hyperparasitaemia, severe malaria caused by *P. knowlesi* occurs across a wide spectrum of parasitaemia. Relatively low parasite counts, ≥15,000 parasites/μl carry a high risk of severe disease (Willmann et al., 2012; Cooper et al., 2020). Severe *P. knowlesi* malaria is characterised by one or more of the WHO criteria for severe malaria including; anaemia, acute kidney injury, acute and late-onset respiratory distress, hypotension, jaundice, and metabolic acidosis (Cox-Singh et al., 2008; Daneshvar et al., 2009; Cox-Singh et al., 2010; Barber et al., 2011; World-Health-Organization, 2013; Grigg et al., 2018). Indeed, until the discovery of zoonotic malaria caused by *P. knowlesi*, severe malaria was the preserve of *P. falciparum* and severe malaria guidelines written for *P. falciparum* infection. With few tools to study the pathways to severe malaria and the absence of a comparator disease, assigning clinical cause and effect in malaria was roadblocked.

Even so characterising and untangling the combined contribution of human host response and pathogen to disease presentation and outcome is inherently complex, in the literal sense. The human race survives and often thrives in a harsh microbial world (Numbers, 2011). Of the many microbes only a few are pathogenic, and even-then not uniformly so. Host innate immune function is key to infectivity with co-factors, including age, co-morbidity and co-evolution influencing disease outcomes. Such disease determinants are poorly defined, yet critical to understand, as witnessed in the ongoing Coronavirus pandemic (Mills et al., 2015; Loy et al., 2017; Mandl et al., 2018; Petersen et al., 2020). Human host diversity and response to infection, including response to infection with potentially virulent malaria parasites, is outside of the scope of this article. Rather we focus on determining clinically relevant *Plasmodium* spp. genetic diversity and propose a model system to test for association between parasite genetic diversity and clinical outcome (**Figure 1**).

It is also perhaps worth noting here that in general, disease phenotype precision remains a limiting factor in genome-wide associations studies (GWAS) and an area that lags behind available technologies (MacRae, 2019). No different is the study of parasite genetic characters associated with disease

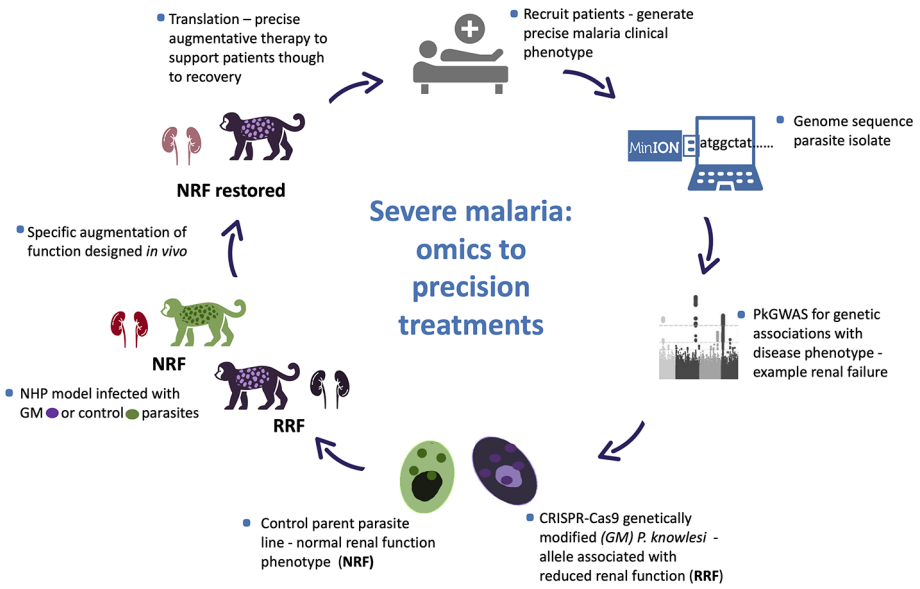


FIGURE 1 | *Plasmodium knowlesi* - proposed translational and dynamic model system for severe malaria. Precise clinical phenotyping malaria patients coupled with parasite genome sequencing "bed-side". Parasite genome wide association studies (pGWAS) with markers of disease – an example of renal impairment in severe *P. knowlesi* infection is used here. Associating alleles are superimposed onto an experimental parasite genetic background to produce genetically modified parasites (GM). GM parasites phenotyped *in vitro* and *in vivo* (non-human primate -NHP, model) for matched disease phenotype. NHP model used to develop strategies to specifically reduce the impact of the clinical phenotype for translation into therapies for patients with severe malaria.

progression (pGWAS) in malaria where disease phenotype precision is currently lacking. No matter how sophisticated the model system, mathematical, *in silico*, *in vitro*, or *in vivo*, the outputs can only be as precise as the input data.

Modelling malaria is especially difficult because malaria parasites have co-evolved with their vertebrate hosts each exerting selective forces on the other in a dynamic dance for survival (Loy et al., 2018). That dynamic is complicated further because malaria parasites are eucaryotic with a relatively large genome, 20–40 mega bases organised into 14 chromosomes (Gardner et al., 2002; Carlton et al., 2008; Pain et al., 2008; Ansari et al., 2016). Designing experiments to identify the drivers of pathogenesis, of parasite virulence and disease cause and effect are challenging.

The advent of "omics" may better inform models for malaria through multiple data generation platforms; genomics, transcriptomics and proteomics (Pinheiro et al., 2015; Campino et al., 2018; Benavente et al., 2019; Lindner et al., 2019). Even with these tools all too often, information is extrapolated. Rodent models and experimental lines that are unable to capture clinically relevant parasite diversity are much better characterised for markers of parasite virulence than diverse contemporary clinical isolates (Plewes et al., 2018). The value of supporting sophisticated forward genetic screens on laboratory isolates with clinical isolate genotyping was demonstrated in a recent study on *P. falciparum* gene clusters involved in erythrocyte invasion (Campino et al., 2018). Invasion phenotypes generated from crossing two experimental lines and phenotype-associated deletions were compared with long-read sequence data available from a small number of clinical isolates where indels in the same large locus were identified supporting invasion pathway variation in nature. Ignoring the impact of natural pathogen diversity on disease progression and virulence creates an inexcusable vacuum when analysing data for parasite association with disease severity.

P. knowlesi is an adaptable and naturally diverse parasite. A genetic study on clinical isolates of *P. knowlesi* identified an association between certain haplotypes of a short polymorphic fragment (~885bp) of the *Plasmodium knowlesi normocyte binding protein (Pknbp)xa* on chromosome 14 and continuous data on markers of disease severity (Ahmed et al., 2014). In addition to disease association, the fragment was dimorphic, clinical isolates clustered into one of two distinct genotypes at that locus, begging the question how far the dimorphism extended across the *Pknbpxa* 9578bp gene and chromosome 14.

Harnessing the power of next-generation sequencing seemed the obvious choice to take this work forward within the caveat that the clinical isolates of *P. knowlesi* available to study were small volume (<1mL) frozen whole blood. Undeterred and as proof of concept, we developed a method to deplete human DNA and concentrate parasite DNA in the samples. We produced *P. knowlesi* genome sequences from six clinical isolates using massively parallel Illumina short-read sequencing platforms (Pinheiro et al., 2015). The move from genetics to genomics for clinical isolate characterization unlocked a wealth of information. Subsequent analyses found that the *Pknbpxa* dimorphism extended along the gene and chromosome 14. Indeed, SNPs associated with the dimorphism were found on all chromosomes and involved more than half of all genes in *P. knowlesi* parasites isolated from patients. The work

demonstrated that *P. knowlesi* isolated from human infections in Sarawak, Malaysian Borneo are one of two distinct genotypes.

Both pieces of work unlocked hidden genome-wide characters in clinical isolates of *P. knowlesi*. The studies reinforced the idea that pathogen genome sequence data extracted from clinically well characterised infections provides a valuable resource for studies on the role of pathogen diversity in virulence and disease outcome.

P. knowlesi – A Model for Malaria

P. knowlesi was first described in a long tailed macaque in the 1930s (Knowles and Gupta, 1932). Early work demonstrated that *P. knowlesi* was an adaptable parasite and experimental lines were developed and maintained in rhesus macaques, *Macaca mulatta*, to model for malaria. The *P. knowlesi* – rhesus macaque malaria model was used extensively for studies on malaria antigenic variation, vaccine development, parasite invasion, and biology, recently reviewed (Butcher and Mitchell, 2018; Galinski et al., 2018; Pasini et al., 2018). Traditionally *P. knowlesi* was not favoured as a model for disease, pathophysiology, mostly because the *P. knowlesi* in *Macaca mulatta* was particularly aggressive and not representative of human malaria caused by *P. falciparum*. A view supported in more recent work on cytokine responses in *M. mulatta* experimentally infected with *P. knowlesi* where a dampened response, and if anything, an anti-inflammatory response was observed in this model, a response that is uncharacteristic of human-host *Plasmodium* infections (Praba-Egge et al., 2002). *P. falciparum* malaria and indeed *P. knowlesi* clinical infections, are characterised by vigorous pro- and anti-inflammatory responses depending on age and endemicity (Cox-Singh et al., 2011; Farrington et al., 2017). Taken together there was little support for the utility of *P. knowlesi* in *M. mulatta* as an *in vivo* model for severe malaria. *P. knowlesi* in other experimental non-human primates (NHP's) produces a disease more representative of human malaria and it is surprising that this opportunity to model severe malaria has not been taken forward (Langhorne and Cohen, 1979; Ozwara et al., 2003; Onditi et al., 2015). Unfortunately lack of support for using *P. knowlesi* to model for severe malaria is compounded by evolutionary distance. *P. knowlesi* and *P. falciparum* occupy distinct phylogenetic clades and phylogenetic distance is often used to argue against using *P. knowlesi* to model for *P. falciparum*. Evolutionary distance continues to be used to question the validity of comparing *P. knowlesi* with *P. falciparum* malaria, yet they are member-species of the same genus – by definition they are closely related. In practice, evolutionary distance often over-rides biological and comparable clinical characters and *P. knowlesi* is more often favourably viewed as a model for the phylogenetically closer yet phenotypically quite distinct *P. vivax* (Moon et al., 2013; Mohring et al., 2019; Verzier et al., 2019).

Neither *P. falciparum* nor *P. vivax* is permissive in intact experimental NHP hosts and to date, representative heterologous translational models for malaria are not available to interrogate pathways to pathology and to develop augmentative therapies. Consequently, the treatment and management of patients severely ill with malaria remain imprecise and generally supportive. We argue that clinical data collected from patients with naturally acquired severe *P. knowlesi* coupled with homologous laboratory adapted, well characterised and genetically adaptable experimental lines of *P. knowles*i can be exploited to discover the parasite drivers of severe malaria. Laboratory adapted lines of *P. knowlesi* are permissive in a range of NHP hosts, including olive baboons and common marmosets (Langhorne and Cohen, 1979; Ozwara et al., 2003; Onditi et al., 2015). Some of these *in vivo* models exhibit clinical characters representative of severe malaria caused by *P. falciparum* and, importantly, contemporary clinical descriptions of severe malaria caused by *P. knowlesi* (Cox-Singh et al., 2008; Daneshvar et al., 2009; Cox-Singh et al., 2010; Daneshvar et al., 2018). Notwithstanding NHP models are of ethical concern, expensive and valid only if the information obtained significantly advances knowledge, which often is not the case. Experimental lines of *P. knowlesi* even if modelled *in vivo* are effectively research silos lacking the power to inform clinical disease caused by genetically diverse contemporary wild-type zoonotic parasites (Ahmed et al., 2014; Assefa et al., 2015; Divis et al., 2015; Pinheiro et al., 2015).

Clinical descriptions of *P. knowlesi* malaria portray a spectrum of disease from uncomplicated – to severe and fatal infections and can be compared phenotypically with *P. falciparum* malaria (Cox-Singh et al., 2008; Daneshvar et al., 2009; Cox-Singh et al., 2010; Cox-Singh et al., 2011; Rajahram et al., 2012; Ahmed et al., 2014; Barber et al., 2018a; Barber et al., 2018b).

The duality of *P. knowlesi* as an adaptable experimental model and human pathogen offers a unique opportunity to develop a comprehensive representative translational model system for malaria informed by same-species clinical disease.

Two important advances enhance the utility of *P. knowlesi* as a model for disease. The first is the adaptation of an experimental line of *P. knowlesi* to *in vitro* growth in human erythrocytes (Moon et al., 2013). The second is transfection technology. *P. knowlesi* in macaque erythrocytes was already shown to be more amenable to transfection, meaning genetic modification, than experimental lines of *P. falciparum* (Kocken et al., 2002). The human erythrocyte adapted line is similarly receptive to transfection and indeed CRISPR-Cas9 targeted genetic modification technology, genome editing, has been developed for *P. knowlesi* (Moon et al., 2013; Mohring et al., 2019). These technologies together with genome sequence data, generated from clinical isolates, will facilitate the introduction of clinically relevant alleles of *P. knowlesi* into experimental lines for *in vitro* characterisation and the unique opportunity to take this work forward *in vivo* (Cox-Singh and Culleton, 2015).

A long journey to cause and effect harnessing omics, genetically modified parasites and comprehensive model systems to properly ascribe parasite virulence to malaria pathophysiology while possible is a long game, difficult, time-consuming and expensive. However, failure to make this effort is to perpetuate acceptance of clinical and

therapeutic blind-spots, imprecise and generally supportive treatment and management for severe malaria, that perhaps is only acceptable if there is no alternative.

In the first instance, the ability to genome sequence *Plasmodium* species isolated from clinically well-characterised malaria patients will facilitate *Plasmodium* Genome-Wide Associations Studies (*p*GWAS) and identify virulence gene candidates. We show how short and long-read *Plasmodium* genome sequence data can be generated from fresh or archived frozen samples held in the many malaria research centres worldwide. *Plasmodium* genome sequence outputs over time and space will facilitate the construction of a substantial genetic reference resource, based on diverse wild-type parasites isolated from patients, that will inform model systems (Milner et al., 2012; Ahmed et al., 2014; Pinheiro et al., 2015; Auburn et al., 2018; Campino et al., 2018; Divis et al., 2018; Otto et al., 2018; Su et al., 2019; Siao et al., 2020). Until now genome sequence data generated from clinical samples was more feasible using Illumina massively parallel short-read sequencing.

As highlighted in the *P. falciparum* invasion gene study described in an earlier section, prohibitively expensive and otherwise impractical long-read genome sequencing data from clinical isolates were required to validate study findings (Campino et al., 2018). We have already developed a method to extract *P. knowlesi* DNA from archived small volume clinical isolates suitable for Illumina short-read sequencing (Pinheiro et al., 2015). To overcome limitations of short-read genome sequencing that are problematic for multiple repeat regions and multiple gene families in *Plasmodium* spp., the method was further adapted for Oxford Nanopore MinION long-read sequencing, third-generation sequencing, that is accessible, affordable, mobile and suitable for low yield DNA samples (**Figure 2**). Briefly 200–400ul samples of archived whole blood from *P. knowlesi* patients was rapidly thawed and immediately diluted in 50mL cold PBS. The suspension was mixed gently before recovering parasites and any contaminating white blood cells (WBC's) by centrifugation: 2,000 x g; 20 minutes; 4°C (pellet 1). The supernatant was transferred to a fresh tube and centrifugation repeated to maximise parasite recovery (pellet 2). The pellets were combined and resuspended in 1.2mL of cold PBS. Surviving host WBCs, a source of human DNA (hDNA) contamination, were removed using magnetic beads coated with antibodies to the ubiquitous WBC surface marker CD45 (Dynabeads^{TM} CD45, Invitrogen^{TM}). Dynabeads were prepared as per manufacturers' instruction and 100ul bead suspension added to the 1.2mL pellet suspension followed by incubation at 4°C with rotation for 30 minutes. WBC's bound to the beads were removed by placing in a magnetic field for two minutes.

The human WBC depleted eluate was carefully removed and transferred to a fresh Eppendorf tube and centrifuged at 2,000 x *g* for 20 min to recover the parasite enriched pellet (PEP). The PEP was suspended in 200ul PBS for DNA extraction (QIAamp DNA Blood Mini Kit, QIAGEN). Recovered DNA was eluted in 150ul of Buffer AE. Percent hDNA depletion and Plasmodium DNA recovery were determined using quantitative PCR (qPCR) (Klaassen et al., 2003; Divis et al., 2010). *P. knowlesi* qPCR *ct* values negatively correlated with genome coverage ($p = 0.0375$). *P. knowlesi* DNA enriched samples (post hDNA depletion) from isolates with a starting parasitaemia of <40,000 per ul had low parasite DNA yield and sequence coverage.

Twenty-one (21) samples from 15 different patients, median parasitaemia 193,600 parasites/ul (IQR 127,875 – 321,750; min 20,656; max 794,063) with >90% hDNA depletion were taken forward to PCR-free rapid barcoding library preparation (Oxford Nanopore, SQK-RBK004). SQK-RBK004 library preparation is suitable for small yield DNA samples in the region of 400ng and includes a tagmentation step that generates read lengths normally distributed around a mean length of 4,500 kb. Of 21 library preparations 13 (62%) had >10x genome sequence coverage and six of these >30x coverage. Coverage of 100x was achieved especially when >1 sequencing library was prepared per isolate.

For the first time it is possible to generate long-read *Plasmodium* genome sequence data from small clinical samples from malaria patients. Samples that are archived or collected prospectively can be sequenced in a cost-effective and time-efficient manner anywhere. The importance of this capability is the opportunity to move

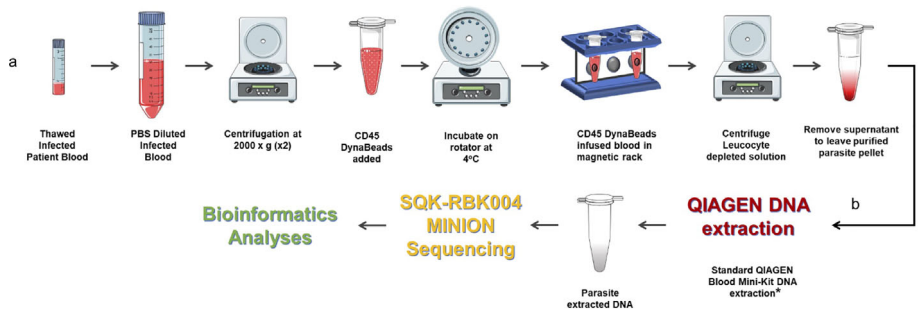


FIGURE 2 | Protocol for processing low volume frozen whole blood samples from patients with malaria in preparation for third generation genome sequencing. **(A)** Human leucocyte depletion and **(B)** sequencing pipeline. *Human DNA was quantified using qPCR (Klaassen et al., 2003) and a standard curve derived from Human Genomic Control DNA (Applied Biosystems®, TaqMan®). In the absence of pure control parasite DNA, cycle threshold (*ct*) values from *P. knowlesi* qPCR (Divis et al., 2010) were normalised by volume and used to estimate parasite DNA enrichment following human DNA depletion. *Parts of the figure are conceptualised and adapted using Servier Medical Art, Servier:* https://smart.servier.com.

forward from a necessary dependence on *Plasmodium* genome sequence generated from experimental lines to genome sequence generated from clinical samples with matched clinical data. The methodology we describe is particularly applicable to *P. knowlesi* infections that tend to be single genotype and reach relatively high parasitaemia. The methods can be applied to, albeit, relatively uncommon single genotype *P. falciparum* infections. Although multiple genotype infections present a limiting factor to the methods described here, the potential to generate valuable genome-wide information on even a small number of clinical isolates to inform studies on *P. falciparum* virulence should not be overlooked.

Subsequent *p*GWAS on long read sequence data from clinical isolates with matched high quality continuous clinical and laboratory data, relative to particular clinical manifestations of severe malaria, will help unravel the contribution of parasite diversity to virulence. In addition to accessibility and field application of low-cost real-time sequencing in-house, Nanopore MinION long-read sequencing can resolve important multiple gene family members, including the *P. knowlesi kirs* and *SICAvars* (Pain et al., 2008; Pinheiro et al., 2015; Lapp et al., 2018). These and other gene clusters encode surface antigens that are implicated in malaria parasite virulence and are difficult to sequence, formerly requiring expensive sequencing platforms equally prohibitive in cost and quantity of input DNA required (Campino et al., 2018).

Our particular interest is to use MinION sequence data from clinical isolates of *P. knowlesi* in *p*GWAS studies. We will analyse matched continuous clinical data predictive of precise characteristics of severe malaria to identify candidate alleles implicated in virulence to take forward in functional studies. CRISPR-Cas9 technology developed for *P. knowlesi* (Mohring et al., 2019) will facilitate locus-specific gene editing to superimpose clinically relevant alleles onto experimental lines and offer the opportunity for allele-specific phenotyping *in vitro*. Genetically modified lines with *in vitro* phenotypic characters that carry a high suspicion of involvement in parasite virulence and following exhaustive experimentation will be deemed suitable to take forward *in vivo* for clinical phenotyping and translational research (**Figure 1**).

Our immediate goal is to promote third-generation genome sequencing and capacity strengthening in bioinformatics for routine genetic studies on clinical malaria in endemic countries. The outputs will create a repository that captures diversity and information on multiple gene families hitherto outside the remit of all but large centres mostly working on model parasites. Our long-term vision is to develop a precise experimental model system for severe malaria pathophysiology informed by clinical infections and culminating in *in vivo* disease phenotyping and translational research. A model that, for the first time, will have the power to characterise parasite allele-specific cause and effect. A model system that exploits the utility of *P. knowlesi*, a laboratory model, and *P. knowlesi* that is responsible for naturally acquire human disease.

Not the end of the story or perfect by any standard but our sequencing capability represents a significant step forward towards creating the means to understand malaria pathophysiology and to inform the rational design and development of adjunctive therapies for patients with severe malaria.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of St. Andrews. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JC-S and CD conceived the work and wrote the manuscript. DRO did sample processing, sequencing, and bioinformatics. All authors contributed to the article and approved the submitted version.

FUNDING

DRO is supported by the Wellcome Trust ISSF award 204821/Z/16/Z. Bioinformatics and computational biology analyses were supported by the University of St Andrews Bioinformatics Unit (AMD3BIOINF), funded by Wellcome Trust ISSF award 105621/Z/14/Z. The sample BioBank was compiled with informed consent (Medial Research Council, www.mrc.ac.uk, grant G0801971). Genome sequencing was supported by Tenovus Scotland (T16/03).

