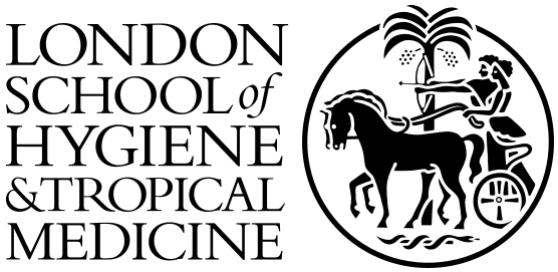
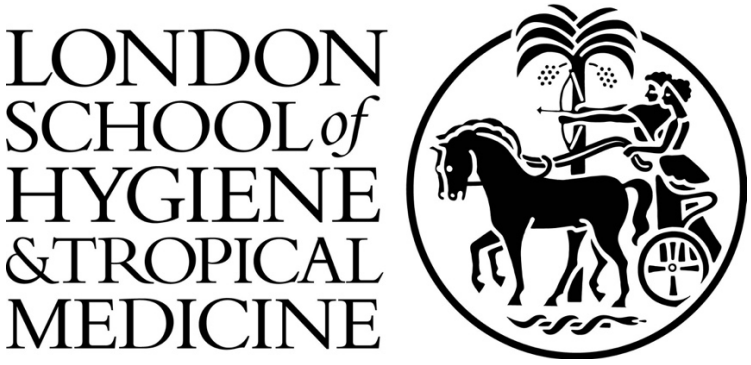




Zika virus surveillance in human and mosquito populations in Cabo Verde – exploring molecular and serological tools for the surveillance of emerging infectious diseases

Daniel Ward

Thesis submitted in accordance with the requirements for the degree of
Doctor of Philosophy of the
University of London
April 2022

Department of Infection Biology

Faculty of Infectious and Tropical Diseases

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by The Bloomsbury Colleges PhD Studentships

Research group affiliation(s): Professor Taane Clark
Professor Susana Campino

I, Daniel Ward, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:  Date: 20/04/2022

Additional publications

Throughout the duration of this PhD scholarship, I have contributed to other manuscripts which are not included in this thesis, listed here:

Phelan, J., Deelder, W., Ward, D. *et al.* COVID-profiler: a webserver for the analysis of SARS-CoV-2 sequencing data. *BMC Bioinformatics* **23,** 137 (2022). https://doi.org/10.1186/s12859-022-04632-y

Da Veiga Leal, S., Ward, D., Campino, S. *et al.* Drug resistance profile and clonality of *Plasmodium falciparum* parasites in Cabo Verde: the 2017 malaria outbreak. *Malar J* **20,** 172 (2021). https://doi.org/10.1186/s12936-021-03708-z

Higgins M, Ravenhall M, Ward D, Phelan J, Ibrahim A, Forrest MS, *et al.* PrimedRPA: primer design for recombinase polymerase amplification assays. Bioinformatics. **35**, 682–4 (2019). https://doi.org/10.1093/bioinformatics/bty701

Abstract

The first confirmed cases of Zika virus disease in the Americas were described in Northeast Brazil in May 2015, following the introduction from French Polynesia. Five months later, the first confirmed cases of autochthonous transmission and Zika congenital syndrome were reported in Cabo Verde, Africa. Following the outbreak, human and entomological samples across endemic regions in Cabo Verde were collected to provide insights into outbreak dynamics. In this thesis, I profile the *Aedes aegypti* mosquito population on Cabo Verde, the primary vector of Zika virus transmission. This work investigates their susceptibility to insecticides, through the targeted sequencing of a cross-section of entomological samples collected in Praia, the capital city. The analysis revealed two *Aedes aegypti* mosquitoes with detectable levels of Zika virus. Building on these findings, I applied molecular techniques to profile the vectors' virome and proceeded to enrich and sequence whole-genome data for Zika virus. The resulting sequence data, combined with a global Zika sequence dataset, were placed phylogenetically into the broader epidemiological context of the 2015-2016 Zika epidemic, revealing two discrete introductions of Zika that occurred from the Americas to Cabo Verde in this period. In parallel, I describe the findings of a sero-epidemiological study based on a cross-sectional cohort of human participants, sampled shortly after the cessation of Zika transmission. By comparing a panel of arbovirus serological assays and using a multivariate logistic modelling approach, key risk-factors for Zika seropositivity were determined. This work was followed up by an investigation on how virus serological diagnostics may be improved. I initially formulate a novel *in-silico* meta-analysis pipeline, which may guide the selection of specific antigenic targets for immunoassay development. This approach is applied in the context of another emerging infectious disease, SARS-CoV-2. Expanding on this, I report a refined, novel molecular methodology for the translation of *in-silico* reverse antigen design techniques into scalable immunoassays. Overall, this thesis describes the application of molecular and serological techniques to understand the dynamics of the Zika outbreak in Cabo Verde. I develop novel methodologies to improve on current serological tools, translating them to address current challenges in the surveillance of emerging infectious diseases.

Acknowledgements

I would like to thank Taane Clark, for giving me the opportunity to study in the group for which I am incredibly grateful. I would like to thank Susana Campino for her guidance over the past four years, both professionally and personally. Your inextinguishable passion for science, which I have eagerly adopted, made my studies a truly enjoyable experience.

The London School is a magnificent place in which to work, entirely due to the amazing people that inhabit it. Having the privilege of working with such an exceptional group of people has enriched my life significantly. The members of the Clark/Campino group are like no other. Through friendship, support for one another and a shared passion for science, the group remains a wonderful, positive collective in which to work. It is a pleasure to work with all of you. Matt, without your support I would not have completed this PhD. Amy and Pepi, our time together has left me with an unbreakable respect for both of you, as scientists and friends. Thank you for a great four years. I'll get the pints in.

I have always believed my ability stems solely from the sum of the people around me. Jody and Ernest, having the privilege of learning from you both has been inspiring. Your always-giving attitude is unparalleled and is an asset to anybody around. Fernanda, Christian and Martin, you took a feral student and turned him in to a molecular biologist. I have had so much fun learning from all of you. Archie, Patrick, Gigi, Aline, Avi, Blem, and all regular attendees of the Pumphandle Bar. Thank you for such a great time, it's been so much fun.

My mum and brothers, Nick and Alex, have set a precedent for familial support, which I truly appreciate. To the finest bunch in London and the West Mids, I have the pleasure of calling my friends, particularly Sam, Tim and Billy.

I have been privileged to study for my PhD, on a subject matter I find interesting. I acknowledge that, while I have enjoyed doing the work contained in this thesis, I have not forgotten its purpose and meaning. Far as I am from its reach, my thoughts go out to all of those affected by Zika virus, particularly the families who have been affected by congenital Zika syndrome. I hope, somehow our collective efforts can contribute to the alleviation and reduction of the morbidity of the 2015-2016 Zika epidemic and any future outbreaks.

Table of Contents

Abbreviations

AMPs	Anti-microbial peptides
AUC	Area under curve
bp	Base pairs
BOB	Blockade of binding assay
BS	Bootstrap support
CDC	Centres for Disease Control and Prevention
CHIKV	Chikungunya virus
DENV	Dengue virus
DAB	Double antigen binding
DNA	Deoxyribonucleic acid
E	Envelope protein
EBOV	Ebola virus
ELISA	Enzyme linked immunosorbent assay
EID	Emerging infectious diseases
EIAV	Equine infectious anaemia virus
FP	French Polynesian
GST	Glutathione S-transferase
HPD	Highest Posterior Density
IBs	Inclusion bodies
Ig	Immunoglobulin
IEDB	Immune Epitope Database
JEV	Japanese encephalitis virus
kb	Kilobase pair
LAMP	Loop mediated isothermal amplification
LFAs	Lateral Flow Assays
mAb	monoclonal antibody
MAC-ELISA	IgM antibody capture enzyme-linked immunosorbent assay
MAYV	Mayaro virus
Mb	Megs base pairs
MCS	multiple cloning site
MFI	Median fluorescent intensity

mg	Milligram
ML	Maximum-likelihood
ml	Millilitre
mM	Millimolar
NAT	Nucleic acid testing
NGS	Next generation sequencing
nM	Nanomolar
NOIDs	Notifiable infectious diseases
NS	Non-structural protein
OD	Optical density
ODU	Optical density unit
pAb	Polyclonal antibody
PAHO	Pan American Health Organisation
PCR	Polymerase chain reaction
PHEIC	Public Health Emergency of International Concern
prM	precursor membrane
PRNT	Plaque reduction neutralization test
proMED	Program for Monitoring Emerging Diseases
QC	Quality control
RDT	Rapid Diagnostic Test
RNA	Ribonucleic acid
RPA	Recombinase polymerase amplification
rRNA	Ribosomal RNAs
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SD	Standard deviation
SNP	Single nucleotide polymorphism
UK-PHRST	UK's Public Health Rapid Support Team
WNV	West Nile virus
WGS	Whole genome sequencing
WHO	World Health Organization
WHOPES	WHO pesticide evaluation scheme
WTA	Whole transcriptome amplification
YFV	Yellow fever virus
ZIKV	Zika virus

Chapter One

Introduction

Hand-drawn dot-maps to satellite imagery. Left: The illustration drawn in 1854 by John Snow geospatially mapping cholera cases in London (originally published in 1854 by C.F.). **Right:** Excerpt of satellite imagery depicting a hospital in South America. The imagery is used to track attendance through the use of image recognition (Nsoesie, E et al. 2015).

1 - Introduction

1.1 Emerging infectious diseases (EIDs)

Emerging infectious diseases (EIDs) pose a substantial threat to the world's animal and plant populations, their morbid effects have destructively punctuated humanity's history thus far. A shift in the evolutionary or environmental compartment in which an organism resides, results in the emergence of once inconsequential or new pathogens to the epidemiological forefront. This emergence or re-emergence occurs through newly imparted pathogenicity, virulence, host availability, or a zoonosis establishing autochthonous transmission in a novel host or vector population. Looking back, the usage of the phrase 'emerging [infectious] disease' started around 1960s, effectively coined when Maurer *et al.* (1962) published their research on the, then emerging, equine piroplasmosis [1]. Since then, notable outbreaks have warranted its

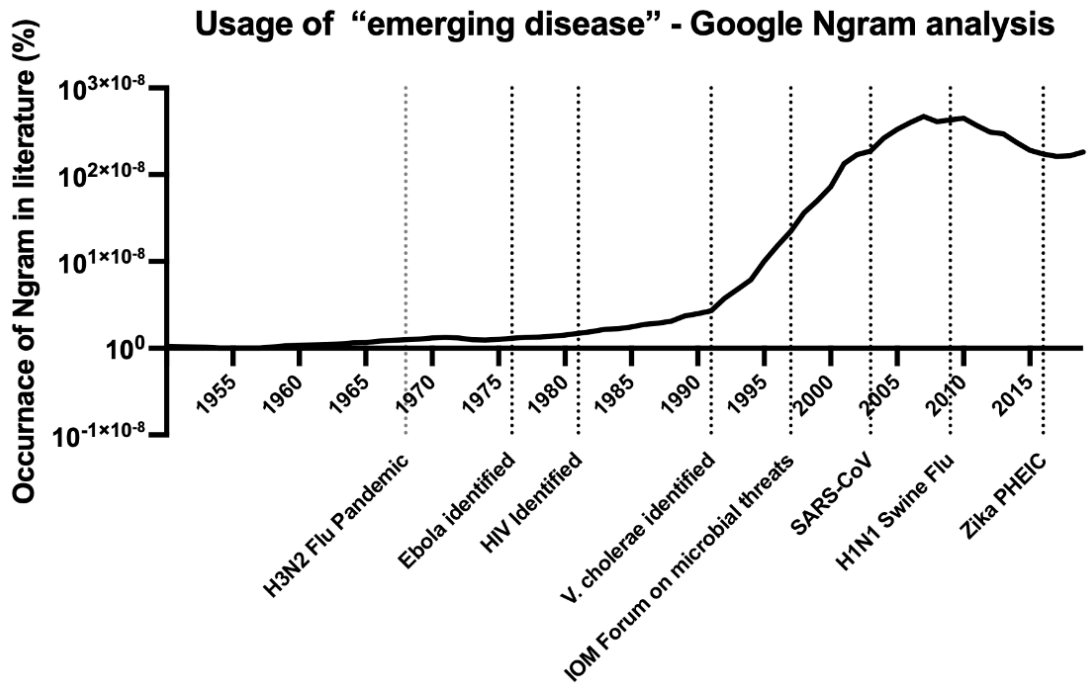


Figure 1: Usage of the phrase "emerging [infectious] disease" increases in the late 20th century to 2019 with a marked increase coinciding the characterisation novel human pathogens. The Ngram database is generated intermittently and has not been updated since 2019.

continual use, such as the H3N2 influenza pandemic in 1968, the characterisation of Ebola virus in 1976, the emergence of HIV in 1981 and SARS-CoV in 2003 (**Figure 1).**

Between years 1980 and 2007, 87 new human pathogens were identified. Of these the majority were single stranded RNA viruses (51%), infamous for their characteristic error-prone replication which acts to speed up evolutionary processes. Of the ~1400 known human pathogens, as few as 7% of those are obligate human pathogens, notably, HIV-1, *Plasmodium falciparum* and variola virus (smallpox), which all have their origins rooted in historic mammalian hosts [3-4], with HIV-1 having shifting more recently [4]. The remaining 93% are transmitted from animal or sapronotic reservoirs [5]. Animal pathogens are therefore of central interest in the surveillance of EIDs.

A *spill-over* infection refers to a transmission event where a novel host, outside of a pathogen's current range, comes into contact with the pathogen and is infected. While a new host may become infected, it is not a guarantee that it will establish onward transmission within the novel population, as is the case with many common zoonoses. This self-limiting transmission chain is called a *dead-end* infection, an example of which is in cases of rabies virus infection, where very few recorded examples of human-human transmission have been recorded [6–8]. While most spill-over events result in dead-ends, there are several prominent cases which do not. The Ebola virus (EBOV) outbreak in West Africa between years 2013 and 2016 resulted in 28,616 suspected or confirmed Ebola virus disease cases and 11,323 deaths, and started from a suspected spill-over infection from a bat [9]. More recently, the RNA virus SARS-CoV-2, another suspected zoonosis, has caused in excess of > 6 million deaths, and $12.5 trillion in estimated global financial burden [10].

The Institute of Medicine's committee on EIDs (1992) published a summary of 'Microbial threats to the United States', in which it was summarised some of the driving forces in disease emergence **(Table 1)** [11]. Aside from 'microbial adaptation', most factors can be attributed to human activity, where new or continued contact with zoonotic reservoirs that would have otherwise remained isolated is exacerbated by economic and technological development, globalisation and localised societal factors driving changes in human behaviour. One of the many examples of this paradigm at work is with the emergence of Zika virus (ZIKV) in 2015-2016, where dense cities populated with all-naïve hosts living alongside the thriving urban-adapted *Aedes* (*Ae.) aegypti* vector encouraged the sustained transmission of a novel, highly virulent lineage of ZIKV introduced from another continent [12, 13].

Table 1 –Factors driving pathogen emergence with examples of EIDs. Excerpt from Institute of Medicine's committee on EIDs (1992) report [11].

Factor	Related Diseases
Human demographics and behaviour	**Ebola (EBOV), Zika (ZIKV)**
Technology and industry	**Vaccine derived polio**
Economic development and land use	***Plasmodium knowlesi* malaria**
International travel and commerce	**Influenzas, coronavirus**
Microbial adaptation and change	**MTB*, MRSA**, Dengue**
Breakdown of public health measures	**Ebola, cholera**

**Mycobacterium tuberculosis* ***Methicillin-resistant *Staphylococcus auras*

1.1.1 Surveillance of EIDs

A central tool in the prevention of EIDs are surveillance programmes, which are implemented at various levels. At the top, global public health reporting infrastructures are critical in reporting the occurrence of infectious disease cases globally, such as the Program for Monitoring Emerging Diseases (proMED), a publicly-available internet messaging board

established in 1994 for conducting the global reporting of infectious disease outbreaks [14] and the World Health Organizations (WHO's) FluNet, a global web-based tool for influenza virological surveillance established in 1997 [15]. At a smaller scale, there exists national and/or academic consortia set up for the surveillance and response to infectious disease outbreaks. For example, the UK's Public Health Rapid Support Team (UK-PHRST), a specialist team with rapid response capabilities to identify and inform on the prevention of disease outbreaks, set up in the wake of the 2013-16 EBOV outbreak [16]. The foci of surveillance networks can take many forms, while the aforementioned have a primarily host-oriented approach, it is of equal importance to focus on vector surveillance through xenomonitoring [17], insecticide resistance monitoring [18], and environmental pathogen sampling [19].

Effective control strategy depends on specific and sensitive data acquisition techniques. In this section, three key disciplines used in the surveillance of EIDs will be discussed, which are summarised **(Figure 2)**. The first, case reporting, was central to one of the earliest examples of an epidemiological study. Its implementation precipitated the removal of the pump handle from a cholera contaminated well by John Snow in 1854 [20]. Now, similar methods can be driven by *big data* analytical techniques, for instance, by researchers using data from Google searches as an early warning of possible outbreaks, through the identification of abnormal increases in disease-specific symptom search engine queries [21]. The second mode covered is serology, the act of testing patient serum, or other bodily fluids, for antibodies or antigen. With the wide scope of serological techniques available for infectious disease research, their implementation can provide insights into the history of infection in an individual. Molecular techniques, the final area, are built on the principle of detecting or characterising pathogen nucleic acids and take a myriad of forms. From semi-targeted and meta-sequencing approaches, such as those used in the identification of SARS-CoV [22], the amplicon sequencing techniques employed

in the surveillance of EBOV and ZIKV outbreaks [24, 25], to the qPCR and RT-qPCR techniques used daily around the world as gold standard diagnostics for SARS-CoV-2.

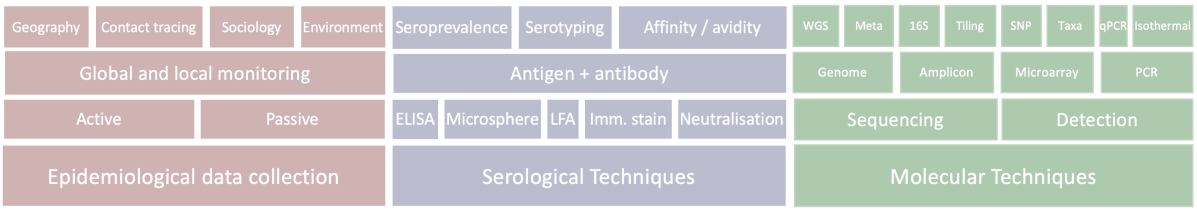


Figure 2- Example techniques in epidemiological data and case reporting, serological and molecular data collection. Three of the disciplines covered in this introduction.

1.1.2 Epidemiological data collection

Epidemiological data collection, case reporting, lays at the foundation of any endeavour in controlling infectious diseases. With its many forms and applications, epidemiological surveillance can be summarised into two principal approaches, passive and active. Presently, in the UK, 33 diseases and their 61 etiological agents are categorised as notifiable infectious diseases (NOIDs), which imparts a statutory duty to all medical practitioners to notify the local council or local health protection team of suspected cases. A weekly report is produced by the UK Health Security Agency (UKHSA) on NOIDs, detailing geographical locale and the frequency of infections. This passive data collection, on the back of primary health care services, facilitates the evaluation and implementation of effective control strategy across the UK. Passive surveillance can be augmented in the 21st century, by big data analytical techniques, on top of the example given previously, involving the analysis of Google data [21], satellite imagery can be used to model and predict the spread of pandemic influenza in developing countries, where passive surveillance networks are unestablished due to the insufficiency of resources and international support [25]. On the global scale, organisations

such as the WHO or Centres for Disease Control and Prevention (CDC) collect passive data on a myriad of communicable diseases. This mode of surveillance is the most used due to its cost-effectiveness and economic reasons, however, the detail of the data collected is usually sparce or incomplete, due to the limited time commitment healthcare workers can dedicate to such tasks, therefore, these systems are often bolstered with incentives. Across Guinea, Liberia, and Sierra Leone, the CDC, WHO and numerous governmental and non-governmental organisation have worked to install an integrated functional surveillance system for the EID, Ebola [26]. These efforts have strengthened epidemiologic and data management capacity, enabling the quick response to future public health emergencies. It is in this example, a well-illustrated intersection between passive and active surveillance lies. Passive surveillance invariably instigates active surveillance, where an increase in cases of a notified disease leads to the activation of focused and specific resources. In this case, reported EBOV cases instigate the prompt deployment of teams, fixed on case identification, contact tracing, and containment.

Active surveillance does not only facilitate the monitoring and containment of EIDs. With the increased granularity at which it is usually undertaken, it can increase the fundamental understanding of the etiological agents behind them. In the context of Dengue virus (DENV) research, studies on the dynamics and immunology of infection, can involve the recruitment of participants during local surveys, or through associations with regional medical centres, where the targeted population has a vested interest in controlling disease morbidity. Detailed information is gathered, covering their health, socio-demographic data and in many cases, biological samples for screening. It was found, in the case of Dengue, that active surveillance studies captured significantly more cases when compared to passive surveillance, at times, between 10 to 21-fold higher than the caseloads detected by a national reporting system [27]. Alarming as this differential may be, improving national surveillance or sustaining the logistics

and funding behind active studies is, in many cases, economically prohibitive, and so, a balance between these methodologies must be struck. Sentinel surveillance embodies just that, where a scaled-down, specialised and often incentivised operation is sustained for the monitoring of EIDs. In the WHO's Global Strategy for Dengue Prevention and Control (2012-2020), sentinel surveillance was an integral part of the proposed framework [28]. With this, the economic burden of tracking Dengue incidence is spread, while still hitting the key deliverables of active surveillance studies.

In many endemic settings, several arbovirus species with similar symptoms are in autochthonous transmission at any one time. Reporting accurate epidemiological data for DENV or any other disease, requires a diagnostic strategy that is specific, which can effectively differentiate Dengue from other common diseases. It is therefore essential to augment practices commonly used in passive surveillance, diagnostics based on symptoms or patient self-reports, with methodologies that offer enhanced specificity and sensitivity, such as serological and molecular assays.

1.1.3 Serological techniques

Diphtheria, tuberculosis and typhoid were all a blight of 19th century life in Europe. Across England and Wales, these three diseases were responsible for the death of 0.1% of all children aged 1 to 15 years [29]. One of the first serological tests, the Widal test (1897), used antibody agglutination to indicate typhoid seroconversion [30]. With the application and development of techniques such as these, diseases were characterised, monitored, and immunised against, to the point that now, in Europe, all three pose a comparatively reduced risk to public health. Throughout the 20th century serological tools were developed, aiding in the identification of novel viral pathogens. One example of which, is found in the isolation of ZIKV, first reported

in 1947 [31]. Only in 1952 did Dick *et al.* characterise the suspected novel virus by determining Zika's unique serological specificity, comparing the neutralising activity of convalescent Zika sera with Dengue and yellow fever using *in-vivo* passage techniques. Around this period, research led by Coons *et al.* (1941), centred on antibody fluorophore conjugation, paved the way for 80 years of immunoassay development [32], where today, the enzyme linked immunosorbent assay (ELISA) and its variants form a central part of disease surveillance strategy.

One of the many uses of immunoassays today, in the context of EIDs, is to monitor the prevalence of seroconversion in a population (seroprevalence). Monitoring seroprevalence allows epidemiologists to track the emergence of diseases geospatially and temporally. A key aspect of this practice is the detection of discrete antibody isotypes using indirect ELISA techniques. These methodologies are capable of differentiating, for example, captured immunoglobulin M (IgM) and G (IgG), through isotype specific secondary antibody conjugated with a radioactive label, fluorophore or a chemiluminescent or colorimetric reporter. This association yields a quantitative measure of an antibody-bound analyte. As is the case with many primary antibody responses to infection, an initial increase in pathogen-specific IgM titres precedes, often by greater than two weeks, the IgG response [32][33]. Recognizing this differential response can infer active/recent, convalescent or in some cases, secondary infections. These methodologies have been applied recently in the surveillance of DENV in combination with molecular assays to identify active, *primary* and *post-primary* infections in a multiple-serotype endemic settings, using a differential in RT-qPCR, IgM and IgG ELISA assays to determine the chronology of infections [35].

Another format of ELISA is used in the detection of an etiological agent, or the components of it, within a given serological sample. These 'direct' and 'sandwich' ELISA work on the principle of antigen immobilisation, either through covalent or non-covalent interactions with a solid phase, or immobilisation using a 'capture antibody'. The antigen is then detected using a detection antibody conjugate, comparable with those used for indirect ELISA. Importantly, the antibodies used in both immobilisation (sandwich) or detection, can be either monoclonal (mAb) or polyclonal (pAb), depending on the application. Where mAb reagents offer high specificity, they can, in some cases, result in reduced sensitivity, as they operate within a limited space in which it can react with an antigen, as opposed to polyclonal reagents, which employ a diverse array of epitopes for capture and/or detection of the antigen. While antigen detection is an efficient means of disease surveillance, this methodology has a major limitation when compared to antibody detection. Namely, antigen is usually cleared by the immune system promptly with the resolution of the infection. This limits the effective sensitivity of the assay to a narrow window, the duration of the infection only.

A crucial application of serologic assays in EID control is in HIV diagnostics and surveillance. Since its emergence in 1983, there have been four generations of serological HIV diagnostics. The first two generations employed IgG antibody detection methodologies using a panel of HIV-1/2 antigens, with the addition of IgM detection in the third generation, increasing the assay's effective sensitivity from a timeframe of six to three weeks post-infection [36]. In the fourth generation, antigen and antibody detection methodologies were combined, reducing the 'test-negative' window to two weeks.

The development of immunoassay techniques in disease surveillance advanced significantly in 1989 with a patent filed by Becton Dickinson and Company for a capillary flow assay, today

known as lateral flow assays [37]. This format, also referred to as a rapid diagnostic test (RDT), is used as a central component of control strategies worldwide, for the surveillance of Dengue [38], malaria [39], SARS-CoV-2 [40] and many more. One hundred and twenty five years after the deployment of the Widal test, in the control of SARS-CoV-2, over 384 million antigen capture lateral flow assays (LFAs) have been used in a single year to help curb SARS-CoV-2 transmission [41]. RDTs enable the point-of-care analysis of samples, which can drive decision making both in the clinic and in field settings.

Building on the fundamental mechanics of an ELISA, microsphere technologies, often referred to under the brand name 'Luminex', enable the analysis of up to 500 analytes in a single assay. This methodology employs microscopic 'beads', each with an individually distinguishable fluorescent signature. These beads substitute the solid phase of an ELISA (the polystyrene plate medium) with a pool of microspheres, each of which is capable of binding antigen/antibody, with a range of covalent and non-covalent 'coupling' options. On incubation with an analyte, the presence of antigen/antibody is quantitatively analysed by means of a fluorescently conjugated reporter. This technology is not only used in detecting pathogen antigens and the host responses to them, but a myriad of other biomarkers such as cytokines, cancer markers and hormones. While this technology holds great promise, the initial capital investment and cost-per assay, means that these technologies are unavailable in many settings for sustained EID surveillance.

Microarray platforms can determine the reactivity of sera, to an array of hundreds of targets in a single assay. Through this, the dissection of responses to an entire proteome can be achieved [42]. Like microsphere technology, however, this process has a large initial capital investment.

Numerous additional technologies for serological research exist, on top of those outlined here. When optimised an applied resourcefully, they offer a versatile and efficient means of pathogen detection on the population level.

1.1.4 Molecular techniques

There exists an ever-growing armoury of molecular tools at the disposal of clinicians and researchers in the perpetual competition between humans and their parasites. In the past forty years, these technologies have revolutionised bioscience and healthcare. At their foundation, molecular assays rely on the detection of nucleic acids and their products to facilitate specific diagnostics and disease surveillance. Initially, restriction digestion and southern blot assays were applied to diagnose human genetic diseases [43], now, these techniques are employed as gold standard diagnostics in the detection of EIDs, most notably, SARS-CoV-2.

The polymerase chain reaction (PCR) and its derivatives, applied in diagnostic settings, is capable of detecting pathogens with unparalleled sensitivity and specificity. In this thesis, two of its most powerful applications are covered, high throughput RT-qPCR (reverse-transcriptase quantitative PCR) and PCR coupled with amplicon sequencing. In arbovirus endemic regions, molecular diagnostics are capable of differentiating pathogens with high specificity and sensitivity. In the case of DENV surveillance, they are capable of differentiating the four serotypes, which would otherwise only be possible through costly and laborious neutralisation assays [44]. As outlined in the aforementioned WHO strategy, this information lies at the centre of DENV surveillance [28]. Multiplex qPCR assays can detect numerous pathogens in a single reaction, enabling the high throughput screening of thousands of samples over a short period. A further iteration of PCR technology includes loop mediated isothermal amplification (LAMP) and recombinase polymerase amplification (RPA), two isothermal assays that use

alternative amplification technologies for the detection of pathogen nucleic acids. These assays have a particular edge against conventional techniques due to their ease of use and potential application outside of a laboratory environment, thereby enabling point of care diagnostics both in clinics and in the field [45].

Amplicon sequencing has evolved with the advent of next generation sequencing (NGS) technologies into a powerful methodology. This technique has laid the foundation for numerous other procedures, including 16S and tiling amplicon genomic sequencing. The latter is used frequently to sequence whole viral genomes from complex samples, for use in genomic epidemiology studies, a discipline which is now key in the control of EIDs. Through methodologies applied in this discipline, the real-time tracking of viral evolution and transmission dynamics is possible, which is covered further in **Chapter 3**. Also covered in that chapter, is the application of meta sequencing approaches, which like amplicon sequencing, is enabled by high-yield second and third generation sequencing technologies. These methodologies can facilitate the bulk-sequencing of complex samples, without enrichment. Unlike amplicon sequencing, this method does not require *a-priori* knowledge of pathogen sequences and has been used previously for the discovery of novel pathogens.

1.2 The emergence of Zika virus

The 2015-16 ZIKV outbreak instigated the most recent Public Health Emergency of International Concern (PHEIC), called by the WHO in February 2016 in response to a prolonged substantial epidemic spreading rapidly across the Americas, and the causal association of pre-natal ZIKV infection and microcephaly. Since then, 49 countries across the region have reported 583,451 suspected cases of ZIKV infection [46]. ZIKV is a positive-sense single stranded RNA virus, with a 10 kb non-segmented genome. It belongs to the genus

Flavivirus, which it shares with, amongst others, an array of common human arboviruses. These include Dengue virus (DENV), West Nile virus (WNV) and yellow fever virus (YFV), all of which are transmitted by the *Ae. aegypti* mosquito vector [47]. Naturally these viruses bear a similar profile of endemicity (**Figure 3)**, which is complicated further by a common clinical presentation.

Before the 2015 ZIKV outbreak, *Flavivirus* infections were understood to cause only infrequent severe disease, with infections predominantly self-limiting. Of the *Flavivirus* species*,* DENV results in the highest morbidity worldwide, with 1.14 million disability-adjusted life years lost in 2013, and an estimated 9,221 deaths per year [48]. Severe Dengue or Dengue haemorrhagic fever, characterised by plasma leaking, fluid accumulation, respiratory distress, severe bleeding, or organ impairment, occurred in 0.4% of cases in the Americas in 2015 [49]. Unlike Dengue, Zika, since its discovery in Uganda in 1947, had only warranted trivial investigation from the scientific community [31]; it rarely caused disease in the majority of those infected. With the 2015 ZIKV outbreak in Brazil and the definition of congenital Zika syndrome, the condition which encompasses the congenital birth defects associated with pre-natal Zika infection, Zika became an essential topic for biomedical research.

Zika congenital syndrome effects between 1.0-4.9% of children who are infected prenatally [47][48]. The manifestations of Zika congenital syndrome can be varied, but are commonly characterised by five key birth defects: (i) severe microcephaly, where the skull in some cases has partially collapsed; (ii) decreased brain tissue, characterised by a distinct pattern of brain damage and subcortical calcifications, (iii) macular scarring and focal pigmentary retinal mottling; (iv) muscle contractures; and (v) hypertonia – restricted body movement soon after birth [52]. To further compound the morbid influence of Zika emergence, retrospective studies

have associated the spread of ZIKV with an increase in the incidence of Guillain-Barré Syndrome, an autoimmune mediated neurological disease causing nerve damage, muscle weakness and paralysis [50][51].

With the rapid emergence and intercontinental spread of Zika over recent years, WHO, Pan American Health Organisation (PAHO), CDC and other research and health organisations have advocated the need to understand the epidemiology of the disease. The 2016 PHEIC, coupled with the 2016 Rio de Janeiro Olympic games, attracted significant media attention [55], resulting in an injection of funding for Zika research. Despite these efforts, the implications of the 2015-16 American Zika outbreak are far reaching. For example, thousands of children have been afflicted with severe disability ranging from cranial abnormalities to severe movement and learning difficulties. The majority of these children reside in low to middle income countries, where the high level of funding required to meet their needs may not be available. Given the current outlook on climate change, and the increasing presence of the *Ae. aegypti* mosquito, ZIKV has the potential to spread further and wider than seen during the previous outbreak.

"Overall, the global risk assessment has not changed. ZIKV continues to spread geographically to areas where competent vectors are present. Although a decline in cases of ZIKV infection has been reported in some countries, or in some parts of countries, vigilance needs to remain high." - **WHO Zika Situation Report (2017)** [56].

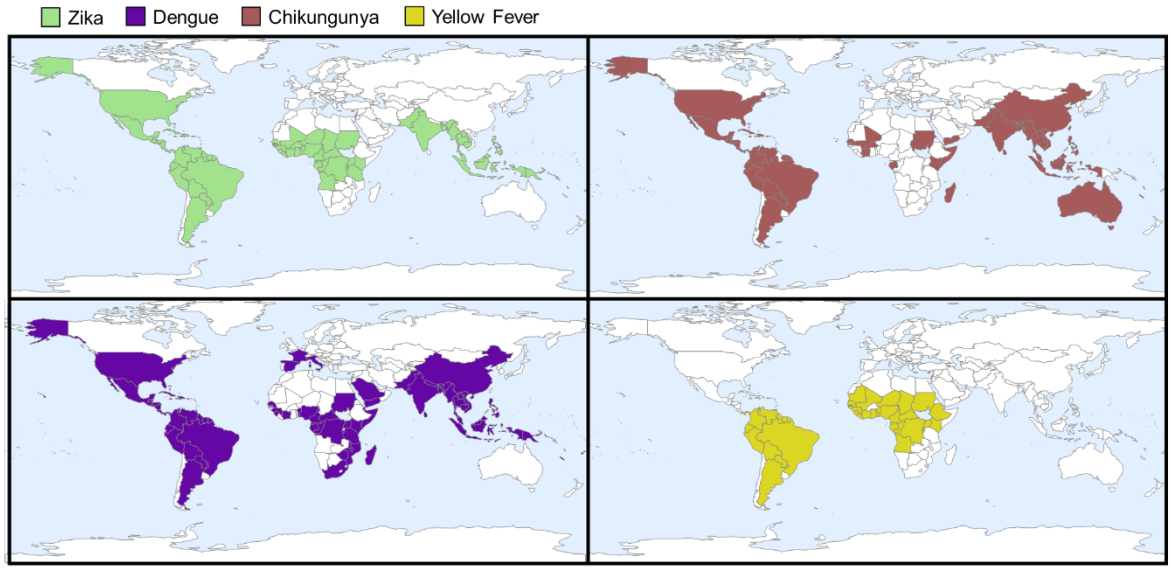


Figure 3. Countries with a history of ZIKV, DENV, CHIKV, and YFV transmission. Data was sourced from the WHO, CDC and Leta *et al.* (2018).

1.2.1 Current diagnostic and surveillance strategy

On the 23rd of March 2016, WHO published *"Laboratory testing for Zika virus infection - Interim guidance"* (**Figure 4**) – a document providing guidance to healthcare professionals and laboratory technicians on testing strategies prioritised for patients in high-risk groups: symptomatic and asymptomatic pregnant women with possible exposure to ZIKV. One of the primary obstacles to effective Zika diagnosis is the crossover between *Flavivirus* species in initial patient presentation. Moreover, other endemic non-Flaviviral *Ae. aegypti* transmitted arboviruses such as Chikungunya virus (CHIKV) further increase diagnostic complexity, which like *Flavivirus* species, result in the characteristic high fever, joint/muscle pain and cephalgia. This issue is illustrated in **Figure 3**, a geographic plot of countries that have reported transmission of the above-mentioned arboviruses. In 2015, before the ZIKV outbreak was declared, Brazil had reported consistent autochthonous transmission of CHIKV and DENV, which ultimately led to a delay in the identification of Zika cases. The first notified case of

ZIKV infection was recorded in May 2015 – years after the suspected introduction of ZIKV to Brazil [57].

Due to the transient period of viremia exhibited during ZIKV infection, diagnostic strategies are modified depending on time of presentation relative to disease onset. Although viral RNA persistence in patient urine and semen has been reported to outlast that in haematological specimens, during the course of a typical infection, viral clearance occurs around 7 days after disease onset [55, 56]. Patients presenting with onset of symptoms ≤ 7 days will be subject to nucleic acid testing (NAT). In this application of NAT, a multiplex RT-qPCR protocol may be used with a pan-flavivirus multiplex primer-probe panel, or individual panels testing ZIKV, DENV, YFV, and WNV sequentially. Samples collected from patients who present ≥ 7 days after disease onset will undergo both NAT and serological diagnostics. These assays consist of IgM antibody capture enzyme-linked immunosorbent assays (MAC-ELISA), coated individually with either ZIKV, DENV, YFV or WNV whole-cell lysate antigen. The principal obstacle to this methodology is the inherent serological cross reactivity which is exhibited by *Flavivirus* species.

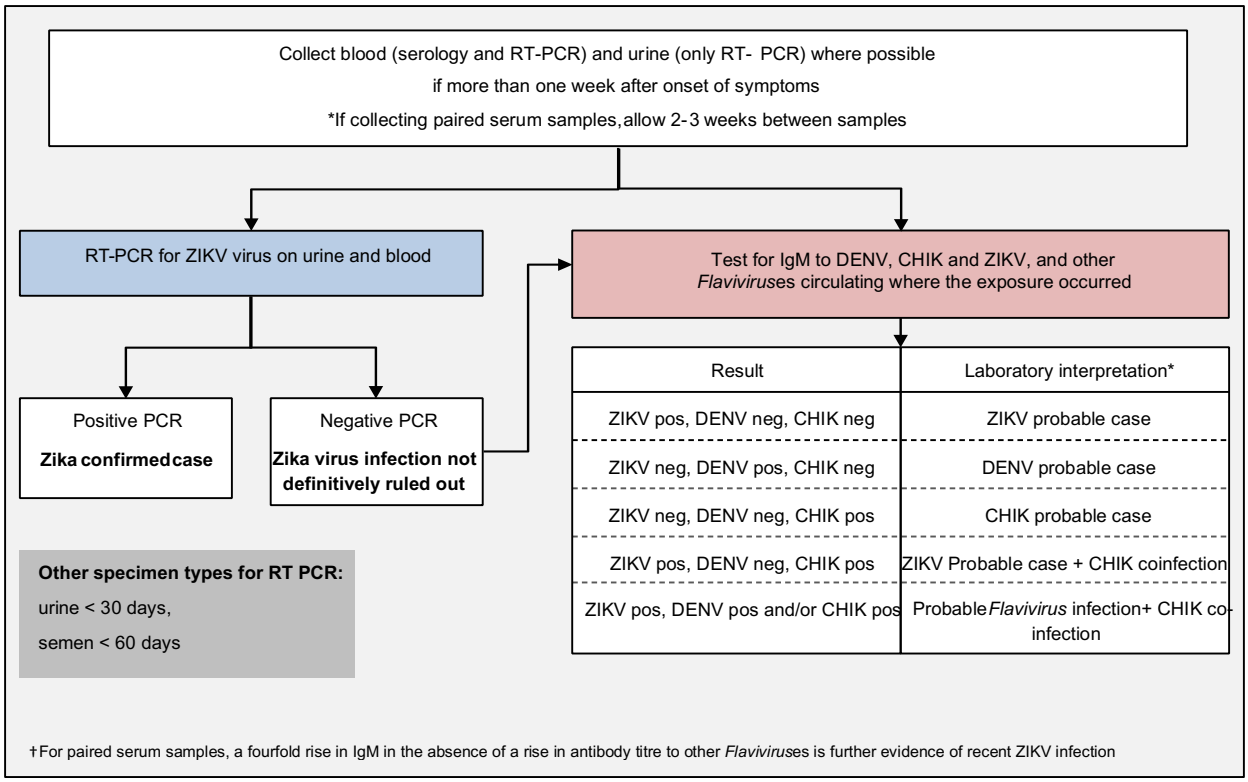


Figure 4. Testing algorithm for suspected cases of arbovirus infection more than one week after onset of symptoms – Laboratory testing for ZIKV infection (WHO 2016).

There is high genomic sequence identity shared between ZIKV, DENV, WNV and YFV, especially in surface exposed immunodominant regions such as the envelope protein (E) and non-structural protein 1 (NS1). Therefore, the antibody response elicited during infection with ZIKV or DENV species bears a high level of inter-*Flavivirus* cross reactivity, resulting from a high frequency of common antibody epitopes. This confounds antigen based serological diagnostics, as a patient who tests positive for Zika IgM, may have been infected with any combination of other *Flavivirus*, particularly in regions where there are multiple endemic species. That said, the MAC-ELISA diagnostic will determine if there has been a recent infection, as the presence of anti-*Flavivirus* IgM is generally indicative of recent *Flavivirus* exposure. With a positive serological assay, the 'gold standard' diagnostic may then be performed: a plaque reduction neutralization test (PRNT). This technique however is inherently technical and requires specialised facilities and trained technicians. Furthermore, the capacity

for the PRNT to be optimised for high throughput clinical sample processing is non-existent, due to the frequent incubation periods and a labour-intensive methodology. This technique, however costly, is required for an accurate diagnosis after viral clearance, providing vital information to future parents on the possible outcome of their pregnancy.

1.2.2 Humoral response to ZIKV infection

Primary *Flavivirus* infection results in a typical naïve antibody response (**Figure 5 - left**). Anti-ZIKV IgM titres increase 3-5 days after symptom onset, which begins to wane after 10 days, but will remain detectable for up to 20 days. A primary IgG response then forms shortly after that of IgM, which continues to increase for months post-infection [60]. On secondary infection, a prompt specific memory IgG response is reared, with little or no production of IgM [61].

In cases where there have been multiple distinct *Flavivirus* species infections (a common example entailing historical DENV infection followed by a recent ZIKV infection), the secondary memory IgG response is subject to a phenomenon known as antigenic 'original sin'. In this scenario, a DENV memory IgG response will be reared to a primary ZIKV infection (see **Figure 3 - right**). While this DENV memory IgG population may mature into a more specific ZIKV infection antibody response with time, the initial response period will exhibit substantial levels of crossreactivity to the *original Flavivirus* infection.

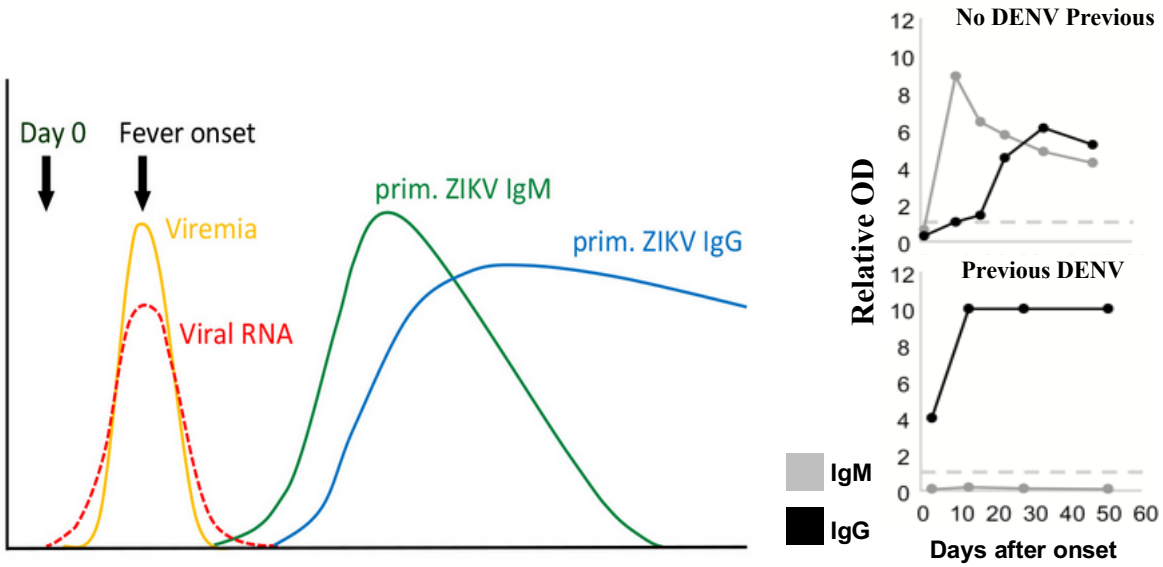


Figure 5. **(left)** A typical primary ZIKV infection antibody response. **(right)** ZIKV antibody response in patients with and without DENV previous infection, sourced from Barzon, L. *et al* [61].

1.2.3 Current Zika diagnostics

Besides the pressing requirement for post-convalescence diagnostics in the context of pregnancy and pre/post-natal care, specific serological techniques are also required for disease surveillance studies. Currently there are limited commercial solutions for non-cross reactive Zika serological assays. The leading commercial assay is produced by Euroimmun (PerkinElmer, Inc), an ELISA based on a recombinant ZIKV non-structural protein 1 (NS1) antigen, capable of detecting both IgM and IgG [62–65]. Despite its apparent strength in escaping cross-reactive signal, the Euroimmun assay remains inaccessible to many research groups and is unfavourable for large cross-sectional studies, due to its high cost (£1k per 96 well plate unit).

It is therefore apparent that alternative strategies should be developed to perform this task. An assay for such an application should be high-throughput, specific and cost-effective, with

minimal logistical requirements for transit and implementation. Several groups have attempted to address this problem. Tsai *et al.*, (2017) used recombinant whole-gene ZIKV and DENV NS1 protein ELISA in series, to differentially tease apart antibody responses. Multiplex microsphere immunoassays such as that demonstrated by Basile *et al.*, (2013) show great high-throughput potential, but are limited by the specificity of the antigen coupled to the microsphere beads. Other strategies such as LFAs based on monoclonal antibody viral antigen capture have been developed, but are limited by the relatively quick clearance of viral antigen from the host, thereby restricting the detection period [68]. Another solution, the a blockade of binding assay (BOB), employs label-conjugated monoclonals known to bind a virus-specific epitope in a competitive assay, which indicates the presence of Zika-specific antibodies in the analyte polyclonal response [69]. Computational methods of discriminating cross-reactive antibody responses have also been demonstrated, using models based on data surrounding antibody response specificity to heterologous *Flavivirus* antigens, antibody avidity and antibody isotype, to differentiate responses [70].

1.2.4 *Aedes aegypti*: a highly competent vector

The principal vector for ZIKV in the Americas is the *Ae. aegypti* mosquito. Also responsible for the transmission of DENV, CHIKV and YFV, *Ae. aegypti* has become well adapted for life in urbanised regions. These mosquitoes thrive in both fresh and stagnated water, in vessels ranging from barrels to bottle-caps, and colonise human residencies. A preference to feeding on human blood and a tendency to feed multiple times during an egg-laying cycle imparts this particular vector with a remarkable efficiency in pathogen transmission [69,70]. Furthermore, the wide distribution of this vector has bestowed a capacity to spread rapidly, *Flavivirus*es and other associated arboviruses [73].

1.2.5 Vector surveillance and control

In 2016, for deployment across the Americas, the WHO laid out a "Vector control" operations framework for the control of ZIKV [74]. These measures targeted all stages of the *Ae. aegypti* life cycle. Strategies targeting eggs, larvae and pupae were centred around the removal of breeding grounds. Small water receptacles such as discarded tires and plastic containers were removed in community led clean up campaigns. Large receptacles for domestic water storage were advised to be covered and treated with larvicidal compounds. Larger bodies such as ponds or reservoirs were subject to the introduction of larvivorous fish or other larvivorous aquatic organisms. To address the mature, adult mosquitoes requires the use of chemical methodologies, targeted residual spraying, and space spraying. Such spraying involves the application of appropriate insecticides to *Ae. aegypti* resting sites, such as walls, furniture, dark, moist, and enclosed spaces and to a lesser extent, around houses. Space spraying involves the application of insecticides validated by the WHO pesticide evaluation scheme (WHOPES) to larger spaces. WHOPES recommends insecticides are selected based on the susceptibility of local mosquitoes to specific compounds.

The selective pressure exerted on mosquito populations through the continual use of insecticides gives rise to insecticide resistance, which reduces the capacity of insecticides to eliminate mosquitoes. This resistance is screened by WHOPES, primarily using bioassays, which involves the rearing of mosquitoes in bottles coated with a known amount of insecticide, and subsequent assessment of insecticidal efficacy. These bottle bioassays are labour intensive taking weeks to perform. Therefore, molecular methodologies have been developed as an approach to monitoring insecticide resistance through genomic analyses.

As discussed previously, targeted molecular assays require *a-priori* knowledge of targets for screening. The wide applicability of insecticides across several mosquito species results in the conservation of genes associated with resistance, and so, loci associated with these phenotypes are well described [75–77]. Combined with xenomonitoring techniques, the practice of molecular screening of resistance associated loci can facilitate the accurate assessment of the resistance profile and infectious status of vector populations in a region. These data are essential to the control of vector-borne EIDs and understanding the transmission dynamics of arboviruses as their inevitable spread coincides with the change in the global climate.

1.3 Zika in Cabo Verde

Before the end of 2015, across the Americas, 11 countries reported PCR positive ZIKV cases, including the USA. Moreover, evidence of further intercontinental transmission was found when positive PCR assays were reported in Cabo Verde, located off the coast of West Africa, 500 km west from Senegal. Historically, Cabo Verde has strong intercontinental associations, after colonisation in 1456 by the Portuguese Empire. These links to this day result in an influx of travellers from both Portugal and Brazil, with regular direct flights to both countries, as well as the USA and Africa. Furthermore, Cabo Verde is reliant on the importation of agricultural goods and commodities from neighbouring countries. These factors may increase the risk of introducing novel pathogens and vectors to the island. The first outbreak of any *Flavivirus* species occurred on Cabo Verde in 2009, a DENV-3 epidemic, culminating in 20,914 reported cases and 4 cases of severe Dengue with mortality [78]. The 2015-16 Cabo Verdean outbreak was the first large ZIKV outbreak with confirmed cases of microcephaly in Africa. The Cabo Verde Ministry of Health (Ministério da Saúde) reported 7,580 suspected cases between October 2015 and May 2016 with 18 cases of Zika congenital syndrome – 63% of which were

in Praia on Santiago Island, with cases peaking on week 47 of 2015 [79, 80]. The initial diagnoses of ZIKV on Cabo Verde were confirmed by molecular techniques. Further cases were reported based on the identification of symptomatic patients. The ZIKV outbreak on Cabo Verde is an overlooked topic in the ZIKV forum with very little research undertaken. Its position between continents places the archipelago in a crucial position for EID transmission, and should therefore be a key topic for research and the prevention of future outbreaks.

1.4 Outline of thesis

In this thesis, the range of the techniques discussed above are integrated in a multi-disciplinary analysis that describes the origin and dynamics of the ZIKV outbreak on Cabo Verde, reporting on several of the key facets which drove the outbreak in 2016. This is followed up with the description of an ongoing development of methodologies through which current serological tools can be improved upon for applications in EID surveillance in the laboratory.

In **Chapter 2**, *Ae. aegypti* populations on Cabo Verde are characterised through the application of sensitive *Flavivirus* xenomonitoring and amplicon sequencing techniques, as a part of a collaborative piece of research. The incidence of ZIKV infection in the vector population in Praia is reported and their susceptibility to insecticides is described. These findings are built on in **Chapter 3**, applying metagenomic sequencing techniques to characterise the RNA virome of a ZIKV infected mosquito. Enrichment and whole-genome sequencing techniques are applied to describe the introduction of ZIKV to Cabo Verde, placing the outbreak in a global context using phylogeographic reconstruction.

Approaching the outbreak from the perspective of the host, in Chapter Four, a cross section of the human population in Praia is probed for evidence of *Flavivirus* infection using a panel of serological assays. These data are analysed, combined with the metadata collected with each sample, to build a model yielding risk-factors of *Flavivirus* seroconversion.

Noticing a necessity for the improvement of the tools available for the serological surveillance of *Flavivirus* species, I apply *in-silico* techniques to guide the selection of antigenic peptides in Chapter 5. Throughout the disruption caused by the SARS-CoV-2 pandemic, I re-focused my efforts, to build further on our *in-silico* analyses, through which, our own ZIKV antigen panel in enhanced, and a web-tool for performing similar analysis on SARS-CoV-2 is produced.

Finally, I explore the ways in which these analyses might be translated, *in-vitro,* to serve as specific and scalable surveillance tools. I describe my progress in the development of a novel molecular assembly and expression technique, which could ultimately form the basis of low-cost diagnostics. The final chapter (Discussion – Chapter 7) discusses the over-arching strengths and limitations of my work and provides a framework for future work.

Together, this thesis describes the application of current methodologies in describing the emergence of the Asian lineage of ZIKV in an African population, covering two of the key facets of arbovirus transmission. In noticing one of the key challenges in completing such research, I make efforts to solve them, in an exploration of the molecular and serological techniques for diagnostic design.

1.5 Aims & objectives

Chapter 2

Profiling disease vector populations is central to outbreak control efforts. Two key components in building such a profile are the screening of vector populations for infection with human pathogens and the characterisation of vector germline mutations conferring resistance to insecticides. I believe that by producing such a profile, covering the vector population on Cabo Verde after the 2015-2016 ZIKV outbreak, the understanding of the role *Ae. aegypti* mosquitoes play in future *Flavivirus* outbreaks will be increased, facilitating both preventative and outbreak control efforts.

To assess the prevalence of Zika and Dengue virus infection in an entomological dataset collected across Praia, the capital and greatest region of human arbovirus infection of Cabo Verde. To achieve this, I will:

- Extract nucleic acids from ~800 entomological samples.
- Validate and implement a dengue and Zika specific multiplex RT-qPCR.

To profile the same entomological dataset, identifying key mutations present in the population, that may confer insecticide resistance and reveal how *Ae. aegypti* colonised Cabo Verde. To achieve this, I will:

- Target genomic loci associated with insecticide resistance and geographic distribution, amplify, and sequence them.
- Analyse the prevalence of resistance associated SNPs in the Cabo Verdean *Ae. aegypti* population.
- Place mitochondrial sequences of Cabo Verdean *Ae. aegypti* in a global context, analysing these data using phylogenetic inferences.

Chapter 3

Employing genomic epidemiology to understand the origins of epidemics, and the transmission that sustains them, enhances both preventative and control measures. These techniques have been applied regularly throughout the 2015-2016 Zika outbreak in South America.

Given the positioning of Cabo Verde archipelago, between the African and South American continents, and the distinct ZIKV lineages circulating within them respectively, understanding the origin of the outbreak on Cabo Verde will make a significant contribution to the understanding of ZIKV transmission. Despite the continued autochthonous transmission of the African Zika lineage across the continental west coast, it is hypothesised that the circulating lineage of Zika in Cabo Verde was of South American origin (Asian lineage).

To understand the geographical origins and the approximate time of introduction of Zika to Cabo Verde. To achieve this, I will:

- Apply target-enrichment techniques to complex RNA isolates to amplify ZIKV genomic fragments for whole-genome assembly.
- Build a temporal and geographic phylogeny to place the genomic data into an epidemic context, revealing the origin of the Cabo Verdean Zika outbreak.

Chapter 4

Sero-epidemiological studies can bolster outbreak control efforts, both during the outbreak and in passive surveillance applications. I believe that in describing the seroprevalence of

arboviruses in human populations across Cabo Verde in a post-epidemic crossectional analysis, the understanding of arbovirus dynamics will be enhanced, building a knowledgebase for application in future outbreaks.

To provide a crossectional description of arbovirus seroconversion in a cohort of Cabo Verdean participants. To achieve this, I will:

- Apply serological assays for ZIKV, DENV, YFV, CHIKV and WNV to obtain data for seroconversion on the Cabo Verdean cohort.
- Model risk-factors and biological correlates of infection.
- Assess and reduce the effects of serological cross reactivity in or analyses.

Chapter 5

Improving serological assays for the detection of specific *Flavivirus* responses is a vital step in the future of disease control efforts. While the SARS-CoV-2 pandemic has hindered furthering these aims in-vitro, the exploration of these issues *in-silico* is possible through the analysis of antigenic protein structures, their traits, and inferred specificities. I hypothesise that it is possible to improve the specificity of the current generation of Zika serological diagnostics through applying novel reverse-diagnostic design techniques.

To capture and understand the determinants of antigen cross-reactivity. To achieve this, I will:

- Analyse *Flavivirus* antigen sequence and three-dimensional structural data.

- Combine different data types employing consensus approach to choosing regions likely to be specific to a given virus.

To express candidate peptides and assess their viability for use in Zika diagnostics. To achieve this, I will:

- Generate constructs for protein expression in *E. coli* expression systems.
- Purify recombinant proteins and validate them for use in both ELISA and Luminex assay formats.

Chapter 6

A workflow through which *in-silico* antigen design processes can be translated into viable assays for *Flavivirus* surveillance is required. Currently, the cost to synthesise such peptides is prohibitive to most applications. I hypothesise that in applying advanced molecular techniques to mitigate the common obstacles to expressing small peptides in *E. coli* expression systems, a novel workflow may lay the foundation for a sustainable and specific surveillance network.

To design a medium-high throughput workflow for peptide antigen expression in *E. coli.* To achieve this, I will:

- Explore cloning techniques for optimising short antigen expression in *E. coli* expression systems.
- Test recombinant peptides as antigens for *Flavivirus* antibody detection.