REFERENCES

Ahmed, A. M., Pinheiro, M. M., Divis, P. C., Siner, A., Zainudin, R., Wong, I. T., et al. (2014). Disease progression in Plasmodium knowlesi malaria is linked to variation in invasion gene family members. *PloS Negl. Trop. Dis.* 8 (8), e3086. doi: 10.1371/journal.pntd.0003086

Ansari, H. R., Templeton, T. J., Subudhi, A. K., Ramaprasad, A., Tang, J., Lu, F., et al. (2016). Genome-scale comparison of expanded gene families in Plasmodium ovale wallikeri and Plasmodium ovale curtisi with Plasmodium malariae and with other Plasmodium species. *Int. J. Parasitol.* 46 (11), 685–696. doi: 10.1016/j.ijpara.2016.05.009

Assefa, S., Lim, C., Preston, M. D., Duffy, C. W., Nair, M. B., Adroub, S. A., et al. (2015). Population genomic structure and adaptation in the zoonotic malaria parasite Plasmodium knowlesi. *Proc. Natl. Acad. Sci. U.S.A.* 112 (42), 13027–13032. doi: 10.1073/pnas.1509534112

Auburn, S., Benavente, E. D., Miotto, O., Pearson, R. D., Amato, R., Grigg, M. J., et al. (2018). Genomic analysis of a pre-elimination Malaysian Plasmodium vivax population reveals selective pressures and changing

transmission dynamics. *Nat. Commun.* 9 (1), 2585. doi: 10.1038/s41467-018-04965-4

Barber, B. E., William, T., Jikal, M., Jilip, J., Dhararaj, P., Menon, J., et al. (2011). Plasmodium knowlesi malaria in children. *Emerg. Infect. Dis.* 17 (5), 814–820. doi: 10.3201/eid1705.101489

Barber, B. E., Rajahram, G. S., Grigg, M. J., William, T., and Anstey, N. M. (2017). World Malaria Report: time to acknowledge Plasmodium knowlesi malaria. *Malar. J.* 16 (1), 135. doi: 10.1186/s12936-017-1787-y

Barber, B. E., Grigg, M. J., Piera, K. A., William, T., Cooper, D. J., Plewes, K., et al. (2018a). Intravascular haemolysis in severe Plasmodium knowlesi malaria: association with endothelial activation, microvascular dysfunction, and acute kidney injury. *Emerg. Microbes Infect.* 7 (1), 106. doi: 10.1038/s41426-018-0105-2

Barber, B. E., Russell, B., Grigg, M. J., Zhang, R., William, T., Amir, A., et al. (2018b). Reduced red blood cell deformability in Plasmodium knowlesi malaria. *Blood Adv.* 2 (4), 433–443. doi: 10.1182/bloodadvances.2017013730

Benavente, E. D., Gomes, A. R., De Silva, J. R., Grigg, M., Walker, H., Barber, B. E., et al. (2019). Whole genome sequencing of amplified Plasmodium knowlesi DNA from unprocessed blood reveals genetic exchange events between Malaysian Peninsular and Borneo subpopulations. *Sci. Rep.* 9 (1), 9873. doi: 10.1038/s41598-019-46398-z

Butcher, G. A., and Mitchell, G. H. (2018). The role of Plasmodium knowlesi in the history of malaria research. *Parasitology* 145 (1), 6–17. doi: 10.1017/S0031182016001888

Campino, S., Marin-Menendez, A., Kemp, A., Cross, N., Drought, L., Otto, T. D., et al. (2018). A forward genetic screen reveals a primary role for Plasmodium falciparum Reticulocyte Binding Protein Homologue 2a and 2b in determining alternative erythrocyte invasion pathways. *PloS Pathog.* 14 (11), e1007436. doi: 10.1371/journal.ppat.1007436

Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L., Lorenzi, H., Caler, E., et al. (2008). Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature* 455 (7214), 757–763. doi: 10.1038/nature07327

Chin, A. Z., Maluda, M. C. M., Jelip, J., Jeffree, M. S. B., Culleton, R., and Ahmed, K. (2020). Malaria elimination in Malaysia and the rising threat of Plasmodium knowlesi. *J. Physiol. Anthropol.* 39 (1), 36. doi: 10.1186/s40101-020-00247-5

Cooper, D. J., Rajahram, G. S., William, T., Jelip, J., Mohammad, R., Benedict, J., et al. (2020). Plasmodium knowlesi Malaria in Sabah, Malaysi-2017: Ongoing Increase in Incidence Despite Near-elimination of the Human-only Plasmodium Species. *Clin. Infect. Dis.* 70 (3), 361–367. doi: 10.1093/cid/ciz237

Cox-Singh, J., and Culleton, R. (2015). Plasmodium knowlesi: from severe zoonosis to animal model. *Trends Parasitol.* 31 (6), 232–238. doi: 10.1016/j.pt.2015.03.003

Cox-Singh, J., Davis, T. M., Lee, K. S., Shamsul, S. S., Matusop, A., Ratnam, S., et al. (2008). Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. *Clin. Infect. Dis.* 46 (2), 165–171. doi: 10.1086/524888

Cox-Singh, J., Hiu, J., Lucas, S. B., Divis, P. C., Zulkarnaen, M., Chandran, P., et al. (2010). Severe malaria - a case of fatal Plasmodium knowlesi infection with post-mortem findings: a case report. *Malar. J.* 9:10. doi: 10.1186/1475-2875-9-10

Cox-Singh, J., Singh, B., Daneshvar, C., Planche, T., Parker-Williams, J., and Krishna, S. (2011). Anti-inflammatory cytokines predominate in acute human Plasmodium knowlesi infections. *PloS One* 6 (6), e20541. doi: 10.1371/journal.pone.0020541

Daneshvar, C., Davis, T. M., Cox-Singh, J., Rafa'ee, M. Z., Zakaria, S. K., Divis, P. C., et al. (2009). Clinical and laboratory features of human Plasmodium knowlesi infection. *Clin. Infect. Dis.* 49 (6), 852–860. doi: 10.1086/605439

Daneshvar, C., William, T., and Davis, T. M. E. (2018). Clinical features and management of Plasmodium knowlesi infections in humans. *Parasitology* 145 (1), 18–31. doi: 10.1017/S0031182016002638

Divis, P. C., Shokoples, S. E., Singh, B., and Yanow, S. K. (2010). A TaqMan real-time PCR assay for the detection and quantitation of Plasmodium knowlesi. *Malar. J.* 9, 344. doi: 10.1186/1475-2875-9-344

Divis, P. C., Singh, B., Anderios, F., Hisam, S., Matusop, A., Kocken, C. H., et al. (2015). Admixture in Humans of Two Divergent Plasmodium knowlesi Populations Associated with Different Macaque Host Species. *PloS Pathog.* 11 (5), e1004888. doi: 10.1371/journal.ppat.1004888

Divis, P. C. S., Duffy, C. W., Kadir, K. A., Singh, B., and Conway, D. J. (2018). Genome-wide mosaicism in divergence between zoonotic malaria parasite subpopulations with separate sympatric transmission cycles. *Mol. Ecol.* 27 (4), 860–870. doi: 10.1111/mec.14477

Farrington, L., Vance, H., Rek, J., Prahl, M., Jagannathan, P., Katureebe, A., et al. (2017). Both inflammatory and regulatory cytokine responses to malaria are blunted with increasing age in highly exposed children. *Malar. J.* 16 (1), 499. doi: 10.1186/s12936-017-2148-6

Feachem, R. G. A., Chen, I., Akbari, O., Bertozzi-Villa, A., Bhatt, S., Binka, F., et al. (2019). Malaria eradication within a generation: ambitious, achievable, and necessary. *Lancet* 394 (10203), 1056–1112. doi: 10.1016/S0140-6736(19)31139-0

Galinski, M. R., Lapp, S. A., Peterson, M. S., Ay, F., Joyner, C. J., LE Roch, K. G., et al. (2018). Plasmodium knowlesi: a superb in vivo nonhuman primate model of antigenic variation in malaria. *Parasitology* 145 (1), 85–100. doi: 10.1017/S0031182017001135

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., et al. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 419 (6906), 498–511. doi: 10.1038/nature01097

Grigg, M. J., William, T., Barber, B. E., Rajahram, G. S., Menon, J., Schimann, E., et al. (2018). Age-Related Clinical Spectrum of Plasmodium knowlesi Malaria and Predictors of Severity. *Clin. Infect. Dis.* 67 (3), 350–359. doi: 10.1093/cid/ciy065

Hussin, N., Lim, Y. A., Goh, P. P., William, T., Jelip, J., and Mudin, R. N. (2020). Updates on malaria incidence and profile in Malaysia from 2013 to 2017. *Malar. J.* 19 (1), 55. doi: 10.1186/s12936-020-3135-x

Jiram, A. I., Ooi, C. H., Rubio, J. M., Hisam, S., Karnan, G., Sukor, N. M., et al. (2019). Evidence of asymptomatic submicroscopic malaria in low transmission areas in Belaga district, Kapit division, Sarawak, Malaysia. *Malar. J.* 18 (1), 156. doi: 10.1186/s12936-019-2786-y

Klaassen, C. H., Jeunink, M. A., Prinsen, C. F., Ruers, T. J., Tan, A. C., Strobbe, L. J., et al. (2003). Quantification of human DNA in feces as a diagnostic test for the presence of colorectal cancer. *Clin. Chem.* 49 (7), 1185–1187. doi: 10.1373/49.7.1185

Knowles, R., and Gupta, B. M. D. (1932). A Study of Monkey-Malaria, and Its Experimental Transmission to Man. *Ind. Med. Gaz* 67 (6), 301–320.

Kocken, C. H., Ozwara, H., van der Wel, A., Beetsma, A. L., Mwenda, J. M., and Thomas, A. W. (2002). Plasmodium knowlesi provides a rapid in vitro and in vivo transfection system that enables double-crossover gene knockout studies. *Infect. Immun.* 70 (2), 655–660. doi: 10.1128/iai.70.2.655-660.2002

Langhorne, J., and Cohen, S. (1979). Plasmodium knowlesi in the marmoset (Callithrix jacchus). *Parasitology* 78 (1), 67–76. doi: 10.1017/s0031182000048599

Lapp, S. A., Geraldo, J. A., Chien, J. T., Ay, F., Pakala, S. B., Batugedara, G., et al. (2018). PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C correction and manual annotation of the SICAvar gene family. *Parasitology* 145 (1), 71–84. doi: 10.1017/S0031182017001329

Lee, K. S., Cox-Singh, J., Brooke, G., Matusop, A., and Singh, B. (2009a). Plasmodium knowlesi from archival blood films: further evidence that human infections are widely distributed and not newly emergent in Malaysian Borneo. *Int. J. Parasitol.* 39 (10), 1125–1128. doi: 10.1016/j.ijpara.2009.03.003

Lee, K. S., Cox-Singh, J., and Singh, B. (2009b). Morphological features and differential counts of Plasmodium knowlesi parasites in naturally acquired human infections. *Malar. J.* 8, 73. doi: 10.1186/1475-2875-8-73

Liew, J. W. K., Mahpot, R. B., Dzul, S., Abdul Razak, H. A. B., Ahmad Shah Azizi, N. A. B., Kamarudin, M. B., et al. (2018). Importance of Proactive Malaria Case Surveillance and Management in Malaysia. *Am. J. Trop. Med. Hyg.* 98 (6), 1709–1713. doi: 10.4269/ajtmh.17-1010

Lindner, S. E., Swearingen, K. E., Shears, M. J., Walker, M. P., Vrana, E. N., Hart, K. J., et al. (2019). Transcriptomics and proteomics reveal two waves of translational repression during the maturation of malaria parasite sporozoites. *Nat. Commun.* 10 (1), 4964. doi: 10.1038/s41467-019-12936-6

Loy, D. E., Liu, W., Li, Y., Learn, G. H., Plenderleith, L. J., Sundararaman, S. A., et al. (2017). Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. *Int. J. Parasitol.* 47 (2-3), 87–97. doi: 10.1016/j.ijpara.2016.05.008

Loy, D. E., Plenderleith, L. J., Sundararaman, S. A., Liu, W., Gruszczyk, J., Chen, Y. J., et al. (2018). Evolutionary history of human Plasmodium vivax revealed by genome-wide analyses of related ape parasites. *Proc. Natl. Acad. Sci. U.S.A.* 115 (36), E8450–E8459. doi: 10.1073/pnas.1810053115

MacRae, C. A. (2019). Closing the 'phenotype gap' in precision medicine: improving what we measure to understand complex disease mechanisms. *Mamm. Genome* 30 (7-8), 201–211. doi: 10.1007/s00335-019-09810-7

Mandl, J. N., Schneider, C., Schneider, D. S., and Baker, M. L. (2018). Going to Bat(s) for Studies of Disease Tolerance. *Front. Immunol.* 9, 2112. doi: 10.3389/fimmu.2018.02112

Mills, C. D., Ley, K., Buchmann, K., and Canton, J. (2015). Sequential Immune Responses: The Weapons of Immunity. *J. Innate Immun.* 7 (5), 443–449. doi: 10.1159/000380910

Milner, D.A. Jr., Vareta, J., Valim, C., Montgomery, J., Daniels, R. F., Volkman, S. K., et al. (2012). Human cerebral malaria and Plasmodium falciparum genotypes in Malawi. *Malar. J.* 11, 35. doi: 10.1186/1475-2875-11-35

Mohring, F., Hart, M. N., Rawlinson, T. A., Henrici, R., Charleston, J. A., Diez Benavente, E., et al. (2019). Rapid and iterative genome editing in the malaria parasite Plasmodium knowlesi provides new tools for P. vivax research. *Elife* 8, 1–29. doi: 10.7554/eLife.45829

Moon, R. W., Hall, J., Rangkuti, F., Ho, Y. S., Almond, N., Mitchell, G. H., et al. (2013). Adaptation of the genetically tractable malaria pathogen Plasmodium knowlesi to continuous culture in human erythrocytes. *Proc. Natl. Acad. Sci. U.S.A.* 110 (2), 531–536. doi: 10.1073/pnas.1216457110

Noordin, N. R., Lee, P. Y., Mohd Bukhari, F. D., Fong, M. Y., Abdul Hamid, M. H., Jelip, J., et al. (2020). Prevalence of Asymptomatic and/or Low-Density Malaria Infection among High-Risk Groups in Peninsular Malaysia. *Am. J. Trop. Med. Hyg.* 103 (3), 1107–1110. doi: 10.4269/ajtmh.20-0268

Numbers, M. B. (2011). Microbiology by numbers. *Nat. Rev. Microbiol.* 9 (9), 628–628. doi: 10.1038/nrmicro2644

Onditi, F. I., Nyamongo, O. W., Omwandho, C. O., Maina, N. W., Maloba, F., Farah, I. O., et al. (2015). Parasite accumulation in placenta of non-immune baboons during Plasmodium knowlesi infection. *Malar. J.* 14:118. doi: 10.1186/s12936-015-0631-5

Otto, T. D., Bohme, U., Sanders, M., Reid, A., Bruske, E. I., Duffy, C. W., et al. (2018). Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres. *Wellcome Open Res.* 3, 52. doi: 10.12688/wellcomeopenres.14571.1

Ozwara, H., Langermans, J. A., Maamun, J., Farah, I. O., Yole, D. S., Mwenda, J. M., et al. (2003). Experimental infection of the olive baboon (Paplio anubis) with Plasmodium knowlesi: severe disease accompanied by cerebral involvement. *Am. J. Trop. Med. Hyg.* 69 (2), 188–194. doi: 10.4269/ajtmh.2003.69.188

Pain, A., Bohme, U., Berry, A. E., Mungall, K., Finn, R. D., Jackson, A. P., et al. (2008). The genome of the simian and human malaria parasite Plasmodium knowlesi. *Nature* 455 (7214), 799–803. doi: 10.1038/nature07306

Pasini, E. M., Zeeman, A. M., Voorberg-VAN DER Wel, A., and Kocken, C. H. M. (2018). Plasmodium knowlesi: a relevant, versatile experimental malaria model. *Parasitology* 145 (1), 56–70. doi: 10.1017/S0031182016002286

Petersen, E., Koopmans, M., Go, U., Hamer, D. H., Petrosillo, N., Castelli, F., et al. (2020). Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *Lancet Infect. Dis.* 20 (9), e238–e244. doi: 10.1016/S1473-3099(20)30484-9

Pinheiro, M. M., Ahmed, M. A., Millar, S. B., Sanderson, T., Otto, T. D., Lu, W. C., et al. (2015). Plasmodium knowlesi genome sequences from clinical isolates reveal extensive genomic dimorphism. *PloS One* 10 (4), e0121303. doi: 10.1371/journal.pone.0121303

Plewes, K., Turner, G. D. H., and Dondorp, A. M. (2018). Pathophysiology, clinical presentation, and treatment of coma and acute kidney injury complicating falciparum malaria. *Curr. Opin. Infect. Dis.* 31 (1), 69–77. doi: 10.1097/QCO.0000000000000419

Praba-Egge, A. D., Montenegro, S., Cogswell, F. B., Hopper, T., and James, M. A. (2002). Cytokine responses during acute simian Plasmodium cynomolgi and Plasmodium knowlesi infections. *Am. J. Trop. Med. Hyg.* 67 (6), 586–596. doi: 10.4269/ajtmh.2002.67.586

Raja, T. N., Hu, T. H., Kadir, K. A., Mohamad, D. S. A., Rosli, N., Wong, L. L., et al. (2020). Naturally Acquired Human Plasmodium cynomolgi and P. knowlesi Infections, Malaysian Borneo. *Emerg. Infect. Dis.* 26 (8), 1801–1809. doi: 10.3201/eid2608.200343

Rajahram, G. S., Barber, B. E., William, T., Menon, J., Anstey, N. M., and Yeo, T. W. (2012). Deaths due to Plasmodium knowlesi malaria in Sabah, Malaysia: association with reporting as Plasmodium malariae and delayed parenteral artesunate. *Malar. J.* 11, 284. doi: 10.1186/1475-2875-11-284

Shearer, F. M., Huang, Z., Weiss, D. J., Wiebe, A., Gibson, H. S., Battle, K. E., et al. (2016). Estimating Geographical Variation in the Risk of Zoonotic Plasmodium knowlesi Infection in Countries Eliminating Malaria. *PloS Negl. Trop. Dis.* 10 (8), e0004915. doi: 10.1371/journal.pntd.0004915

Siao, M. C., Borner, J., Perkins, S. L., Deitsch, K. W., and Kirkman, L. A. (2020). Evolution of Host Specificity by Malaria Parasites through Altered Mechanisms Controlling Genome Maintenance. *mBio* 11 (2), 1–6. doi: 10.1128/mBio.03272-19

Singh, B., and Daneshvar, C. (2013). Human infections and detection of Plasmodium knowlesi. *Clin. Microbiol. Rev.* 26 (2), 165–184. doi: 10.1128/CMR.00079-12

Singh, B., Kim Sung, L., Matusop, A., Radhakrishnan, A., Shamsul, S. S., Cox-Singh, J., et al. (2004). A large focus of naturally acquired Plasmodium knowlesi infections in human beings. *Lancet* 363 (9414), 1017–1024. doi: 10.1016/S0140-6736(04)15836-4

Su, X. Z., Lane, K. D., Xia, L., Sa, J. M., and Wellems, T. E. (2019). Plasmodium Genomics and Genetics: New Insights into Malaria Pathogenesis, Drug Resistance, Epidemiology, and Evolution. *Clin. Microbiol. Rev.* 32 (4), 1–29. doi: 10.1128/CMR.00019-19

Sutherland, C. J., Tanomsing, N., Nolder, D., Oguike, M., Jennison, C., Pukrittayakamee, S., et al. (2010). Two nonrecombining sympatric forms of the human malaria parasite Plasmodium ovale occur globally. *J. Infect. Dis.* 201 (10), 1544–1550. doi: 10.1086/652240

Verzier, L. H., Coyle, R., Singh, S., Sanderson, T., and Rayner, J. C. (2019). Plasmodium knowlesi as a model system for characterising Plasmodium vivax drug resistance candidate genes. *PloS Negl. Trop. Dis.* 13 (6), e0007470. doi: 10.1371/journal.pntd.0007470

White, N. J. (2018). Anaemia and malaria. *Malar. J.* 17 (1), 371. doi: 10.1186/s12936-018-2509-9

William, T., Menon, J., Rajahram, G., Chan, L., Ma, G., Donaldson, S., et al. (2011). Severe Plasmodium knowlesi malaria in a tertiary care hospital, Sabah, Malaysia. *Emerg. Infect. Dis.* 17 (7), 1248–1255. doi: 10.3201/eid1707.101017

Willmann, M., Ahmed, A., Siner, A., Wong, I. T., Woon, L. C., Singh, B., et al. (2012). Laboratory markers of disease severity in Plasmodium knowlesi infection: a case control study. *Malar. J.* 11, 363. doi: 10.1186/1475-2875-11-363

World-Health-Organization (2013). *Management of Severe Malaria - A Practical Handbook. 3rd ed* (Geneva: Worldhealth Organization).

World-Health-Organization (2019). *World Malaria Report 2019* (Geneva: Licence: CC BY-NC-SA 3.0 IGO.).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

Preliminary Leucoycte Depletion Experiments

Òf̀ìífì̀ì là ńrií, a ò rí́ òkodoro; ḿbọ́, baba gba-n-gba — ***All we see is shadows, not clarity; but clarity will come, the father of all candor***

Yorùbá adage

A.1 Equipment and Reagents

A.1.1 Equipment

Table A.1: List of equipment utilised for protocol development, optimisation and implementation

Equipment	Full Product Name	Manufacturer	Catalog Number
0.2 mL PCR tube	0.2 mL Thin Wall PCR Tubes with Flat Cap	Axygen	PCR-02-C
1.5 mL Eppendorf tubes	1.5 mL MaxyClear Snaplock Microcentrifuge Tube	Axygen	MCT-150-C
1000 µL pipette tips	1000 µL Maxymum Recovery® Universal Fit FilterTips	Axygen	TF-1000-L-R-S
100 µL pipette tips	100 µL Maxymum Recovery® Universal Fit Filter Tips	Axygen	TF-100-L-R-S
10 µL pipette tips	10 µL Maxymum Recovery® Filter Tips	Axygen	TF-400-L-R-S
15 mL centrifuge tubes	15 mL Cellstar Tubes	Cellstar	E150335N
200 µL pipette tips	200 µL Maxymum Recovery® Universal Fit Filter Tips	Axygen	TF-200-L-R-S
200 µL wide-bore pipette tips	200 µL Universal Fit Filter Tips, Wide-Bore	Axygen	TF-205-WB-R-S
20 µL pipette tips	ep Dualfilter T.I.P.S®0.5 - 20 µL PCR	Eppendorf	30078527
50 mL centrifuge tubes	50 mL Cellstar Tubes	Cellstar	E19103E9
AMPure XP beads	AMPure XP, 5 mL	Beckman Coulter	A63880
DynaBeads	Dynabeads™ CD45	Invitrogen	11153D
Lo-Bind Tube	Nonstick, RNase-free Microfuge Tubes, 1.5 mL	Ambion	AM12450
Magnetic Stand	6-Tube Magnetic Separation Rack	New England Biolabs	S1506S

Continued on next page

Table A.1 – *Continued from previous page*

Equipment	Full Product Name	Manufacturer	Catalog Number
MinION	MinION Mk1b	Oxford Nanopore Technologies	None
MinKNOW	MinION Software	Oxford Nanopore Technologies	None
NanoDrop	NanoDrop 2000	ThermoFisher	ND-2000
Plasmodipur	Plasmodipur 8011	Europroxima	8011FILTER10U
R10 flowcell_version 1	FLO-MIN110	Oxford Nanopore Technologies	Various
R10 flowcell_version 2	FLO-MIN111	Oxford Nanopore Technologies	Various
R9.4.1 flowcell	FLO-MIN106D	Oxford Nanopore Technologies	Various
RotorGene Q	Rotor-Gene Q	Qiagen	None

A.1.2 Reagents

Table A.2: List of reagents and chemicals utilised for protocol development, optimisation and implementation

Reagents	Full Product Name	Manufacturer	Catalog Number
BSA	Bovine Serum Albumin	Sigma	A2058
DNA Blood Mini Kit	QIAamp DNA Blood Mini Kit	Qiagen	51104
EDTA	EDTA (0.5 M), pH 8.0, RNase-free	Invitrogen	AM9260G
Ethanol	Ethanol Absolute	VWR Chemicals	20821.33
Histopaque-1119	Histopaque®-1119	Sigma	11191
NaCl	NaCl (5 M), RNase-free	Invitrogen	AM9760G
Nuclease-free water	Nuclease-Free Water (not DEPC-Treated)	Ambion	AM9937

Continued on next page

Table A.2 – *Continued from previous page*

Equipment	Full Product Name	Manufacturer	Catalog Number
PBS	PBS, pH 7.4	Gibco(trademark)	10010-023
Proteinase K	Protease	Qiagen	1016330
Saponin	Saponin	Sigma	47036
SQK-RAD002	SQK-RAD002 Sequencing library	Oxford Nanopore Technologies	None
SQK-RBK004	SQK-RBK004 Sequencing library	Oxford Nanopore Technologies	None
TaqMan human control	TaqMan™ Control Genomic DNA (human)	Applied Biosystems	4312660
TaqMan Mastermix	TaqMan™ Universal PCR Master Mix	Applied Biosystems	4304437
Tris-HCl	UltraPure™ 1 M Tris-HCl Buffer, pH 7.5	Gibco(trademark)	15567-027

A.2 Starting volumes of Whole Blood and final volume of eluted DNA

Table A.3: The starting whole blood volumes and the final volumes the extracted DNA is eluted in

Exp.	Conditions	Whole Blood Vol. (µL)		Eluted DNA Vol. (µL)	
		Control	CD45 Treated	Control	CD45 Treated
1	**Saponin_CD45_vs_Donor**	200	200	200	200
2	**Saponin_CD45_vs_Simulated**	200	200	100	100
3	**Saponin_CD45_vs_Simulated_2**	200	200	100	100
4	**Saponin_CD45_Histopaque_vs_Simulated**	200	200	100	100
5	**Saponin_CD45Sedimentation_vs_Simulated_a**	200	200	100	100
6	**Saponin_CD45Sedimentation_vs_Simulated_b**	200	200	100	100
7	**Inverse_Saponin_CD45_vs_sks265**	50	200	100	100
8	**Inverse_Saponin_CD45_vs_sks078**	200	250	100	100
9	**Inverse_Saponin_CD45_vs_sks367**	50	210	150	150
10	**Inverse_Saponin_CD45_vs_sks134**	50	600	150	150
11	**CD45_Plasmodiupur_vs_sks269**	50	550	150	150
12	**CD45_vs_sks074**	50	450	150	150

Whole blood volumes for the control sample and the subsequently leucocyte depleted sample. As experiments proceeded, the volume of the Control aliquoted from the total whole blood decreased in favour of normalisation as part of downstream analyses. Once extraction was complete, the DNA was eluted in stated volumes of nuclease-free water or Buffer AE for long term storage.

A.3 Raw qPCR cycle thresholds calculations

Table A.4: Raw human DNA average cycle thresholds from preliminary leucocyte depletion methods

Exp.	Conditions	Measured Cycled times (Ct)											
		Control			*Saponin*			*Saponin + CD45 DynaBeads*			*Histopaque 200*		
		1	**2**	**Avg.**	**1**	**2**	**Avg.**	**1**	**2**	**Avg.**	**1**	**2**	**Avg.**
1	**Saponin_CD45_vs_Donor**	20.73	21.80	21.27	21.49	21.79	21.64	31.48	31.16	31.32	N.A	N.A	N.C
2	**Saponin_CD45_vs_Simulated**	27.57	32.66	30.12	26.89	26.18	26.54	30.39	30.18	30.29	N.A	N.A	N.C
3	**Saponin_CD45_vs_Simulated_2**	29.66	31.96	30.81	26.69	27.29	26.99	30.34	33.24	31.79	N.A	N.A	N.C
4	**Saponin_CD45_Histopaque_vs_Simulated**	19.11	20.20	19.66	20.26	20.26	20.26	28.95	NTC	28.95	21.03	21.05	21.04
5	**Saponin_CD45Sedimentation_vs_Simulated_a**	20.97	21.13	21.05	23.23	23.59	23.41	49.78	34.97	42.38	N.A	N.A	N.C
6	**Saponin_CD45Sedimentation_vs_Simulated_b**	20.20	20.34	20.27	19.38	19.53	19.46	20.83	19.86	20.35	N.A	N.A	N.C
7	**Inverse_Saponin_CD45_vs_sks265**	20.11	20.69	20.40	N.A	N.A	N.C	20.05	19.37	19.71	N.A	N.A	N.C
8	**Inverse_Saponin_CD45_vs_sks078**	17.44	17.88	17.66	N.A	N.A	N.C	20.34	20.69	20.52	N.A	N.A	N.C
9	**Inverse_Saponin_CD45_vs_sks367**	20.75	20.56	20.66	N.A	N.A	N.C	19.66	19.85	19.76	N.A	N.A	N.C
10	**Inverse_Saponin_CD45_vs_sks134**	20.60	21.28	20.94	N.A	N.A	N.C	18.76	19.28	19.02	N.A	N.A	N.C
11	**CD45_Plasmodiupur_vs_sks269**	19.17	19.07	19.12	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C
12	**CD45_vs_sks074**	20.29	19.66	19.98	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C

Raw cycle threshold (Ct) values for measuring human DNA content in DNA extracted during preliminary experiments in order to determine extent of leucocyte depletion. Here, leucocyte depletion is determined by measuring the δ change value between the normalised control sample and the sample treated with the experiment conditions. The Ct values were normalised using the average values described above and the percentage of input method as described in subsection 2.4.9 and presented in Table 2.6.