1.6 References

1. Maurer, F. D. Equine piroplasmosis--another emerging disease. *J. Am. Vet. Med. Assoc.* **141**, 699–702 (1962).
2. Babkin, I. V & Babkina, I. N. The origin of the variola virus. *Viruses* **7**, 1100–1112 (2015).
3. York, A. On the origin of Plasmodium falciparum. *Nat. Rev. Microbiol.* **16**, 393 (2018).
4. Chakrabarti, L. *et al.* Sequence of simian immunodeficiency virus from macaque and its relationship to other human and simian retroviruses. *Nature* **328**, 543–547 (1987).
5. Woolhouse, M. & Gaunt, E. Ecological Origins of Novel Human Pathogens. *Crit. Rev. Microbiol.* **33**, 231–242 (2007).
6. Winkler, W. G., Fashinell, T. R., Leffingwell, L., Howard, P. & Conomy, J. P. Airborne Rabies Transmission in a Laboratory Worker. *JAMA* **226**, 1219–1221 (1973).
7. Houff, S. A. *et al.* Human-to-Human Transmission of Rabies Virus by Corneal Transplant. *N. Engl. J. Med.* **300**, 603–604 (1979).
8. Fekadu, M. *et al.* Possible human-to-human transmission of rabies in Ethiopia. *Ethiop. Med. J.* **34**, 123—127 (1996).
9. Coltart, C. E. M., Lindsey, B., Ghinai, I., Johnson, A. M. & Heymann, D. L. The Ebola outbreak, 2013–2016: old lessons for new epidemics. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160297 (2017).
10. Lytras, S., Xia, W., Hughes, J., Jiang, X. & Robertson, D. L. The animal origin of SARS-CoV-2. *Science (80-.).* **373**, 968–970 (2021).
11. Oaks Jr, S. C., Shope, R. E. & Lederberg, J. Emerging infections: microbial threats to health in the United States. (1992).
12. Wilke, A. B. B. *et al.* Proliferation of Aedes aegypti in urban environments mediated

by the availability of key aquatic habitats. *Sci. Rep.* **10**, 12925 (2020).

13. Faria, N. R. *et al.* Zika virus in the Americas: Early epidemiological and genetic findings. *Science* **352**, 345–349 (2016).
14. Yu, V. L. & Madoff, L. C. ProMED-mail: An Early Warning System for Emerging Diseases. *Clin. Infect. Dis.* **39**, 227–232 (2004).
15. Flahault, A. *et al.* FluNet as a Tool for Global Monitoring of Influenza on the Web. *JAMA* **280**, 1330–1332 (1998).
16. Raftery, P., Hossain, M. & Palmer, J. An innovative and integrated model for global outbreak response and research - a case study of the UK Public Health Rapid Support Team (UK-PHRST). *BMC Public Health* **21**, 1378 (2021).
17. Cameron, M. M. & Ramesh, A. The use of molecular xenomonitoring for surveillance of mosquito-borne diseases. *Philos. Trans. R. Soc. B* **376**, 20190816 (2021).
18. Campos, M. *et al.* Surveillance of Aedes aegypti populations in the city of Praia, Cabo Verde: Zika virus infection, insecticide resistance and genetic diversity. *Parasites and Vectors* **13**, (2020).
19. Lang, A. S., Kelly, A. & Runstadler, J. A. Prevalence and diversity of avian influenza viruses in environmental reservoirs. *J. Gen. Virol.* **89**, 509–519 (2008).
20. Snow, J. *On the mode of communication of cholera*. (John Churchill, 1855).
21. Yang, S., Ning, S. & Kou, S. C. Use Internet search data to accurately track state level influenza epidemics. *Sci. Rep.* **11**, 4023 (2021).
22. Drosten, C. *et al.* Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1967–1976 (2003).
23. Faria, N. R. *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8**, 97 (2016).
24. Arias, A. *et al.* Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* **2**, (2016).
25. Randhawa, N. *et al.* Fine scale infectious disease modeling using satellite-derived data. *Sci. Rep.* **11**, 1–11 (2021).
26. McNamara, L. A. Ebola Surveillance—Guinea, Liberia, and Sierra Leone. *MMWR Suppl.* **65**, (2016).
27. Vitale, M. *et al.* A comparison of passive surveillance and active cluster-based surveillance for dengue fever in southern coastal Ecuador. *BMC Public Health* **20**, 1–10 (2020).
28. Organization, W. H. Global strategy for dengue prevention and control 2012-2020.

(2012).

29. Gale, A. H. A century of changes in the mortality and incidence of the principal infections of childhood. *Arch. Dis. Child.* **20**, 2 (1945).
30. Thompson, W. G. The Clinical Value Of The Widal Test For Enteric Fever. *Br. Med. J.* 1775–1777 (1897).
31. Dick, G. W. ., Kitchen, S. . & Haddow, A. . Zika Virus (I). Isolations and serological specificity. *Trans. R. Soc. Trop. Med. Hyg.* **46**, 509–520 (1952).
32. Coons, A. H., Creech, H. J. & Jones, R. N. Immunological properties of an antibody containing a fluorescent group. *Proc. Soc. Exp. Biol. Med.* **47**, 200–202 (1941).
33. Tardei, G. *et al.* Evaluation of immunoglobulin M (IgM) and IgG enzyme immunoassays in serologic diagnosis of West Nile virus infection. *J. Clin. Microbiol.* **38**, 2232–2239 (2000).
34. Rabe, I. B. *et al.* Interim guidance for interpretation of Zika virus antibody test results. *Morb. Mortal. Wkly. Rep.* **65**, 543–546 (2016).
35. Biggs, J. R. *et al.* A serological framework to investigate acute primary and post-primary dengue cases reporting across the Philippines. *BMC Med.* **18**, 1–14 (2020).
36. Alexander, T. S. Human immunodeficiency virus diagnostic testing: 30 years of evolution. *Clin. Vaccine Immunol.* **23**, 249–253 (2016).
37. W, R. R. & G, B. T. Solid phase assay employing capillary flow. (1989).
38. Peeling, R. W. *et al.* Evaluation of diagnostic tests: dengue. *Nat. Rev. Microbiol.* **8**, S30–S37 (2010).
39. Bell, D. & Peeling, R. W. Evaluation of rapid diagnostic tests: malaria. *Nat. Rev. Microbiol.* **4**, S34–S38 (2006).
40. Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A. & Kozlakidis, Z. Considerations for diagnostic COVID-19 tests. *Nat. Rev. Microbiol.* **19**, 171–183 (2021).
41. Torjesen, I. Covid-19: How the UK is using lateral flow tests in the pandemic. *BMJ* **372**, (2021).
42. Jiang, H. *et al.* SARS-CoV-2 proteome microarray for global profiling of COVID-19 specific IgG and IgM responses. *Nat. Commun.* **11**, 3581 (2020).
43. Kan, Y. W., Golbus, M. S. & Dozy, A. M. Prenatal diagnosis of α-thalassemia: clinical application of molecular hybridization. *N. Engl. J. Med.* **295**, 1165–1167 (1976).
44. Johnson, B. W., Russell, B. J. & Lanciotti, R. S. Serotype-specific detection of dengue viruses in a fourplex real-time reverse transcriptase PCR assay. *J. Clin. Microbiol.* **43**,

4977–4983 (2005).

45. Silva, S. J. R. da, Pardee, K. & Pena, L. Loop-mediated isothermal amplification (LAMP) for the diagnosis of Zika virus: a review. *Viruses* **12**, 19 (2019).
46. WHO. Zika Cumulative Cases. *PAHO -WHO* www.paho.org/hq/index.php?option=com_content&view=article&id=12390&Itemid=42090&lang=en (2018).
47. Muktar, Y., Tamerat, N. & Shewafera, A. Aedes aegypti as a Vector of Flavivirus. *J. Trop. Dis.* **04**, 1–7 (2016).
48. Stanaway, J. D. *et al.* The global burden of dengue: an analysis from the Global Burden of Disease Study 2013. *Lancet. Infect. Dis.* **16**, 712–723 (2016).
49. WHO. Number of Reported Cases of Dengue and Severe Dengue (SD) in the Americas, by Country. *PAHO WHO* http://www.who.int/mediacentre/factsheets/2017-cha-dengue-cases-mar-27-ew-11.pdf (2015).
50. Cauchemez, S. *et al.* Association between Zika virus and microcephaly in French Polynesia, 2013–15: a retrospective study. *Lancet* **387**, 2125–2132 (2016).
51. Jaenisch, T. *et al.* Risk of microcephaly after Zika virus infection in Brazil, 2015 to 2016. *Bull. World Health Organ.* **95**, 191–198 (2017).
52. Rasmussen, S. A., Jamieson, D. J., Honein, M. A. & Petersen, L. R. Zika Virus and Birth Defects — Reviewing the Evidence for Causality. *N. Engl. J. Med.* **374**, 1981–1987 (2016).
53. dos Santos, T. *et al.* Zika Virus and the Guillain–Barré Syndrome — Case Series from Seven Countries. *N. Engl. J. Med.* **375**, 1598–1601 (2016).
54. Cao-Lormeau, V.-M. *et al.* Guillain-Barré Syndrome outbreak associated with Zika virus infection in French Polynesia: a case-control study. *Lancet* **387**, 1531–1539 (2016).
55. The Telegraph. Rio Olympics: which athletes have withdrawn over Zika fears? https://www.telegraph.co.uk/sport/0/rio-olympics-which-athletes-have-withdrawn-over-zika-fears/ (2016).
56. WHO | Zika situation report. *WHO* (2017).
57. Lowe, R. *et al.* The Zika Virus Epidemic in Brazil: From Discovery to Future Implications. *Int. J. Environ. Res. Public Health* **15**, (2018).
58. Gourinat, A.-C., O'Connor, O., Calvez, E., Goarant, C. & Dupont-Rouzeyrol, M. Detection of Zika virus in urine. *Emerg. Infect. Dis.* **21**, 84–6 (2015).
59. Atkinson, B. *et al.* Detection of Zika Virus in Semen. *Emerg. Infect. Dis.* **22**, 940

(2016).

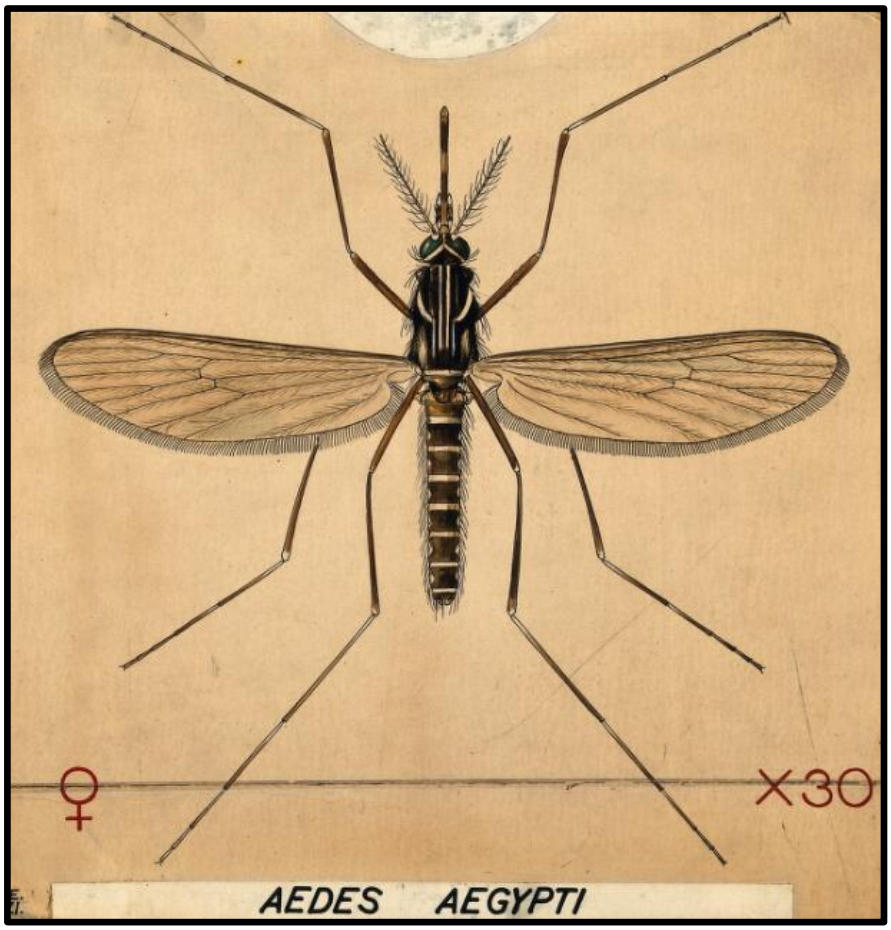
60. Keasey, S. L. *et al.* Antibody Responses to Zika Virus Infections in Environments of Flavivirus Endemicity. *Clin. Vaccine Immunol.* **24**, e00036-17 (2017).
61. Barzon, L. *et al.* Virus and Antibody Dynamics in Travelers With Acute Zika Virus Infection. *Clin. Infect. Dis.* **66**, 1173–1180 (2018).
62. Granger, D. *et al.* Serologic Testing for Zika Virus: Comparison of Three Zika Virus IgM-Screening Enzyme-Linked Immunosorbent Assays and Initial Laboratory Experiences. *J. Clin. Microbiol.* **55**, 2127–2136 (2017).
63. Huzly, D., Hanselmann, I., Schmidt-Chanasit, J. & Panning, M. High specificity of a novel Zika virus ELISA in European patients after exposure to different flaviviruses. *Eurosurveillance* **21**, 30203 (2016).
64. Steinhagen, K. *et al.* Serodiagnosis of Zika virus (ZIKV) infections by a novel NS1-based ELISA devoid of cross-reactivity with dengue virus antibodies: a multicohort study of assay performance, 2015 to 2016. *Euro Surveill.* **21**, (2016).
65. Kadkhoda, K., Gretchen, A. & Racano, A. Evaluation of a commercially available Zika virus IgM ELISA: specificity in focus. *Diagn. Microbiol. Infect. Dis.* **88**, 233–235 (2017).
66. Tsai, W.-Y. *et al.* Distinguishing Secondary Dengue Virus Infection From Zika Virus Infection With Previous Dengue by a Combination of 3 Simple Serological Tests. *Clin. Infect. Dis.* **65**, 1829–1836 (2017).
67. Basile, A. J. *et al.* Multiplex Microsphere Immunoassays for the Detection of IgM and IgG to Arboviral Diseases. *PLoS One* **8**, e75670 (2013).
68. Bosch, I. *et al.* Rapid antigen tests for dengue virus serotypes and Zika virus in patient serum. *Sci. Transl. Med.* **9**, eaan1589 (2017).
69. Balmaseda, A. *et al.* Antibody-based assay discriminates Zika virus infection from other flaviviruses. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8384–8389 (2017).
70. Rönnberg, B. *et al.* Compensating for cross-reactions using avidity and computation in a suspension multiplex immunoassay for serotyping of Zika versus other flavivirus infections. *Med. Microbiol. Immunol.* **206**, 383–401 (2017).
71. Edman, J. D., Strickman, D., Kittayapong, P. & Scott, T. W. Female Aedes aegypti (Diptera: Culicidae) in Thailand rarely feed on sugar. *J. Med. Entomol.* **29**, 1035–8 (1992).
72. Scott, T. W. *et al.* Detection of Multiple Blood Feeding in Aedes aegypti (Diptera: Culicidae) During a Single Gonotrophic Cycle Using a Histologic Technique. *J. Med.*

Entomol. **30**, 94–99 (1993).

73. Scott, T. W. & Takken, W. Feeding strategies of anthropophilic mosquitoes result in increased risk of pathogen transmission. *Trends Parasitol.* **28**, 114–121 (2012).
74. Organization, W. H. *Vector control operations framework for Zika virus*. (2016).
75. Weill, M. *et al.* The unique mutation in ace-1 giving high insecticide resistance is easily detectable in mosquito vectors. *Insect Mol. Biol.* **13**, 1–7 (2004).
76. Davies, T. G. E., Field, L. M., Usherwood, P. N. R. & Williamson, M. S. A comparative study of voltage-gated sodium channels in the Insecta: implications for pyrethroid resistance in Anopheline and other Neopteran species. *Insect Mol. Biol.* **16**, 361–375 (2007).
77. Douris, V. *et al.* Resistance mutation conserved between insects and mites unravels the benzoylurea insecticide mode of action on chitin biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14692–14697 (2016).
78. Dia, I. *et al.* Insecticide susceptibility of Aedes aegypti populations from Senegal and Cabo Verde Archipelago. *Parasit. Vectors* **5**, 238 (2012).
79. Lourenço, J. *et al.* Epidemiology of the Zika Virus Outbreak in the Cabo Verde Islands, West Africa. *PLoS Curr.* (2018) doi:10.1371/currents.outbreaks.19433b1e4d007451c691f138e1e67e8c.
80. Kindhauser, M. K., Allen, T., Frank, V., Santhana, R. S. & Dye, C. Zika: the origin and spread of a mosquito-borne virus. *Bull. World Health Organ.* **94**, 675-686C (2016).
81. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).
82. Alberts, B. *et al.* The generation of antibody diversity. in *Molecular Biology of the Cell. 4th edition* (Garland Science, 2002).
83. Leta, S. *et al.* Global risk mapping for major diseases transmitted by Aedes aegypti and Aedes albopictus. *Int. J. Infect. Dis.* **67**, 25–35 (2018).

Chapter Two

Surveillance of *Aedes aegypti* populations in the city of Praia, Cabo Verde



60% of the world's population will be at risk of Dengue infection by 2080 (Messina, J. P. er al. 2019). *Ae. aegypti* (Yellow Fever mosquito) drawing with pen and ink (A.J.E. Terzi 1872). *Ae. aegypti* is a vector for Zika, Dengue, Yellow Fever and Chikungunya.

2.1 RESEARCH PAPER COVER SHEET

SECTION A – Student Details

Student ID Number	**1603403**	**Title**	Mr
First Name(s)	Daniel		
Surname/Family Name	Ward		
Thesis Title	Zika virus surveillance in human and mosquito populations in Cabo Verde – exploring molecular and serological tools for the surveillance of emerging infectious diseases		
Primary Supervisor	Prof. Taane Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	**Parasites & Vectors**		
When was the work published?	**September 2020**		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	**N/A**
Please list the paper's authors in the intended authorship order:	N/A
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	**DW: Implementation of lab work; entomological sample preparation, DNA extraction, molecular analyses including phylogenetic inferences. DW co-wrote the manuscript with MC and RFM.**

SECTION E

Student Signature [redacted]

Date 20/04/2022

Supervisor Signature [redacted]

Date April 20, 2022

Campos *et al. Parasites Vectors* (2020) 13:481
https://doi.org/10.1186/s13071-020-04356-z

Parasites & Vectors

RESEARCH Open Access

Surveillance of *Aedes aegypti* populations in the city of Praia, Cape Verde: Zika virus infection, insecticide resistance and genetic diversity

Monica Campos[1†], Daniel Ward[1†], Raika Francesca Morales[1†], Ana Rita Gomes[2], Keily Silva[3], Nuno Sepúlveda[1,4], Lara Ferrero Gomez[3], Taane G. Clark[1,5] and Susana Campino[1*]

Abstract

Background: *Aedes* spp. are responsible for the transmission of many arboviruses, which contribute to rising human morbidity and mortality worldwide. The *Aedes aegypti* mosquito is a main vector for chikungunya, dengue and yellow fever infections, whose incidence have been increasing and distribution expanding. This vector has also driven the emergence of the Zika virus (ZIKV), first reported in Africa which spread rapidly to Asia and more recently across the Americas. During the outbreak in the Americas, Cape Verde became the first African country declaring a Zika epidemic, with confirmed cases of microcephaly. Here we investigate the prevalence of ZIKV and dengue (DENV) infected *Ae. aegypti* mosquitoes in the weeks following the outbreak in Cape Verde, and the presence of insecticide resistance in the circulating vector population. Genetic diversity in the mosquito population was also analysed.

Methods: From August to October 2016, 816 *Ae. aegypti* mosquitoes were collected in several locations across Praia, Cape Verde, the major hot spot of reported ZIKV cases in the country. All mosquitoes were screened by reverse transcription PCR for ZIKV and DENV, and a subset ($n = 220$) were screened for knockdown insecticide resistance associated mutations in the voltage gated sodium channel (*VGSC*) gene by capillary sequencing. The mitochondrial *NADH dehydrogenase subunit 4* (*nad4*) gene was sequenced in 100 mosquitoes. These data were compared to 977 global sequences in a haplotype network and a phylogenetic tree analysis.

Results: Two *Ae. aegypti* mosquitoes were ZIKV positive (0.25%). There were no SNP mutations found in the *VGSC* gene associated with insecticide resistance. Analysis of the *nad4* gene revealed 11 haplotypes in the Cape Verdean samples, with 5 being singletons. Seven haplotypes were exclusive to Cape Verde. Several of the remaining haplotypes were frequent in the global dataset, being present in several countries (including Cape Verde) across five different continents. The most common haplotype in Cape Verde (50.6 %) was also found in Africa and South America.

Conclusions: There was low-level Zika virus circulation in mosquitoes from Praia shortly after the outbreak. The *Ae. aegypti* population did not appear to have the *kdr* mutations associated with pyrethroid resistance. Furthermore,

*Correspondence: Susana.campino@lshtm.ac.uk
†Monica Campos, Daniel Ward and Raika Francesca Morales contributed equally to this work
†Lara Ferrero Gomez, Taane G. Clark and Susana Campino are joint senior authors
[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK
Full list of author information is available at the end of the article

haplotype and phylogenetic analyses revealed that Cape Verde *Ae. aegypti* mosquitoes are most closely related to those from other countries in Africa and South America.
Keywords: *Aedes aegypti*, Zika, Cape Verde, *kdr*, *nad4*

Background

In recent decades, there has been a rise in the emergence and re-emergence of epidemic arboviral diseases, including those caused by yellow fever (YFV), dengue (DENV), chikungunya (CHIKV) and Zika (ZIKV) viruses [1–3]. More than 2.5 billion people in over 100 countries are at risk of contracting dengue [4], and the Asian strain of the ZIKV has spread throughout the Americas [5–7]. CHIKV has also reached the Americas and has undergone recent rapid spread [8, 9]. Outbreaks of yellow fever in unvaccinated individuals have been reported in the Americas and in Africa and there is a risk it is imported into Asia [10]. The *Aedes aegypti* mosquito is the main vector responsible for the transmission of DENV, CHIKV, YFV and ZIKV worldwide due to its highly anthropophilic behaviour and close proximity with the human environment [11]. This mosquito species thrives in both fresh and stagnated water, in vessels ranging from barrels to bottle caps, and it colonises human residencies. An anthropophilic preference to feeding on human blood and a tendency to feed multiple times during an egg-laying cycle, imparts this particular vector with a remarkable efficiency in pathogen transmission [12–14]. *Aedes aegypti* mosquitoes are currently found in 188 countries and territories, putting an estimated 3 billion people at risk of the aforementioned and future-emerging arboviral diseases [15].

Cape Verde is an archipelago in West Africa comprising of ten volcanic islands located 550 km west from Senegal. *Aedes aegypti* mosquitoes were first detected in 1931 on the island of São Vicente and subsequently spread to the other islands [16, 17]. There are no records of other species of *Aedes* vector, such as *Aedes albopictus* in the region, nevertheless the presence of the non-vector species such as *Aedes caspius* has been recorded [17]. Mitochondrial DNA sequencing analysis of *Ae. aegypti* mosquitoes from Africa indicated a possible West African origin of the Cape Verdean population [18]. With heavy human and goods trans-Atlantic traffic coming in and out of São Vicente, particularly during the 16th to 17th centuries, it is possible that the *Ae. aegypti* population from Cape Verde was imported from regions of the West African coast and also contributed to the New World population. Mitochondrial DNA sequencing analysis among mosquito populations from different geographical locations is a well-described method of determining mosquito ancestry, as well as for the analysis of genetic diversity [19–21]. With the increase in international travel, there is a latent threat that new strains of arboviruses or new vectors are introduced worldwide, and it is important to study the population genetic diversity of *Ae. aegypti*. Cape Verde continues to be a strategic trans-Atlantic route linking particularly West Africa countries with Europe and the Americas [22].

The first arboviral outbreak in Cape Verde occurred in 2009, most likely originating from neighbouring countries in West Africa [23]. This outbreak resulted in 21,137 reported DENV suspected cases and four registered deaths. Eight out of the nine inhabited islands in the archipelago were affected, with Santiago Island, where the city capital of Praia is located, reporting the greatest number of cases [21]. *Aedes aegypti* was identified as the vector, where DENV-3 and DENV-4 virus strains were detected in mosquitoes from Cape Verde and Senegal [23, 24]. A second arboviral outbreak in Cape Verde, caused by ZIKV, occurred from October 2015 to July 2016, shortly following its establishment in Brazil, where the first congenital ZIKV microcephaly cases were reported. The Cape Verdean Ministry of Health officially declared the Zika virus epidemic on 2 November 2015, becoming the first African country to register an epidemic for this virus. There were 7589 suspected cases of ZIKV infection and 18 microcephaly cases officially recorded in Cape Verde, making it the first African country to report ZIKV-associated microcephaly cases [25]. It is possible that the outbreak was caused by importation of strains circulating in the Americas, given the comparable clinical consequences, timing and traveling from those regions [25].

Following the DENV and ZIKV outbreaks in Cape Verde, vector control measures were reinforced which included public education efforts and use of the insecticides temephos (organophosphate) and deltamethrin (pyrethroid), targeting larva and adult mosquitoes, respectively, as well as adulterated diesel and the mosquito fish *Gambusia* sp. for larval control [26].

Pyrethroids, the primary choice of control against adult *Ae. aegypti*, target the voltage gated sodium channel (VGSC) [27]. Knockdown resistance (*kdr*) occurs when an amino acid substitution on the *VGSC* gene reduces the binding affinity of the pyrethroids. The evolutionary selection of mutations in the *VGSC* gene that

confer pyrethroid resistance have been described [27]. F1534C is the most common *kdr* mutation, detected in Asia, Africa and the Americas [28]. Several other mutations have been reported, including the V1016G mutation found in Asia [29], and V1016I particularly detected in Latin America and Africa [30]. Co-mutation confers higher levels of resistance, such as the triple mutations of F1534C/V1016I/ S989P that confer extreme resistance, as detected in Myanmar [31]. These mutations have not been detected in *Ae. aegypti* collected in Cape Verde between 2007 and 2014 [18, 32, 33]. However, insecticide susceptibility assays have revealed resistance to dichlorodiphenyltrichloroethane (DDT), the carbamate propoxur, pyrethroids (deltamethrin, cypermethrin) and to the organophosphate temephos [32, 33].

With ongoing issues surrounding the deployment of certain *Flavivirus* vaccines and ineffective antiviral treatment [34], the prevention and control of ZIKV and DENV diseases rely on ongoing surveillance of arboviruses in mosquito vectors and vector control. While insecticides are still the primary measure used for vector control, it is crucial to identify acquired insecticide resistance at an early stage for control efforts to remain effective. Here we aimed to perform the surveillance in *Ae. aegypti* mosquitoes collected in Praia, Cape Verde, just after the ZIKV outbreak in 2015, investigating the presence of ZIKV and DENV infections and screening for *kdr* mutations. In addition, we also assessed the local genetic diversity, phylogeny and ancestry of the *Ae. aegypti* mosquito population by analysing mitochondrial sequences of the *nad*4 gene and comparing them with a global dataset.

Methods

Field-collected mosquitoes

A total of 816 *Ae. aegypti* adult mosquitoes were collected across 27 locations in the capital city of Praia located in Santiago Island, Cape Verde, for seven weeks from 17 August to 5 October 2016 (Fig. 1). The locations were centered in two areas: Tira Chapeu ($n=309$ mosquitoes, within 500 m of GPS coordinates 14° 55.207′ N, 23° 32.323′ W) and Plateau ($n=445$ mosquitoes, within 500 m of GPS coordinates 14° 55.371′ N, 23° 30.334′ W). These two locations are approximately 2.5 km apart. For 62 mosquitoes we could not assign location. Ten BG-Sentinel-2 traps with lure odour bait were placed once a week for 24 h at the selected locations. The collected samples were taken to the University of Jean Piaget laboratory for identification to the species level, and determination of sex and gonadotropic status, using a stereomicroscope and the taxonomic key of mosquitoes in Cabo Verde [16]. After each specimen was identified, all the *Ae. aegypti* mosquitoes collected were individually immersed in 300 µl of RNAlater Stabilization Solution (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA) and stored at −20 °C until they were transported to the London School of Hygiene and Tropical Medicine.

Extraction of DNA and RNA and virus detection

Individual mosquitoes in 1.5 ml microcentrifuge tubes were centrifuged for 2 min at 1300× *rpm* and supernatant was removed. Samples were then washed two times with 1 ml phosphate saline buffer (PBS) and resuspended in 300 ml of RLT buffer containing β-mercaptoethanol (10 µl/ml). Cells were disrupted using Tissue Ruptor II (Qiagen, Hilden, Germany) at speed 3 for 60 s. DNA and RNA of each single mosquito were extracted using Qiagen AllPrep DNA/RNA 96 Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. The quantity of DNA in each sample was measured using the Qubit 2.0 fluorimeter HS DNA Kit (Thermo Fisher Scientific, Waltham, MA, USA). The DNA quantity for each sample was variable with < 1 µg obtained per sample. cDNA was synthesized and amplified from RNA using QuantiNova Probe RT-PCR Kit (Qiagen) following the manufacturer's protocol. Individual mosquitoes were screened using primers and probes to detect the presence of DENV (DENV1-3 and DENV-4) and ZIKV (NS5) (Additional file 1: Table S1). Primers/probes for ZIKV were modified from Grubaugh et al. [7] to take into account the genetic diversity in Asian and American ZIKV samples using data from the NCBI. Each probe contained a fluorescent reporter dye 6-carboxyfluorescein (FAM) at the 5'-end and the Black Hole Quencher 1 at the 3'-end.

Mosquito *kdr* and *nad*4 sequencing

The primers used to amplify exons 21 and 31 of the *VGSC* gene, which encode for domain II subunit 5 and domain III subunit 6, respectively [31], as well as the mitochondrial *nad*4 gene [35] are described in Additional file 1: Table S2. Polymerase chain reaction (PCR) was carried out in a total volume of 25 µl using standard protocols with 1× reaction buffer, 200 µM dNTP, 0.5 µM forward primer, 0.5 µM reverse primer, 0.02 U/µl DNA polymerase, 17.5 µl water and 1.5 µl DNA sample (1–5 ng). The PCR thermocycler profile consisted of: 1 cycle at 98 °C for 30 s, 30 cycles at 98 °C for 10 s, 56 °C (*VGSC* gene) or 59 °C (*nad*4 gene) for 30 s, 72 °C for 30 s and a final cycle at 72 °C for 2 min. PCR products (10 µl each) were detected by 1% agarose gel electrophoresis in TAE buffer, stained using GelRed (Cambridge Bioscience, Cambridge, UK). The samples were sequenced by capillary sequencing. The resulting DNA sequences have been submitted to GenBank under the accession numbers MT721877-MT721961.

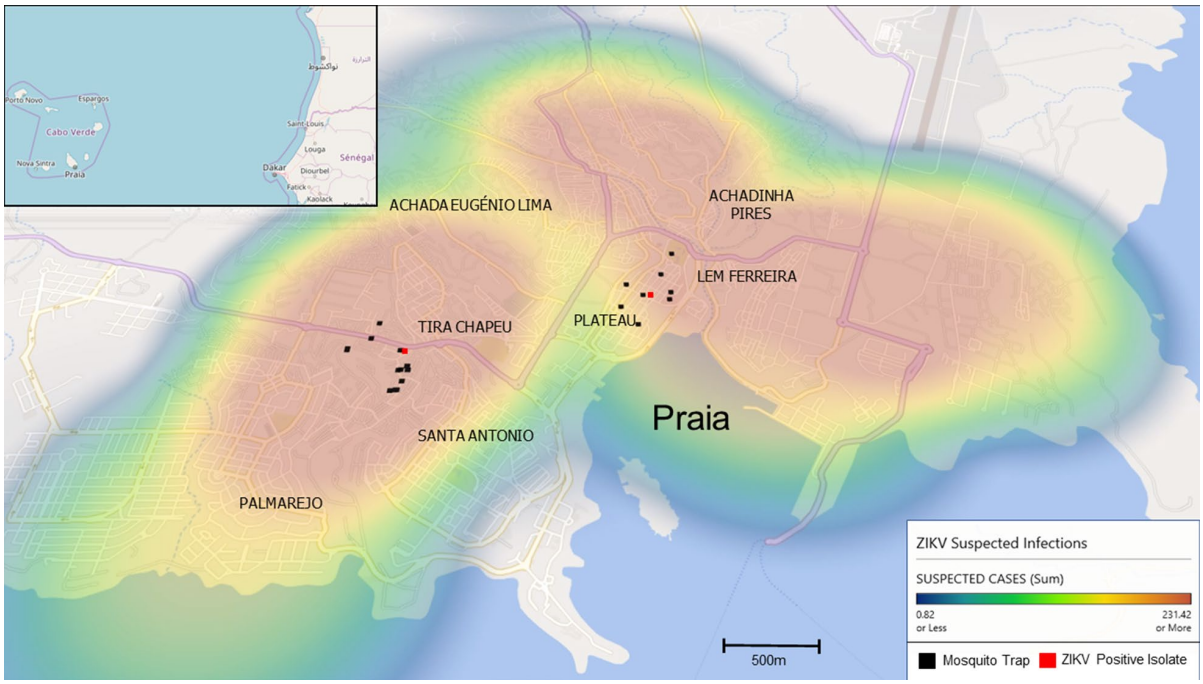


Fig. 1 Map showing mosquitoes sampling sites in Praia city, Santiago Island. Sample sites are indicated by black points. Sampling sites with ZIKV positive mosquitoes are indicated in red. Heat map indicates the number of ZIKV suspected infections in a given area

Mosquito haplotype and phylogenetic analysis

Sequences were trimmed and edited using Geneious (version 11.0) [36]. BLASTn was used to confirm the taxonomic identification. Our *nad*4 sequences were added to a dataset of available *Ae. aegypti nad*4 sequences ($n=1101$) downloaded from GenBank. Sequences with > 25% missing data and mislabelled as other *Aedes* species were excluded. The sequences were aligned using MAFFT (v7.450). The multiple sequence alignment (MSA) was visualized and trimmed using AliView (v1.23). The trees in Additional file 2: Figure S1 and Additional file 3: Figure S2 were inferred using IQTREE (v1.6.12) with automatic selection of the best-fit. For sequences which had collection date information available, TempEst (v1.5.1) [37] was used to investigate the temporal signal and 'clocklikeness' of our molecular phylogeny. To investigate the genetic diversity of the *Ae. aegypti* mosquitoes, a haplotype network was constructed using the R-package *PEGAS* [38]. Here, *nad*4 sequences from Cape Verde [present study: Tira Chapeu ($n=23$), Plateau ($n=62$); other study collected during 2007 and 2010 in City of Praia ($n=7$) [18] and other countries [NCBI GenBank, total ($n=977$); Asia ($n=607$), South America ($n=296$); North America ($n=17$); Africa ($n=56$); and Europe ($n=1$)] were analysed. Sequences with > 25% missing sequence data were omitted. A neighbour-joining clustering method was chosen to infer the phylogeny [39]. The evolutionary distances were computed using the number of differences method. Branch lengths were inferred using the same units as those of the evolutionary distances. MEGA X (v10.1.6) [40] and FigTree (v1.4.2) software tools were used to manipulate the resulting tree. Haplotype diversity (h), nucleotide diversity (π) and the neutral mutation Tajima's D were calculated using the R-package *PEGAS*.

Results

Prevalence of arbovirus infections

All mosquito samples were DENV negative and two mosquitoes were positive for ZIKV infection (0.25%). The two ZIKV positive *Ae. aegypti* mosquitoes were collected from different locations at different times: (i) 24 August 2016 at Plateau (14° 55.229′ N, 23° 30.500′ W); and (ii) 5 October 2016 at Tira Chapéu (14° 54.327′ N, 23° 31.285′ W) (Fig. 1).

Knockdown resistance (*kdr*) mutations

Two hundred out of 816 *Ae. aeygpti* mosquitoes were selected randomly and screened for *kdr* mutations in exon 21 to check for the V1016G/I and S989P mutations and were also screened for exon 31, to investigate

the presence of the F1534C mutation. Sequences were edited, trimmed and poor-quality sequences were excluded from the dataset, resulting in a total of 124 sequences from exon 21 and 133 sequences from exon 31. Nucleotide sequences were compared with 200 or 245 other sequences available in GenBank for exons 21 and 31, respectively. The V1016G/I, S989P and F1534C mutations associated with insecticide resistance were not observed. Furthermore, no other single nucleotide polymorphisms (SNP) were detected.

Population genetics and phylogenetic analysis of the *nad4* gene

The alignment of 85 high quality mitochondrial *nad4* gene sequences (291 bp length) from Cape Verde sourced *Ae. aegypti* mosquitoes revealed the presence of 20 SNPs and 11 haplotypes (5 singletons, i.e. haplotypes found in only one sample each; Table 1). A combined analysis of the Cape Verdean and other countries sequences ($n=977$) revealed a total of 182 haplotypes, of which 144 were singletons, with a nucleotide diversity (π) of 0.0048 and a haplotype diversity (h) of 0.71, similar to previous reports in Cape Verde ($\pi=0.002$ and $h=0.609$) [41]. The Tajima's D value was negative (-1.91), but not statistically significant (P-value = 0.056).

African samples ($n=56$) had the highest number of singletons (55.8% of the haplotypes) and the most frequent haplotype (XI, 20.9%) was also present in the other global populations (Fig. 2, Table 1). Asia ($n=607$) had the lowest frequency of singletons (3.5%) and two main haplotypes. The most frequent haplotype in Asia (XLVI, 42.5%) was also the most frequent in South America (XLVI, $n=296$, 24.0%) and was present in Africa (XLVI, 4.7%) (Fig. 2, Table 1). The most frequent haplotype in Cape Verde (IX, 50.6%) was also found in Africa and South America (Fig. 2, Table 2). Haplotype XI, present in all populations and the most frequent in Africa, was the second most common (17.3%) in Cape Verde, followed by haplotype VIII (14.8%), which is unique to the country. Other *nad4* sequences from Cape Verde mosquitoes ($n=7$) collected during 2007 and 2010 [41] were also included in the analysis and show that these samples share 5 haplotypes with our Cape Verdean samples. The three most frequent haplotypes are the same for both collections and 6 haplotypes were unique from Cape Verde (Table 2).

A median-joining network, excluding singletons and low frequent haplotypes (< 1%), shows a core haplotype (XI) that includes samples from across all continents and connects with clusters from predominantly Asia and South America (Additional file 4: Figure S3). Using the same dataset, but including all sequences, we constructed a phylogenetic tree (total branch length of 325.06; Fig. 3), which showed that Cape Verde *Ae. aegypti* samples are more related to samples from Africa and South America. Across our 85 Cape Verdean sequences (and 7 previously published), we observed 5 distinct clusters (denoted I to V) with high inter-cluster diversity. Cluster I included 2 Cape Verdean samples and consisted primarily of South American isolates paired with a single Asian isolate. Cluster II included 22 Cape Verdean samples, including 1 pre-existing local isolate and all grouping with sequences from the Americas. Cluster III contained 47% of the Cape Verdean sequences and also included 2 previously described local sequences, 2 other African isolates and 15 South American, all of which shared 97% identity. Cluster IV included 18 Cape Verdean samples, including 2 other publicly available samples. This cluster resided within a predominantly African clade with 75% of the isolates originating from that continent. Cluster V contained 7 Cape Verdean isolates, all of which originate from our study. The rest of the clade consisted of Asian ($n=7$), African ($n=3$) and South American ($n=7$) samples. There was strong clustering of our dataset with the other Cape Verdean previously published sequences, with several having 100% identity. Clusters II and V did not contain any samples from Cape Verde previously reported and therefore are novel undescribed Cape Verdean sequences. No correlation was observed between the Cape Verdean haplotype frequency and sampling site (Additional file 2: Figure S1).

Table 1 Mitochondrial *nad4* global haplotype frequency in *Aedes aegypti* mosquitoes including the top five most frequent haplotypes per region

Haplotype/Region	Haplotype frequency (%)				
	Africa	Cape Verde (2016)[a]	North America	South America	Asia
XI	20.9	17.3	17.6	1.0	0.3
V	7.0	8.6	0.0	9.8	1.2
XXII	7.0	0.0	5.9	0.0	0.0
XLVI	4.7	0.0	0.0	24.0	42.5
IX	4.7	50.6	0.0	7.1	0.0
XXIX	0.0	0.0	17.6	5.7	0.0
XXXIX	0.0	0.0	5.9	4.7	0.2
XXXIV	0.0	0.0	5.9	0.7	26.9
VI	0.0	1.2	17.6	0.0	0.0
XLVII	0.0	0.0	0.0	0.0	12.4
LXIII	0.0	0.0	0.0	0.0	4.3

[a] Samples collected in this study

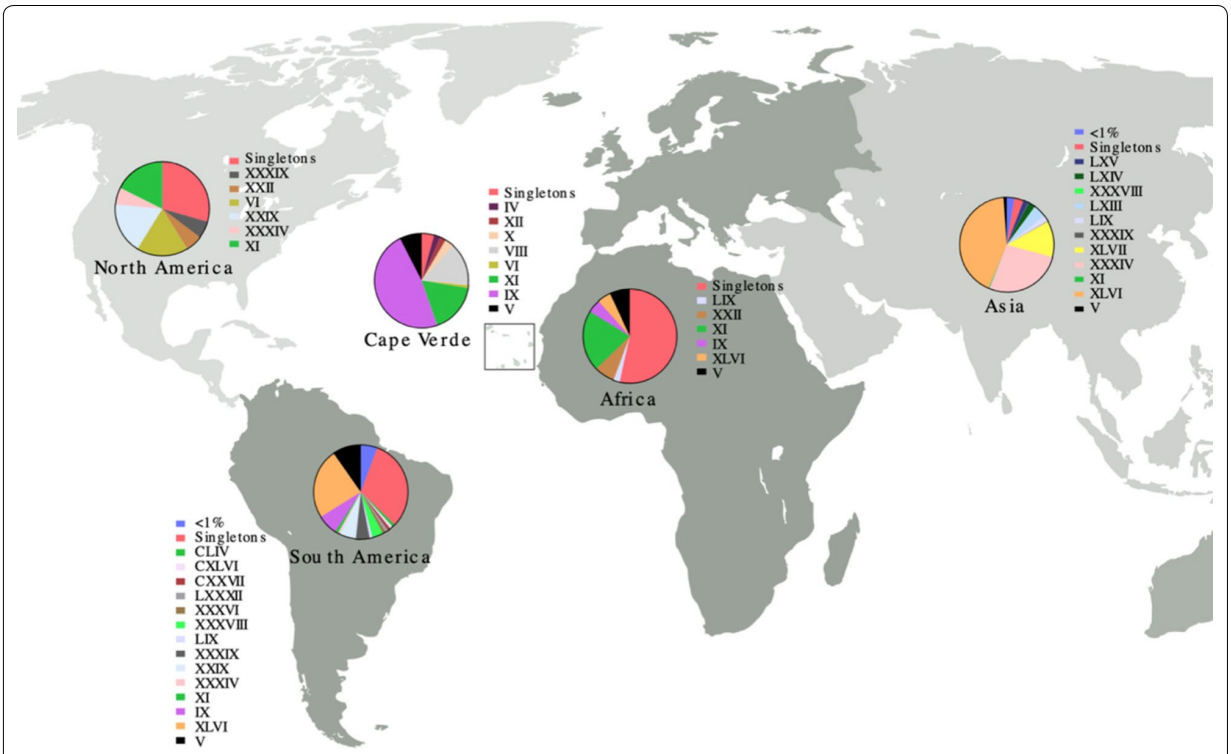


Fig. 2 Global map showing the distribution of the *Aedes aegypti* mitochondrial *nad*4 haplotypes. The colours represent the different haplotypes while the size of the pie chart represents the sample size

Table 2 Mitochondrial *nad*4 haplotype frequency in *Aedes aegypti* mosquitoes from Cape Verde including all haplotypes identified in Cape Verde

Haplotype/Region	Haplotype frequency (%)					
	Cape Verde (2016)[a]	Cape Verde (2007–2010)[b]	Africa	South America	North America	Asia
IX	50.6	28.6	4.7	7.1	0.0	0.0
XI	17.3	14.3	20.9	1.0	17.6	0.3
VIII	14.8	14.3	0.0	0.0	0.0	0.0
V	8.6	0.0	7.0	9.8	0.0	1.2
I	1.2	0.0	0.0	0.0	0.0	0.0
II	1.2	0.0	0.0	0.0	0.0	0.0
III	1.2	0.0	0.0	0.0	0.0	0.0
IV	1.2	14.3	0.0	0.0	0.0	0.0
VI	1.2	0.0	0.0	0.0	17.6	0.0
VII	1.2	0.0	0.0	0.0	0.0	0.0
X	1.2	14.3	0.0	0.0	0.0	0.0
XII	0.0	14.3	0.0	0.0	0.0	0.0

[a] Samples collected in this study

[b] Samples collected in a previous study [18]

The phylogenetic tree also showed across all samples two major distinct clades, indicating the presence of two separate lineages (Additional file 3: Figure S2) and supporting a previous report [41]. These clades contained isolates from all continents and showed no specific geospatial signal. Using TempEst software, there was weak evidence of a temporal signal across the *nad*4 sequence dataset (R^2 value of 0.015), not sufficient for the calibration of an accurate molecular clock. However, this analysis would be improved with greater depth of sampling and availability of date of collection, particularly from Africa, where the sequences only represented 4.4% of the total dataset.

Discussion

The ZIKV outbreak in the Cape Verde archipelago began around October 2015 and lasted until June 2016 with cases mainly reported in in Santiago Island, particularly in Praia, the capital and largest city. We performed an entomological surveillance study to detect the circulation of arboviruses, as this represents a valuable tool to assist in the prevention of further outbreaks, especially in areas with the possible co-circulation of different arboviruses. In our study, we screened 816 *Ae. aegypti* mosquitoes collected in Praia post-outbreak and found a low number of ZIKV-positive (2 in 816) and no DENV-positive mosquitoes. Low levels of ZIKV circulating among

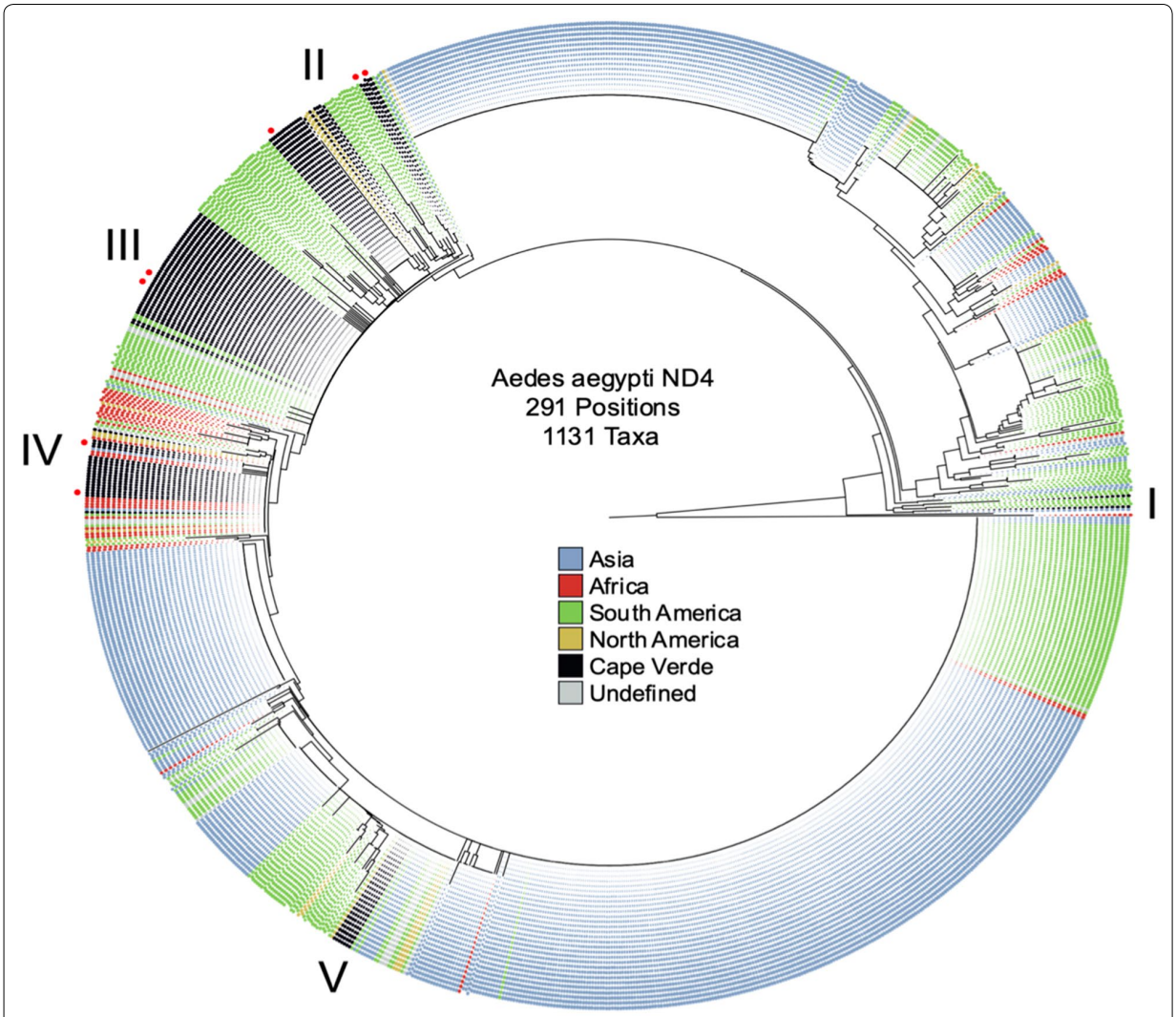


Fig. 3 *Aedes aegypti* phylogenetic tree based on the mitochondrial *nad*4 gene. A total of 291 positions were used from 1131 sequences. Clustering was performed using the neighbour joining method. Leaf colour indicates the continental origin. Grey colour indicates undefined geographical origin. Red markers denote Cape Verdean sequences from Salgueiro et al. [20] (CV1-7). Numerals I-V indicate Cape Verdean clusters

mosquito populations post-outbreak have been reported in other countries. For example, post-outbreak ZIKV was detected in pools of *Ae. aegypti* in southern Mexico (15/55 pools of 472 female mosquitoes) and Rio de Janeiro (3/198 pools of 315 female and 235 male mosquitoes) [42, 43]. In Singapore, *Ae. aegypti* mosquitoes were caught during an ongoing outbreak and the number of ZIKV detected in mosquitoes was low (9/1051) [44]. The low level of ZIKV prevalence in *Ae. aegypti* may be due to various factors: the application of strong vector control during and after the outbreak, the timing of sample collection at the end of the outbreak, differences in vector competence of the *Ae. aegypti* from Cape Verde to transmit the ZIKV strain [45], or to a low detection sensitivity of ZIKV due to degradation of viral RNA.

As the use of insecticides represents currently the primary choice for vector control, it is essential to detect insecticide resistance at an early stage to inform control programmes of the most effective measures. This should be done through active surveillance programmes that monitor insecticide resistance through using bioassays or by detecting associated mutations. Knockdown resistance associated with pyrethroid resistance occurs with certain mutations in the *VGSC* gene. Several *kdr* resistance mutations have been identified, with F1534C being the most widely reported [30], but others are geographically distributed, including V1016G mutation in Asia [29], V1016I in Africa and the Americas, [30, 46], and I1011M is found in the Americas [31]. Other mutations have been detected but only a few have been functionally confirmed to confer resistance, including F1534C, V1016G, I1011M and also S989P and the recent V410L [28, 47–49]. In our study, we found that the mosquito samples screened in Cape Verde did not show any of the previously published insecticide resistance mutations such as V1016G/I, S989P and F1534C. No other non-synonymous mutations were detected in these regions of the gene. Two other studies conducted in *Ae. aegypti* from Cape Verde collected between 2007 and 2014 also did not detect any of these mutations [18, 32]. A recent investigation performed in *Anopheles arabiensis* populations in Praia city, revealed the presence of the *kdr* mutation L1014S at a frequency of 7%, and it was suggested that pyrethroid resistance may arise, sweep through the mosquito population, and affect the process of malaria elimination [50]. Although no *kdr* mutations have been detected in the *Ae. aegypti* population, it is possible that the mosquitoes may have acquired metabolic resistance, frequently associated with the overexpression of enzymes responsible for the insecticide detoxification. The absence of co-relation between insecticide resistance and *kdr* allele frequency has been reported in *Anopheles* and *Aedes* mosquitoes suggesting metabolic mechanisms may also contribute to the resistant phenotype [51, 52]. In Cape Verde, bioassay results have previously shown that mosquitoes collected in 2009 were resistant to DDT while mosquitoes collected in 2012 were already resistant to deltamethrin (pyrethroid), cypermethrin (pyrethroid) and also temephos (organophosphate), but susceptible to malathion (organophosphate) [32, 33] The use of temephos and deltamethrin was reinforced in Cape Verde after the DENV (2009) and ZIKV (2016) outbreaks. Investigations of the changes in the frequency of *kdr* mutations in response to insecticide treatments have shown annual increases in mosquito populations across geographical locations. A study of *An. gambiae* resistance in Burkina Faso detected a significant increase in *kdr* mutation frequency between 2008 and 2010 [53]. Likewise, a survey of *Ae. aegypti* in Venezuela during 2008, 2010 and 2012 reported that the I1016 allele frequency increased from 0.01 to 0.37, and for C1534 from 0.35 up to fixation, both due to selection effects with deltamethrin [54]. Continuous monitoring of mosquitos across Cape Verde is important to obtain a clearer picture of underlying and changing insecticide resistance profiles in the country, thereby understand the emergence and spread of resistance, and inform vector control programmes.

We have also performed haplotype and phylogenetic analysis using sequence data of the mitochondrial *nad*4 gene from Africa, America and Asia. The analysis indicated that the Cape Verdean *Ae. aegypti* are related to those from other countries in Africa and South America, corroborating with historical facts. The origin of the populations of *Ae. aegypti* mosquitoes using mitochondrial genes has been previously investigated. Population genetic analyses conducted with samples from Brazil using the *nad*4 gene among 163 mosquitoes identified two clades sharing haplotypes with populations from West Africa and Asia, and similar results were obtained with the mitochondrial *cytochrome c oxidase subunit 1* (*cox*1) gene [21]. Similar work based on nine microsatellite loci and both *nad*4 and *cox*1 sequences carried out in Bolivia also revealed the existence of two clades, one related with West Africa [55]. Phylogenetic analyses using 34 *nad*4 unique haplotypes from African *Ae. aegypti* mosquitoes and global sequences also revealed 2 clades, with the global haplotypes occurring in both clades [41]. In our results, using a larger dataset of global samples, we also observed two distinct clades contain isolates from all continents, with the Cape Verdean samples located within either clade. A previous study analysing the genetic diversity of samples collected from Cape Verde suggested a West African origin of local mosquitoes [18]. Several of the haplotypes detected in our sample collection were also present in the previously reported Cape Verdean samples. In total,

6 haplotypes were unique to the Cape Verdean mosquitoes while the other haplotypes were present particularly in African and American populations. Historically, *Ae. aegypti* was first detected in 1931 on the island of São Vicente, probably with origin from close African countries. The establishment of trade routes between Europe, Africa and the New World in the 15th century led to the dissemination of *Ae. aegypti* [11, 18]. Following its introduction into the Americas from West Africa *via* slave trade ships between 15th and 18th centuries, *Ae. aegypti* mosquitoes disseminated westwards to the Asia-Pacific region in the late 19th century. Since then, population growth, urbanization and climate change has allowed *Ae. aegypti* to thrive, with suitable foci in 188 countries/territories [1, 15].

History has shown that the opening of travel and trade routes between countries has been accompanied by the spread of mosquitoes and arboviruses, even more so now with the expansion of global air travel. Cape Verde has a strategic location in the middle of the Atlantic, with established air and sea lines and hence heavy traffic coming in and out of the country's four international airports and four international ports. The economy of Cape Verde relies heavily on tourism and in the importation of goods due to the lack of natural resources. The constant flow of human and trade traffic, alongside with an established *Ae. aegypti* population, means that Cape Verde will always be at risk of arboviral importation. Molecular and epidemiological results indicate that DENV was imported to Cape Verde from Senegal [23] and ZIKV from Brazil [25, 33]. There is a great risk that other vector borne diseases, such as CHIKV and West Nile virus, could be imported into Cape Verde, hence control measures, including strengthening mosquito surveillance, are essential.

Conclusions

Our results showed that the populations of *Ae. aegypti* collected in Praia at the end of the ZIKV outbreak displayed a low rate of ZIKV infection. In addition, *kdr* mutations associated with insecticide resistance were not detected. Haplotype and phylogeny analysis revealed unique haplotypes in the Cape Verdean *Ae. aegypti* and also indicated that these mosquitoes are related to populations found in Africa and South America. Since Cape Verde has a strategic location with a constant movement of human and trade traffic, studies on vector and pathogen screening, including early detection of insecticide resistance and screening for arboviruses, should be ongoing to support vector control measures and rapid response to future outbreaks in the country.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13071-020-04356-z.

Additional file 1: Table S1. Primers and probes used to detect Zika (ZIKV) and dengue (DENV) virus. **Table S2.** *Aedes aegypti* mitochondrial *nad*4 and *VGSC* primers for PCR assays.

Additional file 2: Figure S1. Phylogenetic tree inferred using only sequences collected in Cape Verde. The maximum-likelihood phylogeny was inferred using IQTREE with automatic selection of the best-fit model.

Additional file 3: Figure S2. Phylogenetic tree using unique haplotypes per country ($n = 262$). Colouration of leaves indicates isolate continental origin. Taxa included in Moore et al.[44] are annotated in blue (clade 1) and red (clade 2). Novel Cape Verdean Sequences presented in this study are indicated (*) alongside pre-existing publicly available isolates (•). The maximum-likelihood tree was inferred using IQTREE with automatic selection of the best-fit model.

Additional file 4: Figure S3. Haplotype Network based on *Aedes aegypti* mitochondrial *ND4* sequences. The number of circles represents the number of haplotypes found in Cape Verde. The colours represent the countries in the dataset. The size of the circles does not represent the sample size as 62% of the data are from Asia. The haplotypes are connected by a straight line if they differ by a single mutational step. Singletons and haplotypes with low frequency were not included.

Acknowledgements
We would like to thank those in Cape Verde who contributed to this study, including the research group on tropical diseases at GIDTPiaget and Jean Piaget University, the WHO desk in Cape Verde and the Ministry of Health. The MRC-UK eMedLab computing resource was used for bioinformatics and statistical analysis.

Authors' contributions
LFG, TGC and SC designed the study, performed experiments, analysed the data and drafted the manuscript. MC, DW and RFM carried out experiments, analysed the data and drafted the manuscript. KS, ARG and LFG were responsible for collecting the mosquitoes and taxonomy. NS provided expert knowledge and critical feedback during the study design and manuscript preparation. LFG, SC, ARG and TC coordinated the work and critically reviewed the manuscript. All authors read and approved the final manuscript.