N.A - Not Applicable. This refers to conditions not applicable to the corresponding experiment because the conditions were not carried out within that experiment
N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated
NTC - Sample was run using the stated conditions however, qPCR failed due to not surpassing the set threshold.

Table A.5: Further Raw human DNA average cycle thresholds from preliminary leucocyte depletion methods

Exp.	Conditions	Measured Cycled times (Ct)											
		Histopaque 1000			*Plasmodipur + CD45 DynaBeads*			*CD45 DynaBeads*			*Bead Pellet*		
		1	**2**	**Avg.**	**1**	**2**	**Avg.**	**1**	**2**	**Avg.**	**1**	**2**	**Avg.**
1	**Saponin_CD45_vs_Donor**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	28.30	28.73	28.52
2	**Saponin_CD45_vs_Simulated**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	30.62	NTC	30.62
3	**Saponin_CD45_vs_Simulated_2**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	NTC	NTC	N.C
4	**Saponin_CD45_Histopaque_vs_Simulated**	23.73	23.73	23.73	N.A	N.A	N.C	N.A	N.A	N.C	19.58	19.80	19.69
5	**Saponin_CD45Sedimentation_vs_Simulated_a**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	21.19	20.83	21.01
6	**Saponin_CD45Sedimentation_vs_Simulated_b**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	22.89	22.33	22.61
7	**Inverse_Saponin_CD45_vs_sks265**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	21.61	20.72	21.17
8	**Inverse_Saponin_CD45_vs_sks078**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	18.61	18.23	18.42
9	**Inverse_Saponin_CD45_vs_sks367**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	20.32	20.52	20.42
10	**Inverse_Saponin_CD45_vs_sks134**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C
11	**CD45_Plasmodiupur_vs_sks269**	N.A	N.A	N.C	24.40	25.13	24.77	N.A	N.A	N.C	18.34	18.07	18.21
12	**CD45_vs_sks074**	N.A	N.A	N.C	N.A	N.A	N.C	23.00	22.40	22.70	17.62	17.45	17.54

Raw cycle threshold (Ct) values for measuring human DNA content in DNA extracted during preliminary experiments in order to determine extent of leucocyte depletion. Here, leucocyte depletion is determined by measuring the δ change value between the normalised control sample and the sample treated with the experiment conditions. The Ct values were normalised using the average values described above and the percentage of input method as described in subsection 2.4.9 and presented in Table 2.6.

N.A - Not Applicable. This refers to conditions not applicable to the corresponding experiment because the conditions were not carried out within that experiment
N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated
NTC - Sample was run using the stated conditions however, qPCR failed due to not surpassing the set threshold.

Table A.6: Raw *P. knowlesi* DNA average cycle thresholds from preliminary leucocyte depletion methods

Exp.	Conditions	Measured Cycled times (Ct)											
		Control			*Saponin*			*Saponin + CD45 DynaBeads*			*Histopaque 200*		
		1	2	Avg.	1	2	Avg.	1	2	Avg.	1	2	Avg.
1	**Saponin_CD45_vs_Donor**	23.23	23.47	23.23	N.V	NTC	N.C	NTC	NTC	N.C	N.A	N.A	N.C
2	**Saponin_CD45_vs_Simulated**	21.84	N.V	21.84	20.85	21.48	21.17	26.83	34.72	30.78	N.A	N.A	N.C
3	**Saponin_CD45_vs_Simulated_2**	N.V	26.51	26.51	21.44	22.20	21.82	N.V	M.Ct	N.C	N.A	N.A	N.C
4	**Saponin_CD45_Histopaque_vs_Simulated**	17.00	17.68	17.34	18.17	18.55	18.36	23.77	27.44	25.61	22.49	25.50	24.00
5	**Saponin_CD45Sedimentation_vs_Simulated_a**	18.03	17.89	17.96	19.92	20.20	20.06	22.23	23.60	22.92	N.A	N.A	N.C
6	**Saponin_CD45Sedimentation_vs_Simulated_b**	19.82	19.99	19.91	20.83	20.75	20.79	N.V	N.V	N.C	N.A	N.A	N.C
7	**Inverse_Saponin_CD45_vs_sks265**	M.Ct	N.V	N.C	N.A	N.A	N.C	N.V	N.V	N.C	N.A	N.A	N.C
8	**Inverse_Saponin_CD45_vs_sks078**	15.58	15.57	15.58	N.A	N.A	N.C	19.05	18.51	18.78	N.A	N.A	N.C
9	**Inverse_Saponin_CD45_vs_sks367**	N.V	N.V	N.C	N.A	N.A	N.C	N.V	NTC	N.C	N.A	N.A	N.C
10	**Inverse_Saponin_CD45_vs_sks134**	7.67	M.Ct	7.67	N.A	N.A	N.C	N.V	N.V	N.C	N.A	N.A	N.C
11	**CD45_Plasmodiupur_vs_sks269**	N.V	N.V	N.C	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C
12	**CD45_vs_sks074**	20.95	20.20	20.58	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C

Raw cycle threshold (Ct) values for measuring *P. knowlesi* DNA content in DNA extracted during preliminary experiments in order to determine extent of parasite DNA retention. Here, parasite retention is determined by measuring the δ change value between the normalised control sample and the sample treated with the experiment conditions. As a low negative or a positive δ change is desirable to indicate parasite DNA retention and enrichment. δ change values are presented in Table 2.7. The Ct values were normalised using the average values described above and the percentage of input method as described in subsection 2.4.9 and presented in Table 2.7.

N.A - Not Applicable. This refers to conditions not applicable to the corresponding experiment because the conditions were not carried out within that experiment
N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated
NTC - Sample was run using the stated conditions however, qPCR failed due to not surpassing the set threshold.
N.V - Sample was run using the stated conditions however, qPCR report showed no reading; outputting, a blank reading for unspecified reasons

Table A.7: Further Raw *P. knowlesi* DNA average cycle thresholds from preliminary leucocyte depletion methods

Exp.	Conditions	Measured Cycled times (Ct)											
		Histopaque 1000			*Plasmodipur + CD45 DynaBeads*			*CD45 DynaBeads*			*Bead Pellet*		
		1	2	Avg.	1	2	Avg.	1	2	Avg.	1	2	Avg.
1	**Saponin_CD45_vs_Donor**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	NTC	NTC	N.C
2	**Saponin_CD45_vs_Simulated**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	N.V	NTC	N.C
3	**Saponin_CD45_vs_Simulated_2**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	NTC	N.V	N.C
4	**Saponin_CD45_Histopaque_vs_Simulated**	26.23	N.V	26.23	N.A	N.A	N.C	N.A	N.A	N.C	19.13	21.29	20.21
5	**Saponin_CD45Sedimentation_vs_Simulated_a**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	20.07	18.94	19.51
6	**Saponin_CD45Sedimentation_vs_Simulated_b**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	N.V	NTC	N.C
7	**Inverse_Saponin_CD45_vs_sks265**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	N.V	N.V	N.C
8	**Inverse_Saponin_CD45_vs_sks078**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	21.45	20.84	21.15
9	**Inverse_Saponin_CD45_vs_sks367**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	N.V	N.V	N.C
10	**Inverse_Saponin_CD45_vs_sks134**	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C	N.A	N.A	N.C
11	**CD45_Plasmodiupur_vs_sks269**	N.A	N.A	N.C	N.V	N.V	N.C	N.A	N.A	N.C	N.V	N.V	N.C
12	**CD45_vs_sks074**	N.A	N.A	N.C	N.A	N.A	N.C	21.78	20.51	21.15	23.35	22.94	23.15

Raw cycle threshold (Ct) values for measuring *P. knowlesi* DNA content in DNA extracted during preliminary experiments in order to determine extent of parasite DNA retention. Here, parasite retention is determined by measuring the δ change value between the normalised control sample and the sample treated with the experiment conditions. As a low negative or a positive δ change is desirable to indicate parasite DNA retention and enrichment. δ change values are presented in Table 2.7. The Ct values were normalised using the average values described above and the percentage of input method as described in subsection 2.4.9 and presented in Table 2.7.

N.A - Not Applicable. This refers to conditions not applicable to the corresponding experiment because the conditions were not carried out within that experiment
N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated
NTC - Sample was run using the stated conditions however, qPCR failed due to not surpassing the set threshold.
N.V - Sample was run using the stated conditions however, qPCR report showed no reading; outputting, a blank reading for unspecified reasons

A.4 Exploratory Experiments

Table A.8: DNA concentration measured in the outputs of the Saponin Lysis method to determine optimum Saponin concentration

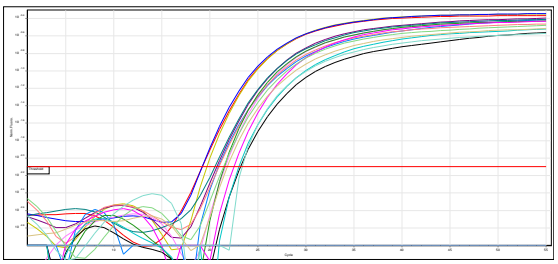
a *Concentration Values calculated by NanoDrop 2000 spectrophotometer*

Saponin Treatment	Concentration (ng/µL)				260/280 Values			
	1	2	3	Avg.	1	2	3	Avg.
0% - Control	14.9	14.7	15.2	14.9	1.73	1.71	1.63	1.69
1%	4.7	4.6	6	5.1	1.63	1.53	1.46	1.54
5%	5.5	5.1	5.1	5.2	1.98	1.47	1.56	1.67
10%	3.5	3.5	3.9	3.6	1.28	1.25	1.33	1.29
20%	4.7	5	4.4	4.7	1.43	1.38	1.46	1.42

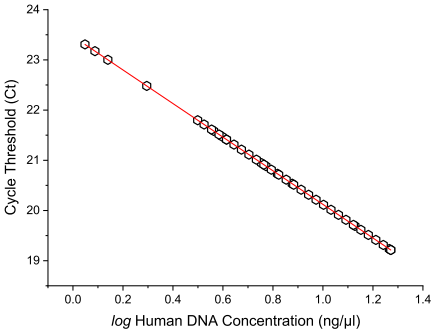
b *Concentration measured using Qubit*

Saponin Treatment	Concentration (ng/µL)		
	1	2	Average
0% - Control	13.00	12.3	12.65
1%	4.84	5.24	5.04
5%	5.12	4.52	4.82
10%	2.34	2.20	2.27
20%	4.60	4.10	4.35

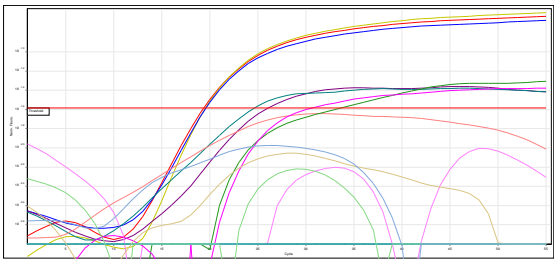
Saponin of varying concentrations were assessed for their effect on human DNA (hDNA) and *P. knowlesi* DNA yield after DNA extraction. A cultured PkA1H1 isolate was taken treated with the stated Saponin concentration and the extracted DNA quantified using NanoDrop 2000 and Qubit platforms



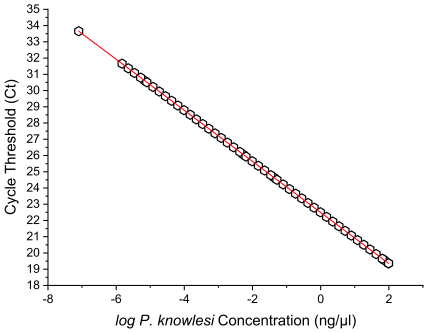
(a) *hDNA Quantitation Analysis Trace*



(b) *hDNA Linear Fit*



(c) *pkDNA Quantitation Analysis Trace*



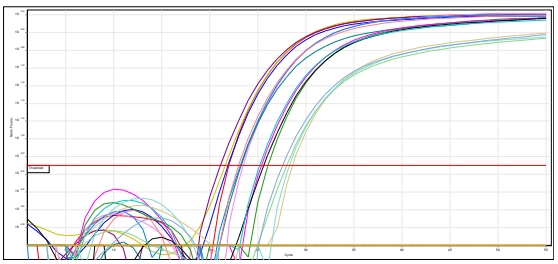
(d) *pkDNA Linear Fit*

Figure A.1: Quantitation analysis traces and plots for hDNA and *pk*DNA generated to optimise Saponin Concentration. Quantitation traces (a,c) generated by the Rotor-Gene Q Series Software after linear fitting (b,d) to the standard calibration plot generated (*see Figure 2.3*). Traces were generated with Dynamic Tube Correction, Slope Correction and a fluorescence threshold of 10 %. Ct values were fit to linear equations described in Figure 2.3.

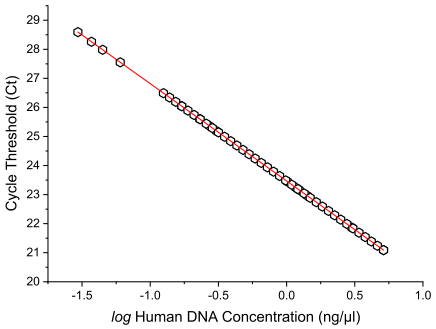
Table A.9: Cycle thresholds from qPCR detection of DNA present in simulated infected Whole Blood after Saponin treatment

Saponin Concentrations	Triplicates	Human DNA (Ct) (a)	Parasite DNA (Ct) (b)
0% - Control	**1**	19.23	19.36
	2	19.69	19.55
	3	19.21	19.59
	Avg.	19.38	19.50
1%	**1**	20.73	26.06
	2	21.58	N.V
	3	20.88	NTC
	Avg.	21.06	26.06
5%	**1**	20.53	24.64
	2	20.94	N.V
	3	21.47	33.66
	Avg.	20.98	29.15
10%	**1**	22.48	30.60
	2	23.17	NTC
	3	23	NTC
	Avg.	22.88	30.60
20%	**1**	21.80	N.V
	2	21.41	N.V
	3	23.31	NTC
	Avg.	22.17	N.C

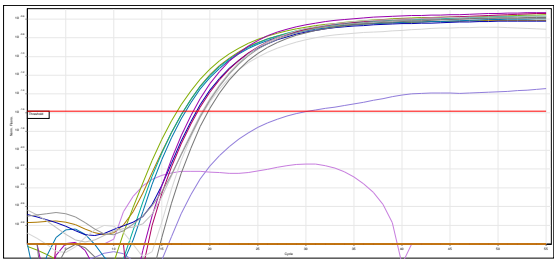
Saponin Concentration refer to concentration of Saponin solution at preparation, prior to being added into the Whole Blood. Once added to 200 µL of simulated Whole Blood, concentrations become 0.1 %, 0.5 %, 1 % and 2 % respectively. Simulated Whole Blood was formulated through the mixture of the Cultured PkA1H1 strain with healthy donor Whole Blood at a 1:2 ratio by volume.



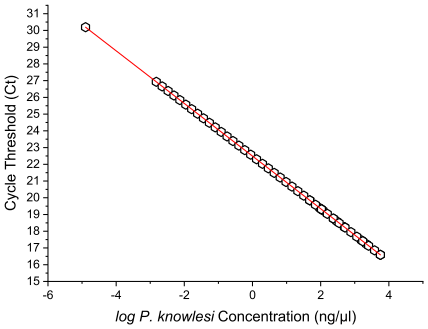
(a) *hDNA Quantitation Analysis Trace*



(b) *hDNA Linear Fit*



(c) *pkDNA Quantitation Analysis Trace*



(d) *pkDNA Linear Fit*

Figure A.2: Quantitation analysis traces and plots for hDNA and *pk*DNA to determine efficiency of formulated *pk*DNA primers. Quantitation traces (a,c) generated by the Rotor-Gene Q Series Software after linear fitting (b,d) to the standard calibration plot generated (*see Figure 2.3*). Traces were generated with Dynamic Tube Correction, Slope Correction and a fluorescence threshold of 10 %. Ct values were fit to linear equations described in Figure 2.3

Table A.10: Cycle threshold values comparing newly formulated pkDNA primers to previously used primers

Isolate	Old pkDNA Primers			New pkDNA Primers		
	1	**2**	**Avg.**	**1**	**2**	**Avg.**
Control	30.2	N.V	30.2	NTC	NTC	N.C
1 :10	17.23	16.59	16.91	17.23	17.46	17.345
1 : 50	18.59	18.25	18.42	18.74	19.38	19.06
1 : 100	19.24	19.32	19.28	19.84	NTC	19.84

Input DNA was extracted from a Cultured PkA1H1 strain isolate with an untreated (Control) concentration of 60.3 ng/µL measured by NanoDrop 2000. A serial dilution was performed resulting in DNA concentrations of 5 ng/µL (1:10), 0.7 ng/µL (1:50) and 0.001 ng/µL (1:100). The Ct values for the hDNA channel are presented here as only the *pk*DNA primers were formulated.

N.V - No value. The sample was run on qPCR however, no Ct value was reported with no specified reason

N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated

NTC - Sample was run using the stated conditions however, qPCR failed due to not surpassing the set threshold.

APPENDIX B

LEUCOCYTE DEPLETION USING CD45 DYNABEADS

Òfìífìí là ńrií, a ò rí òkodoro; ḿbọ́, baba gba-n-gba — ***All we see is shadows, not clarity; but clarity will come, the father of all candor***

Yorùbá adage

B.1 Patient Isolate Information

Table B.1: Patient Isolate information as reported in biobank database

Study Ref.	Date Extracted	Parasitaemia (parasites/µL)		% Life Stage (Rings : Trophs : Schizonts)
		Health Cntr.	*Lab.*	
sks047*	08/02/2008	128544	94245	100 : 0 : 0
sks048*	12/02/2008	85920	78857	49 : 8 : 43
sks050	17/03/2008	-	398675	73 : 16 : 0
sks058*	07/04/2008	148810	186631	12 : 86 : 2
sks070*	11/07/2008	610402	-	-
sks074*	04/09/2008	164280	-	-
sks125	14/12/2008	37468	11592	73 : 15 : 11
sks133	25/11/2008	45240	1689	0 : 78 : 22
sks134	14/11/2008	38024	13563	89 : 11 : 0
sks234	21/06/2009	~25000**	-	-
sks254	25/02/2009	193600	52764	85 : 8 : 1
sks276	29/05/2009	18240	775	100 : 0 :0
sks280	05/05/2009	265200	500	76 : 24 : 0
sks325	30/11/2009	241200	94256	83 : 15 : 1
sks330	03/10/2009	20565	448	92 : 8 : 0
sks331	05/10/2009	236250	25100	98 : 2 : 0
sks333	12/10/2009	127875	11598	94 : 6: 0
sks339*	11/11/2009	321750	52721	78 : 17 : 2
sks344	21/11/2009	794063	63360	100 : 0 : 0
sks367	16/01/2010	35550	15732	100 : 0 : 0

Continued on next page

Table B.1 – *Continued from last page*

Study Ref.	Date Extracted	Parasitaemia (parasites/µL)		% Life Stage (Rings : Trophs : Schizonts)
		Health Cntr.	*Lab.*	

Parasitaemia was recorded at the point of collection at either the hospital or local health centre and later confirmed at the Universiti Malaysia Sarawak (UNIMAS) laboratories. Parasite life stage estimation was carried out in UNIMAS. Percentage life stage does not refer to percentage of whole blood but rather, the percentage of estimated individual parasites within the whole blood volume.

* - Patient isolate with more than one isolate processed through the CD45 DynaBeads leucocyte depletion method

** - Parasitaemia was not measured but inferred from associated case notes

- - Parasitaemia was not available in our BioBank records

Health Cntr. - Hospital or Healthcare centre

Study Ref. - Study reference

Table is presented in alpha-numerical order, not the order of use.

Table B.2: Cultured *P. knowlesi* PkA1H1 experimental strain isolates prepared by Dr Fauzi Muh

Study Reference	%Rings	%Trophozoites	%Schizonts
PkA1H1(c)	60.00	20.00	21.82
PkA1H1(d)	60.00	20.00	21.82
PkA1H1(e)	75.81	20.97	3.23
PkA1H1(f)	75.81	20.97	3.23
PkA1H1(g)	49.18	45.90	4.92
PkA1H1(h)	49.18	45.90	4.92
PkA1H1(i)	49.18	45.90	4.92
PkA1H1(j)	37.70	34.43	27.87
PkA1H1(k)	32.05	33.33	34.62
PkA1H1(l)	43.66	43.66	12.68
PkA1H1(m)	60.78	33.33	5.88
PkA1H1(n)	60.78	33.33	5.88

P. knowlesi parasites were assessed to estimate the proportion of individuals at different major life stages using light microscopy. PkA1H1(a) and PkA1H1(b) were omitted as these were generated by [1] and no such information was available.

B.2 DNA concentrations using NanoDrop 2000 and Qubit Quantification

Table B.3: Qubit DNA concentration quantification of extracted DNA after CD45 DynaBeads leucocyte depletion

Isolates	Strt. Vol. (µL)		Normalised Conc. (ng/µL)			% reduction	
	Ctrl	*CD45 Treated*	*Ctrl*	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*
PkA1H1(c)	10	390	32.84	2.60	10.10	92.08	69.24
PkA1H1(d)	10	390	31.67	2.88	8.66	90.91	72.65
PkA1H1(e)	10	390	34.40	2.92	14.90	91.51	56.68
PkA1H1(f)	10	390	41.34	4.46	15.20	89.21	63.23
PkA1H1(g)	10	390	51.48	3.30	23.20	93.59	54.93
PkA1H1(h)	10	390	44.85	2.48	27.20	94.47	39.35
PkA1H1(i)	10	390	109.98	5.42	45.60	95.07	58.54
PkA1H1(j)	10	390	39.39	18.60	21.20	52.78	46.18
PkA1H1(k)	10	390	31.67	23.60	25.40	25.48	19.79
PkA1H1(l)	10	390	121.68	14.30	44.20	88.25	63.68
PkA1H1(m)	10	390	46.41	4.46	15.90	90.39	65.74
PkA1H1(n)	10	390	109.98	10.00	29.60	90.91	73.09
sks047(a)	50	600	N.V	21.00	N.V	N.C	N.C
sks047(b)	10	900	N.V	47.60	N.V	N.C	N.C
sks048(a)	10	300	N.V	57.40	N.V	N.C	N.C
sks048(b)	10	325	100.10	9.12	14.60	90.89	85.41
sks050	10	480	N.V	93.40	N.V	N.C	N.C
sks058	50	450	N.V	9.90	N.V	N.C	N.C
sks058(b)	10	500	N.V	33.60	N.V	N.C	N.C
sks070(a)	20	370	N.V	13.20	N.V	N.C	N.C

Continued on next page

Table B.3 – *Continued from last page*

Isolates	**Strt. Vol. (µL)**		**Normalised Conc. (ng/µL)**			**% reduction**	
	Ctrl	***CD45 Treated***	***Ctrl***	***CD45 Treated***	***Pellet***	***CD45 Treated***	***Pellet***
sks070(b)	10	450	212.40	13.70	36.80	93.55	82.67
sks125	50	300	N.V	6.52	N.V	N.C	N.C
sks133	50	750	N.V	0.38	N.V	N.C	N.C
sks134	50	460	N.V	0.76	N.V	N.C	N.C
sks254	50	500	N.V	30.60	N.V	N.C	N.C
sks276	10	1080	N.V	3.38	N.V	N.C	N.C
sks280	10	550	N.V	6.08	N.V	N.C	N.C
sks325	10	330	118.14	12.50	N.V	89.42	N.C
sks330	10	180	N.V	9.04	N.V	N.C	N.C
sks331	10	140	N.V	6.78	N.V	N.C	N.C
sks333	50	425	N.V	16.40	N.V	N.C	N.C
sks339(a)	10	190	N.V	116.00	N.V	N.C	N.C
sks339(b)	20	200	N.V	11.40	N.V	N.C	N.C
sks344	10	410	N.V	71.20	N.V	N.C	N.C
sks367	50	260	N.V	5.78	N.V	N.C	N.C

Complete list of isolates that had DNA quantification on the Qubit DNA quantification platform, however, not all isolates were quantified on this platform. Isolates which were not quantified using the Qubit platform and excluded from this table are *sks074(a), sks234, sks074(b), PkA1H1(a,b)*. These isolates were quantified using the NanoDrop2000 platform [Appendix Table B.4]. Raw control DNA concentrations (*not shown*) were normalised for comparison using the equation: $(CD45\ Treated\ Starting\ Volume - Control\ Starting\ Volume) * Measured\ Control\ concentration$. Qubit was commonly only used prior to carry out sequencing, hence only the CD45 Treated sample was available in these instances

Strt. Vol. - Starting volume

Normalised Conc. - Normalised concentration

N.D - Not done. Refers to instances where the isolate's DNA concentration was not quantified using the Qubit platform. In many cases where this occurs, DNA quantification was carried out on the NanoDrop2000 platform

N.C - Not Calculated. Occurs when the values required to carry out a calculation are unavailable to facilitate the calculation

Table is presented in alphabetical order, not the order of use

Table B.4: NanoDrop 2000 DNA concentration quantification of extracted DNA after CD45 DynaBeads leucocyte depletion

Isolates	Starting volume (µL)		Normalised DNA Concentration measured on NanoDrop 2000 (ng/µL)									Average % reduction	
	Control	*CD45 Treated*	*Control*			*CD45 Treated*			*Pellet*			*CD45 Treated*	*Pellet*
			1	**2**	**Avg**	**1**	**2**	**Avg**	**1**	**2**	**Avg**		
PkA1H1(a)	20	180	9.90	9.90	9.90	0.50	0.90	0.70	2.40	2.20	2.30	92.93	76.77
PkA1H1(b)	20	180	16.20	8.10	12.15	1.60	1.00	1.30	1.20	1.70	1.45	89.30	88.07
sks047(a)	50	600	154.80	156.00	155.40	4.30	3.60	3.95	42.40	42.40	42.40	97.46	72.72
sks047(b)	10	900	N.V	N.V	N.C	10.70	10.90	10.80	N.V	N.V	N.C	N.C	N.C
sks048(a)	10	300	N.V	N.V	N.C	12.20	11.30	11.75	N.V	N.V	N.C	N.C	N.C
sks050	10	480	254.40	187.20	220.80	9.40	9.90	9.65	12.40	12.20	12.30	95.63	94.43
sks058	50	450	57.60	62.10	59.85	2.20	1.80	2.00	26.80	27.40	27.10	96.66	54.72
sks058(b)	10	500	N.V	N.V	N.C	3.60	3.30	3.45	N.V	N.V	N.C	N.C	N.C
sks070(a)	20	370	N.V	N.V	N.C	18.20	15.70	16.95	N.V	N.V	N.C	N.C	N.C
sks074(a)	50	450	83.70	82.80	83.25	3.10	2.60	2.85	41.20	41.30	41.25	96.58	50.45
sks074(b)	25	470	78.96	77.08	78.02	1.50	2.30	1.90	28.50	28.00	28.25	97.56	63.79
sks074(c)	10	530	100.70	132.50	116.60	0.30	0.80	0.55	28.20	28.70	28.45	99.53	75.60
sks125	50	300	61.80	60.00	60.90	1.50	1.50	1.50	16.80	16.00	16.40	97.54	73.07
sks133	50	750	118.50	118.50	118.50	0.90	0.90	0.90	56.30	56.10	56.20	99.24	52.57
sks134	50	460	67.16	64.40	65.78	1.00	0.40	0.70	20.70	21.10	20.90	98.94	68.23
sks234	25	650	15.60	20.80	18.20	1.10	0.50	0.80	25.00	24.70	24.85	95.60	-36.54
sks254	50	500	134.00	143.00	138.50	12.20	12.10	12.15	65.00	63.80	64.40	91.23	53.50
sks330	10	180	N.V	N.V	N.C	1.10	N.V	1.10	N.V	N.V	N.C	N.C	N.C
sks331	10	140	42.00	50.40	46.20	2.70	1.90	2.30	24.50	26.90	25.70	95.02	44.37
sks333	50	425	87.55	91.80	89.68	1.70	2.00	1.85	37.50	37.60	37.55	97.94	58.13
sks339(a)	10	190	N.V	N.V	N.C	11.25	N.V	11.25	N.V	N.V	N.C	N.C	N.C
sks339(b)	20	200	N.V	N.V	N.C	23.00	23.40	23.20	N.V	N.V	N.C	N.C	N.C
sks344	10	410	323.90	311.60	317.75	8.90	9.80	9.35	35.80	36.20	36.00	97.06	88.67
sks367	50	260	43.68	39.52	41.60	4.20	2.60	3.40	18.70	18.00	18.35	91.83	55.89

List of isolates with DNA quantification on the NanoDrop 2000. Isolates which were not quantified using the NanoDrop 2000 and excluded from this table are *sks276, sks280, sks048(b), sks070(b), sks325, Cultured_PkA1H1(c-n)*. These isolates were quantified using the Qubit platform [Appendix Table B.3]. Raw control DNA concentrations (*not shown*) were normalised for comparison using the equation: $(CD45\ Treated\ Starting\ Volume - Control\ Starting\ Volume) * Measured\ Control\ concentration.$

N.D - Not done. Refers to instances where the isolate's DNA concentration was not quantified using the NanoDrop 2000 platform. In many cases where this occurs, DNA quantification was carried out on the Qubit quantification platform

N.C - Not Calculated. Occurs when the values required to carry out a calculation are unavailable to facilitate the calculation

Table is presented in alphabetical order, not the order of use.