Funding
This work was funded by an MRC UK and Wellcome Trust ZIKA Rapid Response grant (Ref. MC_PC_15103). TGC received funding from the MRC UK (Grant no. MC_PC_15103, MR/K000551/1, MR/M01360X/1, MR/N010469/1, MR/R020973/1) and BBSRC UK (BB/R013063/1). SC received funding from the Medical Research Council UK grants (MC_PC_15103, MR/R020973/1) and the BBSRC UK (BB/R013063/1).

Availability of data and materials
Data supporting the conclusions of this article are included within the article and its additional files. The newly generated sequences were deposited in the GenBank database under the accession numbers MT721877-MT721961.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. [2] Laboratory of Pathogen-Host Interactions (LPHI), UMR5235, CNRS, Montpellier University, 34095 Montpellier,

France. [3] Universidade Jean Piaget (UniPiaget), Praia, Cabo Verde. [4] Centre of Statistics and Its Applications of University of Lisbon, Lisbon, Portugal. [5] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK.

Received: 3 May 2020 Accepted: 11 September 2020
Published online: 21 September 2020

References

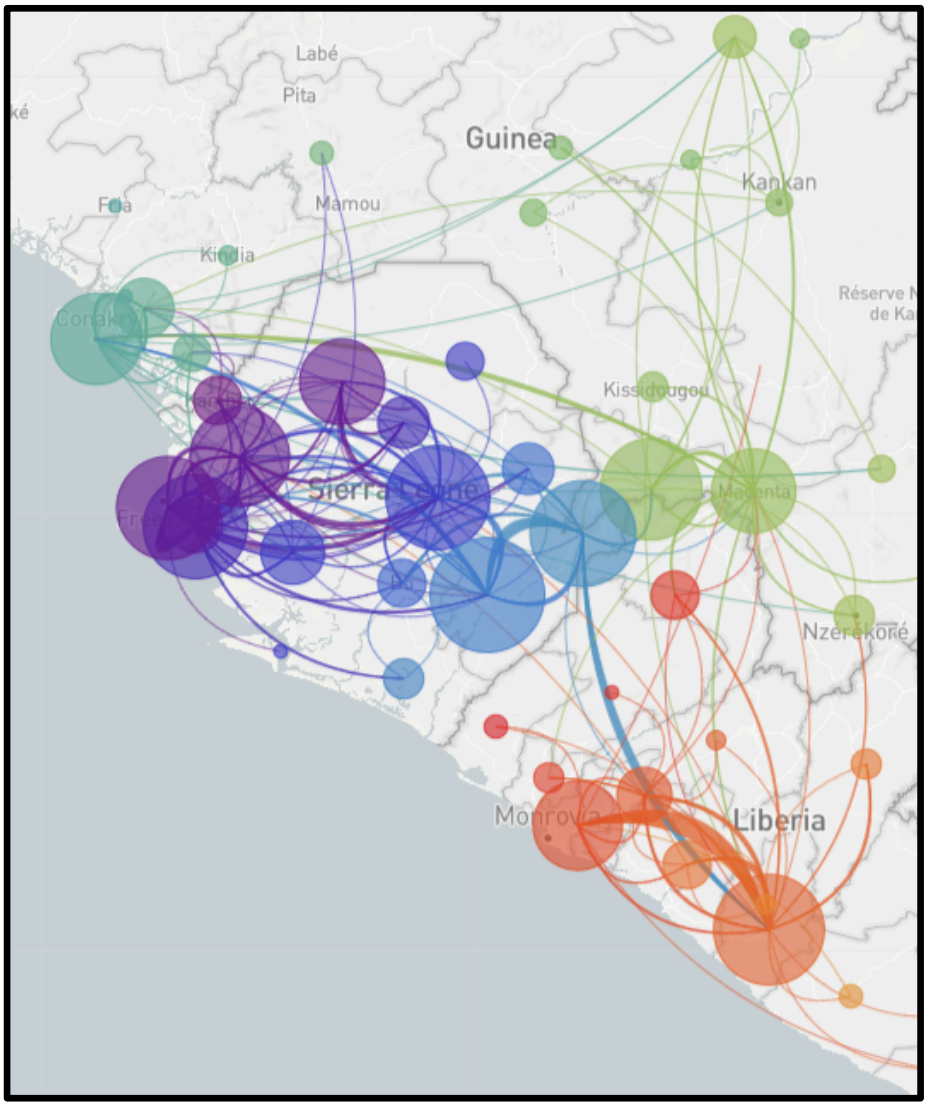
1. Leta S, Beyene TJ, De Clercq EM, Amenu K, Kraemer MUG, Revie CW. Global risk mapping for major diseases transmitted by *Aedes aegypti* and *Aedes albopictus*. Int J Infect Dis. 2018;67:25–35.
2. Weaver SC, Charlier C, Vasilakis N, Lecuit M. Zika, chikungunya, and other emerging vector-borne viral diseases. Annu Rev Med. 2018;69:395–408.
3. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 2013;496:504–7.
4. WHO. Factsheet Vector-borne diseases. Geneva: World Health Organization. 2014. http://www.who.int/kobe_centre/mediacentre/vbdfactsheet.pdf.
5. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546:406–10.
6. Krauer F, Riesen M, Reveiz L, Oladapo OT, Martínez-Vega R, Porgo TV, et al. Zika virus infection as a cause of congenital brain abnormalities and Guillain-Barré syndrome: systematic review. PLoS Med. 2017;14:e1002203.
7. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017;546:401–5.
8. Patterson J, Sammon M, Garg M. Dengue, Zika and chikungunya: emerging arboviruses in the New World. West J Emerg Med. 2016;17:671–9.
9. Costa-da-Silva AL, Ioshino RS, Petersen V, Lima AF, Cunha M, Wiley MR, et al. First report of naturally infected *Aedes aegypti* with chikungunya virus genotype ECSA in the Americas. PLoS Negl Trop Dis. 2017;11:e0005630.
10. Kraemer MUG, Faria NR, Reiner RC, Golding N, Nikolay B, Stasse S, et al. Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015–16: a modelling study. Lancet Infect Dis. 2017;17:330–8.
11. Powell JR, Tabachnick WJ. History of domestication and spread of *Aedes aegypti* - a review. Mem Inst Oswaldo Cruz. 2013;108:11–7.
12. Edman JD, Strickman D, Kittayapong P, Scott TW. Female *Aedes aegypti* (Diptera: Culicidae) in Thailand rarely feed on sugar. J Med Entomol. 1992;29:1035–8.
13. Scott TW, Clark GG, Lorenz LH, Amerasinghe PH, Reiter P, Edman JD. Detection of multiple blood feeding in *Aedes aegypti* (Diptera: Culicidae) during a single gonotrophic cycle using a histologic technique. J Med Entomol. 1993;30:94–9.
14. Scott TW, Takken W. Feeding strategies of anthropophilic mosquitoes result in increased risk of pathogen transmission. Trends Parasitol. 2012;28:114–21.
15. Kraemer MUG, Sinka ME, Duda KA, Mylne AQN, Shearer FM, Barker CM, et al. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. Elife. 2015;4:e08347.
16. Ribeiro H, da Cunha Ramos H, Capela RA, Pires CA. Os mosquitos de Cabo Verde, sistemática, distribuição, bioecologia, e importância médica. Lisbon: Junta de Investigaçoes Cientifícas do Ultramar; 1980.
17. Alves J, Gomes B, Rodrigues R, Silva J, Arez AP, Pinto J, et al. Mosquito fauna on the Cape Verde Islands (West Africa): an update on species distribution and a new finding. J Vector Ecol. 2010;35:307–12.
18. Salgueiro P, Serrano C, Gomes B, Alves J, Sousa CA, Abecasis A, et al. Phylogeography and invasion history of *Aedes aegypti*, the dengue and Zika mosquito vector in Cape Verde islands (West Africa). Evol Appl. 2019;12:1797–811.
19. Gloria-Soria A, Ayala D, Bheecarry A, Calderon-Arguedas O, Chadee DD, Chiappero M, et al. Global genetic diversity of *Aedes aegypti*. Mol Ecol. 2016;25:5377–95.
20. Bennett KL, Shija F, Linton YM, Misinzo G, Kaddumukasa M, Djouaka R, et al. Historical environmental change in Africa drives divergence and admixture of *Aedes aegypti* mosquitoes: a precursor to successful worldwide colonization? Mol Ecol. 2016;25:4337–54.
21. Lima RS Jr, Scarpassa VM. Evidence of two lineages of the dengue vector *Aedes aegypti* in the Brazilian Amazon, based on mitochondrial DNA ND4 gene sequences. Genet Mol Biol. 2009;32:414–22.
22. Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, MacArthur DG, et al. Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics. 2009;183:1065–77.
23. Franco L, Di Caro A, Carletti F, Vapalahti O, Renaudat C, Zeller H, et al. Recent expansion of dengue virus serotype 3 in West Africa. Euro Surveill. 2010;15:19490.
24. Guedes DRD, Gomes ETB, Paiva MHS, de Melo-Santos MAV, Alves J, Gómez LF, et al. Circulation of DENV2 and DENV4 in *Aedes aegypti* (Diptera: Culicidae) mosquitoes from Praia, Santiago Island. Cabo Verde. J Insect Sci. 2017;17:86.
25. Lourenço J, Monteiro ML, Valdez T, Monteiro Rodrigues J, Pybus O, Rodrigues Faria N. Epidemiology of the Zika virus outbreak in the Cabo Verde Islands, West Africa. PLoS Curr. 2018;10.
26. Direção Nacional da Saúde Programa Integrado de Luta Contra as Doenças Transmitidas por Vetores e Problemas da Saúde Associados ao Meio Ambiente. 2015. https://www.minsaude.gov.cv/index.php/documentosite/paludismo-e-luta-integrada-de-vetores/400-manual-da-luta-integrada-de-vetores-cabo-verde/file.
27. Vais H, Williamson MS, Devonshire AL, Usherwood PNR. The molecular interactions of pyrethroid insecticides with insect and mammalian sodium channels. Pest Manag Science. 2001;57:877–88.
28. Moyes CL, Vontas J, Martins AJ, Ng LC, Koou SY, Dusfour I, et al. Contemporary status of insecticide resistance in the major *Aedes* vectors of arboviruses infecting humans. PLoS Negl Trop Dis. 2017;11:e0005625.
29. Brengues C, Hawkes NJ, Chandre F, McCarroll L, Duchon S, Guillet P, et al. Pyrethroid and DDT cross-resistance in *Aedes aegypti* is correlated with novel mutations in the voltage-gated sodium channel gene. Med Vet Entomol. 2003;17:87–94.
30. Saavedra-Rodriguez K, Urdaneta-Marquez L, Rajatileka S, Moulton M, Flores AE, Fernandez-Salas I, et al. A mutation in the voltage-gated sodium channel gene associated with pyrethroid resistance in Latin American *Aedes aegypti*. Insect Mol Biol. 2007;16:785–98.
31. Martins AJ, Lins RM, Linss JGB, Peixoto AA, Valle D. Voltage-gated sodium channel polymorphism and metabolic resistance in pyrethroid-resistant *Aedes aegypti* from Brazil. Am J Trop Med Hyg. 2009;81:108–15.
32. Rocha HDR, Paiva MHS, Silva NM, de Araújo AP, Camacho DDR, da Moura AJF, et al. Susceptibility profile of *Aedes aegypti* from Santiago Island, Cabo Verde, to insecticides. Acta Trop. 2015;152:66–73.
33. Dia I, Diagne CT, Ba Y, Diallo D, Konate L, Diallo M. Insecticide susceptibility of *Aedes aegypti* populations from Senegal and Cape Verde Archipelago. Parasit Vectors. 2012;5:238.
34. The dengue vaccine dilemma. Lancet Infec Dis. 2018;18:123.
35. Seixas G, Salgueiro P, Silva AC, Campos M, Spenassatto C, Reyes-Lugo M, et al. *Aedes aegypti* on Madeira Island (Portugal): genetic variation of a recently introduced dengue vector. Mem Inst Oswaldo Cruz. 2013;108(Suppl. 1):3–10.
36. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:1647–9.
37. Rambaut A, Lam TT, Carvalho LM, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;2:vew007.
38. Paradis E. Pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics. 2010;26:419–20.
39. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406–25.
40. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGAX: Molecular Evolutionary Genetics Analysis across computing platforms. Mol Biol Evol. 2018;35:1547–9.
41. Moore M, Sylla M, Goss L, Burugu MW, Sang R, Kamau LW, et al. Dual African origins of global *Aedes aegypti sl* populations revealed by mitochondrial DNA. PLoS Negl Trop Dis. 2013;7:e2175.
42. Guerbois M, Fernandez-Salas I, Azar SR, Danis-Lozano R, Alpuche-Aranda CM, Leal G, et al. Outbreak of Zika virus infection, Chiapas State, Mexico,

2015, and first confirmed transmission by *Aedes aegypti* mosquitoes in the Americas. J Infect Dis. 2016;214:1349–56.
43. Ferreira-de-Brito A, Ribeiro IP, Miranda RM, Fernandes RS, Campos SS, Silva KAB, et al. First detection of natural infection of *Aedes aegypti* with Zika virus in Brazil and throughout South America. Mem Inst Oswaldo Cruz. 2016;111:655–8.
44. Ho ZJM, Hapuarachchi HC, Barkham T, Chow A, Ng LC, Lee JMV, et al. Outbreak of Zika virus infection in Singapore: an epidemiological, entomological, virological, and clinical analysis. Lancet Infect Dis. 2017;17:813–21.
45. Calvez E, O'Connor O, Pol M, Rousset D, Faye O, Richard V, et al. Differential transmission of Asian and African Zika virus lineages by *Aedes aegypti* from New Caledonia. Emerg Microbes Infect. 2018;7:159.
46. Zardkoohi A, Castañeda D, Lol JC, Castillo C, Lopez F, Rodriguez RM, Padilla N. Co-occurrence of kdr mutations V1016I and F1534C and its association with phenotypic resistance to pyrethroids in aedes aegypti (diptera: culicidae) populations from costa rica. J Med Entomol. 2020;57(3):830–6. https://doi.org/10.1093/jme/tjz241.
47. Sombié A, Saiki E, Yaméogo F, Sakurai T, Shirozu T, Fukumoto S, et al. High frequencies of F1534C and V1016I *kdr* mutations and association with pyrethroid resistance in *Aedes aegypti* from Somgandé (Ouagadougou). Burkina Faso. Trop Med Health. 2019;47:2.
48. Haddi K, Tomé HVV, Du Y, Valbon WR, Nomura Y, Martins GF, et al. Detection of a new pyrethroid resistance mutation (V410L) in the sodium channel of *Aedes aegypti*: a potential challenge for mosquito control. Sci Rep. 2017;7:46549.
49. Du Y, Nomura Y, Satar G, Hu Z, Nauen R, He SY, et al. Molecular evidence for dual pyrethroid-receptor sites on a mosquito sodium channel. Proc Natl Acad Sci USA. 2013;110:11785–90.
50. da Cruz DL, Paiva MHS, Guedes DRD, Alves J, Gómez LF, Ayres CFJ. Detection of alleles associated with resistance to chemical insecticide in the malaria vector *Anopheles arabiensis* in Santiago. Cabo Verde. Malar J. 2019;18:120.
51. Messenger LA, Shililu J, Irish SR, Anshebo GY, Tesfaye AG, Ye-Ebiyo Y, et al. Insecticide resistance in *Anopheles arabiensis* from Ethiopia (2012–2016): a nationwide study for insecticide resistance monitoring. Malar J. 2017;16:469.
52. Ngoagouni C, Kamgang B, Brengues C, Yahouedo G, Paupy C, Nakouné E, et al. Susceptibility profile and metabolic mechanisms involved in *Aedes aegypti* and *Aedes albopictus* resistant to DDT and deltamethrin in the Central African Republic. Parasit Vectors. 2016;9:599.
53. Badolo A, Traore A, Jones CM, Sanou A, Flood L, Guelbeogo WM, et al. Three years of insecticide resistance monitoring in *Anopheles gambiae* in Burkina Faso: resistance on the rise? Malar J. 2012;11:232.
54. Alvarez LC, Ponce G, Saavedra-Rodriguez K, Lopez B, Flores AE. Frequency of V1016I and F1534C mutations in the voltage-gated sodium channel gene in *Aedes aegypti* in Venezuela. Pest Manag Sci. 2015;71:863–9.
55. Paupy C, Le Goff G, Brengues C, Guerra M, Revollo J, Barja Simon Z, et al. Genetic structure and phylogeography of *Aedes aegypti*, the dengue and yellow-fever mosquito vector in Bolivia. Infect Genet Evol. 2012;12:1260–9.

Publisher's Note

Chapter Three

Applying ‘omics techniques in a molecular investigation of Zika virus transmission



Augmenting epidemiological processes with molecular precision. Map depicting inferred routes of transmission of Ebola virus throughout the 2014-2015 outbreak in West Africa (Hadfield *et al.* 2018). Temporal phylogenies are built on sequence data collected from patients in treatment centres. Leaf data and ancestral state reconstruction enables the geospatial

3.1 Introduction

3.1.1 Genomic epidemiology of ZIKV

The first isolation of ZIKV occurred serendipitously, after a routine YFV surveillance project in the Zika forest, Uganda 1947 [1]. A year later, ZIKV was isolated a further time in a mosquito, *Aedes africanus*, caught in the Zika forest [2]. In 1952 the first reports of human ZIKV infections were recorded in Uganda and Tanzania, where neutralising antibodies were detected in 43 serum isolates from across eight study sites [3]. With the confirmation of human seroconversion, wider ZIKV surveillance programmes began, through which, in the proceeding 30 years, ZIKV infections in human populations would be found to be endemic across equatorial Africa and Asia, with studies reporting seroprevalence in Nigeria (1975) [4] , Sierra Leone (1975) [5], Senegal (1978) [6], Malaysia (1969) [7], Central African Republic (1981) [8], Indonesia (1981) [9], and Pakistan (1983) [10]. It is through these studies that samples and culture lines were archived, which were later made available for sequencing and phylogeographic reconstruction. A temporal and geographic summary of these data is shown in **Figure 1**.

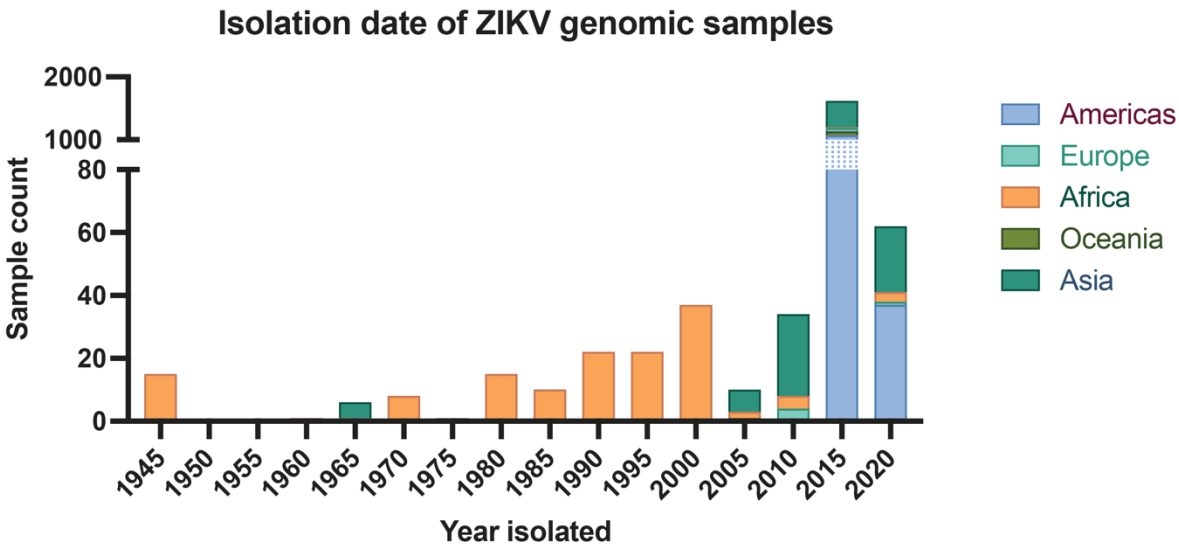


Figure 1. Temporal and geographic metadata associated with 1860 ZIKV whole and partial sequences sourced from NCBI GenBank Virus Database [53] on 10/03/22.

Most retrospectively sequenced isolates, obtained between 1945 and 2000, were sourced from Africa, particularly Uganda, Senegal, and Cote d'Ivoire. In 2007, the first observed outbreak of ZIKV was reported on the Island of Yap, Micronesia, before which, only 14 cases of human ZIKV infection were recorded [11]. RT-qPCR and Sanger sequencing were utilised to confirm ZIKV as the etiological agent. The sequencing data facilitated phylogenetic reconstruction, which indicated that the Micronesian sequence (EU545988.1) was indeed ZIKV, when compared with other *Flavivirus* species sequence data. However, with a limited database of only 4 ZIKV genomes at that time, robust phylogeographic inference was not possible [12]. Five years after the outbreak on Yap, enabled by further ZIKV sequencing endeavours across Southeast Asia, the Micronesian strain was analysed and revealed to cluster with sequences from Cambodia and Malaysia, which indicated the existence of two ZIKV strains, African and Asian [13]. The outbreak on Yap resulted in 185 cases of ZIKV infection with an attack rate of 14.6 per 1000 people. Seroprevalence studies estimated that 73% of the island's population had been recently infected [11].

On October 16th, 2016, nearly 10 years after the outbreak on Yap, Dr Henri-Pierre Mallet, situated on the Pacific Island of French Polynesia, reported an estimate of 400 ZIKV cases in a message on the ProMED surveillance network (ProMED archive number: 20131106.2041959). The outbreak lasted for 6 months (week 41, 2013 to week 14, 2014) with between 30 to 32 thousand reported cases (11.5% of the population). Importantly, a wealth of ZIKV whole-genome sequences were published (n=25), which clustered closely with sequences from Thailand, which is supported by phylogeographic reconstruction (https://nextstrain.org/zika). Further outbreaks were reported shortly after on Easter Island, the Cook Islands and New Caledonia [14,15], from which 81 partial genomes were published,

focusing particularly on the non-structural protein 5 (*NSP5*) gene (~1029 nucleotides), all of which clustered closely with French Polynesian sequences.

A significant effort in research has been invested into the determination of the point of introduction for the Asian lineage of ZIKV to the Americas. The foremost hypotheses centre around large events of international interest due to the increased influx of international travellers. The 2014 FIFA World Cup held in Brazil has long been a prime candidate (12 June to 13 July 2014) [16], with another theory focusing on a canoe race (12 and 17 August 2014), which saw competitors from French Polynesia compete, shortly after the epidemic on their home island [17]. A third hypothesis used dated phylogenies to theorise that introduction may have occurred earlier, during the Confederations Cup tournament (2013), however this took place before cases were reported on French Polynesia [18].

Before the end of 2015, across the Americas, eleven countries reported PCR positive ZIKV cases, including the USA. Moreover, concerning evidence of a causal link between prenatal ZIKV infection and microcephaly have been reported, a finding that led the WHO to declare a Public Health emergency of International Concern (PHEIC) [19]. Retrospective studies would later find an association between microcephaly and ZIKV Asian lineage outbreaks prior to its emergence in the Americas [20]. Further research in to the ZIKV genome and its newly discovered pathogenicity highlighted positions under selective pressure, such as those on the envelope glycoprotein E [21], a mutation on NS1 with a role in immunomodulation (A188V) [22], and an Asian lineage-specific non-synonymous mutation with a putative role in neurotropism (S139N) [23].

The molecular evolution of ZIKV in Americas has been covered thoroughly [18,21,24], in which a consensus on Brazil having served as the main ZIKV exporter on the continent is firmly maintained (**Figure 2**). Reports of the introduction and co-circulation of two or more discrete ZIKV sub-lineages, in confined regions, have been verified with phylogeographic analyses, including within populations in USA, Colombia and Puerto Rico [25–27]. Of further concern, were reports of intercontinental transmission, where positive PCR assay results were found in Cabo Verde, located off the coast of west Africa, 500 km west from Senegal [28]. Historically, Cabo Verde has strong intercontinental associations. The islands were subject to colonial invasion in 1456 by the Portuguese, forging also, close cultural links to Brazil. These links to this day, place Cabo Verde as a travel hub, with regular direct flights to the American continent, Europe and mainland Africa. Furthermore, Cabo Verde is reliant on the importation of agricultural goods and commodities from neighbouring countries. These factors together, may increase the risk of introducing novel pathogens and vectors to the island.

The first outbreak of *Flavivirus* species occurred on Cabo Verde in 2009, a DENV-3 epidemic, resulting in 20,914 reported cases and four cases of severe dengue with mortality events [29]. The 2015-16 Cabo Verdean outbreak was the first large ZIKV outbreak with confirmed cases of microcephaly in Africa. The Cabo Verde Ministry of Health reported 7,580 suspected cases between October 2015 and May 2016 with 18 cases of Zika congenital syndrome, 63% of which were in Praia on Santiago Island, with cases peaking on week 47 of 2015 [30,31].

The genomic epidemiology of the outbreak on Cabo Verde has been described previously, in the analysis of three whole genome sequences obtained from human infections collected in Fogo (11/2015 and April 2016) and Santiago (December 2015) [32]. The resulting phylogeographic reconstruction inferred the introduction of ZIKV to Cabo Verde to have

occurred between June 2014 and August 2015, with the ancestral node indicating the isolates samples were most likely to be of Brazilian origin. Of note, two non-synonymous mutations were identified as being exclusively Cabo Verdean, I756V and T659A. Since the outbreak on Cabo Verde, two further continental African outbreaks have been reported, in Angola (2016) and Guinea-Bissau (2016). Interestingly, the outbreak in Angola was reported to be of American descent (Asian lineage), whereas the strain circulating in Guinea-Bissau was of African origin [33,34].

With the detection of two ZIKV positive *Ae. aegypti* RNA samples, as reported in the previous chapter, two aims are set. First, to sequence the virome of the *Ae. aegypti* mosquito datasets featured in the previous chapter. Secondly, to develop and test advanced enrichment techniques to sequence ZIKV in entomological samples, and analyse them in a global genomic context, elucidating the ancestry, diversity, and dynamics of the ZIKV population on Cabo Verde.

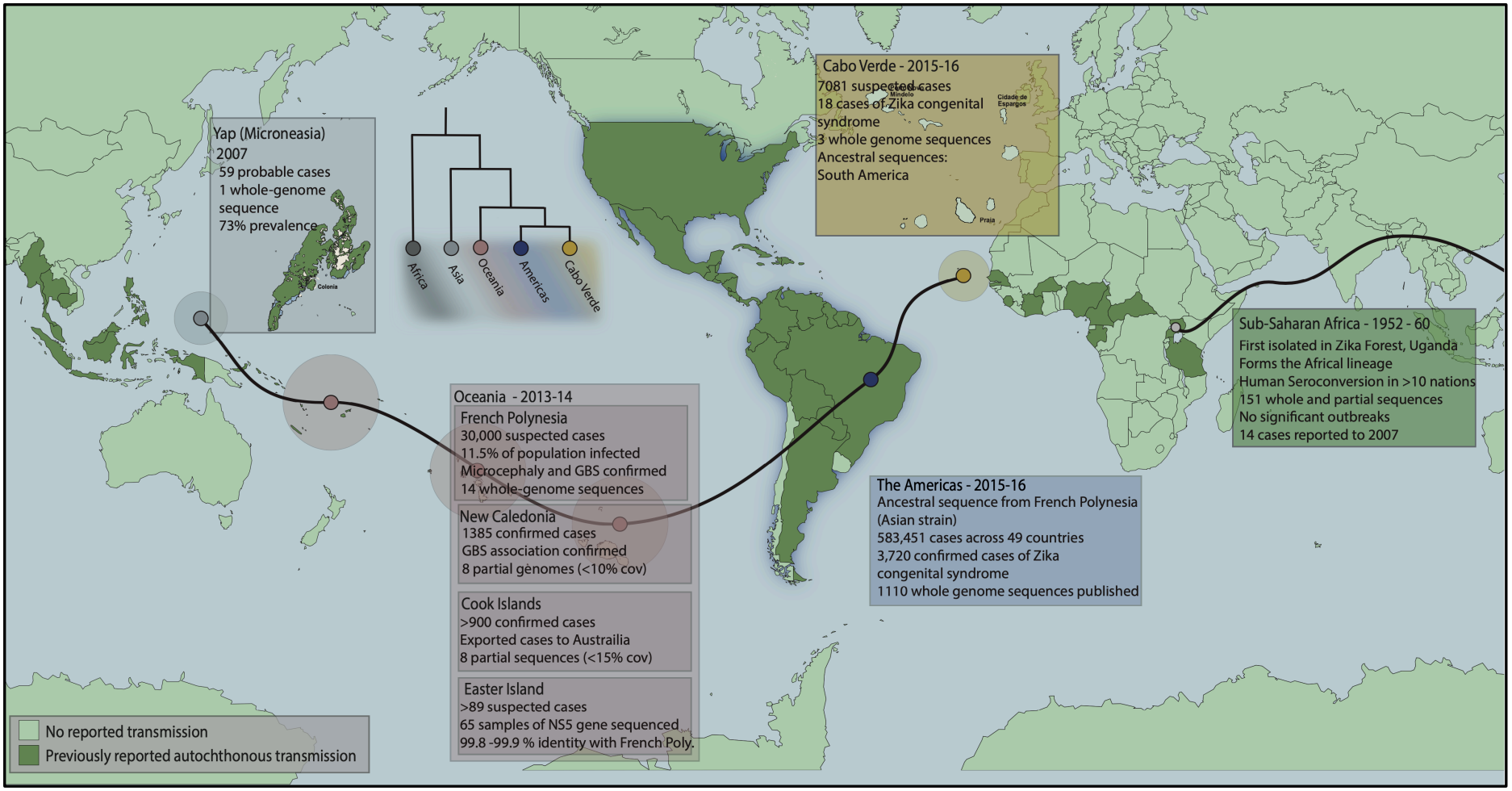


Figure 2. Summary of the emergence of ZIKV worldwide, up to the introduction of ZIKV to Cabo Verde, as detailed in the previous written introductory sections. Genome sequence submission figures were calculated as of 01/03/22. The cladogram shown is for illustration purposes only, the branch lengths are not relative to distance.