B.3 Raw qPCR Cycle Threshold Values after CD45 Treatment

B.3.1 Raw Human DNA Ct Values

Table B.5: Raw human DNA average cycle thresholds after leucocyte depletion using the CD45 DynaBeads

Isolates	Strt vol. (µL)		Measured Cycled times (Ct)								
	Ctrl.	*CD45 Treated*	*Control*			*CD45 Treated*			*Pellet*		
			1	2	Avg.	1	2	Avg.	1	2	Avg.
PkA1H1(a)	20	180	24.54	25.00	24.77	27.96	27.72	27.84	21.83	21.76	21.80
PkA1H1(b)	20	180	25.36	25.18	25.27	31.14	36.25	33.70	23.60	23.57	23.59
PkA1H1(c)	10	390	26.43	25.27	25.85	27.48	27.37	27.43	20.74	20.57	20.66
PkA1H1(d)	10	390	25.80	25.57	25.69	25.74	26.48	26.11	21.08	20.99	21.04
PkA1H1(e)	10	390	25.41	25.28	25.35	27.10	26.30	26.70	20.23	20.01	20.12
PkA1H1(f)	10	390	24.66	25.40	25.03	25.90	26.30	26.10	20.47	20.10	20.29
PkA1H1(g)	10	390	24.00	25.01	24.51	25.26	25.82	25.54	18.85	19.06	18.96
PkA1H1(h)	10	390	24.67	24.40	24.54	25.78	25.22	25.50	18.85	18.76	18.81
PkA1H1(i)	10	390	22.87	23.27	23.07	22.87	23.02	22.95	17.93	17.73	17.83
PkA1H1(j)	10	390	26.22	26.84	26.53	M.Ct	26.12	26.12	20.36	20.18	20.27
PkA1H1(k)	10	390	27.40	26.86	27.13	26.23	25.09	25.66	19.49	19.30	19.40
PkA1H1(l)	10	390	23.10	23.03	23.07	23.46	22.98	23.22	18.75	18.20	18.48
PkA1H1(m)	10	390	24.96	25.01	24.99	28.01	27.71	27.86	19.51	20.48	20.00
PkA1H1(n)	10	390	22.35	23.42	22.89	24.53	23.11	23.82	18.39	18.64	18.52
sks047(a)	50	600	19.06	20.07	19.57	22.39	22.46	22.43	17.55	18.25	17.90
sks047(b)	10	900	22.78	22.37	22.58	21.01	20.91	20.96	17.44	17.51	17.48
sks048(a)	10	300	22.16	21.96	22.06	25.27	25.40	25.34	18.90	19.16	19.03
sks048(b)	10	325	21.83	22.16	22.00	23.50	22.91	23.21	19.19	18.98	19.09
sks050	10	480	20.81	21.39	21.10	22.42	23.01	22.72	18.97	19.51	19.24
sks058	50	450	20.75	21.01	20.88	24.40	24.04	24.22	18.98	18.60	18.79
sks058(b)	10	500	23.03	23.69	23.36	23.21	23.18	23.20	18.96	19.00	18.98
sks070	10	450	20.98	21.31	21.15	19.33	19.60	19.47	18.30	17.98	18.14
sks070(a)	20	370	19.59	19.76	19.68	19.45	20.20	19.83	18.35	18.07	18.21
sks074(b)	25	470	20.83	21.03	20.93	24.73	25.02	24.88	18.37	18.46	18.42

Continued on next page

Table B.5 – *Continued from last page*

Isolates	Strt vol. (µL)		Measured Cycled times (Ct)								
	Ctrl.	*CD45 Treated*	*Control*			*CD45 Treated*			*Pellet*		
			1	**2**	**Avg.**	**1**	**2**	**Avg.**	**1**	**2**	**Avg.**
sks074(c)	10	530	21.36	21.52	21.44	23.37	23.47	23.42	18.16	18.51	18.34
sks125	50	300	20.43	19.86	20.15	24.06	23.74	23.90	19.33	19.60	19.47
sks133	50	750	20.39	20.45	20.42	25.63	25.37	25.50	17.59	17.41	17.50
sks134	50	460	20.31	20.47	20.39	23.88	24.37	24.13	18.79	18.99	18.89
sks234	25	650	21.80	21.66	21.73	21.15	21.58	21.37	18.40	18.53	18.47
sks254	50	500	19.03	19.00	19.02	19.96	20.29	20.13	17.34	16.97	17.16
sks276	10	1080	17.76	17.68	17.72	22.75	22.44	22.60	21.97	23.23	22.60
sks280	10	550	23.41	23.10	23.26	23.94	23.81	23.88	18.72	19.11	18.92
sks325	10	330	22.50	22.73	22.62	24.26	23.94	24.10	19.55	19.40	19.48
sks330	10	180	22.35	22.23	22.29	24.15	23.80	23.98	19.00	18.96	18.98
sks331	10	140	22.39	22.55	22.47	24.56	24.65	24.61	18.67	18.89	18.78
sks333	50	425	20.50	20.25	20.38	21.99	22.94	22.47	20.34	18.75	19.55
sks339(a)	10	190	21.20	21.40	21.30	22.19	21.64	21.92	17.98	18.22	18.10
sks339(b)	20	200	19.46	19.52	19.49	20.28	21.27	20.78	17.77	18.44	18.11
sks344	10	410	20.20	20.01	20.11	21.41	21.99	21.70	18.17	17.82	18.00
sks367	50	260	20.65	20.60	20.63	21.38	21.78	21.58	18.98	19.26	19.12

Ct values reported from the Rotor-Gene Q Series Software from the Yellow gain channel which detects human DNA fluorescence. For each condition an isolate was treated with, the extracted DNA was run in duplicate for each qPCR experiment. Where multiple Whole blood aliquots of a sample were carried out, lettered denotations are given such as sks047(a). PkA1H1 refers to the in-house cultured *P. knowlesi* PKA1H1 strain.

Strt. vol. - Starting volume

Ctrl - Control. Untreated blood acting as isolate and experiment control

CD45 Treated - Whole blood to be leucocyte depleted using CD45 DynaBeads

Pellet - Once the CD45 DynaBeads have been magnetically separated, the resulting Pellet has DNA extraction carried out and assessed

M.Ct - Reported when the calculated fluorescence curve crosses the threshold value at least twice thus a true Ct could not be determined

Table is presented in alphabetical order, not the order of use.

B.3.2 Raw *P. knowlesi* DNA Ct Values

Table B.6: Raw *P. knowlesi* DNA average cycle thresholds after leucocyte depletion and parasite DNA enrichment using the CD45 DynaBeads

Isolates	Strt vol. (µL)		Measured Cycled times (Ct)								
	Ctrl.	*CD45 Treated*	*Control*			*CD45 Treated*			*Pellet*		
			1	2	Avg.	1	2	Avg.	1	2	Avg.
PkA1H1(a)	20	180	19.87	19.91	19.89	17.42	17.24	17.33	19.06	19.26	19.16
PkA1H1(b)	20	180	21.00	21.22	21.11	18.35	19.65	19.00	21.62	21.28	21.45
PkA1H1(c)	10	390	22.06	20.32	21.19	18.43	18.42	18.43	17.15	17.07	17.11
PkA1H1(d)	10	390	20.99	19.42	20.21	17.81	17.91	17.86	17.95	17.62	17.79
PkA1H1(e)	10	390	21.50	21.07	21.29	18.65	18.24	18.45	17.31	17.09	17.20
PkA1H1(f)	10	390	20.27	21.08	20.68	18.55	17.68	18.12	17.93	17.05	17.49
PkA1H1(g)	10	390	21.39	22.89	22.14	18.53	18.24	18.39	16.89	17.05	16.97
PkA1H1(h)	10	390	22.19	21.54	21.87	19.30	19.19	19.25	17.82	17.46	17.64
PkA1H1(i)	10	390	20.07	20.88	20.48	16.86	16.82	16.84	16.35	15.77	16.06
PkA1H1(j)	10	390	21.07	23.21	22.14	15.61	15.01	15.31	16.27	16.55	16.41
PkA1H1(k)	10	390	20.61	20.49	20.55	15.04	15.34	15.19	15.77	16.14	15.96
PkA1H1(l)	10	390	20.73	20.16	20.45	15.89	15.70	15.80	16.86	16.43	16.65
PkA1H1(m)	10	390	25.15	25.26	25.21	17.59	17.51	17.55	21.95	22.60	22.28
PkA1H1(n)	10	390	20.00	21.43	20.72	16.16	15.98	16.07	20.50	22.37	21.44
sks047(a)	50	600	18.86	19.72	19.29	18.44	19.00	18.72	20.74	21.34	21.04
sks047(b)	10	900	M.Ct	23.88	23.88	17.66	17.30	17.48	24.24	N.V	24.24
sks048(a)	10	300	21.18	21.84	21.51	17.36	18.19	17.78	N.V	M.Ct	N.C
sks048(b)	10	325	19.57	19.84	19.71	15.69	15.66	15.68	18.40	18.47	18.44
sks050	10	480	17.92	17.95	17.94	15.28	15.62	15.45	17.56	17.44	17.50
sks058	50	450	20.88	21.05	20.97	20.19	20.29	20.24	22.06	21.77	21.92
sks058(b)	10	500	27.64	M.Ct	27.64	21.77	20.82	21.30	N.V	28.57	28.57
sks070	10	450	22.67	22.69	22.68	18.64	19.10	18.87	20.72	20.30	20.51
sks070(a)	20	370	18.80	18.63	18.72	14.93	15.74	15.34	21.92	23.55	22.74
sks074(b)	25	470	24.87	27.98	26.43	22.00	21.34	21.67	24.21	25.39	24.80
sks074(c)	10	530	22.34	22.70	22.52	18.46	18.62	18.54	21.81	22.16	21.99
sks125	50	300	24.00	24.06	24.03	21.69	20.86	21.28	26.73	26.30	26.52
sks133	50	750	43.66	33.50	38.58	25.23	24.60	24.92	N.V	N.V	N.C
sks134	50	460	N.V	N.V	N.C	27.37	N.V	27.37	NTC	N.V	N.C
sks234	25	650	NTC	NTC	N.C	26.54	31.66	29.10	N.V	N.V	N.C
sks254	50	500	21.56	21.22	21.39	22.86	23.12	22.99	25.14	25.15	25.15
sks276	10	1080	N.V	N.V	N.C	32.21	30.87	31.54	N.V	N.V	N.C

Continued on next page

Table B.6 – *Continued from last page*

Isolates	Strt vol. (µL)		Measured Cycled times (Ct)								
	Ctrl.	*CD45 Treated*	*Control*			*CD45 Treated*			*Pellet*		
			1	**2**	**Avg.**	**1**	**2**	**Avg.**	**1**	**2**	**Avg.**
sks280	10	550	N.V	NTC	N.C	29.69	28.41	28.41	N.V	N.V	N.C
sks325	10	330	20.93	20.80	20.87	17.07	17.14	17.11	20.59	20.04	20.32
sks330	10	180	25.19	24.77	24.98	22.06	21.61	21.84	25.84	24.29	25.07
sks331	10	140	23.62	23.60	23.61	20.45	20.74	20.60	2.39	23.58	12.99
sks333	50	425	24.44	23.46	23.95	20.54	21.50	21.02	N.V	NTC	N.C
sks339(a)	10	190	19.44	19.13	19.29	16.66	16.48	16.57	18.75	18.66	18.71
sks339(b)	20	200	17.76	17.68	17.72	15.50	15.86	15.68	18.41	19.01	18.71
sks344	10	410	18.66	18.23	18.45	15.65	15.50	15.58	18.50	18.24	18.37
sks367	50	260	26.29	29.06	27.68	23.31	24.29	23.80	N.V	N.V	N.C

Ct values reported from the Rotor-Gene Q Series Software from the Green gain channel which detects *[P. knowlesi]* DNA fluorescence. For each condition an isolate was treated with, the extracted DNA was run in duplicate for each qPCR experiment. Where multiple Whole blood aliquots of a sample were carried out, lettered denotations are given such as sks047(a). PkA1H1 refers to the in-house cultured *P. knowlesi* PKA1H1 strain.

Strt. vol. - Starting volume

Ctrl - Control. Untreated blood acting as isolate and experiment control

CD45 Treated - Whole blood to be leucocyte depleted using CD45 DynaBeads

Pellet - Once the CD45 DynaBeads have been magnetically separated, the resulting Pellet has DNA extraction carried out and assessed

M.Ct - Reported when the calculated fluorescence curve crosses the threshold value at least twice thus a true Ct could not be determined

N.A - Not Applicable. This refers to conditions not applicable to the corresponding experiment because the conditions were not carried out within that experiment

N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated

NTC - Sample was run using the CD45 DynaBeads depletion method however, qPCR failed due to not surpassing the set threshold.

N.V - Sample was run using the CD45 DynaBeads depletion method however, qPCR report showed no reading; outputting, a blank reading for unspecified reasons

Table is presented in alphabetical order, not the order of use.

B.4 Normalised qPCR Cycle Threshold Values after CD45 Treatment

B.4.1 Normalised Human DNA Ct Values

Table B.7: Normalised human DNA average cycle thresholds from isolates after leucocyte depletion using the CD45 DynaBeads

Isolate	Adj. fac	Normalised Avg. Ct values			Delta Change		% inputs		% reductions	
		Control	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*
PkA1H1(a)	3.17	21.60	27.84	21.80	-6.24	-0.19	1.32	87.36	98.68	12.64
PkA1H1(b)	3.17	22.10	33.70	23.59	-11.59	-1.48	0.03	35.73	99.97	64.27
PkA1H1(c)	5.29	20.56	27.43	20.66	-6.86	-0.09	0.86	93.93	99.14	6.07
PkA1H1(d)	5.29	20.40	26.11	21.04	-5.71	-0.64	1.91	64.38	98.09	35.62
PkA1H1(e)	5.29	20.06	26.70	20.12	-6.64	-0.06	1.00	95.90	99.00	4.10
PkA1H1(f)	5.29	19.74	26.10	20.29	-6.36	-0.54	1.22	68.76	98.78	31.24
PkA1H1(g)	5.29	19.22	25.54	18.96	-6.32	0.26	1.25	120.13	98.75	-20.13
PkA1H1(h)	5.29	19.25	25.50	18.81	-6.25	0.44	1.31	136.09	98.69	-36.09
PkA1H1(i)	5.29	17.78	22.95	17.83	-5.16	-0.05	2.80	96.90	97.20	3.10
PkA1H1(j)	5.29	21.24	26.12	20.27	-4.88	0.97	3.41	196.51	96.59	-96.51
PkA1H1(k)	5.29	21.84	25.66	19.40	-3.82	2.45	7.10	546.26	92.90	-446.26
PkA1H1(l)	5.29	17.78	23.22	18.48	-5.44	-0.70	2.30	61.75	97.70	38.25
PkA1H1(m)	5.29	19.70	27.86	20.00	-8.16	-0.30	0.35	81.48	99.65	18.52
PkA1H1(n)	5.29	17.60	23.82	18.52	-6.22	-0.92	1.34	53.02	98.66	46.98
sks047(a)	3.58	15.98	22.43	17.90	-6.44	-1.92	1.15	26.43	98.85	73.57
sks047(b)	6.49	16.08	20.96	17.48	-4.88	-1.39	3.40	38.11	96.60	61.89
sks048(a)	4.91	17.15	25.34	19.03	-8.18	-1.88	0.34	27.23	99.66	72.77
sks048(b)	5.02	16.97	23.21	19.09	-6.23	-2.11	1.33	23.13	98.67	76.87
sks050	5.58	15.52	22.72	19.24	-7.20	-3.72	0.68	7.56	99.32	92.44
sks058	3.17	17.71	24.22	18.79	-6.51	-1.08	1.10	47.31	98.90	52.69
sks058(b)	5.64	17.72	23.20	18.98	-5.48	-1.26	2.24	41.64	97.76	58.36
sks070	5.49	15.65	19.47	18.14	-3.81	-2.49	7.12	17.84	92.88	82.16
sks070(a)	4.21	15.47	19.83	18.21	-4.36	-2.74	4.87	14.92	95.13	85.08
sks074(b)	4.23	16.70	24.88	18.42	-8.18	-1.72	0.35	30.40	99.65	69.60

Continued on next page

Table B.7 – *Continued from last page*

Isolate	Adj. fac	Normalised Avg. Ct values			Delta Change		% inputs		% reductions	
		Control	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*
sks074(c)	5.73	15.71	23.42	18.34	-7.71	-2.62	0.48	16.23	99.52	83.77
sks125	2.58	17.56	23.90	19.47	-6.34	-1.90	1.23	26.70	98.77	73.30
sks133	3.91	16.51	25.50	17.50	-8.99	-0.99	0.20	50.46	99.80	49.54
sks134	3.20	17.19	24.13	18.89	-6.94	-1.70	0.82	30.74	99.18	69.26
sks234	4.70	17.03	21.37	18.47	-4.34	-1.44	4.95	36.97	95.05	63.03
sks254	3.32	15.69	20.13	17.16	-4.43	-1.46	4.63	36.30	95.37	63.70
sks276	6.75	10.97	22.60	22.60	-11.63	-11.63	0.03	0.03	99.97	99.97
sks280	5.78	17.47	23.88	18.92	-6.40	-1.44	1.18	36.82	98.82	63.18
sks325	5.04	17.57	24.10	19.48	-6.53	-1.90	1.08	26.71	98.92	73.29
sks330	4.17	18.12	23.98	18.98	-5.85	-0.86	1.73	55.10	98.27	44.90
sks331	3.81	18.66	24.61	18.78	-5.94	-0.12	1.63	92.19	98.37	7.81
sks333	3.09	17.29	22.47	19.55	-5.18	-2.26	2.76	20.91	97.24	79.09
sks339(a)	4.25	17.05	21.92	18.10	-4.86	-1.05	3.44	48.37	96.56	51.63
sks339(b)	3.32	16.17	20.78	18.11	-4.61	-1.94	4.10	26.12	95.90	73.88
sks344	5.36	14.75	21.70	18.00	-6.95	-3.25	0.81	10.53	99.19	89.47
sks367	2.38	18.25	21.58	19.12	-3.33	-0.87	9.92	54.58	90.08	45.42

Average hDNA Ct values were calculated from the Raw hDNA Ct values [Appendix Table B.5]. Aevrage control Ct value were normalised based on a calculated adjustment factor. The adjustment factor, delta change and percentage input values were calculated as described in chapter 2 subsection 2.4.9. Where multiple Whole blood aliquots of a sample were carried out, lettered denotations are given such as sks047(a). PkA1H1 refers to the in-house cultured *P. knowlesi* PKA1H1 strain.
Control - Untreated blood acting as isolate and experiment control
CD45 Treated - Whole blood to be leucocyte depleted depending on the conditions of the experiment
Pellet - DNA extracted from the separated DynaBeads pellet from the associated experiment
Delta change - Calculated by the subtraction of the normalised treat Ct from the normalised Control Ct value to determine the amount of change after CD treatment
% input - Calculated using the equation: $100 * 2^{\delta change}$
Table is presented in alphabetical order, not the order of use.

B.4.2 Normalised *P. knowlesi* DNA Ct Values

Table B.8: Normalised *P. knowlesi* DNA average cycle thresholds from isolates after leucocyte depletion using the CD45 DynaBeads

Isolate	Adj. fac	Normalised Avg. Ct values			Delta Change		% inputs		% reductions	
		Control	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*
PkA1H1(a)	3.17	16.72	17.33	19.16	-0.61	-2.44	65.52	18.43	34.48	81.57
PkA1H1(b)	3.17	17.94	19.00	21.45	-1.06	-3.51	47.97	8.78	52.03	91.22
PkA1H1(c)	5.29	15.90	18.43	17.11	-2.52	-1.21	17.43	43.36	82.57	56.64
PkA1H1(d)	5.29	14.92	17.86	17.79	-2.94	-2.87	13.03	13.72	86.97	86.28
PkA1H1(e)	5.29	16.00	18.45	17.20	-2.45	-1.20	18.36	43.52	81.64	56.48
PkA1H1(f)	5.29	15.39	18.12	17.49	-2.73	-2.10	15.12	23.32	84.88	76.68
PkA1H1(g)	5.29	16.85	18.39	16.97	-1.53	-0.12	34.62	92.31	65.38	7.69
PkA1H1(h)	5.29	16.58	19.25	17.64	-2.67	-1.06	15.76	47.95	84.24	52.05
PkA1H1(i)	5.29	15.19	16.84	16.06	-1.65	-0.87	31.86	54.70	68.14	45.30
PkA1H1(j)	5.29	16.85	15.31	16.41	1.54	0.44	291.72	136.09	-191.72	-36.09
PkA1H1(k)	5.29	15.26	15.19	15.96	0.07	-0.69	105.31	61.97	-5.31	38.03
PkA1H1(l)	5.29	15.16	15.80	16.65	-0.64	-1.49	64.38	35.71	35.62	64.29
PkA1H1(m)	5.29	19.92	17.55	22.28	2.37	-2.36	516.80	19.54	-416.80	80.46
PkA1H1(n)	5.29	15.43	16.07	21.44	-0.64	-6.01	64.15	1.56	35.85	98.44
sks047(a)	3.58	15.71	18.72	21.04	-3.01	-5.33	12.37	2.48	87.63	97.52
sks047(b)	6.49	17.39	17.48	24.24	-0.09	-6.85	93.83	0.87	6.17	99.13
sks048(a)	4.91	16.60	17.78	N.C	-1.17	N.C	44.38	N.C	55.62	N.C
sks048(b)	5.02	14.68	15.68	18.44	-0.99	-3.75	50.27	7.42	49.73	92.58
sks050	5.58	12.35	15.45	17.50	-3.10	-5.15	11.66	2.82	88.34	97.18
sks058	3.17	17.80	20.24	21.92	-2.44	-4.12	18.37	5.75	81.63	94.25
sks058(b)	5.64	22.00	21.30	28.57	0.70	-6.57	162.58	1.05	-62.58	98.95
sks070(a)	5.49	17.19	18.87	20.51	-1.68	-3.32	31.17	10.00	68.83	90.00
sks070(b)	4.21	14.51	15.34	22.74	-0.83	-8.23	56.27	0.33	43.73	99.67
sks074(b)	4.23	22.19	21.67	24.80	0.52	-2.61	143.63	16.41	-43.63	83.59
sks074(c)	5.73	16.79	18.54	21.99	-1.75	-5.19	29.77	2.73	70.23	97.27
sks125	2.58	21.45	21.28	26.52	0.17	-5.07	112.51	2.98	-12.51	97.02
sks133	3.91	34.67	24.92	N.C	9.76	N.C	86593	N.C	-86493	N.C
sks134	3.20	N.C	27.37	N.C	N.C	N.C	N.C	N.C	N.C	N.C
sks234	4.70	N.C	29.10	N.C	N.C	N.C	N.C	N.C	N.C	N.C
sks254	3.32	18.07	22.99	25.15	-4.92	-7.08	3.30	0.74	96.70	99.26
sks276	6.75	N.C	31.54	N.C	N.C	N.C	N.C	N.C	N.C	N.C
sks280	5.78	N.C	29.05	N.C	N.C	N.C	N.C	N.C	N.C	N.C

Continued on next page

Table B.8 – *Continued from last page*

Isolate	Adj. fac	Normalised Avg. Ct values			Delta Change		% inputs		% reductions	
		Control	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*	*CD45 Treated*	*Pellet*
sks325	5.04	15.82	17.11	20.32	-1.28	-4.49	41.05	4.44	58.95	95.56
sks330	4.17	20.81	21.84	25.07	-1.02	-4.25	49.14	5.24	50.86	94.76
sks331	3.81	19.80	20.60	12.99	-0.79	6.82	57.74	11280	42.26	-11180
sks333	3.09	20.86	21.02	N.C	-0.16	N.C	89.66	N.C	10.34	N.C
sks339(a)	4.25	15.04	16.57	18.71	-1.53	-3.67	34.56	7.87	65.44	92.13
sks339(b)	3.32	14.40	15.68	18.71	-1.28	-4.31	41.12	5.03	58.88	94.97
sks344	5.36	13.09	15.58	18.37	-2.49	-5.28	17.83	2.57	82.17	97.43
sks367	2.38	25.30	23.80	N.C	1.50	N.C	282.16	N.C	-182.16	N.C

Average *P. knowlesi* Ct values were calculated from the Raw *P. knowlesi* Ct values [Appendix Table B.6]. Average control Ct value were normalised based on a calculated adjustment factor. The adjustment factor, delta change and percentage input values were calculated as described in chapter 2 subsection 2.4.9. Where multiple Whole blood aliquots of a sample were carried out, lettered denotations are given such as sks047(a). PkA1H1 refers to the in-house cultured *P. knowlesi* PKA1H1 strain.