3.1.2 Technical considerations of viral genomic studies

The sequencing of virus nucleic acids from complex samples presents a set of challenges depending on the organism and the objectives of the study (**Figure 3**). The first consideration is in understanding complexity of the sample, which is derived from the estimated mixture of foreground (useful, on-target) and background (less-useful, off-target) nucleic acids. For instance, when applying meta-sequencing techniques to target the RNA virome from human peripheral blood, the background would make 99% of the acquired reads, comprising of human mRNAs and ribosomal RNAs captured from nucleated blood cells along with bacterial and fungal RNA contaminants, and the foreground (<1%), on-target viral RNA transcripts. This approach is advantageous in that it requires no *a-priori* knowledge of virus sequence and introduces no selection bias, and so can detail the exact proportional representation of sequence in the wider transcriptomic context.

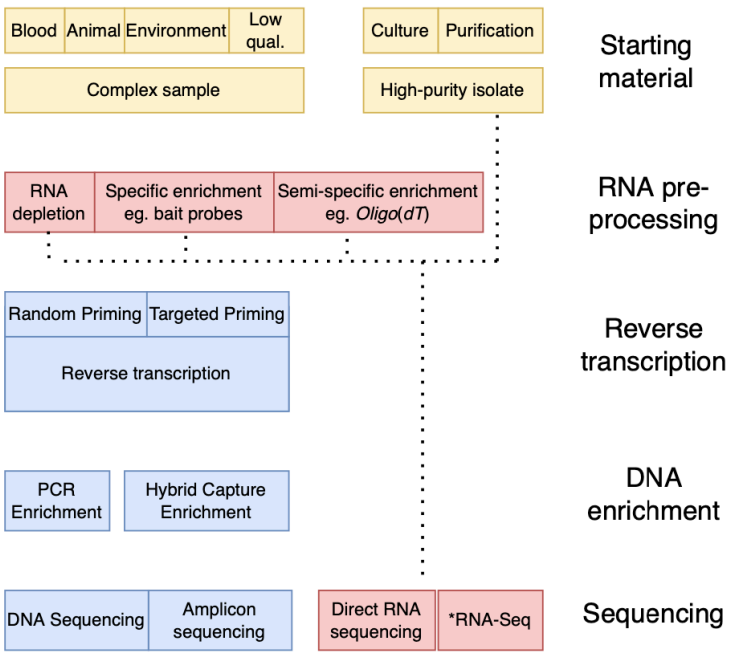


Figure 3. Summary of RNA virus sample processing and sequencing. Hashed line denotes the bypassing of reverse transcription steps for nanopore direct RNA sequencing and RNA-seq library preparation. *RNA-seq workflows require reverse transcription.

However, it may prove costly when scaled up, and the low foreground yield can produce poor results when the objective is whole-genome data acquisition.

A key example demonstrating the traits of sequencing strategies can be found in the genomic study of Ebola virus (EBOV), which exhibits substantial viraemia in acute patients. In theory, this should present sufficient material for sequencing, however, Li *et al.* demonstrated that meta-sequencing confirmed EBOV patients' (RT-qPCR +ve) peripheral blood resulted in only 0.13% of total reads acquired mapping to the EBOV reference [35]. This translated into an average of 1X covering 30% of genome. With the application of EBOV-specific RNA capture/hybridisation enrichment techniques, the yield increased to 19% on-target (EBOV) reads, but still only 47% average EBOV genome coverage. Throughout the 2014-2016 West African EBOV outbreak, molecular epidemiology alongside contact tracing, played a vital role in the surveillance of (emerging) infections. Quick *et al.* implemented a 'tiling amplicon' sequencing approach, which entails the PCR amplification and sequencing of a target genome in short, fixed overlapping segments [36]. This technique has been proven to be economic, sensitive, and quick, and yielded 142 high quality genomes mostly sequenced within days of receiving the sample, enabled by the highly portable Oxford Nanopore sequencing platform, MinION.

Years later, with the announcement of the ZIKV outbreak in South America, the same team produced a primer scheme for the whole genome sequencing of Zika. Consisting of 35 primer pairs, amplifying overlapping regions covering the entire genome in a single multiplexed PCR reaction, the technique included, as before, nanopore technology as the sequencing platform, this time while operating out of a camper van [37,38]. While this exciting project was, in principle very similar to its EBOV predecessor, the resulting data was found to be sub-optimal.

Across 55 sequenced samples in 5 study sites, the average genome coverage (>10-fold) was 20.2%. It was soon evident that unlike EBOV, the majority of ZIKV infections were mild, a product of the characteristic low-level viraemia in most ZIKV infections [39]. This resulted in a reduced quantity of ZIKV template material, which when combined with the inherent fragility of RNA presents a limitation of the tiling amplicon technique implemented, namely, an inability to amplify and enrich low-quality templates.

While meta-sequencing indiscriminately captures all fragments for sequencing, PCR reactions require contiguous, intact RNA transcripts for efficient priming and amplification. An attempt at mitigating this limitation came in the form of shortened amplicon design, similar to 'jackhammering' techniques used in degraded HIV whole-genome sequencing [40]. Where the EBOV scheme produced amplicons of ~1000 bp, the ZIKV scheme used in the ZIBRA study produced ~500 bp amplicons (**Figure 4**), presenting a greater opportunity for successful priming. Regardless of the poor results of the study, the jackhammering-like technique was applied throughout the epidemic and yielded increasingly better results. Four years later, the same researchers behind the EBOV and ZIKV sequencing projects, would provide the foundation for applications to COVID-19. The SARS-CoV-2 ARTIC WGS methodology, applied successfully the multiplex-tiling amplicon sequencing strategy in tandem with third-generation nanopore sequencing technology (and optionally NGS platforms). This approach is currently in its fourth iteration and, along with other similar techniques, has facilitated sequencing of nine million SARS-CoV-2 sequences (https://artic.network/).

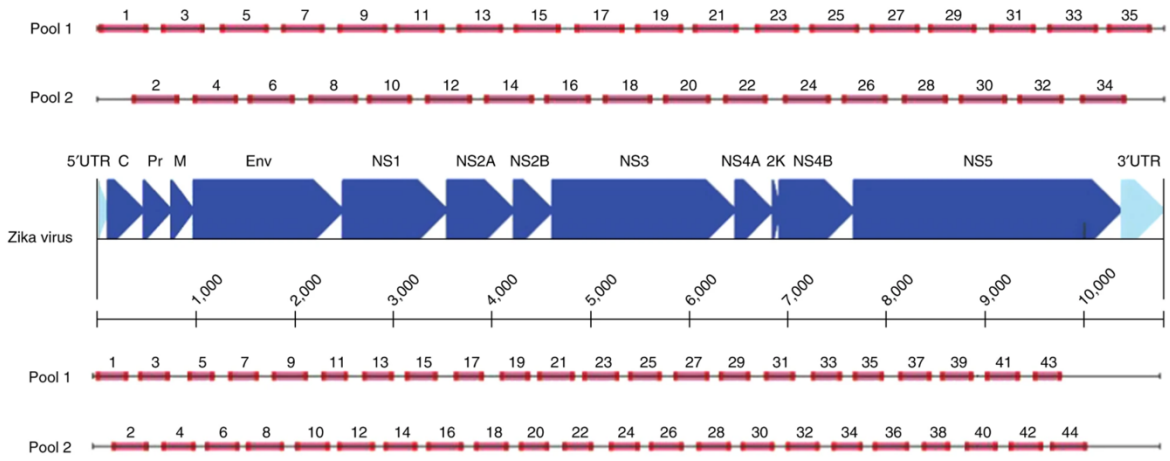


Figure 4. The 'ZIKV_asian' tiling amplicon scheme featured in Quick *et al*. [38]. Thirty-five primers span the entire length of the ZIKV genome in two pools, enabling the trimming of primer sequences from subsequent reads.

3.1.3 Alternative techniques for enrichment of viral sequencing materials

Where hybridising probes and amplification techniques target nucleic acids directly to enrich foreground materials, the prolonged infectivity of virus particles allow for propagation by *in-vitro* culture from relatively low-titre and challenging isolates. This method is favourable, in many cases, as it can yield unlimited quantities of sequencing materials and removes the requirement for nucleic-acid enrichment techniques. Importantly, there are four caveats of this process: (i) virus culture can be challenging, is time consuming and expensive; (ii) culture of BSL 3 and 4 organisms is logistically challenging and poses a significant risk to the user and environment; (iii) unlike PCR methods, the virus must be viable; and (iv) passage of virus in culture may apply or limit selective pressures, driving evolution, and confounding genomic epidemiological analyses.

A further consideration is the gross total amount of biological material available. As noted in the previous chapter, the isolates used here for ZIKV whole genome sequencing are *Ae. aegypti* mosquitoes. Techniques for isolating *Flavivirus spp.* from vectors have been well described [41–44]. Most approaches employ an isolate pooling strategy to increase RNA yields and

processivity, with either silica-based column RNA extraction or TRIzol extraction methodologies. Where there is a limited quantity of nucleic acids, there are a range of pre-amplification techniques available. In the case of whole-transcriptome amplification, random RT-PCR is followed by transcript ligation and whole-genome amplification, using phi29 polymerase with random priming. This technique is said to uniformly amplify all transcripts, with input quantities as little as 1 ng yielding >1 µg of cDNA.

3.2 Methods

Zika RNA whole transcriptome amplification

RNA from one RT-qPCR positive sample identified in the previous chapter was quantified using Qubit™ RNA High Sensitivity (HS) fluorometric reagents (Q32852) as according to manufacturer's instruction and screened using UV spectrophotometry for inhibitory contaminants such as guanidine thiocyanate. A total of 50 ng was used in the initial reverse transcription step of the QIAGEN QuantiTect Whole Transcriptome Kit (207045). The protocol was performed as per manufacturer's instruction except for the following conditions:

Condition	Additional primers added:	
	Reverse Transcription	**phi29 amplification**
1	Random primers	Random primers
2	ZIKV specific + Random	Random primers
3	Random primers	ZIKV specific + Random
4	ZIKV specific + Random	ZIKV specific + Random

Table 1. Conditions for WTA experiments.

The ZIKV specific primers consisted of primer sets 1-4 and 31-35 from the 'tiling amplicon' ZIKV sequencing scheme (**Table S1**). The random primers were hexamers included with the WTA kit. The WTA products were screened for ZIKV DNA contigs using NEB Q5 high-

fidelity polymerase (M0491S) in single primer-pairs and visualised using 1% agarose gel electrophoresis with a 100 bp ladder (N0467S).

Nanopore sequencing of WTA products

The WTA products were quantified using the Qubit™ dsDNA HS (Q32851) fluorometric kit. Library preparation was performed using the RBK-004 rapid-barcoding sequencing kit on an R9.4 MinION flowcell. The library preparation process was modified, as per the Premium whole genome amplification protocol, to mitigate hyperbranched DNA structures created by phi29 polymerase by the addition of T7 endonuclease in an additional clean-up step. The remainder of the RBK-004 protocol was performed as per manufacturer's instructions. A total of 150 fmol was loaded on to the flowcell, which had 548 remaining pores. Sequencing progressed for 13 hours and yielded 1.93 Gb of sequencing data. The reads were basecalled using ONT Bonito v0.5.1 (https://github.com/nanoporetech/bonito) with the dna_r9.4.1_e8_sup@v3.3 super-high accuracy basecalling model. Reads were then trimmed using Porechopv0.2.4 [45], removing barcodes and adapter sequences.

Metagenomic analysis of WTA sequencing data

For the taxonomic classification of WTA reads, Kraken v2.1.2 [46] linked to the NCBI 'nt' database (downloaded 24/11/2021) was used. For the visualisation and quantification of taxa, Recentrifuge (v 1.5.1) software [47] was used. Taxa were quantified and tabulated based of a quality score on >100.

Phylogenetic analysis of equine infectious anaemia virus sequence

All WTA reads were mapped to an equine infectious anaemia virus (EIAV) reference genome (NC_001450) using minimap2 (v 2.18) software [48]. Variant calling and consensus sequence

generation was performed using BCFtools (v 1.3.1) [49]. Positions with < 5-fold coverage were masked using BEDTools (v 2.30.0) [50]. EIAV sequences were downloaded from GenBank, 816 samples were aligned using minimap2 software. Sixty-six samples remained after metadata quality control procedures, and sequences which had coverage of the region sequenced in this study. Phylogenetic inference was performed using IQtree 2.2.0 [51], yielding a maximum-likelihood (ML) tree from the sequence alignment with the best-fit model automatically selected by ModelFinder, with SH-aLRT test and ultrafast bootstrap with 1000 replicates. The resultant tree was visualised using FigTree (v1.4.4) software [52].

Pooled reverse-transcription and tiling-amplicon PCR of ZIKV from Cabo Verdean entomological samples

Five µL of 95 RNA samples were pooled in to a single Eppendorf 1.5 LoBind tube. The pooled samples were concentrated by means of a QIAamp Viral RNA column (52904) and the process was performed to manufacturer's instructions. The sample was eluted into a 30 µL of RNAse free water. Reverse transcription was performed using SuperScript™ IV Reverse Transcriptase (18090010). Twenty-five mM (0.5 mL) of random hexamers (N8080127) were added to 5 µL of pooled RNA (1250 ng total mass), and heated at 65 ºC for 5 minutes. Additional components were added as per manufacturer's instructions, and the reaction was incubated at 23 ºC for 20 minutes and 55 ºC for 30 minutes, followed by inactivation at 80 ºC for 10 minutes. cDNA was stored at -20 ºC for less than 1 week.

For the tiled amplification of the ZIKV genome from cDNA, the ZIKV_asian primer scheme previously mentioned [38] was applied. Q5 polymerase was used with an elongation time of 1 minute and annealing temperature of 60 ºC for 40 cycles. The amplicons were visualised

using a 1% agarose gel with a 100 bp ladder (N0467S). Amplicons were pooled and cleaned using QIAquick PCR Purification Kit (28104) as according to manufacturer's instructions.

Nanopore sequencing of ZIKV amplicons

The cleaned amplicons were quantified and prepared for nanopore sequencing using the Oxford nanopore LSK-109 ligation sequencing kit with the amplicon sequencing protocol on the MIN-FLG001 flowcell, as per manufacturer's instructions. Reads were quality filtered for a Q-score >7, and trimmed using Porechop software. The reads were mapped, using minimap2 software, to a defined Brazilian reference sequence (NC035889) resulting in a 50-fold coverage of 84.9% (**Figure 5**). Positions with zero coverage were masked using BEDTools (v 2.30.0) [50].

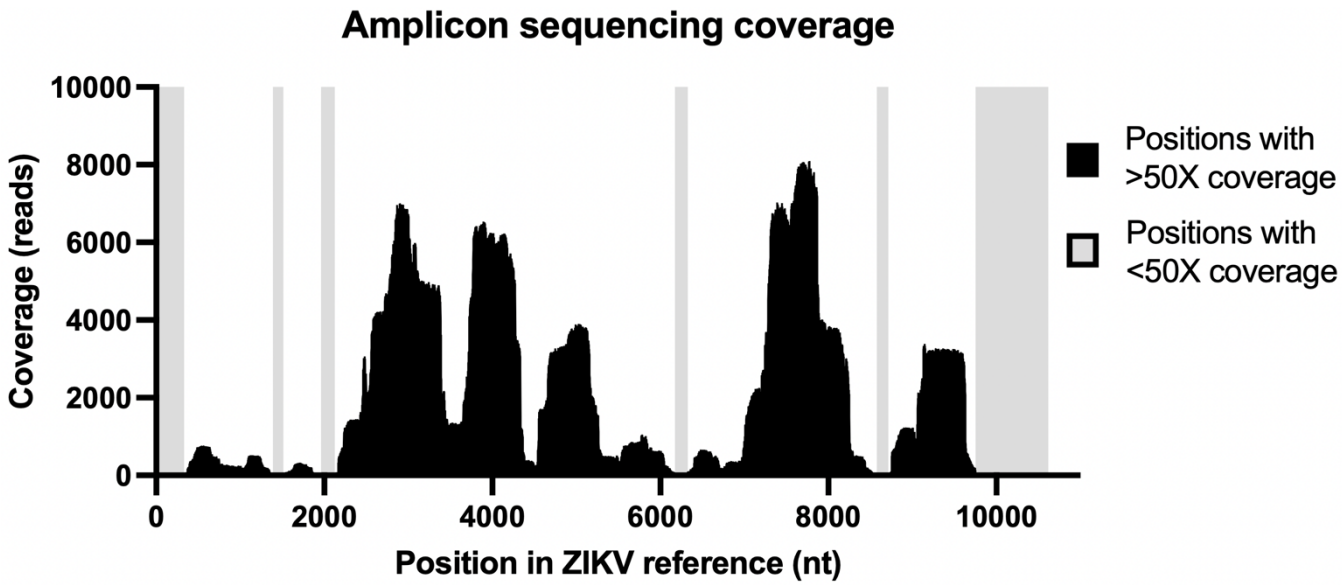


Figure 5. Coverage of ZIKV reference NC035889 after mapping of pooled tiling amplicon sequencing data. Regions with <50-fold coverage are shown in grey. All regions with <50-fold coverage were masked with 'N' positions for downstream analysis.

Bayesian phylogeographic reconstruction of ZIKV dataset

A dataset of 407 ZIKV whole genome sequences was downloaded from NCBI [53] and aligned using MAFFT v7.453 [54]. All sequences were screened for correct date of isolation and collection geolocation. A test phylogeny was performed using IQtree (v 2.2.0) [51], yielding a maximum-likelihood tree from the sequence alignment with the best-fit model automatically selected by ModelFinder, with the SH-aLRT test option and branch robustness quantified by bootstrap sampling of 1000 replicates. This tree was analysed using TempEst (v 1.5.3) [55] for temporal signal. BEAST (v 1.10.4) software [56] was used for temporal inference and discrete ancestral state phylogeographic reconstruction. A strict molecular clock model was applied, with a SRD06 codon-partitioned substitution model. The prior model used was the Skygrid-based coalescent tree distribution [57]. Multiple instances of 100,000,000 length chains were run until convergence. Three replicates were run and converged on the same values. A burn-in of 15% was implemented and various trace metrics analysed using Tracer (v 1.7) software [58]. The resulting phylogeny was visualised and annotated using FigTree (v 3.1) [52]. The Bayesian skygrid population analysis was completed using the Tracer (v 1.7) software.

3.3 Results

In the previous chapter, the findings from the analysis of entomological samples collected following the ZIKV outbreak in Cabo Verde were reported. A total of 816 samples were individually screened for DENV and ZIKV using RT-qPCR using primers and probes (see **Table S3.1**). No DENV and two ZIKV positive samples were detected. The second sample was only detected on the second technical replicate, which suggests the viral titre was on the lower limit of the RT-qPCR assay's sensitivity.

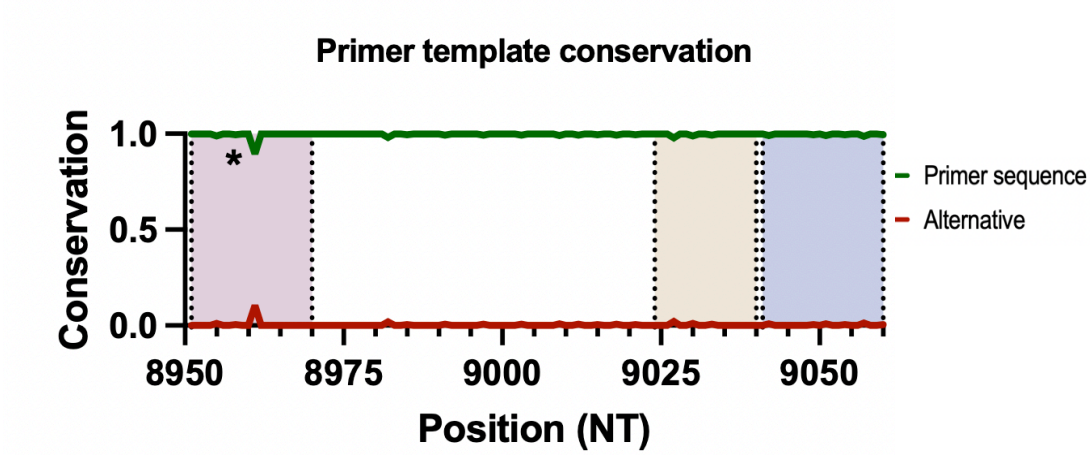


Figure 6. Analysis of ZIKV qPCR primer annealing sites, in the context of 432 ZIKV whole genome sequences. The analysis of nucleotide conservation was performed using JalView [74]. **Red** – forward primer, **yellow** – probe annealing site, **blue** – reverse primer. The asterisk denotes the site where the degenerate 'Y' nucleotide is situated.

To ensure the PCR reagents used were suitable for the detection of isolates from the geographical regions of suspected introduction, the primers and probes were screened against 432 ZIKV whole genome sequences to assess the conservation of annealing sites (**Figure 6**). Conservation across the dataset was high, except for a single site in the forward primer, in which 10.3% of South American sequences had a single alternative allele (T8961C). This variation had been accounted for with the incorporation of degenerate base in the primer sequence (Y).

3.3.1 Whole transcriptome sequencing of ZIKV positive *Aedes aegypti*

Using 50 ng (~10%) of the RNA acquired from the total RNA extraction, on the two ZIKV positive mosquitoes a whole transcriptome amplification (WTA) technique was applied to increase the concentration of nucleic acids available for analysis. The process was applied in 4 conditions: (i) as per manufacturer's instructions; (ii) with the addition of ZIKV specific primers in the reverse transcription step; (iii) with addition of ZIKV specific primers at the

phi29 amplification step; and (iv) a combination of conditions (ii) and (iii). All conditions yielded >2 μg total amplified DNA, which was then cleaned and assayed using 5 ZIKV primer pairs from the multiplex tiling amplicon scheme. Of the 35 individual PCR reactions run across all 4 of the WTA reactions, only four amplicons were observed, two of which were indicated as being significantly greater than the ~500 bp. Following the confirmation of partially successful ZIKV transcript amplification, the WTA reaction libraries for nanopore sequencing were prepared. In total, 5 Gb of reads were generated over a 12-hour sequencing run. Following basecalling and sequencing adapter trimming, a taxonomic labelling pipeline was applied to the reads, using the entire nucleotide database of GenBank as a comparison database. The results are shown in **Table 2** and **Figure 4.**

Organism	**Read count**	**Proportion of total reads (%)**
Eukaryota		
Homo sapiens	48441	17.0
Sus scrofa (boar)	39298	14.0
Aedes aegypti	24505	9.0
Canis lupus familiaris	2273	0.8
Equus caballus (horse)	32	0.0
Other	86857	30.2
Bacteria		
Escherichia coli	28370	10.0
Salmonella spp	11520	4.0
Moraxella osloensis	5866	2.0
Wolbachia spp.	7	0.002
Other	23487	8.0
Virus		
Equine infectious anaemia virus	496	0.2
Human immunodeficiency virus	22	0.008
Orthornavirae (RNA Virus)	148	0.05
Of which *Flavivirus* spp.	16	0.006
Other	424	0.4
Archea/unclassified	13123	4.3
Total	284885	100

Table 2. Summary of taxonomic classification of *Ae. aegypti* WTA sequencing.

First and foremost, no ZIKV reads were detected. While there were *Flavivirus* reads present, these were identified as *Pestivirus*, a virus which causes haemorrhagic syndromes, abortion, and fatal mucosal disease in cattle, sheep, goats, and swine. The greatest proportion of transcripts belonged to *Homo sapiens* (17%) and *Sus scrofa* (14%) and *Ae. aegypti* (9%) (**Figure 7**). The metadata associated with both samples indicated that the mosquitoes had indeed blood-fed prior to collection.

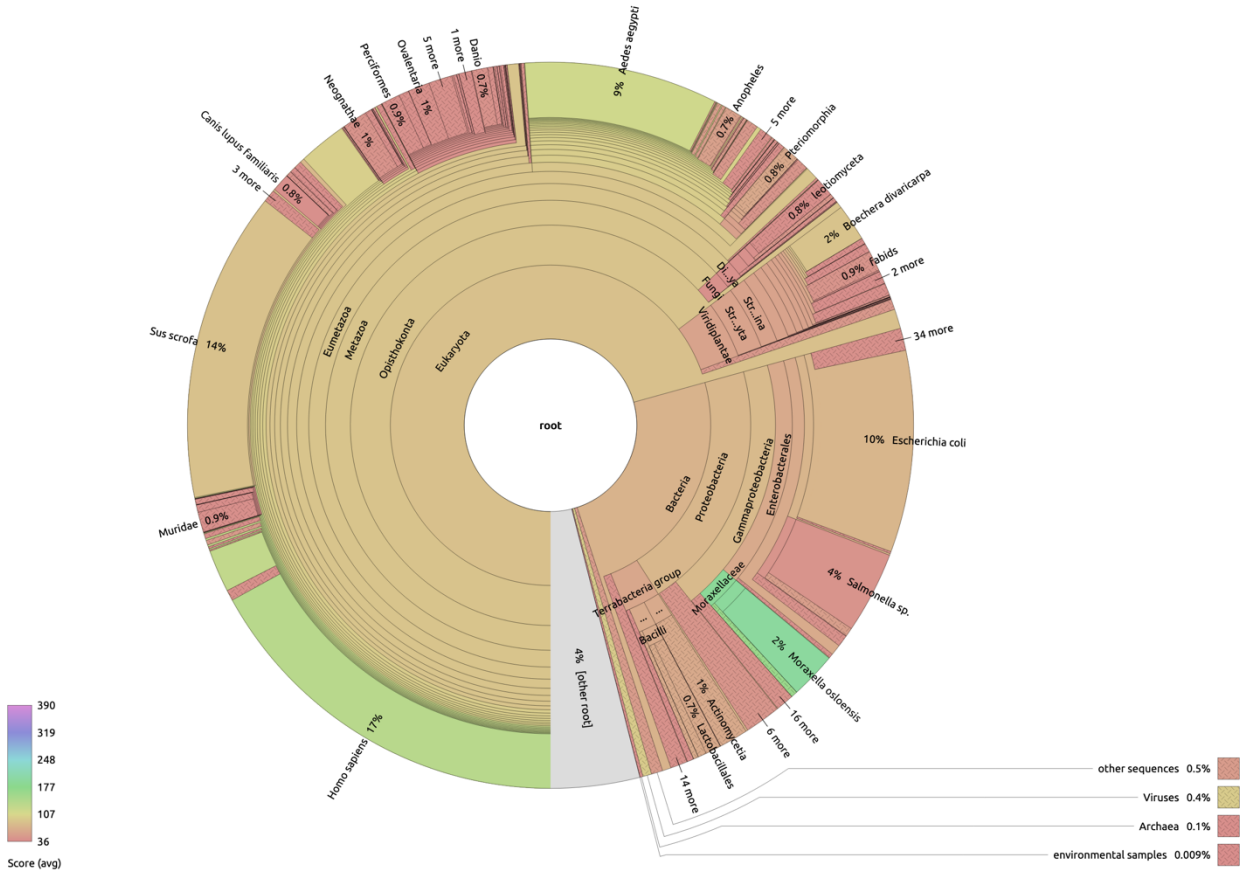


Figure 7. The taxonomic ratios of combined WTA sequencing reads. Classification was performed using Kraken (v 2) software [46], with the full NCBI GenBank nucleotide database (downloaded 24/11/2021).

The most prevalent classification in the virus clade was the Equine infectious anaemia virus (EIAV), which accounted for 45% of all viral reads, and 0.2% of the total reads **(Figure 7)**. EIAV is an arthropod transmitted retrovirus infecting Equidae species, endemic across all continents [59]. While the primary vectors have been reported as insects in the Tabanidae family [60], the role mosquito species play in transmission is still a matter of debate [61–64]. The 496 reads found in this metagenomic analysis were mapped to the reference sequence (NC_001450), which yielded 10-fold coverage of 20.2% the EIAV genome. To scrutinise these data further, the sequences were combined with sequences from isolates with complete geographical metadata from NCBI GenBank (*n*=66), and aligned with the mapped fragment obtained from the metasequencing reads. A phylogenetic tree was built which placed the EIAV sequence obtained from the Cabo Verdean *Ae. aegypti* mosquito into a global geographic context (**Figure 8**). The inference yielded a highly supported tree, consisting of sequences from 9 geographical locations. A monophyletic clade (**A**) (bootstrap support (BS) = 74) consisting exclusively of Chinese EIAV taxa forms the most homogeneous cluster. Separated from clade **A** by a well-supported node (BS=79), a second clade (denoted as **B**; BS=70), contains sequences for Ireland (*n*=23) and Italy (*n*=2). The Cabo Verdean taxa (red) has clustered with a group of sequences from the USA. Although geographic clusters appear, given the limited EIAV sequence data available and the nature of this truncated convenience sampled dataset, there was little evidence from which general geographic inferences can be made. This argument is further supported by the lack of geographic clustering observed in taxa originating from the United Kingdom (blue), where the three taxa have been placed in three separate clades across the tree. A phylogeographic reconstruction of EAIV has been reported previously [65], which despite the lack of available data, inferred introduction and exportation events that indicated events occurring between the USA, Brazil and Ireland between the late 20th and 21st centuries. Given the close ties between Brazil, the USA and Cabo Verde, the close phylogenetic

association observed here is feasible, however further sampling would be required to make any accurate inferences.

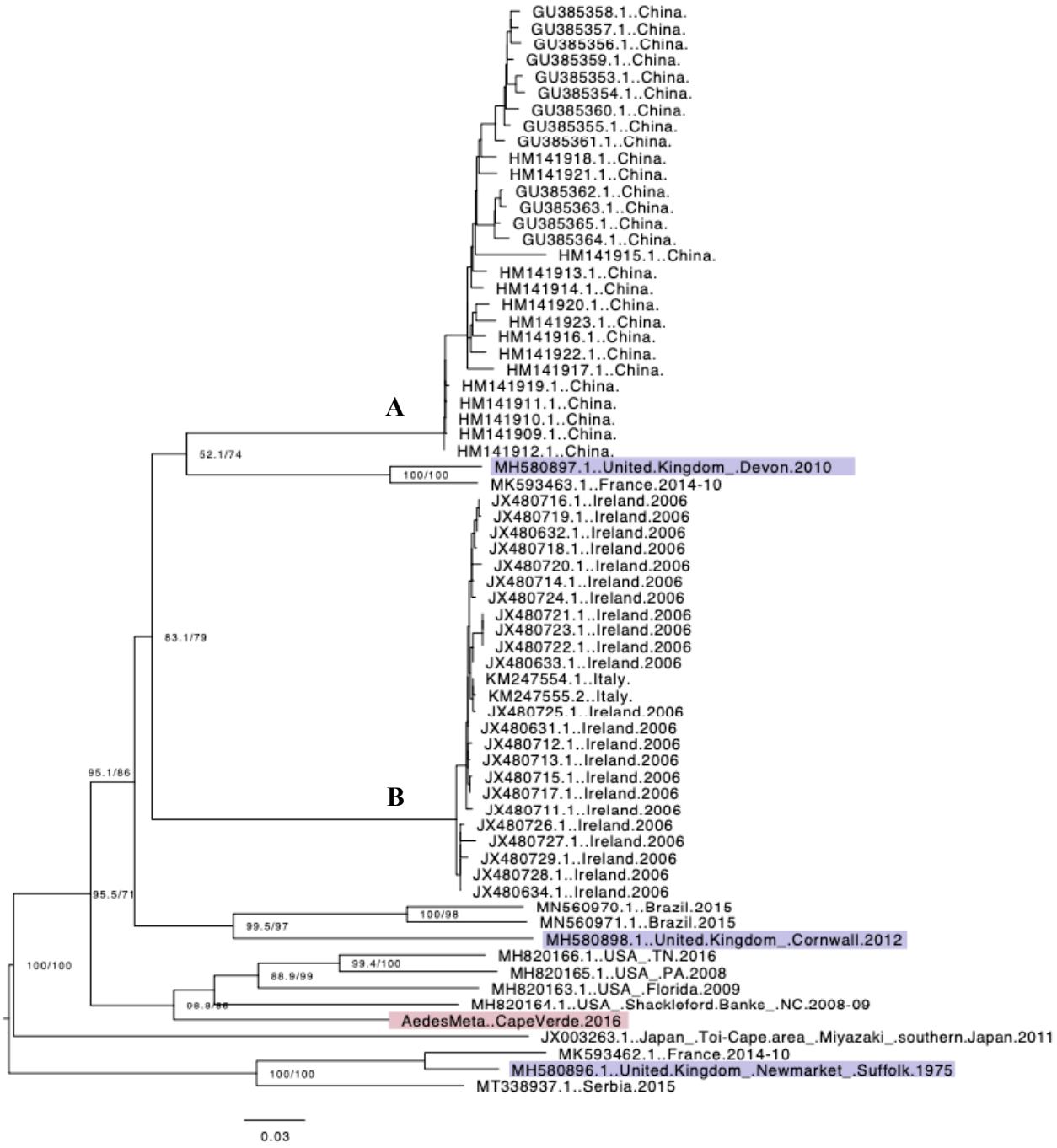


Figure 8. A maximum-likelihood phylogenetic tree inferred from the EIAV dataset (n=66) with the added sequence isolated from the *Ae. aegypti* sample. Numbers in the parentheses indicate (SH-aLRT support (%) / bootstrap support (%)). Sequences highlighted in blue were isolated in the United Kingdom.

3.3.2 Phylogeographic reconstruction of ZIKV in the Americas and Cabo Verde

A single ZIKV genomic sequence was successfully enriched from pooled *Ae. aegypti* RNA by way of multiplex tiling PCR amplification and Oxford Nanopore sequencing. The mosquitoes in the sequenced pool were collected between 17/08/2016 and 25/10/2016 from 18 sites across the Praia municipality. The resultant ZIKV consensus sequence from the mosquito pool was combined across 407 ZIKV whole-genome sequences, including 17 ancestral sequences from French Polynesia, and 3 sequences from Cabo Verde human isolates [32]. This dataset has incorporated date and location of isolation alongside the genomic data, enabling geographical ancestral state reconstruction and temporal inferences to be made. Before building the temporal inference, a maximum likelihood (ML) tree to test for temporal signal in the dataset was inferred (**Figure 9**).

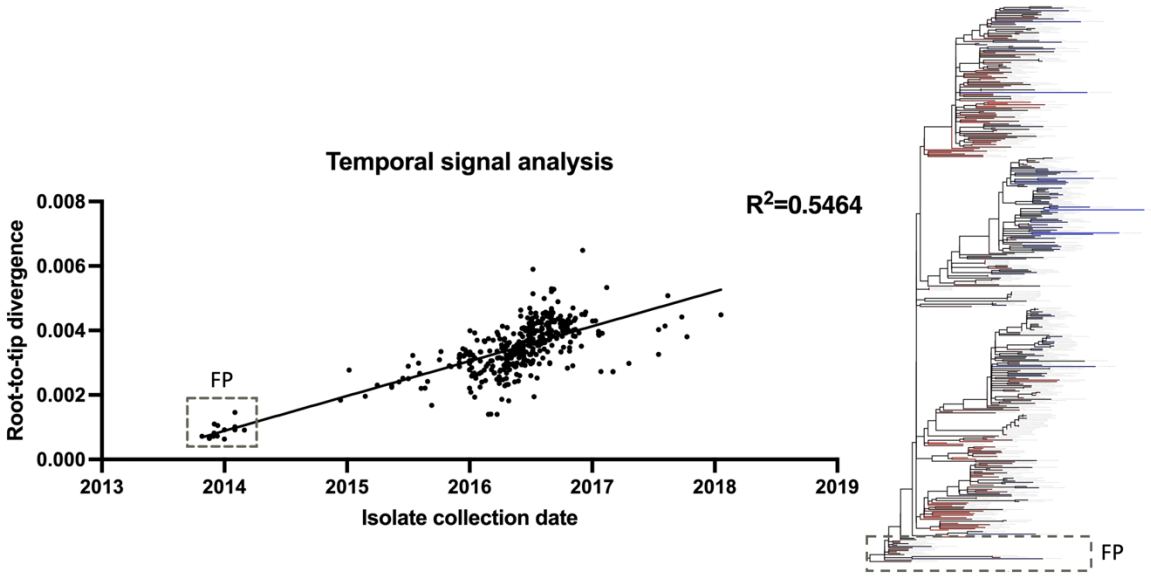


Figure 9. A temporal signal analysis illustrating relationship between genetic divergence and time (temporal signal) and the uniformity of rates across the tree. Analysis was performed using TempEST software. **Left.** Scatterplot of root-to-tip divergence of 407 taxa against collection date, with linear regression. **Right.** An ML tree of 407 taxa with the heuristic residual mean squared (RMSE) colouring of branches. Red = negative RMSE Blue = positive RMSE. FP = French Polynesian taxa.

This analysis confirmed, in concordance with studies of a similar nature, that the mutation rate of the ZIKV dataset was indeed well correlated with its sampling, which enabled the temporal inference of subsequent phylogenies with a strict molecular clock viable (R^2=0.55) (**Figure 7**, left) [18,21,24]. The French Polynesian (FP) strains are highlighted (**Figure 7**), denoting the ancestral clade on which the tree is rooted, which is established practice in the literature given FP's putative role in the introduction of ZIKV to the Americas. **Figure 7** (right) shows the guide tree on which the temporal analysis was performed. The colouration signifies RMSE deviation of branches/taxa from the strict molecular clock. While the divergence shown here is acceptable, the uniformity within clades is nonetheless apparent, which may indicate a differential in selective pressure exerted on these sub-populations or mislabelling of sample isolation dates.

The Bayesian phylogeographic reconstruction of the 408 taxa is shown (**Figure 10**), a summary geographic projection of which is summarised (**Figure 11**). The overall topology of the phylogeny is typical of that which can be found in the literature, where there are between three to four clades, with Brazilian isolates at the basal node of each inferring the exportation and dissemination of discrete ZIKV sub-lineages throughout the Americas over the course of 2015-2016 (posterior = > 0.9) [18,21,24]. The phylogeny also includes ancestral sequences to the American outbreak, isolated in FP. The positioning at the root of the tree infers that the introduction of ZIKV most likely originated in FP, as reported previously. Estimates in the literature as to the approximate introduction of ZIKV to Brazil range from December 2013 and February 2014 [17], which is in concordance with our own estimates (11/2013 – 04/2014 95% HPD; posterior = 1.0).

The ancestry of Cabo Verdean sequences is shown (**Figure 10**; red taxa), where two discrete introduction events are inferred, a subsection of the relevant clade is depicted (**Figure 10, top**). The first introduction (**Figure 7C**) was estimated to have occurred between 05/2014 and 09/2015 (HPD 95%; posterior = 1.0), which is in concordance with previous findings (i.e., 05/2014 and 09/2015; 95% HPD) [32]. The ancestral state at the root of the Cabo Verdean clade is indicated as being of Brazilian ancestry, which is in line with previous findings. Our analysis revealed a second, novel introduction event, which is inferred to be of Puerto Rican origin (**Figure 7D**). The closest taxa to the Cabo Verdean sequence collected in this study is Puerto Rican (collected 12/2015). The ancestral state reconstruction supports this premise (posterior – 0.5). The estimated date of divergence of the ancestral Puerto Rican node is between 05/2015 and 09/2015 (95%HPD; posterior = 0.5), which implies introduction of the second lineage shown here, circulating within the *Ae. aegypti* population sampled between

17/08/2016 and 25/10/2016. This introduction occurred significantly later than the lineage responsible for the primary outbreak. The topology of the Puerto Rican sub-clade indicates a further exportation of ZIKV to the United States Virgin Islands, which is 100 km from Puerto Rico. The observation of multiple ZIKV genotypes exported from discrete locations, circulating in similar geographic and temporal space has been well described and is illustrated well (**Figure 10**) [27,66]**.** A key example of this can be seen in the separation of two Puerto Rican clusters across two clades in this tree: clade two (grey) and three (blue) (**"PR" - Figure 10**). The range of isolate collection dates reported in Puerto Rican sequences in clade three is between 20/10/2016 and 20/01/2017. In clade two, a significant overlap is observed in these dates (01/12/2015 to 03/01/2017), both of which are well sampled, which indicates the presence of multiple discrete lineages circulating simultaneously.

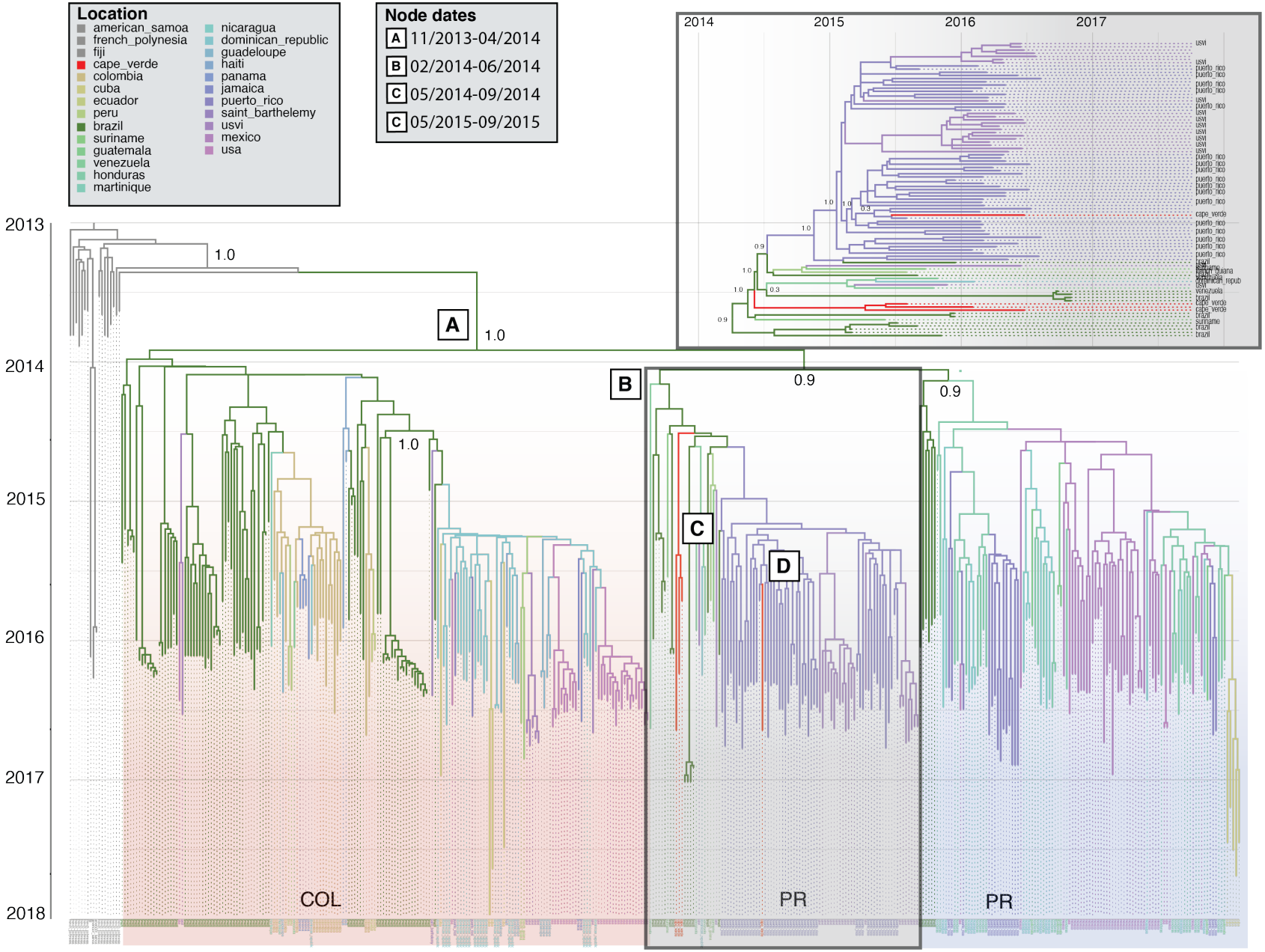


Figure 10. Temporal and phylogeographic reconstruction of 408 ZIKV whole-genome sequences using BEAST software. Node dates indicate the 95% HPD range of geographic ancestral nodes. COL: Colombia, PR: Puerto Rico. Sequences isolated in Cabo Verde are highlighted in red. The tree was coloured and rooted and reordered, increasing by node using FigTree software.

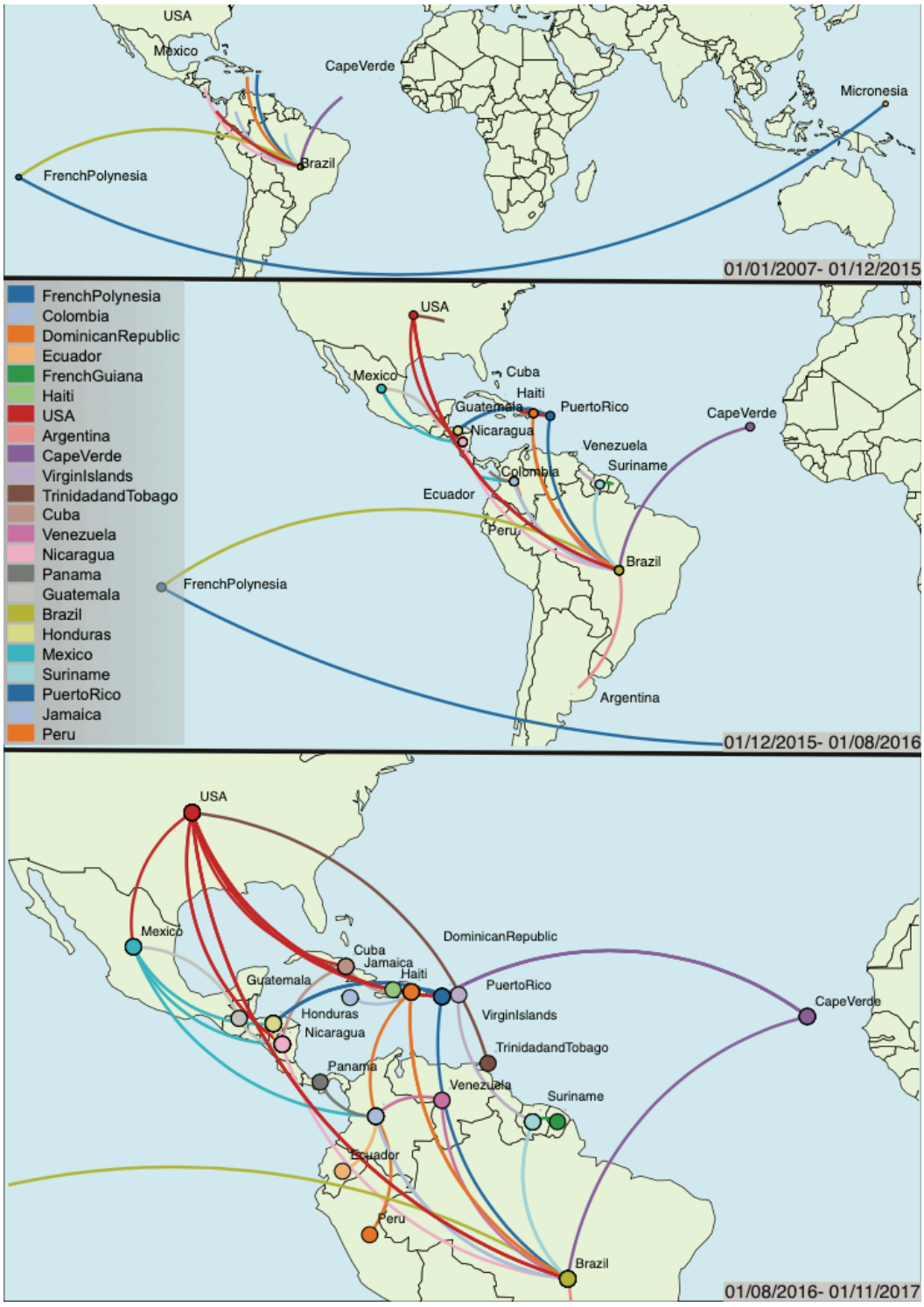


Figure 11. Spatiotemporal summary of the phylogeographic reconstruction (**Figure 10**) illustrates multiple introductions of ZIKV to Cabo Verde. The maps were rendered using SPREAD3 software [75] using the location and node dates. Temporal frames were inferred from the 95% HPD.

The spread of ZIKV across the Americas, as inferred by the temporal phylogeographic reconstruction, is summarised as three phases (**Figure 9**). For this analysis, a further ancestral sequence was added, Micronesia, which illustrates the introduction of ZIKV to FP. In Phase 1, the introduction of ZIKV to FP from Micronesia and subsequently Brazil is shown, where exportation events indicate transmission occurring within six American countries and another to Cabo Verde. In phase two, ten countries have established transmission, including Cabo Verde. In addition to this, secondary transmission events have occurred, through which Columbia plays a central role, where our analysis infers transmission of ZIKV of Colombian descent to Mexico, Panama, Ecuador, and the Dominican Republic. Furthermore, multiple introduction events have occurred into Puerto Rico from both Brazil and Honduras, spawning the two discrete lineages observed (**Figure 10**). In phase three, numerous secondary (non-Brazilian) introduction events have occurred, most notably to the USA, which is inferred to have had 6 discrete introduction events, (in chronological order) from Brazil, Mexico, Nicaragua, Dominican Republic and Puerto Rico. It is important to understand, however, the nature of this dataset, in that sequences designated as USA may be imported infections. Something that was emphasised in Florida, where up to February 2016, 735 travel-associated cases of ZIKV were recorded, whereas fewer than 224 locally-acquired cases (across both Texas and Florida) were reported by the CDC, which reported a basic reproductive rate significantly lower than Central and South American countries at that time ($R_0 = 0.16$) [67,68].

The introduction of ZIKV to Colombia is inferred here to have occurred between 09/2014 and 03/2015 (95% HPD; posterior = 1.0), which is concordant to the interval previously reported (01/2015 – 04/2015 95%HPD)[27]. These inferences, supported by both those in the literature, and our own analysis indicates that the introduction of ZIKV to Cabo Verde occurred before that to Colombia. On the 8th of October 2015, Colombia declared its first outbreak of nine

patients, the first South American country outside of Brazil to report autochthonous transmission. Retrospective analyses on archived samples have identified ZIKV in human serum from June 2015 [69]. Three days prior to the data coming out of Colombia, Cabo Verde reported suspected ZIKV transmission which was later verified by sequencing at Institut Pasteur de Dakar. The epidemiological curve summarising the number of cases in the ZIKV outbreaks in Brazil, Colombia and Cabo Verde is shown (**Figure 12**). The peak of the outbreak in Cabo Verde occurred fifteen weeks prior to that of Colombia.

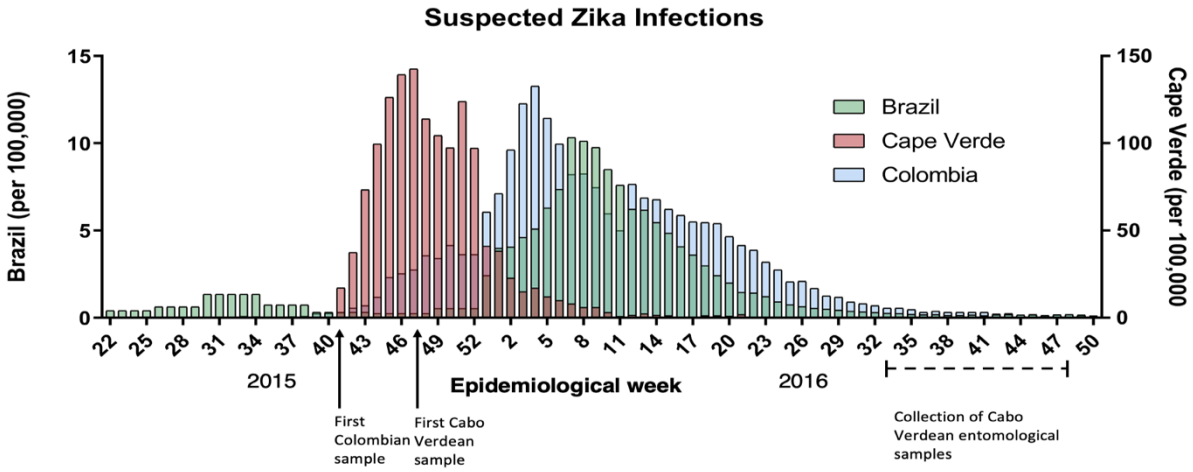


Figure 12. Epidemiological curve of ZIKV cases per 100,000 people in Brazil, Colombia, and Cabo Verde. Data was obtained from PAHO. **Left Y axis:** Brazil and Colombia cases per 100,000 population. **Right Y axis:** Cabo Verde cases per 100,000 population. Markers indicate the first genomic sequences isolated for Colombia and Cabo Verde, as well as our own, captured from *Ae. aegypti.*

3.3.3 Analysis of ZIKV sequence data – SNPs and phylogenetic characteristics

Further analysis of single nucleotide polymorphism (SNP) data based on variant positions in the 408-sequence alignment used to infer the phylogeny presented (**Figure 10**) is shown (**Figure 13**). In this analysis the dissection of 'clade defining' variants are illustrated. For

example, the variants shown in positions A1038G and C2517T are found almost exclusively to be associated with the ancestral FP lineage. Interestingly the C2517T mutation is detected two further times in this dataset, in sequences isolated in Nicaragua and Honduras approximately 2.5 years following its last detection. The C1261T mutation is indicated to be unique to the first Colombian clade, however the second phylogenetically discrete lineage does not possess this SNP, implying a distinct ancestral history to the co-circulating strains. The Cabo Verdean sequence here, as reported in the previous section, clusters strongly with sequences from Puerto Rico but also shares a close common ancestor with the Cabo Verdean clade (**Figure 10, top**). The SNP data displayed at positions T209C, T642C, G1050T, A1362G, T5100C, T6629A and T7258C, justify the placement of the novel Cabo Verdean sample with the Puerto Rican clade. Shared amongst the Cabo Verdean clade and Puerto Rico are SNPs in the positions T3456C and C7258T exclusively, and C9296T which also occurs in the first Colombian clade. The novel Cabo Verdean sequence possesses two unique non-synonymous mutations, V650L (E protein) and G1691W (NS3). The S139N (prM), non-synonymous mutation with a purported role in viral neurotropism is indicated (**Figure 13**) as being fixed throughout the dataset, which is in concordance with previous reports [23]. Likewise, the A188V (NS1) mutation is also fixed in this population.

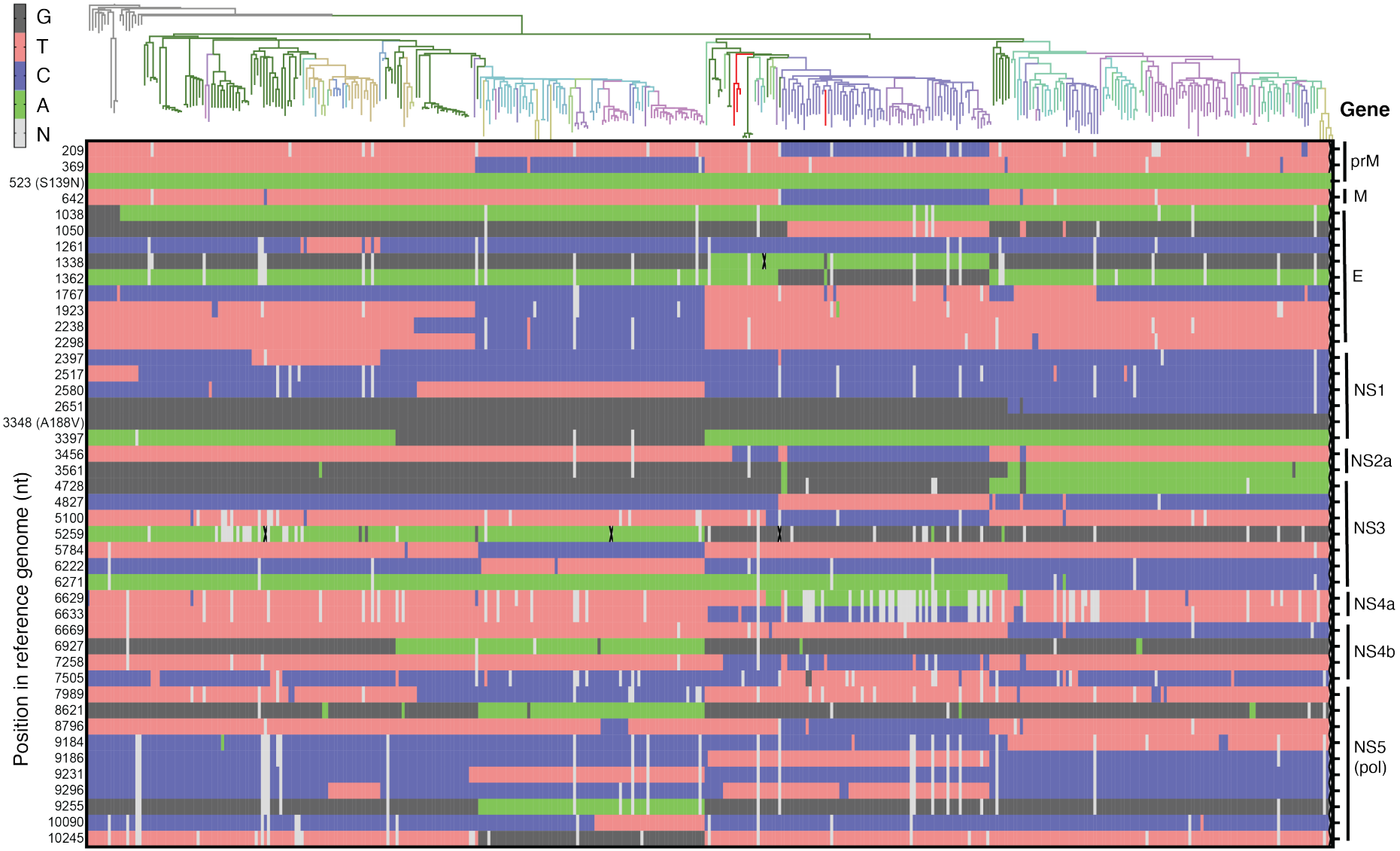


Figure 13. Clade-defining SNPs of the phylogenetic reconstruction. 45 variant positions extracted from the 408 aligned ZIKV whole-genome sequences used to build the phylogeny featured in **Figure 10**. The sequences were sorted to match the order of the phylogeny. The gene in the ZIKV polyprotein in which each position is found is shown on the right. S139N and A188V have been added for illustrative purposes.