Control - Untreated blood acting as isolate and experiment control

CD45 Treated - Whole blood to be leucocyte depleted depending on the conditions of the experiment

Pellet - DNA extracted from the separated DynaBeads pellet from the associated experiment

Delta change - Calculated by the subtraction of the normalised treat Ct from the normalised Control Ct value to determine the amount of change after CD treatment

% input - Calculated using the equation: $100 * 2^{\delta change}$

N.C - Not calculated. Here the condition is carried out however due to different factors, the average cycle threshold could not be calculated

Table is presented in alphabetical order, not the order of use.

B.5 Statistical Tests

B.5.1 Normality Tests

Table B.9: Normality tests for Ct values of each treatment

Shapiro-Wilk Test Normality test		
Isolate	***p-value***	**Conclusion**
Control - hDNA	0.08795	p >0.05, the data are normally distributed
CD45 Treated - hDNA	0.02299	p <0.05, the data are not normally distributed
Pellet - hDNA	0.0009134	p <0.05, the data are not normally distributed
Control - *pk*DNA	7.38E-06	p <0.05, the data are not not normally distributed
CD45 Treated - *pk*DNA	0.0001358	p <0.05, the data are not not normally distributed
Pellet - *pk*DNA	0.2732	p >0.05, the data are normally distributed

Normality test was implemented in R, using the *shapiro.test* function. Each dataset includes all samples (n=40) however, the *pk*DNA control and pellet subsets contain samples with no recorded Ct. These samples were not removed prior to the normality test however, the blank values were omitted by the *shapiro.test* function.

Appendix C

Programming scripts and code generated

Òfìífìí là ńrií, a ò rí òkodoro; ḿbọ́, baba gba-n-gba — ***All we see is shadows, not clarity; but clarity will come, the father of all candor***

Yorùbá adage

C.1 Development commands to assess the impact of basecaller and demultiplexer

```
################################################################
# Basecalling command for Albacore
sf5_read_fast5_basecaller.py -i fast5/ -t $THREAD_COUNT -s
    Basecalled_Output -f FLO-MIN106 -k SQK-RBK004 --barcoding -o fast5,
    fastq -q 0
### where "-i" represents the input folder, "-s" is the folder to save
    output, "-f" and "-k" are the flowcell and library preparation used
     respectively.
################################################################
# Basecalling command for Guppy
guppy_basecaller --flowcell FLO-MIN106 --kit SQK-RBK004 --fast5_out -i
    Raw_Fast5_folder/ -s Basecalled_Output/ -v -r -q 0 --
    qscore_filtering --num_callers 5 --cpu_threads_per_caller 2
### where "--flowcell" and "--kit" indicate the flowcell type and
    library preparation used, "--fast5_out" asks for additional outputs
     as FAST5 files, "-i" is the input folder holding the FAST5 files
    and "-s" is the output folder to save the basecalled files. Further
     options like "-r" ensure the command is carried out recursively
    through all sub-folders of the input folder, "-q" asks for reads to
     be placed in a single FASTQ file and "--qscore_filtering" ensures
    that the basecalled reads are filtered based on the quality
    threshold (default: Q7). "--num_callers" and "--
    cpu_threads_per_caller" are for parallelising the command for
    greater processing speed where num_callers asks for five parallel
    commands, with each command having 2 CPU cores to run.
################################################################
# Guppy demultiplexing:
guppy_barcoder -i Basecalled_Data/ -s Demultiplex/
    GuppyDemultiplex_Output --barcode_kits "SQK-RBK004" -t
    $THREAD_COUNT -q 0
### where "-i" is the input folder of Guppy basecalled reads, "-s" is
    the output folder, "--barcode_kits" is the library preparation kit
    used. "-t" is the number of threads to use to speed up processing
    and "-q" once again outputs all reads of an identified barcode into
     a single FASTQ file
################################################################
# qcat demultiplexing
cat Basecalled/pass/* | qcat -b ~/Demultiplex/qcat_output --detect-
    middle -t $THREAD_COUNT --trim -k RBK004
### where "-b" indicates the output folder to place demultiplexed reads
    , "--detect_middle" checks for barcodes within reads, thereby
```

```
    limiting the chance of chimeric reads and "-t" again states the
    number of CPU processing threads to utilise. "--trim" allows for
    trimming the barcode sequence from the identified reads, while "-k"
     indicates the library preparation kit used.
################################################################
# Porechop:
porechop -i Demultiplex/isolate.fastq -o AdapterRemoved/isolate.fastq
    --verbosity 2 --threads $THREAD_COUNT
### where "-i" is the input fastq file and "-o" is the output file to
    be created. "--verbosity" refers to the level of information to be
    printed for the user "--threads" again is the number of processors
    to use for the command.
################################################################
# fastqc:
fastqc -t $THREAD_COUNT -o output_folder input.fastq
### where "-t" is the number of processors to use and "-o" is the
    output folder to place the output plots
################################################################
```

Code C.1: Commands utilised to carry out basecalling and demultiplexing of sequenced data.

C.2 Pipeline commands for data processing

C.2.1 Basecalling, Demultiplexing, Adapter Removal and Alignment

```
#!/bin/bash
#### guppy command ####
path_to_guppy_package_folder/guppy_basecaller -i ${raw_files_folder} --
    save_path ${output_folder} --flowcell FLO-MIN106 --kit SQK-RBK004 -
    r -v -q 0 --qscore_filtering -x auto
#### demultiplexing command ####
cat path_to_basecalled_data/pass/* | qcat -b ${output_folder} --detect-
    middle -t THREADS --trim -k \textquoteleft{}SQK-RBK004\
    textquoteright{} --guppy
# gzip the demultiplexed fastqs generated
pigz --best path_to_demultiplexed_data_folder/*.fastq
#### porechop adapter removal ####
# porechop would need to be in $PATH for this to work and this is run
    in an iterative loop
```

```
porechop -i single_fastq.gz_file -o path_to_output/isolate_ID.fastq.gz
    --verbosity 2 --threads THREADS --format fastq.gz
#### alignment using Minimap2 and Samtools ####
# align
minimap2 -ax map-ont Human_reference.fasta input.fastq -t THREADS |
    samtools view -@ THREADS -b - | samtools sort -@ THREADS -o
    isolateVsHumanRef.bam -
# extract unmapped reads
samtools view --threads THREADS -f 4 -b isolateVsHumanRef.bam >
    path_to_unmapped_folder/isolateVsHumanRef_unmapped.bam
bedtools bamtofastq -i path_to_unmapped_folder/
    isolateVsHumanRef_unmapped.bam -fq path_to_unmapped_folder/
    isolateVsHumanRef_unmapped.fastq
```

Code C.2: A series of minimum working examples of the commands used to carry out basecalling, demultiplexing, adapter removal and alignment of ONT sequence data generated from *P. knowlesi*-infected whole blood. Basecalling was done using Guppy *v*4.0.15, demultiplexing with Qcat *v.1.1.0*, adapter removal with porechop *v0.2.4*, and alignment with minimap2 *v*2.17

C.2.2 *De novo* genome assembly and decontamination

```
#!/bin/bash
# this script assumes all the software are installed in the $PATH
#### flye assembly command ####
flye --nano-raw input_reads.fasta/q --genome-size genome_size --out-dir
     output_folder --threads THREADS
#### blobtools commands ####
# blobtools involves multiple steps. first alignment with minimap2
minimap2 -ax map-ont assembly.fasta
    path_to_reads_used_to_generate_assembly.fastq -t THREADS | samtools
     view -@ THREADS -b - | samtools sort -@ THREADS -o
    path_to_output_folder/isolate_readsVsAssembly.bam -
# step 2 -- blast
blastn -task megablast -query assembly.fasta -db nt -outfmt \
    textquoteleft{}6 qseqid staxids bitscore std scomnames sscinames
    sblastnames sskingdoms stitle\textquoteright{} -evalue 1e-20 -out
    BLAST_output/isolate_vs_nt.out -num_threads THREADS
# step 3 -- blobtools
blobtools create -i assembly.fasta -b isolate_readsVsAssembly.bam -t
    isolate_vs_nt.out -o blobtools_Output/isolate_folder # create a
    JSON database
```

```
blobtools view -i blobtools_Output/isolate_folder/blobDB.json -o
    blobtools_Output/isolate_folder/ # create a blobtools table of
    terms
blobtools plot -i blobtools_Output/isolate_folder/blobDB.json -o
    blobtools_Output/isolate_folder/ # plot blobplots
blobtools covplot -i blobtools_Output/isolate_folder/blobDB.json -c
    blobtools_Output/isolate_folder/isolate_readsVsAssembly.bam.cov -o
    blobtools_Output/isolate_folder/ --max 1e03  # draw coverage plots
### Decontamination ###
# take out the lines in the blobtable that are not the header,
    Apicomplexa, no-hits and undefined
grep -E \textquoteleft{}Apicomplexa|#|no-hit|undef\textquoteright{}
    blobtools_Output/isolate_folder/blobDB.table.txt > blobtools_Output
    /isolate_folder/clean.blobDB.table.txt
# make a list of the contigs we want to keep in our fasta
cut -f 1,1 blobtools_Output/isolate_folder/clean.blobDB.table.txt >
    blobtools_Output/isolate_folder/nodes.txt
# take off the header in the nodes list
grep -v \textquoteleft{}^#\textquoteright{} blobtools_Output/
    isolate_folder/nodes.txt > blobtools_Output/isolate_folder/
    node_names.txt
# cleaning the FASTA..."
for fas in $inputFol/* ;
do
ls $fas
fname=$(basename "$fas")
echo $fname
# call the fasta tool for the cleaning of the fasta
python3 path_to_Scripts_folder/FastaTool.py assembly.fasta
    blobtools_Output/isolate_folder/node_names.txt > clean_assembly.
    fasta
```

Code C.3: Representative working examples of the commands used to carry out *de novo* genome assembly using Flye *v*2.8.1 and decontamination was done using blobtools *v*1.0.1

```
#!/usr/bin/env python3
from Bio import SeqIO
import sys
ffile = SeqIO.parse(sys.argv[1], "fasta") # read in the assembly file
header_set = set(line.strip() for line in open(sys.argv[2]))  # set the
     header_set for each line in the file containing contaminated
    contig node names
for seq_record in ffile:  # for each contig in the assembly FASTA file,
     check if the header matches the node, if it does, remove the
    contig
try:
```

```
header_set.remove(seq_record.name)
print(seq_record.format("fasta"))
except KeyError:
continue
if len(header_set) != 0:
print(len(header_set),\textquoteleft{}of the headers from list were not
     identified in the input fasta file.\textquoteright{}, file=sys.
    stderr)
```

Code C.4: A custom python script to search for the headers of the identified contaminated contigs and remove them from the input FASTA file.

C.2.3 Draft Genome Polishing and Correction

Racon Polishing

```
#!/bin/bash
#### Racon polishing ####
# Script to take through data for racon polishing. Will carry out racon
     polishing through 4 rounds in order to take this through to Medaka
    .
racon unmapped_reads_isolate.fastq Isolate_readsVsCleanAssembly.sam
    clean_assembly.fasta -t THREADS > isolate_Assembly_iteration1.fasta
       # Do Racon iteration one
minimap2 -ax map-ont isolate_iteration1.fasta unmapped_reads_isolate.
    fastq > path_to_hold_intermediate_racon/isolate_readsVsRacon1.sam -
    t THREADS     # aligning output of iteration 1 to the raw reads
# racon iteration two
racon unmapped_reads_isolate.fastq isolate_readsVsRacon1.sam
    isolate_Assembly_iteration1.fasta -t THREADS >
    isolate_Assembly_iteration2.fasta # Do racon iteration two
# cleaning up. converting and sorting sam file and zipping it
samtools view --threads THREADS -bS isolate_readsVsRacon1.sam >
    isolate_readsVsRacon1.bam
samtools sort --threads THREADS -o isolate_readsVsRacon1_sorted.bam
    isolate_readsVsRacon1.bam
mv ${iterVread}/${fname}_readsVsRacon21_sorted.bam ${iterVread}/${fname
    }_readsVsRacon21.bam
# racon iteration three
minimap2 -ax map-ont isolate_Assembly_iteration2.fasta
    unmapped_reads_isolate.fastq > path_to_hold_intermediate_racon/
```

```
    isolate_readsVsRacon2.sam -t THREADS  # aligning output of
    iteration 2 to the raw reads
racon unmapped_reads_isolate.fastq isolate_readsVsRacon2.sam
    isolate_Assembly_iteration2.fasta -t THREADS >
    isolate_Assembly_iteration3.fasta # Do racon iteration three
# cleaning up. converting and sorting sam file and zipping it
samtools view --threads $THREADS -bS isolate_readsVsRacon2.sam >
    isolate_readsVsRacon2.bam
samtools sort --threads $THREADS -o isolate_readsVsRacon2_sorted.bam
    isolate_readsVsRacon2.bam
mv isolate_readsVsRacon2_sorted.bam isolate_readsVsRacon2.bam
# racon iteration four
minimap2 -ax map-ont isolate_Assembly_iteration3.fasta
    unmapped_reads_isolate.fastq > path_to_hold_intermediate_racon/
    isolate_readsVsRacon3.sam -t THREADS  # aligning output of
    iteration 3 to the raw reads
racon unmapped_reads_isolate.fastq isolate_readsVsRacon3.sam
    isolate_Assembly_iteration3.fasta -t THREADS >
    isolate_Assembly_iteration4.fasta
# cleaning up. converting and sorting sam file and zipping it
samtools view --threads THREADS -bS isolate_readsVsRacon3.sam >
    isolate_readsVsRacon3.bam
samtools sort --threads THREADS -o isolate_readsVsRacon3_sorted.bam
    isolate_readsVsRacon3.bam
mv isolate_readsVsRacon3_sorted.bam isolate_readsVsRacon3.bam
#### Medaka consensus ####
medaka_consensus -i unmapped_reads_isolate.fastq -d
    path_toisolate_Assembly_iteration4.fasta -o output_folder/isolate -
    t THREADS
#### Pilon Correction ####
# carry out alignment of reads against racon/medaka output
bwa index medaka_output/isolate_consensus.fasta
bwa mem medaka_output/isolate_consensus.fasta path_to_illumina_reads/
    isolate/reads_1.fastq path_to_illumina_reads/isolate/reads_2.fastq
    -t THREADS | samtools view -@ THREADS -b - | samtools sort -@
    THREADS -o isolateVsPolishedAssem_0.bam -
samtools index isolateVsPolishedAssem_0.bam
# pilon
pilon -Xmx120G --genome medaka_output/isolate_consensus.fasta --bam
    isolateVsPolishedAssem_0.bam --threads THREADS --outdir
    pilon_output_folder/isolate_iteration0/ --output isolate_ID --
    tracks --fix all,circles
# interation two
# index the output of round 1
bwa index pilon_output_isolate_iteration0.fasta
# align the round 1 correction with the short reads
bwa mem pilon_output_isolate_iteration0.fasta path_to_illumina_reads/
    isolate/reads_1.fastq path_to_illumina_reads/isolate/reads_2.fastq
    -t THREADS | samtools view -@ THREADS -b - | samtools sort -@
```

```
    THREADS -o isolateVsPolishedAssem_1.bam -
samtools index isolateVsPolishedAssem_1.bam
# pilon round two
pilon -Xmx120G --genome pilon_output_isolate_iteration0.fasta --bam
    isolateVsPolishedAssem_1.bam --threads THREADS --outdir
    pilon_output_folder/isolate_iteration1/ --output isolate_ID --
    tracks --fix all,circles
# iteration three
# index the output of round 2
bwa index pilon_output_isolate_iteration1.fasta
# align the round 2 correction with the short reads
bwa mem pilon_output_isolate_iteration1.fasta path_to_illumina_reads/
    isolate/reads_1.fastq path_to_illumina_reads/isolate/reads_2.fastq
    -t THREADS | samtools view -@ THREADS -b - | samtools sort -@
    THREADS -o isolateVsPolishedAssem_2.bam -
samtools index isolateVsPolishedAssem_2.bam
# pilon round 3
pilon -Xmx120G --genome pilon_output_isolate_iteration1.fasta --bam
    isolateVsPolishedAssem_2.bam --threads THREADS --outdir
    pilon_output_folder/isolate_iteration2/ --output isolate_ID --
    tracks --fix all,circles
```

Code C.5: A shell script to carry out alignment and subsequent polishing of draft *de novo* assemblies using Minimap2 and Racon, respectively

C.2.4 Apicoplast and Mitochondrial Sequence Extraction and Deletion Scripts

Extraction

```
# Script to extract identified reads or contigs from a fasta file based
    on the read name/ID in a txt file
# The script is designed to take in a list of read names and then
    search for these in the fasta file, extract them and save in a new
    fasta file
# Note that this does not delete the sequences from the original fasta
    file
# Usage: python extractFasta.py readsListTxtFile file.fasta output.
    fasta
from Bio import SeqIO
import sys
```

```
readsFile = open(sys.argv[1], 'r')
#readsFile = open(sys.argv[1], 'r')
fastaFile = sys.argv[2]
outputFile = open(sys.argv[3], 'w')
wanted = set()
with readsFile as f:
for line in f:
line = line.strip()
if line != "":
wanted.add(line)
fasta_sequences = SeqIO.parse(open(fastaFile),'fasta')
with outputFile as i:
for seq in fasta_sequences:
if seq.id in wanted:
SeqIO.write([seq], i, "fasta")
readsFile.close()
outputFile.close()
```

Code C.6: A python script to extract indicated reads/contigs from the input FASTA file. The script requires a text file with a list of read/contig names to extract. However, the script only copies the reads/contigs into the output. It does not delete them from the input file.

Deletion

```
#!/usr/bin/env python3
# Script to delete identified reads or contigs from a fasta file based
    on the read name/ID in a txt file
# The script is designed to take in a list of read names and then
    search for these in the fasta file, delete them then save remainder
     of the fasta file in a given path
# usage : deleteFasta.py input.fasta list_ofReads.txt > filtered.fasta
from Bio import SeqIO
import sys
ffile = SeqIO.parse(sys.argv[1], "fasta")
header_set = set(line.strip() for line in open(sys.argv[2]))
for seq_record in ffile:
try:
header_set.remove(seq_record.name)
except KeyError:
print(seq_record.format("fasta"))
continue
if len(header_set) != 0:
```

```
print(len(header_set),'of the headers from list were not identified in
    the input fasta file ',header_set, file=sys.stderr)
```

Code C.7: A python script to delete indicated reads/contigs from the input FASTA file. The script requires a text file with a list of read/contig names which the script parses the file for and deletes where matches occur. The FASTA file exlcuding the reads/contigs indicated in the text file is given as an output.

de novo Prokaryotic Assembly, Circularisation and Annotation

```
# Command to carry out canu for apicoplast and mitochondrial
canu -p $PREFIX -d $OUTPUT genomeSize=$size -nanopore-raw $INPUT_reads
    -useGrid=False -maxMemory=$MEMORY -maxThreads=$THREADS
  ## the genome size for apicoplasr used was 0.035m while 0.0006m was
      used for mitochondrial assemblies
# Command to carry out circlator for apicoplast and mitochondrial
circlator all --data-type nanopore-corrected --bwa-opts "-x ont2d" --
    merge_min_id 85 --merge_breaklen 1000 $assembled_CanuOutput
    $canu_CorrectedReads.fastq.gz $outputFolder --threads $THREADS
  ## The corrected reads produced as part of the Canu assembly process
      are used by Circlator
# Command to carry out Prokka annotation of the assembled and
    circularised sequences
prokka --outdir $OUTPUT_folder --prefix $ISOLATE_ID $INPUT_assembly --
    cpus $THREADS --force --compliant --centre UoSA --addgenes --
    locustag $ISOLATE_id_PKCLINC --kingdom Mitochondria/Bacteria --
    proteins $REFERENCE_proteins.fasta
  ## Apicoplasts are plastids which are classed as cyanobacteria.
```

Code C.8: A set of shell commands used to carry out *de novo* assembly of apicoplast and mitochondrial reads extracted from the adapter-removed readset of the sequence isolates. A command to implement assembly using Canu is followed by the command utilised for circularisation with Circlator. Annotation of successful assembled and circularised genomes is subsequently implemented using Prokka.

C.3 Repeatmasking commands

```
#!/bin/bash
# Script to carry out repeatmasking of identified assemblies. Multiple
    steps are being taken with different python scripts
aPATH=$1 # set path to assemblies to mask
sPATH=$2  # set path to output directory
scpts=$3 # set path to repeatmasking scripts location
ocfta=$4 # set path to one code to find them all directory holding the
    scripts
trapsi="" # ONLY manually fill if transposonPSI in your conda
    environment fails. Enter path to transposonPSI perl script
THREADS=32  # set number of threads
LOGFF=$sPATH/Logs
mkdir -p $LOGFF
##################################################
# step 1 - rename contigs incrementally. there are downstream steps
    which may break if the contig ID are too long
echo "First step is to rename the contigs to incremental numbers"
python path_to_scripts/rename_fasta.py -i assembly_to_mask.fasta -o
    output_folder/Renamed_isolate.fasta --pre contig_
##################################################
# Step 2 - repeatmodeling
path_to_scripts/repeatModeler_step.py isolateID Renamed_isolate.fasta
echo "RepeatModeler complete, outputs are saved in ${sPATH}/
    RepeatModeler"
##################################################
# step 3 - classify censor results
# ask if they have uploaded to censor and downloaded results. it must
    be done on FIREFOX due to line wrapping issues with Chrome/Chromium
     browsers
# must name the results "Censor_results" and place in RM* folder in the
     repeatmodeler directory
# Then run censor classification step with:
while true; do
read -r -p 'You need to have carried out the Censor analysis steps
    above. Have you done that?: ' censoree
case "$censoree" in
[Yy][Ee][Ss]|[Yy]) # Yes or Y (case-insensitive).
echo "Beginning the censor classification..."
cd path_to_RepeatModeler_output/isolateID/RM*;
dos2unix Censor_results;
path_to_scripts/classify_Censor.py isolateID Renamed_isolate.fasta
cd; break ;; *) # Anything else (including a blank) is invalid.
;; esac
done
##################################################
# step 4 - CD-HIT
# concatenate all isolate censor outputs into one file to run CDHIT to
    remove redundancy
```

```
cat path_to_RepeatModeler_output/*/RM*/isolate.fa.censor >>
    path_to_RepeatModeler_output/CombinedIsolates.censor.fa;
# run CDHIT
cd-hit-est -i path_to_RepeatModeler_output/CombinedIsolates.censor.fa -
    o path_to_CDHIT/Pk_RepeatLib.censor.fa -c 1.0 -n 10 -d 0 -g 1 -M
    2000 -T THREADS
################################################
# step 5 - repeatmasker step
path_to_scripts/repeatmasker_step.py $i path_to_CDHIT/Pk_RepeatLib.
    censor.fa output_folder/RepeatMasking/isolate
################################################
# step 6 - one code to find them all
path_to_ocfta/build_dictionary.pl --rm RepeatMasking_folder/isolate.
    fasta.out --fuzzy > RepeatMasking_folder/isolate_fuzzy.txt
path_to_ocfta/one_code_to_find_them_all.pl --rm RepeatMasking_folder/
    isolate.fasta.out --ltr RepeatMasking_folder/isolate_fuzzy.txt --
    fasta
################################################
# step 7 - ltr harvest
cd LTRHarvest_outputFolder/isolateID;
path/to/genometools/bin/gt suffixerator -db Renamed_isolate.fasta -
    indexname isolateID -tis -suf -lcp -des -ssp -sds -dna
path/to/genometools/bin/gt ltrharvest -index isolateID -mintsd 5 -
    maxtsd 100 > isolateID_ltr.out
################################################
# step 8 - transposonPSI
cd TransposonPSI_outputFolder/isolateID;
path_to_transposonPSI/transposonPSI.pl Renamed_isolate.fasta nuc
################################################
#step 9 - removing redundancy
cd LociToMask_output/isolateID
path_to_scripts/generate_mask_loci.py isolateID Renamed_isolate.fasta
    RepeatMasking_output/isolateID LTRHarvest_output/isolateID/
    isolateID_ltr.out TransposonPSI_outputFolder/isolateID/isolateID.
    fasta.TPSI.allHits.chains.bestPerLocus
################################################
# step 10 - mask the assemblies based on the loci gffs
echo
echo "beginning masking of identified repeats..."
for i in $aPATH/*;
do bmk=$(basename "$i");
echo $bmk;
bmkr=${bmk%.*};
  echo $bmkr;
  outr=$sPATH/MaskedAssemblies;
  mkdir -p $outr;
  bedtools maskfasta -fi Renamed_isolate.fasta -fo
      MaskedAssembly_output/isolateID_masked.fasta -bed LociToMask/
      isolateID/isolateID.transposable_elements.gff3 -soft 2
```

Code C.9: A bash script written to control the identification and masking of repetitive elements in draft *de novo* assemblies. The script calls multiple other shell and python scripts to carry out the specific steps of the repeatmasking pipeline. The other scripts associated with repeatmasking can be found in the github repository for this project.

C.4 Quality Assessment commands

```
#!/bin/bash
#### assembly-stats ####
# an example command used to generate crude metrics
assembly-stats assembly.fasta > assembly_stats.txt
#### busco ####
# an example command used for busco assessments
singularity exec -H path_to_working_directory path_to_busco_container
    busco -i assembly.fasta -l plasmodium_odb10 --out isolateID --
    out_path output_folder -f -m geno -c THREADS --long
#### quast ####
# an example command used for quast assessments
quast -t THREADS assembly.fasta -o output_folder -r reference.fasta -g
    reference.gff3 --large -f --circos
#### pomoxis ####
# an example command used for pomoxis
assess_assembly -r reference.fasta -i assembly.fasta -H -t THREADS
#### qualimap ####
# an example command used to generate qualimap reports
minimap2 -ax asm5 reference_chromomsome.fasta isolate_chromosome.fasta
    -t THREADS | samtools view -@ THREADS -b - | samtools sort -@
    THREADS -o isolate_chromosome_vs_reference.bam -;
qualimap bamqc -bam isolate_chromosome_vs_reference.bam -hm 7 -nt 8 -nw
     800 --outdir output_folder -outformat PDF:HTML -outfile isolate.
    pdf --java-mem-size=16G;
#### agat ####
# an example command used to carry out agat statistical assessments
agat_sq_stat_basic.pl -i assembly_annotation.gff3 -g reference.fasta -o
     isolate_basicStats.tab
agat_sp_statistics.pl -i assembly_annotation.gff3 -g reference.fasta -d
     -p -o isolate_fullStats.tab
```

Code C.10: Representative commands used to carry out quality assessments of *de novo* assembled genomes as the pipeline progressed

C.5 Structural Variation commands

```
# BCF one-liner for filtering called variants.
bcftools view -i '(SVTYPE = "INS"||SVTYPE = "INV"||SVTYPE = "DUP"||
    SVTYPE = "DEL"||SVTYPE = "TRA"||SVTYPE = "BND") && ABS(SVLEN) > 49
    && ABS(SVLEN) < 100000 && INFO/RE >= 8' input_variant_file.vcf >
    filtered_variant_file.vcf

# sort the vcf output
vcf-sort filtered_variant_file.vcf > filtered_and_sorted_variant_file.
    vcf

# Annotate the filtered and sorted file
vcfanno -p THREADS -ends -lua configFile.lua lua_config_gff_sv.conf
    filtered_and_sorted_variant_file.vcf > annotated_Variants.vcf

# run stats on the annotated file
SURVIVOR stats annotated_variant_file.vcf 49 -1 -1
    annotated_Survivor_stats.txt
```

Code C.11: Commands for analysing structural variants called using both the alignments-based and assembly-based variant calling approaches. Alignment-based variant calling was achieved using the Oxford Nanopore structural variation pipeline (ONTSV), while the assembly-based variant calling was carried out on Assemblytics. Commands below were used after variant calling for manipulating, analysisng annotating the called variants.

Appendix D

Pipeline Development Appendix

Òfìifìì là ńrií, a ò rí òkodoro; ḿbọ́, baba gba-n-gba — ***All we see is shadows, not clarity; but clarity will come, the father of all candor***

Yorùbá adage

D.1 Confirmation of human contamination in contigs of PKNH-guided assembled genomes

Table D.1: Contaminated contigs and their top hits. List of the contaminated contigs identified in each isolate of the PKNH-guided assembly dataset by Blobtools as well as the top hits of each contig's alignment against the NCBI nucleotide database

Assembly Isolate ID	Contaminated contigs	Nucleotide database top hit(s)
Cultured PkA1H1	contig_168	Fragment of Human mitochondrion
sks047	contig_24	Fragment of Human clone BAC JH1
	contig_28	Human mitochondrion (full length)
	contig_29	Fragment of Human clone BAC JH4
	contig_39	Fragment of Human chromosome 16
	contig_48	Pan troglodytes BAC clone CH251-326D17
	contig_51	*Homo sapiens* chromosome 19, cosmid R29827
sks048	None	
sks050	None	
sks058	contig_124	*Homo sapiens* clone BAC JH4
	contig_174	Human mitochondrion (full length)
	contig_182	Fragment of *Homo sapiens* 3 BAC RP11-512E23
sks070	contig_25	Fragment of *Homo sapiens* clone BAC JH4
	contig_28	Fragment of *Homo sapiens* BAC clone RP11-1396O13
	contig_52	*Homo sapiens* isolate KK23 mitochondrion (2x full-length copies)
sks074	contig_102	Fragment of *Homo sapiens* chromosome 19 clone CTD-2037K13
	contig_104	Fragment of Eukaryotic synthetic construct chromosome 19
	contig_138	*Homo sapiens* isolate A1YU119 mitochondrion (full length)

Continued on next page

Table D.1 – *Continued from previous page*

Assembly Isolate ID	Contaminated contigs	Nucleotide database top hit(s)
	contig_143	Fragment of *Homo sapiens* FOSMID clone COR2A-DD0002JBCNU_D1 from chromosome 8
	contig_159	Fragment of *Homo sapiens* clone BAC JH10 and JH18
	contig_177	Fragment of *Homo sapiens* fosmid clone XXFOS-88839D6 from 4
	contig_40	Fragment of *Homo sapiens* clone BAC JH2
	contig_99	Fragment of *Homo sapiens* clone BAC JH4
sks078	contig_166	Fragment of *Homo sapiens* mitochondrion
	contig_178	Fragment of *Homo sapiens* clone BAC JH6
sks078	contig_299	Fragment of Eukaryotic synthetic construct chromosome 22
	contig_31	Fragment of *Homo sapiens* BAC clone CH17-161C24
	contig_312	Fragment of Human DNA sequence from clone RP11-426M5
	contig_39	Fragment of Human DNA sequence from clone RP11-164K15
	contig_4	Fragment of Human DNA sequence from clone RP11-297D8
	contig_40	Fragment of Eukaryotic synthetic construct chromosome 22 (65 % query cover)
	contig_43	Fragment of Eukaryotic synthetic construct chromosome 18
	contig_44	Fragment of *Homo sapiens* clone BAC JH4
sks125	contig_214	Fragment of *Homo sapiens* clone BAC JH4
	contig_233	Human mitochondrion (full length)
	contig_60	Fragment of *Homo sapiens* BAC clone RP11-1396O13
sks325	contig_4	*Homo sapiens* isolate BAL38 mitochondrion (full length)
sks331	contig_184	Fragment of *Homo sapiens* BAC clone RP11-1396O13
	contig_190	Fragment of *Homo sapiens* clone BAC JH11
	contig_197	Fragment of Pan troglodytes BAC clone CH251-704E21
	contig_227	*Homo sapiens* mitochondrion (full length)
	contig_118	Fragment of *Homo sapiens* clone CH17-275P10
	contig_119	Fragment of *Homo sapiens* clone BAC JH1

Continued on next page

Table D.1 – *Continued from previous page*

Assembly Isolate ID	Contaminated contigs	Nucleotide database top hit(s)
sks333	contig_144	Fragment of Eukaryotic synthetic construct chromosome 19
	contig_154	Fragment of *Homo sapiens* chromosome 16
	contig_202	Fragment of Human DNA sequence from clone RP11-426M5
	contig_215	Fragment of *Homo sapiens* BAC clone RP11-575B4 from 4
	contig_216	Fragment of *Homo sapiens* BAC clone RP11-351I12 from 7
	contig_238	Fragment of *Homo sapiens* BAC clone CTD-2205P10 from 4
	contig_253	Fragment of *Homo sapiens* clone BAC JH4
	contig_282	Fragment of *Homo sapiens* clone CH17-275P10 from chromosome 1
	contig_294	Fragment of *Homo sapiens* chromosome 1 clone CH17-358B19
	contig_315	Fragment of *Homo sapiens* clone BAC JH1
	contig_325	Fragment of Eukaryotic synthetic construct chromosome 22
sks333	contig_348	Fragment of *Homo sapiens* BAC clone RP11-525B5 from 2
	contig_351	Fragment of Human DNA sequence from clone RP11-140B17 on chromosome 10
	contig_354	full-length *Homo sapiens* isolate CHN094 haplogroup B4g1a mitochondrion
	contig_356	Fragment of *Homo sapiens* genomic DNA, chromosome 11q clone:RP11-676F20
	scaffold_279	Fragment of *Homo sapiens* BAC clone RP11-1247B7 from chromosome 2
sks339	contig_25	Fragment of *Homo sapiens* clone BAC JH4 and JH11
	contig_32	*Homo sapiens* mitochondrion (full length)
sks344	None	

Table D.2: Numerical statistics of the assemblies generated using the default settings of the Canu assembler on the isolate sequence data available.

Archived Sample ID	Guppy				Qcat			
	Coverage^x	Length	Contigs	N50	Coverage^x	Length	Contigs	N50
PKNH*	*Unknown*	24395979	19	2162603	*Unknown*	24395979	19	2162603
PKNOH**	*Unknown*	24771595	28	1832627	*Unknown*	24771595	28	1832627
sks047	38.65x	22190642	155	415287	32.85x	21928317	187	309965
sks048	36.47x	22654972	139	403454	40.71x	22936358	128	464190
sks058	19.55x	20590970	235	225461	21.16x	20486600	284	199147
sks070	47.29x	22929810	79	736026	52.56x	23210905	74	602720
sks125	11.89x	18943259	468	87185	12.89x	19151256	463	95523
sks339	131.83x	23424519	66	753469	141.11x	23478381	69	597401

Six out of 18 isolate sequences successfully assembled. Coverage refers to overall genome coverage at the start of the assembly based on a 24 Mb estimated genome length. Length is the total base pair assembly length and N50 is the minimum contig length that contains at least 50 % of the genome. Omitted isolates which failed de novo assembly are: *sks071*, *sks074*, *sks078*, *sks125*, *sks231*, *sks254*, *sks280*, *sks330*, *sks331*, *sks333*, *sks343*, *sks344*. All failed due to lack of coverage depth

* - The *P. knowlesi* PKNH reference genome generated using Sanger sequencing [1]

** - The *P. knowlesi* PKNOH genome generated using PacBio sequencing and HiC reads [2]

x - Total input read coverage based on the 24megabase estimate genome length. However Canu only uses longest 40x of reads

D.2 Assessing *de novo* genome assemblers

D.2.1 Assessing the Canu assembler

D.2.2 Assessing the Redbean assembler

Table D.3: Parameters for iterative assemblies using Redbean.

Iteration	Condition	Length (l)	*k*-mer size (-p)	Alignment subsampling (-AS)	*k*-mer frequency filter (-K)	Similarity (-s)	Minimum Read Depth (-e)
1	**Preset**	1000	21	4	0.05	0.5	N/A
2	**Preset**	2000	21	4	0.05	0.5	N/A
3	***k*-mer change**	2000	*default*	*default*	*default*	*default*	*default*
4	***k*-mer + Min. read change**	2000	*default*	*default*	*default*	*default*	2
5	***k*-mer + Min. read + subsampling change**	2000	*default*	2	*default*	*default*	2
6	***k*-mer change**	2000	19	*default*	*default*	*default*	*default*
7	***k*-mer + Min. read change**	2000	19	*default*	*default*	*default*	2
8	***k*-mer + Min. read + subsampling change**	2000	19	2	*default*	*default*	2
9	***k*-mer change**	2000	20	*default*	*default*	*default*	*default*
10	***k*-mer + Min. read change**	2000	20	*default*	*default*	*default*	2
11	***k*-mer + Min. read + subsampling change**	2000	20	2	*default*	*default*	2
12	***k*-mer change**	2000	21	*default*	*default*	*default*	*default*
13	***k*-mer + Min. read change**	2000	21	*default*	*default*	*default*	2
14	***k*-mer + Min. read + subsampling change**	2000	21	2	*default*	*default*	2
15	***k*-mer change**	2000	22	*default*	*default*	*default*	*default*
16	***k*-mer + Min. read change**	2000	22	*default*	*default*	*default*	2
17	***k*-mer + Min. read + subsampling change**	2000	22	2	*default*	*default*	2
18	***k*-mer change**	2000	23	*default*	*default*	*default*	*default*
19	***k*-mer + Min. read change**	2000	23	*default*	*default*	*default*	2
20	***k*-mer + Min. read + subsampling change**	2000	23	2	*default*	*default*	2
21	***k*-mer change**	2000	24	*default*	*default*	*default*	*default*
22	***k*-mer + Min. read change**	2000	24	*default*	*default*	*default*	2
23	***k*-mer + Min. read + subsampling change**	2000	24	2	*default*	*default*	2

Iterative processes carried out to determine the most applicable parameters which generates the most appropriate assembly for the sequence data available. Minimum alignment length (-l) acts as a form of filtering with a default length of 2000 bp. Alignment subsampling (-AS) aids in saving memory; default is 4. Low coverage data benefit from a reduction of the default subsampling rate value. *K*-mer frequency filter (-K) removes high frequency kmers; default is 2. Similarity (-s) is the minimum kmer or aligned matched length; default is 0.05. Minimum read depth (-e) is the minimum read depth coverage of the kmer/aligned region [3]

APPENDIX E

WHOLE GENOME SEQUENCING AND *de novo* ASSEMBLY

Òfìifĩì là ńrií, a ò rí òkodoro; ḿbọ́, baba gba-n-gba — ***All we see is shadows, not clarity; but clarity will come, the father of all candor***

Yorùbá adage

E.1 Alternative Whole Genome Sequencing Protocols

E.1.1 Whole Genome Sequencing of *Plasmodium knowlesi* DNA using SQK-RAD002 protocol

The rapid sequencing kit (SQK-RAD002) was developed to be a faster method to prepare DNA samples to be sequenced on the MinION sequencing platform. To begin, 7.5 µL of eluted DNA containing 200 ng total genomic DNA was prepared. The concentration was measured using the Qubit fluorometer as described (subsubsection:*Qubit Quantification, page 61*). The 7.5 µL of DNA was transferred to a 0.2 mL PCR tube, to which 2.5 µL

of the fragmentation mix (FRM) was added. The resulting 10 µL library was mixed by inversion, spun down and placed in a thermocycler to incubate at 30°C for one minute followed by 75°C for one minute. To the 10 µL of DNA, one µL of rapid adapter (RAD) was added before gentle flicking. The priming mix for the flowcell was generated by combining 480 µL of the running buffer (RBF) and 520 µL of nuclease water. The library was subsequently loaded as described in subsubsection: *SQK-RBK004 Library preparation, page 170*.

E.1.2 Multiplexed Whole Genome Sequencing of *Plasmodium knowlesi* DNA using SQK-RBK001 protocol

The rapid barcoding kit (SQK-RBK001) was developed to be a multiplexed variant of the SQK-RAD002 protocol, thereby allowing faster library preparation of multiple DNA samples to be sequenced on the same flowcell, labelled with unique barcodes. Once again, 7.5 µL of eluted DNA containing 200ng total genomic DNA was prepared for multiple isolates. Each 7.5 µL DNA sample was transferred to a 0.2 mL before 2.5 µL of fragmentation mix barcodes (1-12) was added and mixed. The tubes were placed in a thermocycler at 30°C for one minute followed by 75°C for one minute. Following this, the barcoded samples were pooled into one LoBind microfuge tube and the AMPure XP bead wash step was carried out (subsubsection:*AMPure XP beads wash step, page 169*). On completion, 10 µL of DNA eluted in Tris-HCl Buffer was transferred into a LoBind microfuge tube. Following this, 2 µL of the rapid adapter (RAD) was added and mixed by flicking. 0.2 µL of Blunt/TA Ligase Master Mix was then added and mixed before incubation at room temperature for 5 minutes. After this, the sequencing library was prepared by adding 35 µL of the reaction buffer (RBF), 25.5 µL of the loading beads (LLB), 2.5 µL of nuclease-free water and the 12 µL of pooled DNA. The flowcell priming mix was prepared by mixing 576 µL of reaction buffer (RBF) with 624 µL of nuclease-free water and priming was performed as described in subsubsection: *SQK-RBK004 Library preparation, page 170*.

E.1.3 Whole Genome Sequencing of *Plasmodium knowlesi* DNA using SQK-PBK004/LWB001 protocol

The low molecular weight barcoding kit (SQK-LWB001) was developed to allow for sequencing of multiplexed low input genomic samples by carrying out amplification with Polymerase Chain Reaction (PCR). However, this protocol was superceded by the PCR barcoding kit (SQK-PBK004) which carries out the same procedure. Thus, the SQK-LWB001 sequencing kit was used while following the SQK-PBK004 protocol. For each sample to be sequenced, 100ng of genomic DNA eluted in 55.5 µL was first taken through end repair and dA-tailing of fragmented DNA using the NEBNext E7442 protocol. Briefly, 3 µL of End-prep enzyme mix and 6.5 µL of End-repair reaction buffer were added to the eluted DNA, mixed by pipetting and placed in a thermocycler for 5 minutes at 20°C, followed by 5 minutes at 65°C. Following this, each sample was washed using the AMPure XP beads as described in subsubsection: *AMPure XP beads wash step, page 169*. At the final elution, the DNA was eluted in 16 µL of nuclease-free water into a LoBind microfuge tube.

PCR adapter ligation and amplification began with the addition of 10 µL barcode adapters (BCA) and 25 µL Blunt/TA Ligase mastermix to 15 µL end-prepped DNA (per sample). The mixture was incubated at room temperature for 10 minutes before carrying out another AMPure XP wash step. However, 20 µL of AMPure XP beads were added to each DNA mixture then incubated at room temperature for 5 minutes with gentle rotation. The tubes were placed on a magnetic stand, separating the beads from the eluate. The supernatant was discarded and 500 µL of freshly made ethanol (70 %) was added. The ethanol was promptly removed and the wash step repeated once more. The tubes were spun and placed on the magnetic rack, with excess ethanol removed. The tubes were removed from the magnetic rack and the pellets resuspended in 21 µL of nuclease-free water. After this, the tubes were subsequently incubated for 2 minutes at room temperature before being placed in the magnetic stand until the eluate was clear. The eluate was transferred into a LoBind microfuge tube and the concentration of the DNA within, was quantified using the Qubit fluorometer.

Following this, an adapted DNA PCR was done using the volumes and conditions in

Table E.1: Volumes and program conditions to generate the PCR mastermix for use in the SQK-LWB001/PBK004 sequencing run for 5 isolates.

a *Mastermix Volumes used*

Volumes (µL)	sks070	sks133	sks134	sks276	sks367
Adapter ligated DNA	9	15	15	3.5	2
Nuclease-free Water	15	9	9	20.5	22
Barcode primers (LWB 01-12)	1	1	1	1	1
Taq 2x Mastermix	25	25	25	25	25
Total	50	50	50	50	50

b *PCR Cycling temperatures and conditions*

Cycle Step	Temperature (°C)	Time	Number of cycles
Initial denaturation	95	3 mins	1
Denaturation	95	15 secs	14
Annealing	56	15 secs	14
Extension	65	6 mins	14
Final Extension	65	6 mins	1
Hold	4	∞	

c *Final volumes sequenced*

DNA sample	Volume in pooled DNA library (ul)
sks070	1.7
sks133	2.7
sks134	1.2
sks276a	2.7
sks276b	2.2
sks367a	1
sks367b	1.4

(a) Volumes for the adapted ligated DNA was calculated based on the concentration measured by the Qubit fluorometer. Each sample was made up to be equivalent to ~0.2 ng/µL in 50 µL, hence duplicates for sks276 and sks367 were added, resulting in a total of 7 DNA samples. PCR cycling temperatures are presented (b) and the final volumes pooled into one sample to be sequence are shown (c).

Table E.1. After this, the AMPure XP wash step was repeated. Here, 30 µL of AMPure XP beads were added to each amplified DNA sample and incubated for 5 minutes at room temperature and placed in a magnetic stand. The supernatant was removed, discared and 200 µL of 70 % ethanol was added to each tube. The added ethanol was removed and the ethanol wash repeated once more. The tubes were spun down and excess ethanol removed before the resuspending the pellets in 10 µL of 10 mmol of Tris-HCl buffer. The mixture was incubated for two minutes at room temperature and placed in a magnetic rack until the eluate was clear. One µL of each sample's eluate was removed for DNA concentration quantification using the Qubit fluorometer. From the remaining 10 µL of each sample, varying volumes were pooled into a single LoBind microfuge tube to result in ~100 ng per sample [Table E.1c].

1 µL of rapid adapter (RAP) was added to the tube containing the pooled DNA samples before incubation for 5 minutes at room temperature and then stored in ice. Following this, 34 µL of SQB, 25.5 µL of LB, 2.6 µL nuclease-free water and the DNA library (12.9 µL) were mixed. The flowcell priming mix was prepared and the flowcell was primed as described in subsubsection: *SQK-RBK004 Library preparation, page 170*, and sequencing was started.

E.2 Patient Isolate Sequencing Information

Table E.2: Isolates sequenced on the ONT MinION sequencing platform.

Seq. Exp.	Date	Isolates	Isolate Repeat*	Replicate in Run**	*P. knowlesi* Cluster	Barcode	Flowcell and Kit
1	Nov-2017	sks201	-	-	2	-	R9.4.1; SQK-RAD002
2	Dec-2017	sks343	_	-	1	1	R9.4.1; SQK-RBK001
		sks231			2	2	
		sks074	i		1	3	
		sks071	-		1	4	
		sks344	ii		2	5	
3	Nov-2018	sks078	-	_	N.A	1	R9.4.1; SQK-RBK004
		sks074	ii		1	2	
		sks254	_	a	N.A	3	
		sks254		b	N.A	4	
4	Dec-2018	sks047	i	-	1	5	R9.4.1; SQK-RBK004
		sks058	i		1	6	
		sks331	_		2	7	
		sks333			1	8	
5	Dec-2018	sks125	-	-	1	9	R9.4.1; SQK-RBK004
		sks280			2	10	
		sks330			2	11	
		sks339	i	a	1	12	
		sks339	i	b	1	1	
6	Mar-2019	sks070	i	a	2	1	R9.4.1; SQK-RBK004
		sks070	i	b	2	2	
		sks339	ii	a	1	3	
		sks339	ii	b	1	4	
7	Apr-2019	sks047	ii	a	1	5	R9.4.1; SQK-RBK004
		sks047	ii	b	1	6	
		sks048	i	a	2	7	
		sks048	i	b	2	8	
		sks058	ii	-	1	9	

Continued on next page

Table E.2 – *Continued from previous page*

Seq. Exp.	Date	Isolates	Isolate Repeat*	Replicate in Run**	*P. knowlesi* Cluster	Barcode	Flowcell and Kit
8	Jun-2019	sks070	i	_	2	6	R9.4.1; SQK-LWB001/ PBK004
		sks133	_		N.A	7	
		sks134			1	8	
		sks276		a	N.A	9	
		sks367		b	1	10	
		sks276		a	N.A	11	
		sks367		b	1	12	
9	Jun-2019	sks070	i	_	2	6	R9.4.1; SQK-LWB001/ PBK004
		sks133	_		N.A	7	
		sks134			1	8	
		sks276		a	N.A	9	
		sks367		b	1	10	
		sks276		a	N.A	11	
		sks367		b	1	12	
10	Dec-2019	sks050	_	a	2	1	R10.1; SQK-RBK004
		sks050		b	2	2	
		sks074	iii	-	1	3	
		sks344	ii	a	2	4	
		sks344	ii	b	2	5	
11	Dec-2019	sks048	ii	a	2	6	R10.1; SQK-RBK004
		sks048	ii	b	2	7	
		sks070	ii	a	2	8	
		sks070	ii	b	2	9	
		sks325	_	_	2	10	
		Cultured PkA1.H-1			-	11	
12	Jul-2020	Cultured PkA1.H-1	-	-	-	12	R9.4.1; SQK-RBK004
13	Sep-2020	Cultured PkA1.H-1	_	_	_	1	R9.4.1; SQK-RBK004
		Cultured PkA1.H-1				2	
		Cultured PkA1.H-1				3	
		Cultured PkA1.H-1				4	
		Cultured PkA1.H-1				5	
		Cultured PkA1.H-1				6	

Continued on next page

Table E.2 – *Continued from previous page*

Seq. Exp.	Date	Isolates	Isolate Repeat*	Replicate in Run**	*P. knowlesi* Cluster	Barcode	Flowcell and Kit
14	Sep-2020	Cultured PkA1.H-1	–	–	–	7	R10.3; SQK-RBK004
		Cultured PkA1.H-1				8	
		Cultured PkA1.H-1				9	
		Cultured PkA1.H-1				10	
		sks070	iii		2	11	
15	Sep-2020	Cultured PkA1.H-1	–	–	–	1	R9.4.1; SQK-RBK004
		Cultured PkA1.H-1				2	
		Cultured PkA1.H-1				3	
		Cultured PkA1.H-1				4	
		Cultured PkA1.H-1				5	
16	Sep-2020	Cultured PkA1.H-1	–	–	–	1	R9.4.1; SQK-RBK004
		Cultured PkA1.H-1				2	
		Cultured PkA1.H-1				3	

Information for all isolates sequenced using the Oxford Nanopore Technologies (ONT) MinION platform including the date sequenced, sequencing kits, flowcells and barcodes used. The *P. knowlesi* normocyte-binding (Xa) dimorphic cluster which the isolates belong to are also provided where available.

* - Patient isolate with more than one isolate within our biobank processed through the CD45 DynaBeads leucocyte depletion method where *i, ii or iii* indicate the repeat number of that patient isolate within our biobank i.e. sks070 had 3 isolates in our biobank and the first isolate sks070(i) was sequenced in March 2019.

** - Patient isolates which have replicates within the same sequencing experiment and given different unique barcodes. These are signified in the run as 'Isolate_ID'(a,b). After sequencing, the barcodes are removed and the sequence reads combined together.

E.3 Statistical Tests

E.3.1 Normality Tests

Shapiro-Wilk test on raw isolate sequence data

Table E.3: Normality test on the adapter removed sequenced reads from all isolates.

Dataset	DF	Statistic	*p-value*	Conclusion
Avg. Starting Conc.	20	0.76887	3.07E-04	Reject normality
Yield	25	0.84937	0.00171	Reject normality
Read count	25	0.86051	0.00281	Reject normality
Avg. read length	25	0.92335	0.0611	Normally distributed

Isolate sequence reads after demultiplexing by Qcat and adapter removal by Porechop were checked for normality. The starting concentration, sequence length yield, read count and average read length for each isolate was calculated using assembly-stats.

E.3.2 Correlation Tests

Spearman's Rank correlation test on raw isolate sequence data

Table E.4: Correlation test on the adapter removed isolate sequence data's starting concentration, yield, read count and average read length.

Dataset		Avg. Starting Conc.	Yield	Read count	Average read
Avg. Starting Conc.	*Spearman Corr.*	–	-0.13083	-0.34887	0.70226
	p-value	–	0.58247	0.13166	5.56E-04
Yield	*Spearman Corr.*	-0.13083	–	0.93538	-0.3
	p-value	0.58247	–	7.24E-12	0.14511
Read count	*Spearman Corr.*	-0.34887	0.93538	–	-0.55385
	p-value	0.13166	7.24E-12	–	0.00407
Average read	*Spearman Corr.*	0.70226	-0.3	-0.55385	–
	p-value	5.56E-04	0.14511	0.00407	–

Rejection of normal distribution after Shapiro-Wilk test resulted in non-parametric Spearman's test for correlation. Each dataset was test against others for correlation.

Spearman's Rank correlation test on alignment percentage mapping to human reference genome

Table E.5: Correlation test on the relationship of the mapping percentage to the input read length and starting concentration.

Dataset		Total Length	% mapped	Avg. Starting Conc.
Total Length	*Spearman Corr.*	–	0.517	-0.131
	p-value	–	0.008	0.582
% mapped	*Spearman Corr.*	0.517	–	-0.611
	p-value	0.008	–	0.004
Avg. Starting Conc.	*Spearman Corr.*	-0.131	-0.611	–
	p-value	0.582	0.004	–

The relationship between mapping percentage of isolate reads to the human GRCh38p.13 reference genome and the corresponding input base pair length and starting concentration was assessed using a two-tailed Spearman's ranked sum test with a 95 % confidence level.

Spearman's Rank Sum correlation test of the assembly improvement steps carried out after de novo genome assembly

Table E.6: Correlation test showing the statistical relationship between the input coverage and the BUSCO scores of assembly improvement steps.

		Inp. Cov.	Assembly	Racon	Medaka	Pilon	Annotated
Inp. Cov.	*Spm. Corr.*	–	0.84709	0.95165	0.96484	0.40769	0.58746
	p-value	–	0.00013	5.342E-07	1.316E-07	0.18571	0.01485
Assembly	*Spm. Corr.*	0.84709	–	0.95270	0.89329	0.41138	0.70925
	p-value	0.00013	–	4.238E-07	2.138E-05	0.17262	0.00289
Racon	*Spm. Corr.*	0.95165	0.95270	–	0.97802	0.40769	0.66227
	p-value	5.342E-07	4.238E-07	–	1.937E-08	0.18571	0.00588
Medaka	*Spm. Corr.*	0.96484	0.89329	0.97802	–	0.37062	0.59846
	p-value	1.316E-07	2.138E-05	1.937E-08	–	0.20833	0.01316
Pilon	*Spm. Corr.*	0.40769	0.41138	0.40769	0.37062	–	0.97234
	p-value	0.18571	0.17262	0.18571	0.20833	–	0.00119
Annotated	*Spm. Corr.*	0.58746	0.70925	0.66227	0.59846	0.97234	–

Continued on next page

Table E.6 – *Continued from previous page*

	Inp. Cov.	Assembly	Racon	Medaka	Pilon	Annotated
p-value	0.01485	0.00289	0.00588	0.01316	0.00119	–

A non-parametric test of the correlation between all variables was carried out. Input coverage was used to determine if there was a relationship between how much coverage an isolate possesses and the BUSCO score given at each improvement step
The relationship of each step against its preceding step was also assessed.
Inp. Cov. - Input Coverage
Spm. Corr. - Spearman Correlation

Correlation matrix of annotated draft genomes

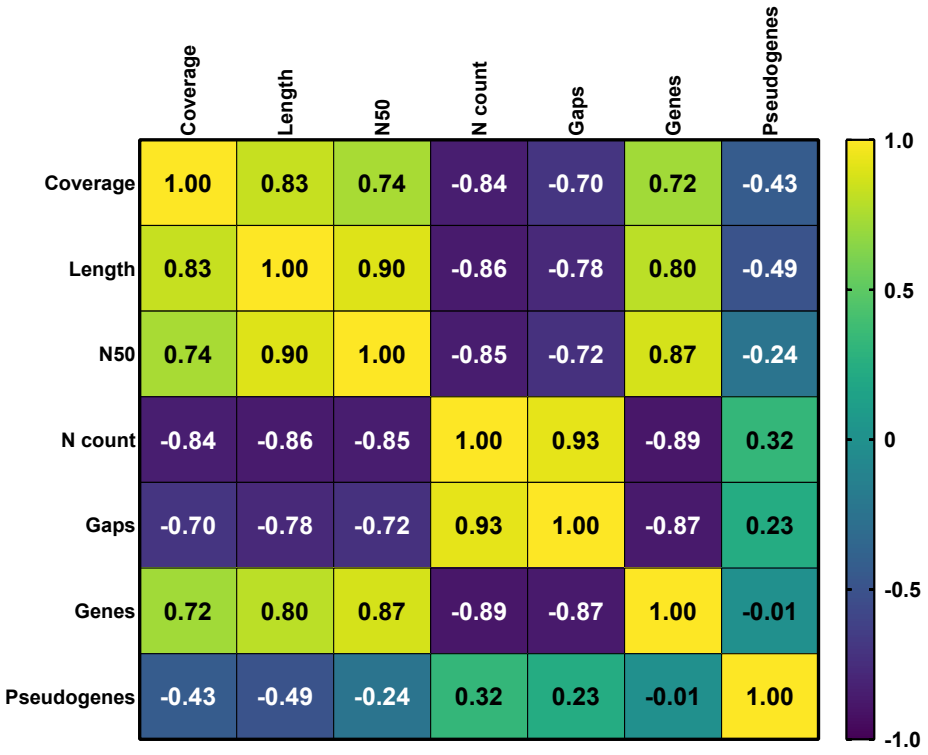


Figure E.1: Spearman's R non-parametric correlation matrix for annotated genomes. Correlation test was carried out with a one-tailed Spearman's R correlation test on all variables described in Table 4.7. Coverage refers to the input read coverage used for *de novo* assembly.

E.4 Pre-processing and Data preparation

E.4.1 Sequencing Outcomes

Table E.7: Descriptive statistics of sequencing yields from each sequencing experiment done.

Exp.	Total Length (bp)	Number of reads	Average read length (bp)	Longest read (bp)
Nov_2017	28591851	6207	4606.39	187510
Dec_2017	92910752	20678	4493.22	49107
Nov_2018	4283219014	1575593	2718.48	72132
Dec_2018_a	7821054005	2694399	2902.71	62007
Dec_2018_b	6197948313	2145740	2888.49	99548
Mar_2019	4289520505	781402	5489.52	195866
Apr_2019	2846385975	604052	4712.15	85524
Jun_2019_a	6776747737	3738488	1812.7	19657
June_2019_b	9370992948	4745648	1974.65	24717
Dec_2019_a	2077135110	387516	5360.13	77877
Dec_2019_b	2714018647	540859	5017.98	83417
Jul_2020	343360522	133128	2579.18	63016
Sep_2020_a	153456422	30292	5065.91	67349
Sep_2020_b	1710060173	366483	4666.14	103617
Sep_2020_c	3543070015	699136	5067.78	104860
Sep_2020_d	1682860450	457592	3677.64	73515

After demultiplexing, FASTQ files were produced by Qcat which were in turn assessed using assembly-stats for descriptive statistics.

E.4.2 Alignment of isolate sequence reads against the Human reference genome

Percentage of adapter removed reads which mapped to the human reference genome

Table E.8: Percentage mapped to the human reference genome.

Isolate	% mapped to human reference
StAPkA1H1	20.23
sks047+	64.73
sks048+	17.69
sks050+	19.07
sks058+	52.23
sks070	56.77
sks071	14.09
sks074	78.92
sks078	84.93
sks125	63.12
sks133	90.9
sks134	96.81
sks201	16.13
sks231	6.61
sks254	97.75
sks276	98.9
sks280	96.94
sks325	17.87
sks330	74.24
sks331	56.45
sks333	80.31
sks339	35.72
sks343	17.69
sks344	27.35
sks367	98.33

Mean	55.34

Adapter removed reads were aligned against the human GRCh38p.13 reference genome and mapping was assessed using samtools flagstat. Mapped reads and unmapped reads were extracted from the aligned BAM files.

Assembly-stats metrics for reads which did not map to the human reference genome

Table E.9: Descriptive statistical metrics for reads which did not map to the human reference genome.

Isolates	Length (bp)	Read Count	Average read length (bp)	Max. read length (bp)
StAPkA1H1	5,507,420,672	1,200,976	4585.79	104604
sks047+	1,732,422,152	527,238	3285.84	65406
sks048+	1,586,507,137	317,816	4991.9	85289
sks050	422,946,909	80,169	5275.69	67040
sks058+	980,437,211	389,004	2520.38	63003
sks070+	2,625,210,379	822,454	3191.92	104244
sks071	11,674,265	3,128	3732.18	36273
sks074	262,245,722	63,831	4108.44	51998
sks078	182,134,102	77,113	2361.91	31382
sks125	363,569,832	171,826	2115.92	37916
sks133	258,073,652	177,184	1456.53	9853
sks134	106,340,974	74,032	1436.42	9614
sks201	26,229,350	5,644	4647.3	187510
sks231	14,530,938	3,445	4217.98	47931
sks254	32,005,550	22,992	1392.03	23155
sks276	43,680,738	36,486	1197.19	13353

Continued on next page

Table E.9 – *Continued from previous page*

Isolates	Length (bp)	Read Count	Average read length (bp)	Max. read length (bp)
sks280	17,171,792	13,960	1230.07	24671
sks325	111,305,028	26,264	4237.93	61699
sks330	59,037,093	26,266	2247.66	64355
sks331	297,992,896	132,658	2246.32	36773
sks333	435,072,936	148,150	2936.71	54604
sks339+	3,751,790,716	989,551	3791.41	99299
sks343	13,928,798	3,216	4331.09	40391
sks344	367,238,148	69,046	5318.75	66461
sks367	63,753,693	36,743	1735.12	14087

Unmapped reads were extracted from minimap2 produced alignment files using samtools and bedtools. The resulting FASTQ files were assessed using assembly-stats.

E.5 *De novo* genome assembly and quality assessments

E.5.1 Descriptive Statistics

Table E.10: Assembly-stats descriptive metrics of the raw outputs of Flye *de novo* assembler.

Isolates	Inp. Cov.	Ilma. Cov.	Assem. Len. (bp)	# Contigs	Avg. Contig Len. (bp)	Max. Contig Len. (bp)	N50
StAPkA1H	226.09	-	24,153,685	73	330,872	2,015,618	688,328
sks047	71.12	30	23,573,757	100	235,738	1,753,326	555,528
sks048	65.13	105	24,495,376	74	331,019	2,230,719	816,993
sks050	17.36	96	23,863,945	114	209,333	1,349,527	421,291
sks058	40.25	99	22,912,954	192	119,338	731,780	302,992
sks070	107.77	-	24,090,440	97	248,355	1,812,524	738,227
sks071	0.48	-	–	–	–	–	–
sks074	10.77	7	22,086,753	142	155,541	915,191	286,577
sks078	7.48	-	17,771,775	330	53,854	290,080	80,529
sks125	14.93	-	20,551,487	338	60,803	379,761	106,443
sks133	10.59	-	1,938,250	116	16,709	44,677	17,295
sks134	4.37	-	172,992	12	14,416	22,017	17,379
sks201	1.08	-	340,145	11	30,922	56,968	41,777
sks231	0.60	-	12,020	1	12,020	12,020	12,020
sks254	1.31	-	25,309	1	25,309	25,309	25,309
sks276	1.79	-	29,354	3	9,785	11,914	8,892
sks280	0.70	-	–	–	–	–	–
sks325	4.57	-	18,831,797	247	76,242	349,015	99,061
sks330	2.42	-	4,484,837	125	35,879	131,596	39,125
sks331	12.23	6	19,831,850	355	55,864	611,237	99,655
sks333	17.86	-	22,074,117	213	103,634	693,342	258,560
sks339	154.02	5	24,026,467	65	369,638	3,027,069	995,217
sks343	0.57	-	5,952	1	5,952	5,952	5,952
sks344	15.08	-	23,428,618	83	282,273	1,648,428	740,638
sks367	2.62	-	715,225	37	19,330	30,498	21,294

Reads which did not map to the human reference genome *coverage shown* were used as input for the Flye *de novo* assembler. Input coverage was calculated using the length of the PKNH reference genome. Isolates which failed to assembler entirely are highlighted in red while isolates which successfully assembled to a satisfactory length are shown in green. Isolates which assembled that did not reach the sequence length threshold are shown in white.
Inp. Cov. - Input Coverage
Ilma. Cov. - Illumina Coverage
Assem. Len. - Assembly Length
Avg. Contig Len. - Average Contig Length
Max. Contig. Len. - Maximum Contig Length

Table E.11: **Metrics and statistics for *de novo* assemblies from initial production to complete annotation.**

Isolate	Assembly Length (Mb)						Contigs/Chromosomes					
	*Raw**	*Racon***	*Medaka*	*Pilon****	*Masked*	*Annotated*	*Raw**	*Racon***	*Medaka*	*Pilon****	*Masked*	*Annotated*
PKNH Reference	–	–	–	–	–	–	–	–	–	–	–	–
PKA1H1 Reference	–	–	–	–	–	–	–	–	–	–	–	–
StAPkA1H1	24.2	24.3	24.1	–	24.1	24.4	73	73	111	–	105	15
sks047+	23.6	23.7	23.6	23.6	23.6	24.2	100	100	116	116	113	15
sks048+	24.5	24.7	24.6	24.6	24.5	24.8	74	74	94	94	92	15
sks050+	23.9	23.9	23.9	23.9	23.8	24.5	114	114	133	133	129	15
sks058+	22.9	23	22.9	22.9	22.9	24.1	192	191	204	204	203	15
sks070	24.1	24.3	24.1	–	24.1	24.5	97	96	126	–	124	15
sks074+	22.1	22.2	22.2	22.1	22.1	23.4	142	141	146	146	145	15
sks078	17.8	17.8	17.8	–	17.8	21.6	330	328	330	–	329	15
sks125	20.6	20.6	20.6	–	20.5	23	338	335	343	–	342	15
sks325	18.8	18.7	18.8	–	18.7	22.6	247	247	256	–	254	15
sks331+	19.8	19.8	19.8	19.8	19.8	22.9	355	353	367	367	366	15
sks333	22.1	22.1	22.1	–	22.1	23.6	213	205	219	–	217	15
sks339+	24	24.2	24.1	24.1	24	24.2	65	65	78	78	75	15
sks344	23.4	23.5	23.5	–	23.4	23.9	83	82	97	–	91	15

Isolate	Average Contig/Chromosomes Length (Megabases (Mb))						N50 (Mb)					
	*Raw**	*Racon***	*Medaka*	*Pilon****	*Masked*	*Annotated*	*Raw**	*Racon***	*Medaka*	*Pilon****	*Masked*	*Annotated*
PKNH Reference	–	–	–	–	–	–	–	–	–	–	–	2.16
PKA1H1 Reference	–	–	–	–	–	–	–	–	–	–	–	2.19
StAPkA1H1	0.33	0.33	0.22	–	0.23	1.63	0.69	0.69	0.69	–	0.69	2.13
sks047+	0.24	0.24	0.20	0.20	0.21	1.61	0.56	0.56	0.56	0.56	0.56	2.09
sks048+	0.33	0.33	0.26	0.26	0.27	1.65	0.82	0.82	0.78	0.78	0.78	2.21
sks050+	0.21	0.21	0.18	0.18	0.18	1.63	0.42	0.42	0.42	0.42	0.42	2.15
sks058+	0.12	0.12	0.11	0.11	0.11	1.61	0.30	0.31	0.30	0.30	0.30	2.11
sks070	0.25	0.25	0.19	–	0.19	1.64	0.74	0.75	0.74	–	0.74	2.18
sks074+	0.16	0.16	0.15	0.15	0.15	1.56	0.29	0.29	0.29	0.29	0.29	2.10
sks078	0.05	0.05	0.05	–	0.05	1.44	0.08	0.08	0.08	–	0.08	1.98
sks125	0.06	0.06	0.06	–	0.06	1.54	0.11	0.11	0.11	–	0.11	2.11
sks325	0.08	0.08	0.07	–	0.07	1.51	0.10	0.10	0.10	–	0.10	1.89
sks331+	0.06	0.06	0.05	0.05	0.05	1.53	0.10	0.10	0.10	0.10	0.10	2.03
sks333	0.10	0.11	0.10	–	0.10	1.57	0.26	0.26	0.26	–	0.26	2.12
sks339+	0.37	0.37	0.31	0.31	0.32	1.62	1.00	1.01	1.00	1.00	1.00	2.17
sks344	0.28	0.29	0.24	–	0.26	1.59	0.74	0.74	0.74	–	0.74	2.13

Isolates which successfully produced a *de novo* genome assembly from the Flye process [1] were corrected and polished using iterative rounds of Racon [2] and Medaka [3]. Further polishing was implemented for isolates which possessed additional matched Illumina short read sequence data using Pilon [4]. Repetitive and transposable elements were identified and masked in the assemblies before chimeric contigs were broken (*not shown*). Draft assemblies were then annotated using the Companion web server. Metrics were calculated using assembly-stats [5].

* - Assemblies that have been cleaned and decontaminated by blobtools ; ** - Assemblies from the final iteration of Racon

*** - Assemblies from the final iteration of Pilon ; + - Isolates which possessed matched Illumina sequence reads for Pilon correction

E.5.2 BUSCO scores of draft assemblies

Table E.12: BUSCO completeness scores reported for each assembly as quality improvement steps are implemented.

Isolate	Complete (%)				
	Assembly	*Racon*	*Medaka*	*Pilon*	*Annotated*
PKNH Reference	-	-	-	-	97.60
PKA1H1 Reference	-	-	-	-	94.40
StAPkA1H1	68.60	76.30	89.70	-	89.50
sks047	67.20	72.10	85.50	95.90	95.90
sks048	68.80	74.90	85.90	95.70	95.60
sks050	67.20	64.80	69.20	95.90	95.90
sks058	67.70	71.20	82.80	95.90	95.80
sks070	66.20	71.10	83.50	-	83.60
sks074	61.70	60.10	66.20	91.80	91.80
sks078	50.50	50.30	54.00	-	54.10
sks125	62.50	62.80	69.80	-	70.10
sks325	40.40	40.90	40.30	-	40.50
sks331	57.10	57.00	64.00	83.10	83.30
sks333	66.90	66.00	73.80	-	73.80
sks339	68.10	72.60	86.90	93.90	93.90
sks344	66.70	64.50	68.30	-	68.20

The completeness scores reported by BUSCO, determining the similarity and presence of known orthologous sequences for the Plasmodium genus within the draft assemblies. BUSCO was repeated for each step of the quality improvement pipeline, culminating in the final draft annotated genome for each isolate. Only the final genomes for the PKNH and PKA1H1 reference genomes were available.

E.5.3 RepeatMasking

Table E.13: Breakdown of the identified repetitive elements within the draft assemblies which were masked

Isolate	SINEs	LINEs	LTR elements	DNA elements	Unclassified	Total interspersed	Satellites	Simple repeats	Low complexity	% masked
StAPkA1H1	0.04	3.34	2.01	6.51	1.11	13	0.18	5.19	0.85	17.2
sks047	0.03	2.93	1.8	6.31	1.2	12.27	0.17	4.55	0.86	15.68
sks048	0.03	3.5	2.11	8.1	1.23	14.96	0.19	4.89	0.86	18.29
sks050	0.02	3.25	1.95	6.16	1.2	12.58	0.2	4.6	0.88	16.11
sks058	0.03	2.23	1.88	5.63	1.11	10.88	0.14	4.51	0.85	14.38
sks070	0.02	3.31	1.88	6.93	1.27	13.39	0.17	4.67	0.85	16.79
sks074	0.02	2.51	1.32	4.47	0.88	9.21	0.11	4.35	0.83	13.02
sks078	0.02	1.22	0.71	2.43	0.48	4.87	0.06	4.02	0.8	9.11
sks125	0.03	1.71	1.53	4.25	0.85	8.37	0.06	4.14	0.83	11.8
sks325	0.02	1.44	1.1	3.11	0.64	6.3	0.04	3.98	0.84	10.2
sks331	0.02	2.07	1.32	4.5	0.83	8.73	0.08	4.37	0.83	12.57
sks333	0.03	2.36	1.73	5.24	0.99	10.35	0.14	4.34	0.8	13.81
sks339	0.03	3.15	1.77	7.28	1.22	13.44	0.13	4.64	0.83	16.73
sks344	0.02	2.85	1.65	5.84	1.11	11.47	0.18	4.49	0.86	15.04

Metrics were calculated by RepeatMasker for Class I retroelements prior to concatenating RepeatMasker-identified repeat elements with independent tests such as TransposonPSI and LTRHarvest which searched for transmembrane and other classes of LTRs, respectively. Total percentage masked excludes the stretches of X/N $\geq$20, thus accounting for the 0.96 - 2.61 % difference between the '% masked' and the calculated total percentages of the repeat classes.
SINEs - Short interspersed elements
LINEs - Long interspersed elements
LTR elements - Long terminal repeat elements

E.6 Apicoplast and Mitochondrial Assembly and Annotation

E.6.1 Flye assembler derived prokaryotic sequences

Table E.14: Outcome of circularisation of apicoplast sequences generated from the Flye assemblies for each draft isolate assembly

Isolate	Contig Length (bp)	# Contigs	Avg. Contig Length (bp)	Circularised?
PKNH (Ref.)	30638	1	30638	N.A
StAPkA1H1	35485	2	17742.5	No
sks047	–	–	–	–
sks048	34475	1	34475	No
sks050	55521	3	18507	No
sks058	–	–	–	–
sks070	26569	1	26569	No
sks074	–	–	–	–
sks078	–	–	–	–
sks125	–	–	–	–
sks325	43561	1	43561	No
sks331	–	–	–	–
sks333	–	–	–	–
sks339	34297	1	34297	No
sks344	31454	2	15727	No

Apicoplast were extracted from *de novo* assemblies generated by Flye. Circularisation was carried out on Circlator using commands described in subsection 4.3.4. Apicoplast sequences largely failed to resolve contigs and all resolved contigs were unable to be circularised.

Table E.15: Outcome of circularisation of mitochondrial sequences generated from the Flye assemblies for each draft isolate assembly

Isolate	Contig Length (bp)	# Contigs	Avg. Contig Length (bp)	Circularised?	Circularised Length (bp)
PKNH (Ref.)	5957	1	5957	–	–
StAPkA1H1	6075	4	1518.75	No	–
sks047	7132	3	2377.33	No	–
sks048	6136	1	6136	No	–
sks050	11744	1	11744	Yes	6035
sks058	5927	1	5927	No	–
sks070	10500	1	10500	No	–
sks074	5959	1	5959	No	–
sks078	5954	1	5954	No	–
sks125	5937	1	5937	No	6011
sks325	11467	1	11467	Yes	6073
sks331	10504	1	10504	Yes	–
sks333	5940	2	2970	No	–
sks339	13950	2	6975	Yes	6259
sks344	9344	4	2336	No	–

Mitochondrial sequences were extracted from *de novo* assemblies generated by Flye. Circularisation was carried out on Circlator using commands described in subsection 4.3.4. Mitochondrial sequences were present in all isolates although, few were successfully circularised.

E.6.2 *de novo* assembly of prokaryotic sequences by Canu

Table E.16: Outcome of Canu apicoplast genome assembly and circularisation for each patient isolate

Isolates	Cov.	Contig Len. (bp)	# Contigs	Avg. Contig Len (bp)	Circularised?
PKNH (Ref.)	N.A	30638	1	30638	N.A
StAPkA1H1	94.87	29531	1		No

Continued on next page

Table E.16 – *Continued from previous page*

Isolates	Cov.	Contig Len. (bp)	# Contigs	Avg. Contig Len (bp)	Circularised?
sks047	7.16	–	–	–	–
sks048	23.5	40056	2	20028	No
sks050	7.49	–	–	–	–
sks058	6.22	–	–	–	–
sks070	52.5	–	–	–	–
sks074	1.9	–	–	–	–
sks078	0.25	–	–	–	–
sks125	1.52	–	–	–	–
sks325	2.64	–	–	–	–
sks331	1.18	–	–	–	–
sks333	1.68	–	–	–	–
sks339	70.67	40503	1	40503	No
sks344	8.33	–	–	–	–

Reads which align to the reference PKNH apicoplast were extracted and inputted into Canu for a *de novo* apicoplast genome assembly. Apicoplast sequences (top) largely failed to resolve contigs and all resolved contigs were unable to be circularised by Circlator.

Table E.17: Outcome of Canu mitochondrial genome assembly and circularisation for each patient isolate

Isolates	Cov.	Contig Len. (bp)	# Contigs	Avg. Contig Len (bp)	Circularised?
PKNH (Ref.)	N.A	5957	1	5957	N.A
StAPkA1H1	708.35	34048	1	34048	Yes
sks047	290.37	28690	2	14345	Yes
sks048	293.18	17941	1	17941	Yes
sks050	134.4	35600	2	17800	Yes
sks058	206.25	15201	1	15201	Yes
sks070	503.9	–	–	–	–

Continued on next page

Table E.17 – *Continued from previous page*

Isolates	Cov.	Contig Len. (bp)	# Contigs	Avg. Contig Len (bp)	Circularised?
sks074	68.85	17538	1	17538	Yes
sks078	21.2	15709	2	7854.5	No
sks125	54.91	17692	2	8846	No
sks325	61.02	27542	2	13771	Yes
sks331	73.26	10447	1	10447	Yes
sks333	140.18	8405	1	8405	Yes
sks339	577.79	–	–	–	–
sks344	89.63	33857	4	8464.25	No

Reads which align to the reference PKNH mitochondrial sequences were extracted and inputted into Canu for a *de novo* mitochondrial genome assembly. Mitochondrial sequences (bottom) were present in all isolates with most isolates being successfully circularised with Circulator.

E.7 Comparative Genomics

Table E.18: Coverage of each chromosome within the patient and experimental line genome against the PKNH reference genome.

Isolate	Chr 0	Chr 1	Chr 2	Chr 3	Chr 4	Chr 5	Chr 6	Chr 7	Chr 8	Chr 9	Chr 10	Chr 11	Chr 12	Chr 13	Chr 14	Mean
StAPkA1H1	0.33	0.96	1.01	0.96	0.98	0.98	1.00	0.98	0.97	0.97	0.98	0.99	0.98	0.99	0.98	0.98
sks047	0.0001	0.87	0.84	0.80	0.85	0.79	0.84	0.89	0.81	0.88	0.87	0.83	0.92	0.88	0.89	0.85
sks048	0.0001	0.88	0.84	0.80	0.85	0.80	0.85	0.89	0.80	0.88	0.85	0.83	0.92	0.88	0.89	0.85
sks050	0.0002	0.88	0.84	0.80	0.85	0.79	0.85	0.90	0.81	0.88	0.86	0.83	0.92	0.88	0.88	0.85
sks058	0.0001	0.87	0.83	0.79	0.85	0.79	0.82	0.88	0.81	0.87	0.85	0.83	0.92	0.85	0.86	0.85
sks070	0.0001	0.88	0.83	0.80	0.84	0.80	0.84	0.91	0.81	0.88	0.87	0.83	0.92	0.88	0.88	0.86
sks074	0.0001	0.87	0.80	0.75	0.83	0.79	0.81	0.85	0.78	0.85	0.83	0.81	0.91	0.84	0.84	0.82
sks078	0.00	0.78	0.61	0.67	0.75	0.68	0.74	0.75	0.66	0.74	0.66	0.66	0.71	0.71	0.61	0.70
sks125	0.00	0.81	0.72	0.75	0.79	0.69	0.77	0.82	0.75	0.82	0.79	0.76	0.84	0.79	0.76	0.78
sks325	0.00	0.59	0.54	0.46	0.56	0.48	0.45	0.56	0.47	0.55	0.58	0.54	0.62	0.54	0.56	0.54
sks331	0.00	0.81	0.73	0.72	0.80	0.75	0.76	0.82	0.71	0.80	0.71	0.76	0.79	0.73	0.68	0.76
sks333	0.00	0.83	0.77	0.75	0.83	0.77	0.80	0.85	0.78	0.85	0.81	0.81	0.90	0.81	0.83	0.81
sks339	0.00	0.87	0.84	0.81	0.85	0.80	0.87	0.89	0.81	0.88	0.85	0.83	0.92	0.88	0.89	0.86
sks344	0.00	0.87	0.82	0.78	0.84	0.78	0.83	0.88	0.80	0.88	0.85	0.82	0.91	0.86	0.87	0.84

Alignments of each isolate draft genome against the PKNH reference genome was achieved using minimap2 and the subsequent BAM file was assessed with Qualimap. Mean coverage of each chromosome was provided between 0 and 1.

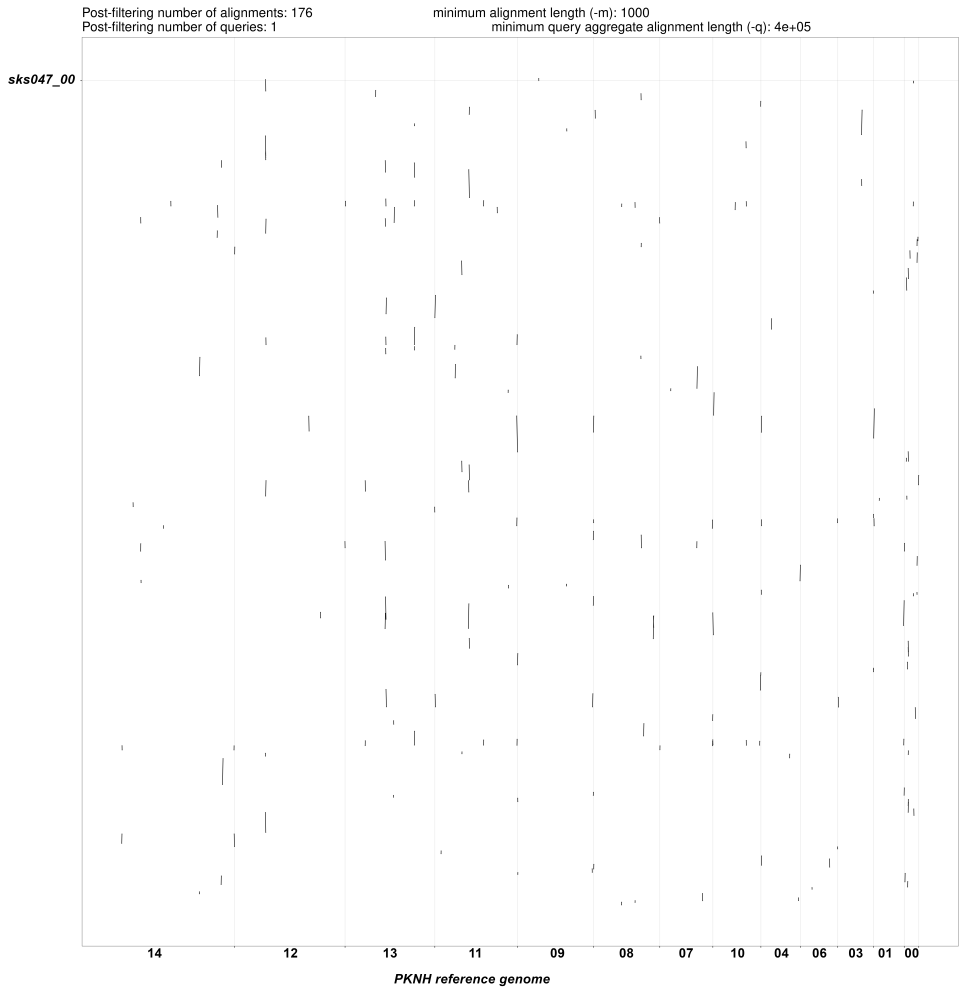


Figure E.2: Alignments of chromosome 00 (bin) for sks047 against the whole PKNH reference genome with a 1 Kb **alignment length filter.** The 'bin' chromosome contain sequence fragments that could not be confidently resolved into a particular chromosome during the scaffolding process. Unplaced contigs in the 'bin' do not show clustering to any PKNH reference chromosome. A similar observation is shown for sks048 (Appendix Figure E.3). In contrast, StAPkA1H1 shows a concentration of sequences aligned to the PKNH 'bin' chromosome 00 (Figure 4.13).

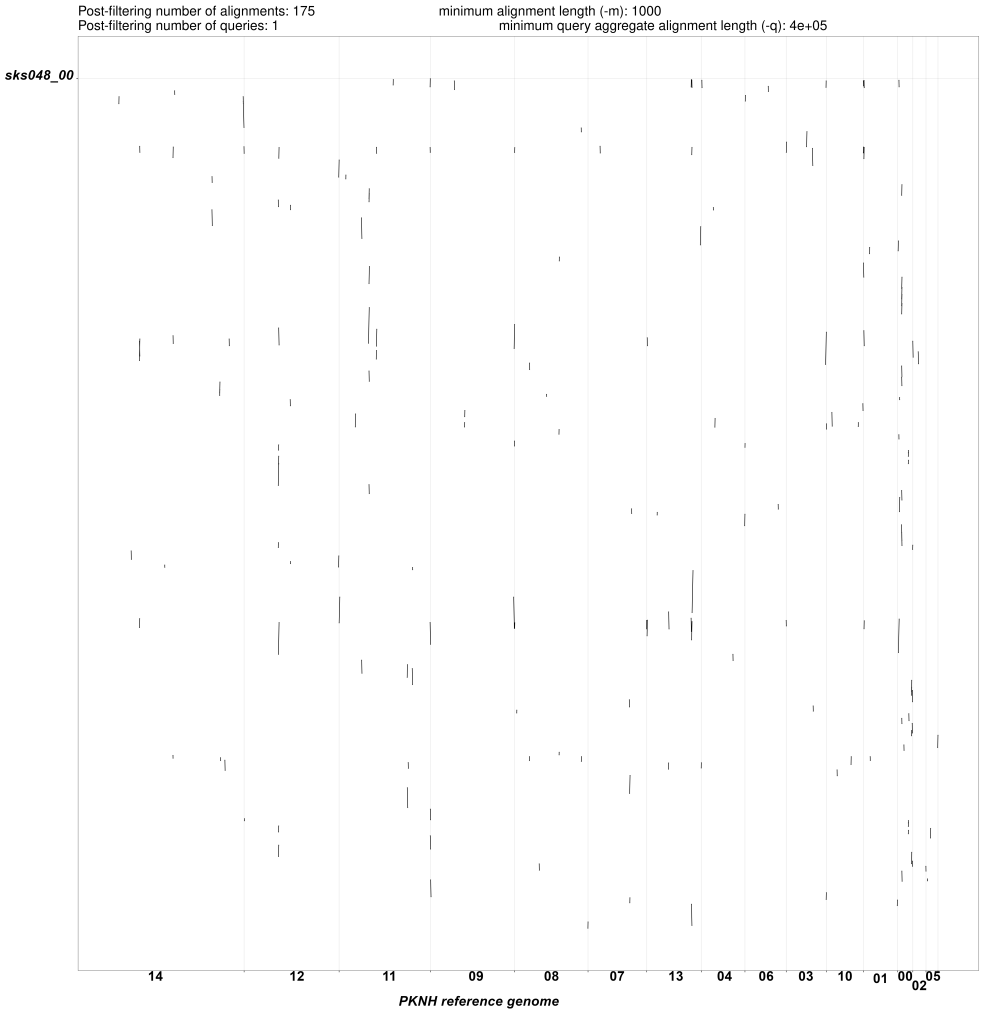


Figure E.3: Alignments of chromosome 00 (bin) for sks048 against the whole PKNH reference genome with a 1 Kb **alignment length filter.** The 'bin' chromosome contain sequence fragments that could not be confidently resolved into a particular chromosome during the scaffolding process. Unplaced contigs in the 'bin' do not show clustering to any PKNH reference chromosome. A similar observation is shown for sks047 (Appendix Figure E.2). In contrast, StAPkA1H1 shows a concentration of sequences aligned to the PKNH 'bin' chromosome 00 (Figure 4.13).

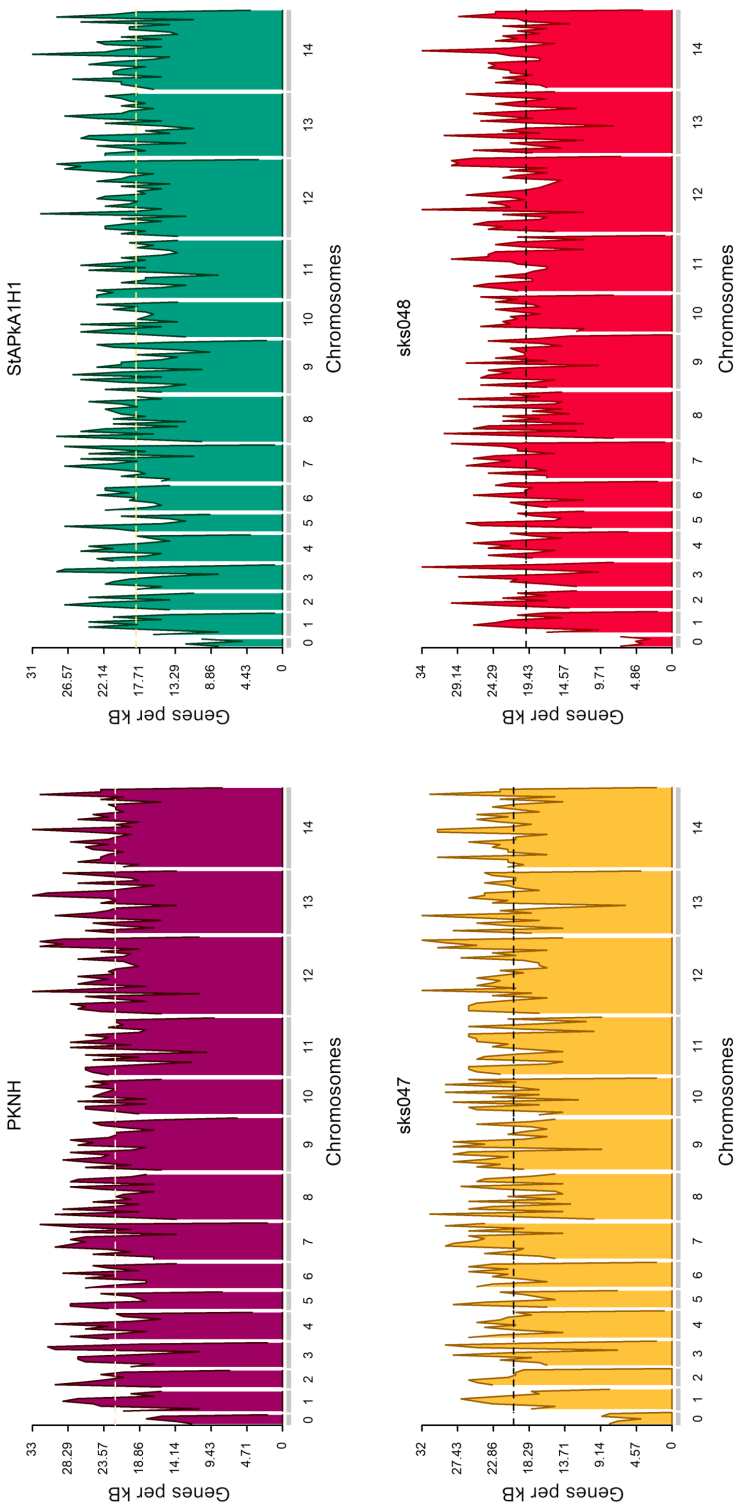


Figure E.4: Gene density plots for the sks050, sks058, sks070, sks074, sks331 and sks339 draft genomes. Gene density was calculated based on the number of identified genes within a sliding window of 100 Kb. Mean density (dashed line) shows the sks050 draft genome to have 20.03 genes per 100 Kb, sks058 to have 19.97, sks070 to have 15.76, sks074 with 18.20 genes and sks078 to have 12.08 genes/100 Kb. Furthermore sks125, sks325, sks331, sks333, sks339 and sks344 had 13.03, 10.12, 16.99, 13.49, 19.47 and 13.08 genes/100 Kb, respectively. Plots were generated using karyoploteR [6].

E.8 Multigene Families of *Plasmodium knowlesi*

This page was intentionally left blank

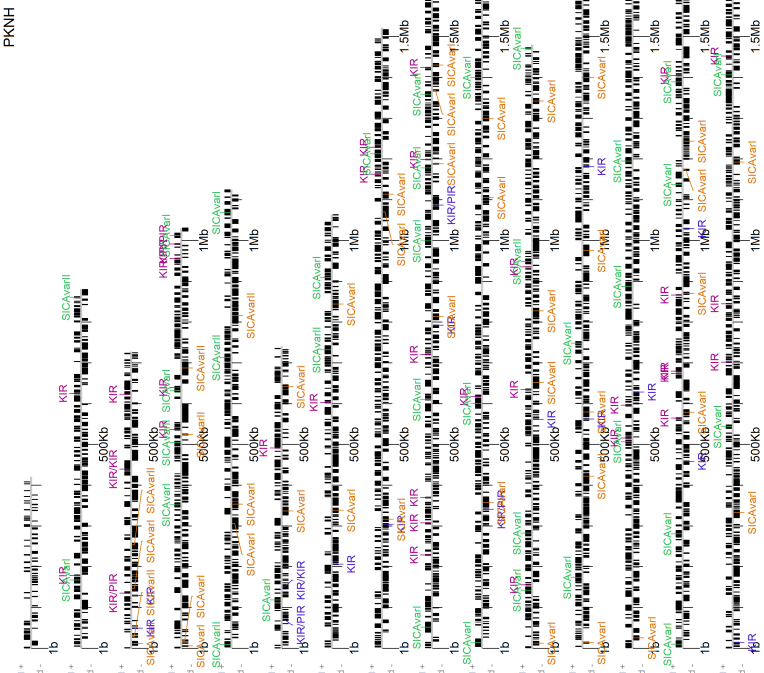


Figure E.5: Genome loci of *SICAvar* and *kir* genes in the PKNH reference genome. Genome annotation was completed on Companion and the resulting General Feature Format (GFF) was parsed for annotated *SICAvar* and *kir* genes. *SICAvar* genes loci are represented in green and orange for the positive and negative strand, respectively. *Kir* genes are shown in plum and purple for the positive and negative strand respectively, per chromosome. Genome loci plots were generated using a custom R script and the karyoploteR package [6].

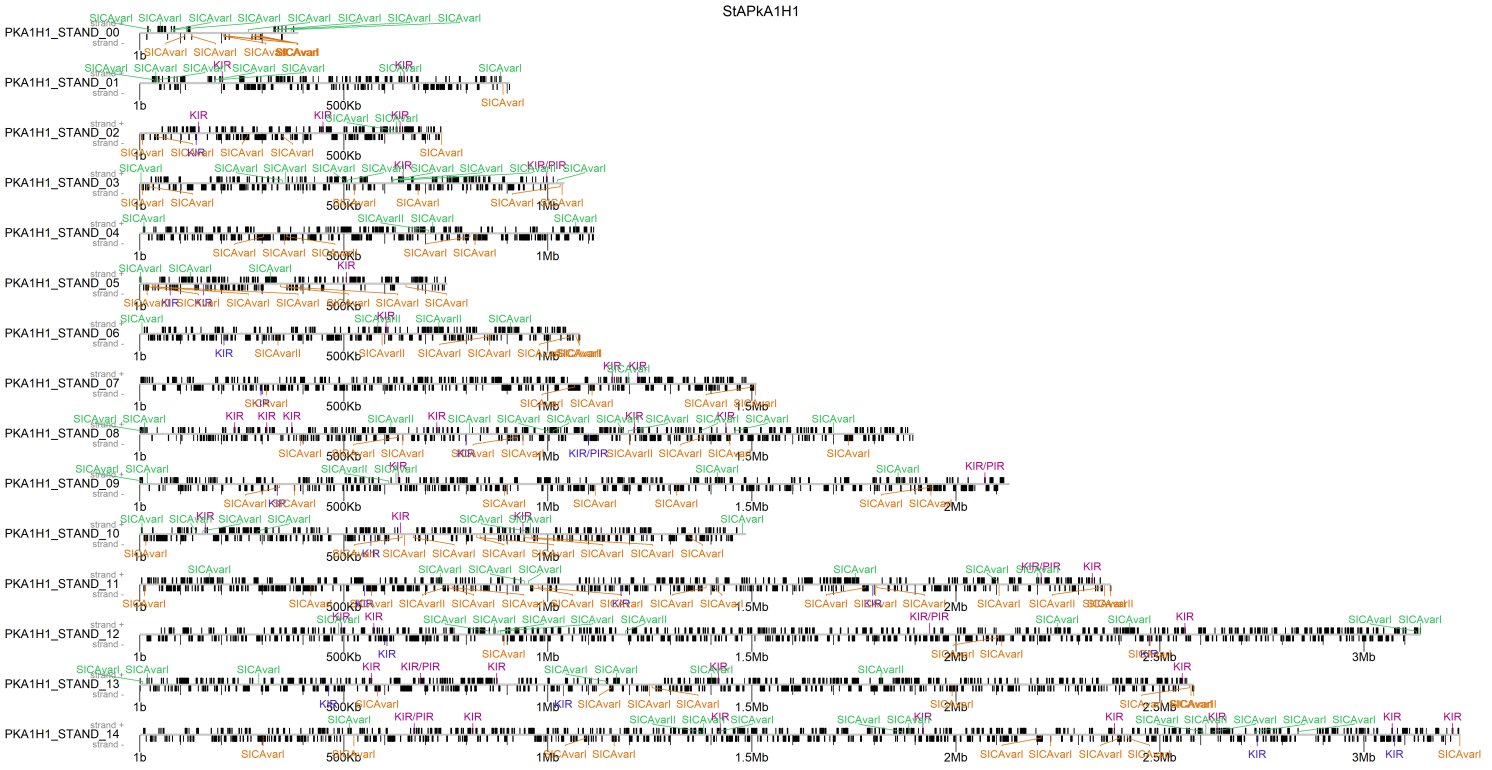


Figure E.6: Genome loci of *SICAvar* and *kir* genes in StAPkA1H1. Genome annotation was completed on Companion and the resulting General Feature Format (GFF) was parsed for annotated *SICAvar* and *kir* genes. *SICAvar* genes loci are represented in green and orange for the positive and negative strand, respectively. *Kir* genes are shown in plum and purple for the positive and negative strand respectively, per chromosome. Genome loci plots were generated using a custom R script and the karyoploteR package [6].

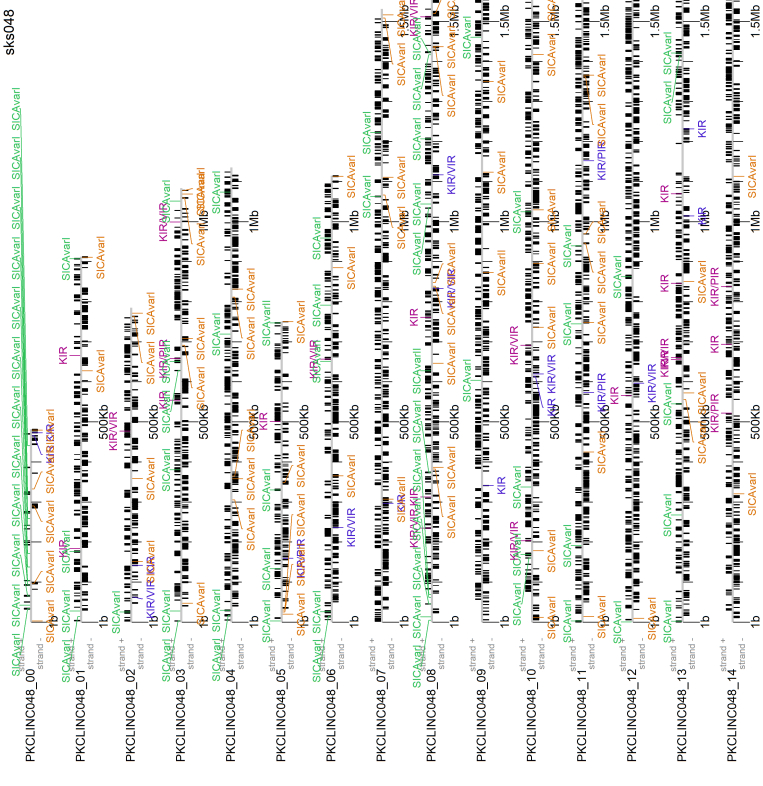


Figure E.7: Genome loci of *SICAvar* and *kir* genes in sks048. Genome annotation was completed on Companion and the resulting General Feature Format (GFF) was parsed for annotated *SICAvar* and *kir* genes. *SICAvar* genes loci are represented in green and orange for the positive and negative strand, respectively. *Kir* genes are shown in plum and purple for the positive and negative strand respectively, per chromosome. Genome loci plots were generated using a custom R script and the karyoploteR package [6].

Figure E.8: Genome loci of *SICAvar* and *kir* genes in sks050. Genome annotation was completed on Companion and the resulting General Feature Format (GFF) was parsed for annotated *SICAvar* and *kir* genes. *SICAvar* genes loci are represented in green and orange for the positive and negative strand, respectively. *Kir* genes are shown in plum and purple for the positive and negative strand respectively, per chromosome. Genome loci plots were generated using a custom R script and the karyoploteR package [6].

Figure E.9: Genome loci of *SICAvar* and *kir* genes in sks058. Genome annotation was completed on Companion and the resulting General Feature Format (GFF) was parsed for annotated *SICAvar* and *kir* genes. *SICAvar* genes loci are represented in green and orange for the positive and negative strand, respectively. *Kir* genes are shown in plum and purple for the positive and negative strand respectively, per chromosome. Genome loci plots were generated using a custom R script and the karyoploteR package [6].

Figure E.10: Genome loci of *SICAvar* and *kir* genes in sks070. Genome annotation was completed on Companion and the resulting General Feature Format (GFF) was parsed for annotated *SICAvar* and *kir* genes. *SICAvar* genes loci are represented in green and orange for the positive and negative strand, respectively. *Kir* genes are shown in plum and purple for the positive and negative strand respectively, per chromosome. Genome loci plots were generated using a custom R script and the karyoploteR package [6].

Figure E.11: Genome loci of *SICAvar* and *kir* genes in sks339. Genome annotation was completed on Companion and the resulting General Feature Format (GFF) was parsed for annotated *SICAvar* and *kir* genes. *SICAvar* genes loci are represented in green and orange for the positive and negative strand, respectively. *Kir* genes are shown in plum and purple for the positive and negative strand respectively, per chromosome. Genome loci plots were generated using a custom R script and the karyoploteR package [6].