3.4 Discussion

In this chapter, the analyses performed describe the virome present in *Ae. aegypti* populations on Cabo Verde shortly after the outbreak in 2016. While a greater number of mosquitoes testing positive for ZIKV was expected, it was believed that the effect of the timing of sampling, coupled with the low rate of transmission at the time was accurately reflected in our results. That said, in this context of low-yield and complex RNA sample analytes, molecular assays are pushed to their upper limit of sensitivity and so the *true* prevalence may be difficult to assess. Our methodologies were comparable to those used previously [43,70,71], however our approach in extracting nucleic acids from only single mosquitoes, and not pools, will be subject to revision in subsequent follow up analyses.

The detection of the EIAV in *Ae. aegypti* samples was not expected. The vectors currently ascribed to EIAV do not emphasise mosquito borne transmission, and therefore, these findings require further investigation. The process of taxonomic classification can produce erroneous results due the vast redundancy between species' on the genomic level, transposable and foreign genomic integrations and incomplete reference sequence databases [72,73]. Furthermore, the confidence score of the classifications was poor in most cases, except for the *H. sapiens* reads. Therefore, the results described in this chapter should be interpreted with these caveats in mind. In doing so, the identification of the numerous animal species in the mosquito sample is brought into question, especially given *Ae. aegypti* mosquito's anthophillic feeding behaviours. Finally, the identification of *Wolbachia* is again, questionable, given the knowledge that these bacteria do not occur naturally in *Ae. aegypti* mosquitoes and *Wolbachia* population control efforts have not been implemented on Cabo Verde. Understanding the origin of these classifications requires further research.

Moving forward, a similar enrichment approach will be applied, combined with ZIKV, here, to EIAV, enabling the acquisition of a whole-genome sequences, which would further enhance our analysis, before reporting our findings. As discussed, with the shortcomings of the taxonomic classification process, further steps are required to validate the origin of the EIAV sequences as it is possible that this may be a historic genomic integration and not viral RNA. These data were acquired using nanopore long-read sequencing, therefore an investigation exploring the context of the EIAV sequence data, for example, verifying the presence of flanking *Ae. Aegypti* genomic DNA sequences, would be a prudent first step.

My aim, in order to confirm the hypothesis of multiple discrete lineages circulating in Cabo Verde, is to isolate at least one further ZIKV genome sequence from our entomological dataset. Pooling multiple samples form a single locale and concentrating the cDNA for multiplex tiling PCR worked well to improve the low success rate of the PCR enrichment of ZIKV RT-qPCR positive samples. Therefore, applying this strategy should enable the acquisition of further ZIKV whole-genome data.

3.5 References

1. Dick GW., Kitchen S., Haddow A. Zika Virus (I). Isolations and serological specificity. Trans R Soc Trop Med Hyg [Internet]. Oxford University Press; 1952 [cited 2018 Aug 14];46:509–20. Available from: https://academic.oup.com/trstmh/article-lookup/doi/10.1016/0035-9203(52)90042-4

2. Dick GWA, Kitchen SF, Haddow AJ. Zika virus (II). Pathogenicity and physical properties. Trans R Soc Trop Med Hyg. 1952;46.

3. Smithburn KC. Neutralizing antibodies against certain recently isolated viruses in the sera of human beings residing in East Africa. J Immunol. Am Assoc Immnol; 1952;69:223–34.

4. Moore D áL, Causey OR, Carey DE, Reddy S, Cooke AR, Akinkugbe FM, *et al.* Arthropod-borne viral infections of man in Nigeria, 1964–1970. Ann Trop Med Parasitol. Taylor & Francis; 1975;69:49–64.

5. Robin Y, Mouchet J. Serological and entomological study on yellow fever in Sierra Leone. Bull Soc Pathol Exot Filiales. 1975;68:249–58.

6. Renaudet J, Jan C, Ridet J, Adam C, Robin Y. A serological survey of arboviruses in the human population of Senegal. Bull Soc Pathol Exot Filiales. 1978;71:131–40.

7. Marchette NJ, Garcia R, Rudnick A. Isolation of Zika virus from Aedes aegypti mosquitoes in Malaysia. Am J Trop Med Hyg. 1969;18.

8. Saluzzo JF, Gonzalez JP, Herve JP, Georges AJ. Serological survey for the prevalence of certain arboviruses in the human population of the south-east area of Central African Republic (author's transl). Bull Soc Pathol Exot Filiales. 1981;74:490–9.

9. Olson JG, Ksiazek TG. Zika virus, a cause of fever in Central Java, Indonesia. Trans R Soc Trop Med Hyg. Elsevier; 1981;75:389–93.

10. Darwish MA, Hoogstraal H, Roberts TJ, Ahmed IP, Omar F. A sero-epidemiological survey for certain arboviruses (Togaviridae) in Pakistan. Trans R Soc Trop Med Hyg. Royal Society of Tropical Medicine and Hygiene; 1983;77:442–5.

11. Duffy MR, Chen T-H, Hancock WT, Powers AM, Kool JL, Lanciotti RS, *et al.* Zika Virus Outbreak on Yap Island, Federated States of Micronesia. N Engl J Med. Massachusetts Medical Society; 2009;360:2536–43.

12. Lanciotti RS, Kosoy OL, Laven JJ, Velez JO, Lambert AJ, Johnson AJ, *et al.* Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. Emerg Infect Dis. Centers for Disease Control and Prevention; 2008;14:1232.

13. Haddow AD, Schuh AJ, Yasuda CY, Kasper MR, Heang V, Huy R, *et al.* Genetic characterization of Zika virus strains: geographic expansion of the Asian lineage. PLoS Negl

Trop Dis. Public Library of Science San Francisco, USA; 2012;6:e1477.

14. Tognarelli J, Ulloa S, Villagra E, Lagos J, Aguayo C, Fasce R, *et al.* A report on the outbreak of Zika virus on Easter Island, South Pacific, 2014. Arch Virol. Springer; 2016;161:665–8.

15. Roth A, Mercier A, Lepers C, Hoy D, Duituturaga S, Benyon E, *et al.* Concurrent outbreaks of dengue, chikungunya and Zika virus infections–an unprecedented epidemic wave of mosquito-borne viruses in the Pacific 2012–2014. Eurosurveillance. European Centre for Disease Prevention and Control; 2014;19:20929.

16. Zhang Q, Sun K, Chinazzi M, Pastore Y Piontti A, Dean NE, Rojas DP, *et al.* Spread of Zika virus in the Americas. Proc Natl Acad Sci U S A [Internet]. National Academy of Sciences; 2017 [cited 2018 Apr 24];114:E4334–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28442561

17. MASSAD E, BURATTINI MN, KHAN K, STRUCHINER CJ, COUTINHO FAB, WILDER-SMITH A. On the origin and timing of Zika virus introduction in Brazil. Epidemiol Infect. Cambridge University Press; 2017;145:2303–12.

18. Faria NR, Azevedo R do S da S, Kraemer MUG, Souza R, Cunha MS, Hill SC, *et al.* Zika virus in the Americas: Early epidemiological and genetic findings. Science [Internet]. American Association for the Advancement of Science; 2016 [cited 2018 Sep 19];352:345–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27013429

19. Mlakar J, Korva M, Tul N, Popović M, Poljšak-Prijatelj M, Mraz J, *et al.* Zika virus associated with microcephaly. N Engl J Med. Mass Medical Soc; 2016;374:951–8.

20. Cauchemez S, Besnard M, Bompard P, Dub T, Guillemette-Artur P, Eyrolle-Guignot D, *et al.* Association between Zika virus and microcephaly in French Polynesia, 2013–15: a retrospective study. Lancet [Internet]. Elsevier; 2016 [cited 2018 Aug 15];387:2125–32. Available from:

https://www.sciencedirect.com/science/article/pii/S0140673616006516?via%3Dihub

21. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, *et al.* Zika virus evolution and spread in the Americas. Nature [Internet]. Nature Publishing Group; 2017;546:411–5. Available from: http://www.nature.com/doifinder/10.1038/nature22402

22. Xia H, Luo H, Shan C, Muruato AE, Nunes BTD, Medeiros DBA, *et al.* An evolutionary NS1 mutation enhances Zika virus evasion of host interferon induction. Nat Commun [Internet]. 2018;9:414. Available from: https://doi.org/10.1038/s41467-017-02816-2

23. Ling Yuan, Xing-Yao Huang C-FQ. A single mutation in the prM protein of Zika virus contributes to fetal microcephaly. Science (80-). 2017;7120.

24. Faria NR, Quick J, Morales I, Theze J, Jesus JG de, Giovanetti M, *et al.* Epidemic establishment and cryptic transmission of Zika virus in Brazil and the Americas. bioRxiv [Internet]. Cold Spring Harbor Laboratory; 2017 [cited 2018 Sep 2];105171. Available from: https://www.biorxiv.org/content/early/2017/03/27/105171

25. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, *et al.* Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature [Internet]. Nature Publishing Group; 2017;546:401–5. Available from: http://www.nature.com/doifinder/10.1038/nature22400

26. Assis FL, Sippert E, Rocha BC, Volkova E, Fares-Gusmao R, Ok S, *et al.* Genomic and Phylogenetic Analysis of Zika Virus Isolates from Asymptomatic Blood Donors in the United States and Puerto Rico, 2016. Am J Trop Med Hyg. The American Society of Tropical Medicine and Hygiene; 2020;102:880.

27. Black A, Moncla LH, Laiton-Donato K, Potter B, Pardo L, Rico A, *et al.* Genomic epidemiology supports multiple introductions and cryptic transmission of Zika virus in Colombia. BMC Infect Dis [Internet]. 2019;19:963. Available from: https://doi.org/10.1186/s12879-019-4566-2

28. Ministério da Saúde. Ministério da Saúde confirma infecção por Vírus Zika no concelho da Praia [Internet]. 2015. Available from: http://www.minsaude.gov.cv/index.php/rss-noticias/912-ministerio-da-saude-confirma-infeccao-por-virus-zika-no-concelho-da-praia

29. Dia I, Diagne C, Ba Y, Diallo D, Konate L, Diallo M. Insecticide susceptibility of Aedes aegypti populations from Senegal and Cape Verde Archipelago. Parasit Vectors [Internet]. BioMed Central; 2012 [cited 2018 Sep 21];5:238. Available from: http://parasitesandvectors.biomedcentral.com/articles/10.1186/1756-3305-5-238

30. Lourenço J, Monteiro M, Tomás T, Monteiro Rodrigues J, Pybus O, Rodrigues Faria N. Epidemiology of the Zika Virus Outbreak in the Cabo Verde Islands, West Africa. PLoS Curr [Internet]. Public Library of Science; 2018 [cited 2018 Apr 27]; Available from: http://currents.plos.org/outbreaks/?p=79551

31. Kindhauser MK, Allen T, Frank V, Santhana RS, Dye C. Zika: the origin and spread of a mosquito-borne virus. Bull World Health Organ [Internet]. 2016 [cited 2018 Apr 27];94:675-686C. Available from: http://www.who.int/entity/bulletin/volumes/94/9/16-171082.pdf

32. Faye O, de Lourdes Monteiro M, Vrancken B, Prot M, Lequime S, Diarra M, *et al.* Genomic Epidemiology of 2015–2016 Zika Virus Outbreak in Cape Verde. Emerg Infect Dis [Internet]. Centers for Disease Control and Prevention; 2020 [cited 2021 Nov 12];26:1084. Available from: /pmc/articles/PMC7258482/

33. Hill SC, Vasconcelos J, Neto Z, Jandondo D, Zé-Zé L, Aguiar RS, *et al.* Emergence of the Asian lineage of Zika virus in Angola: an outbreak investigation. Lancet Infect Dis. Elsevier; 2019;19:1138–47.

34. Musso D, Lanteri MC. Emergence of Zika virus: where does it come from and where is it going to? Lancet Infect Dis. Elsevier; 2017;17:255.

35. Li T, Mbala-Kingebeni P, Naccache SN, Thézé J, Bouquet J, Federman S, *et al.* Metagenomic next-generation sequencing of the 2014 Ebola virus disease outbreak in the

Democratic Republic of the Congo. J Clin Microbiol. Am Soc Microbiol; 2019;57:e00827-19.

36. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, *et al.* Real-time, portable genome sequencing for Ebola surveillance. Nature [Internet]. Nature Publishing Group; 2016 [cited 2018 Sep 22];530:228–32. Available from: http://www.nature.com/articles/nature16996

37. Faria NR, Sabino EC, Nunes MRT, Alcantara LCJ, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil. Genome Med. 2016;8:97.

38. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc [Internet]. 2017;12:1261–76. Available from: http://www.nature.com/doifinder/10.1038/nprot.2017.066

39. Ng DHL, Ho HJ, Chow A, Wong J, Kyaw WM, Tan A, *et al.* Correlation of clinical illness with viremia in Zika virus disease during an outbreak in Singapore. BMC Infect Dis [Internet]. 2018;18:301. Available from: https://doi.org/10.1186/s12879-018-3211-9

40. Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, Teuwen DE, *et al.* 1970s and 'Patient 0'HIV-1 genomes illuminate early HIV/AIDS history in North America. Nature. Nature Publishing Group; 2016;539:98–101.

41. Ayres CFJ, Guedes DRD, Paiva MHS, Morais-Sobral MC, Krokovsky L, Machado LC, *et al.* Zika virus detection, isolation and genome sequencing through Culicidae sampling during the epidemic in Vitória, Espírito Santo, Brazil. Parasit Vectors [Internet]. 2019;12:220. Available from: https://doi.org/10.1186/s13071-019-3461-4

42. Guedes DRD, Gomes ETB, Paiva MHS, Melo-Santos MAV de, Alves J, Gómez LF, *et al.* Circulation of DENV2 and DENV4 in Aedes aegypti (Diptera: Culicidae) mosquitoes from Praia, Santiago Island, Cabo Verde. J Insect Sci [Internet]. 2017;17:86. Available from:

https://doi.org/10.1093/jisesa/iex057

43. Cevallos V, Ponce P, Waggoner JJ, Pinsky BA, Coloma J, Quiroga C, *et al.* Zika and Chikungunya virus detection in naturally infected Aedes aegypti in Ecuador. Acta Trop. Elsevier; 2018;177:74–80.

44. Kosoltanapiwat N, Tongshoob J, Singkhaimuk P, Nitatsukprasert C, Davidson SA, Ponlawat A. Entomological Surveillance for Zika and Dengue Virus in Aedes Mosquitoes: Implications for Vector Control in Thailand. Pathog. . 2020.

45. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb genomics. Microbiology Society; 2017;3.

46. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol [Internet]. 2019;20:257. Available from: https://doi.org/10.1186/s13059-019-1891-0

47. Martí JM. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. PLOS Comput Biol [Internet]. Public Library of Science; 2019;15:e1006967. Available from: https://doi.org/10.1371/journal.pcbi.1006967

48. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics [Internet]. 2018;34:3094–100. Available from: https://doi.org/10.1093/bioinformatics/bty191

49. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, *et al.* Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021;10:giab008. Available from: https://doi.org/10.1093/gigascience/giab008

50. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics [Internet]. 2010;26:841–2. Available from: https://doi.org/10.1093/bioinformatics/btq033

51. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol [Internet]. 2015;32:268–74. Available from: https://doi.org/10.1093/molbev/msu300

52. Rambaut A. FigTree v1. 3.1 (http://tree. bio. ed. ac. uk/software/figtree/) Institute of Evolutionary Biology. Univ Edinburgh, Edinburgh, United Kingdom. 2010;

53. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res. Oxford University Press; 2015;43:D571–7.

54. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol [Internet]. 2013;30:772–80. Available from: https://doi.org/10.1093/molbev/mst010

55. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol [Internet]. 2016 [cited 2019 Nov 25];2:vew007. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27774300

56. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol [Internet]. Oxford University Press; 2018 [cited 2018 Aug 21];4. Available from: https://academic.oup.com/ve/article/doi/10.1093/ve/vey016/5035211

57. Hill V, Baele G. Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model. Mol Biol Evol. Oxford University Press; 2019;36:2620–8.

58. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Susko E, editor. Syst Biol [Internet]. Narnia; 2018 [cited 2019 Nov 25];67:901–4. Available from: https://academic.oup.com/sysbio/article/67/5/901/4989127

59. Leroux C, Cadoré J-L, Montelaro RC. Equine Infectious Anemia Virus (EIAV): what has HIV's country cousin got to tell us? Vet Res. EDP Sciences; 2004;35:485–512.

60. Barros ATM, Foil LD. The influence of distance on movement of tabanids (Diptera: Tabanidae) between horses. Vet Parasitol. Elsevier; 2007;144:380–4.

61. Breaud TP, Steelman CD, Roth EE, Adams Jr W V. Apparent propagation of the equine infectious anemia virus in a mosquito (Culex pipiens quinquefasciatus Say) ovarian cell line. Am J Vet Res. 1976;37:1069–70.

62. Stein CD, Lotze JC, Mott LO. Evidence of Transmission of inapparent (subclinical) Form of Equine Infectious Anemia by Mosquitoes (Psorophora columbiae), and by Injection of the Virus in extremely high Dilution. J Am Vet Med Assoc. Chicago, Ill.; 1943;102.

63. Cupp EW, Kemen MJ. The role of stable flies and mosquitoes in the transmission of equine infectious anemia virus. Proc United States Anim Heal Assoc. 1980. p. 362–7.

64. Williams DL, Issel CJ, Steelman CD, Adams Jr W V, Benton C V. Studies with equine infectious anemia virus: transmission attempts by mosquitoes and survival of virus on vector mouthparts and hypodermic needles, and in mosquito tissue culture. Am J Vet Res. 1981;42:1469–73.

65. Jara M, Frias-De-Diego A, Machado G. Phylogeography of equine infectious anemia virus. Front Ecol Evol. Frontiers; 2020;8:127.

66. Santiago GA, Kalinich CC, Cruz-López F, González GL, Flores B, Hentoff A, *et al.* Tracing the Origin, Spread, and Molecular Evolution of Zika Virus in Puerto Rico, 2016–2017. Emerg Infect Dis. Centers for Disease Control and Prevention; 2021;27:2971.

67. Dinh L, Chowell G, Mizumoto K, Nishiura H. Estimating the subcritical transmissibility of the Zika outbreak in the State of Florida, USA, 2016. Theor Biol Med Model [Internet]. 2016;13:20. Available from: https://doi.org/10.1186/s12976-016-0046-1

68. Statistics and Maps | Zika virus | CDC [Internet]. [cited 2022 Mar 11]. Available from: https://www.cdc.gov/zika/reporting/index.html

69. Pacheco O, Beltrán M, Nelson CA, Valencia D, Tolosa N, Farr SL, *et al.* Zika virus disease in Colombia—preliminary report. N Engl J Med. Mass Medical Soc; 2020;383:e44.

70. Paiva MHS, Guedes DRD, Krokovsky L, Machado LC, Rezende TMT, de Morais Sobral

MC, *et al.* Sequencing of ZIKV genomes directly from Ae. aegypti and Cx. quinquefasciatus mosquitoes collected during the 2015–16 epidemics in Recife. Infect Genet Evol. Elsevier; 2020;80:104180.

71. Parra MCP, Lorenz C, de Aguiar Milhim BHG, Dibo MR, Guirado MM, Chiaravalloti-Neto F, *et al.* Detection of Zika RNA virus in Aedes aegypti and Aedes albopictus mosquitoes, São Paulo, Brazil. Infect Genet Evol. Elsevier; 2022;105226.

72. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci. National Acad Sciences; 2016;113:5970–5.

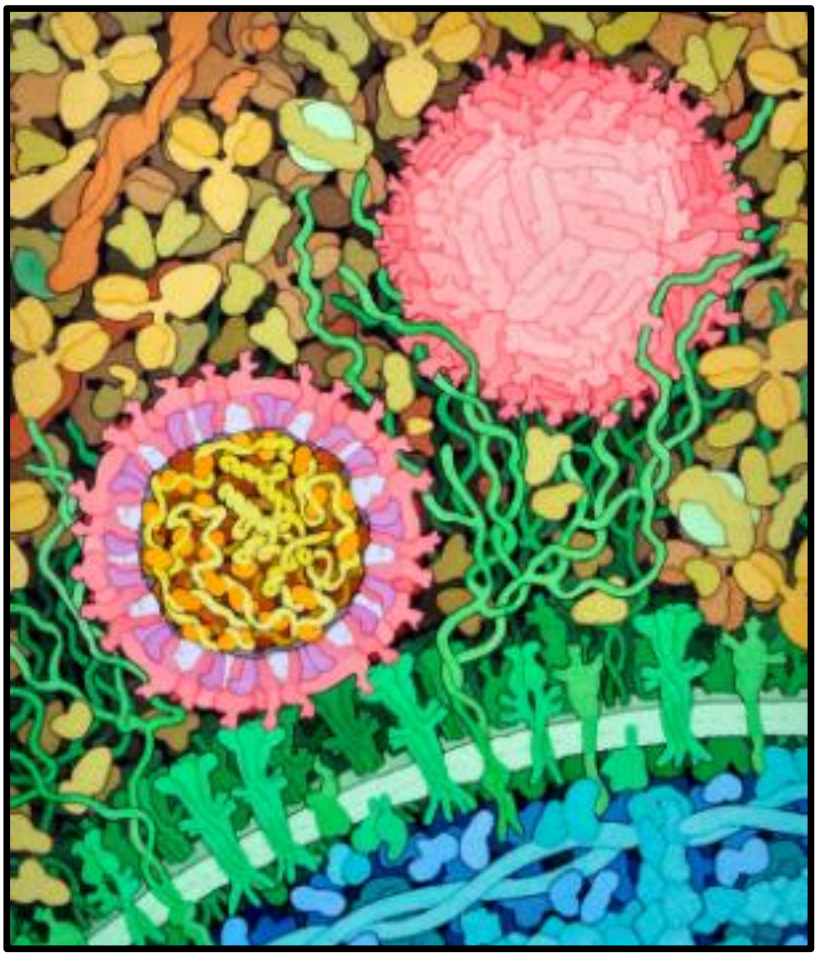
73. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol [Internet]. 2017;18:182. Available from: https://doi.org/10.1186/s13059-017-1299-7

74. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics [Internet]. Narnia; 2009 [cited 2019 Nov 25];25:1189–91. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp033

75. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. Mol Biol Evol. Society for Molecular Biology and Evolution; 2016;33:2167–9.

Chapter Four

Sero-epidemiological study of arbovirus infection following the 2015-2016 Zika virus outbreak in Cabo Verde



"Zika virus is shown in cross section at centre left. On the outside, it includes envelope protein (red) and membrane protein (magenta) embedded in a lipid membrane (light purple). Inside, the RNA genome (yellow) is associated with capsid proteins (orange). The viruses are shown interacting with receptors on the cell surface (green) and are surrounded by blood plasma molecules at the top." – RSCB (Creative Commons Attribution)

4.1 RESEARCH PAPER COVER SHEET

SECTION A – Student Details

Student ID Number	**1603403**	**Title**	Mr
First Name(s)	Daniel		
Surname/Family Name	Ward		
Thesis Title	Sero-epidemiological study of arbovirus infection following the 2015-2016 Zika virus outbreak in Cabo Verde		
Primary Supervisor	Prof. Taane Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	**Scientific Reports**		
When was the work published?	**July 2022**		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	**No**		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	**Samples were collected by ARG. DW designed the experiments, performed all serological and statistical analyses. DW wrote the manuscript in its entirety.**

SECTION E

Student Signature [redacted]

Date 20/04/2022

Supervisor Signature [redacted]

Date April 20, 2022

scientific reports

Check for updates

OPEN

Sero-epidemiological study of arbovirus infection following the 2015–2016 Zika virus outbreak in Cabo Verde

Daniel Ward[1], Ana Rita Gomes[2], Kevin K. A. Tetteh[1], Nuno Sepúlveda[3,4], Lara Ferrero Gomez[5,6], Susana Campino[1,6] & Taane G. Clark[1,6]✉

In November 2015, cases of Zika virus infection were recorded in Cabo Verde (Africa), originating from Brazil. The outbreak subsided after seven months with 7580 suspected cases. We performed a serological survey (n = 431) in Praia, the capital city, 3 months after transmission ceased. Serum samples were screened for arbovirus antibodies using ELISA techniques and revealed seroconverted individuals with Zika (10.9%), dengue (1–4) (12.5%), yellow fever (0.2%) and chikungunya (2.6%) infections. Zika seropositivity was predominantly observed amongst females (70%). Using a logistic model, risk factors for increased odds of Zika seropositivity included age, self-reported Zika infection, and dengue seropositivity. Serological data from Zika and dengue virus assays were strongly correlated (Spearman's $r_s = 0.80$), which reduced when using a double antigen binding ELISA (Spearman's $r_s = 0.54$). Overall, our work improves an understanding of how Zika and other arboviruses have spread throughout the Cabo Verde population. It also demonstrates the utility of serological assay formats for outbreak investigations.

In February 2016, a public health emergency of international concern was declared due to a prolonged Zika virus epidemic (ZIKV) spreading rapidly across the Americas. There were infections implicated in causing microcephaly or congenital Zika syndrome manifestations in 2–50% of children exposed prenatally[1,2]. The spread of ZIKV also coincided with an increase in the incidence of Guillain–Barré syndrome[3,4]. Before the arrival and spread of ZIKV across the Americas[5–7], outbreaks were reported in the Pacific islands of Yap (73% of population exposed)[8], then French Polynesia (2013–2014; 30k cases; 11.5% of population), followed by further cases on New Caledonia, the Cook Islands and Easter Island[9,10]. Phylogenetic analysis revealed that early Brazilian strains are ancestral to a French Polynesian strain (KJ776791), indicating that the ZIKV circulating in the Americas was most likely exported from the Pacific island region[11]. Before the end of year 2015, eleven countries in the Americas had reported PCR positive ZIKV cases. As the cases of ZIKV climbed in South America, a lesser-reported intercontinental transmission event was suspected in November 2015, on Cabo Verde, an archipelago located off the coast of Senegal, Africa, with strong economic, travel and tourism links to Brazil. To date, at least 87 countries worldwide have reported ZIKV transmission[12], with > 580k suspected cases of infection[13].

ZIKV is a positive-sense single stranded RNA virus, with a 10.8 kbp non-segmented genome. It belongs to the genus *Flavivirus,* which includes common human arboviruses such as dengue virus (DENV), West Nile virus (WNV) and yellow fever virus (YFV), which alongside the *Alphavirus* chikungunya (CHIKV), are transmitted by the urban-adapted anthropophilic *Aedes aegypti* mosquito vector. Co-infection of arboviruses has been reported widely[14], including between ZIKV and CHIKV, which is thought to enhance vector transmission[15] and ZIKV pathogenicity[16]. The differential diagnoses of such arboviral infections are complex given their common early-stage presentation in patients. Molecular diagnostics exhibit very high specificity, however, these are only sensitive during the acute stages of infection and can be logistically prohibitive in low-resourced endemic settings[17]. With this, serology has formed an integral role in arbovirus diagnostic and surveillance strategies[18,19], although, viral antigens from related species can result in immune response cross reactivity, which can confound

[1]Department of Infection Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]Université de Montpellier, Montpellier, France. [3]Warsaw University of Technology, Warsaw, Poland. [4]Universidade de Lisboa, Lisbon, Portugal. [5]Universidade Jean Piaget (UniPiaget), Praia, Cabo Verde. [6]These authors contributed equally: Lara Ferrero Gomez, Susana Campino and Taane G. Clark. ✉email: taane.clark@lshtm.ac.uk

assay specificity[20,21]. The use of non-structural proteins, such as *Flavivirus* NS1, in antibody detection assays for disease surveillance has been reported widely, and found to exhibit reduced cross reactivity[22,23]. Further, methods for increasing immunoassay specificity exist. These include blockade-of-binding and double antigen binding (DAB) assays, both of which detect all immunoglobulin (Ig) isotypes and have been shown to reduce non-specific antibody detection, specifically in DENV endemic regions[24–26].

Cabo Verde has experienced numerous recent epidemics, including dengue fever (2009), Zika (2015) and malaria (2017)[27–29]. With respect to ZIKV, in November 2015, public health surveillance in Cabo Verde identified an increase in patients presenting with rashes, conjunctivitis, and myalgia[30]. By the time the outbreak concluded in May 2016, 7580 suspected cases of ZIKV infection and 18 cases of ZIKV congenital syndrome were recorded[31], with 1.4% of the population (size 531k) infected. The estimated reproductive rate (R_0) was 1.9 (95% CI 1.5–2.2)[31]. Praia, the capital city, had the greatest reported rate of ZIKV transmission in the country. Phylogenetic analysis has shown that Cabo Verdean 2015/2016 isolates cluster closely with those sourced from Brazil, suggesting that ZIKV was introduced from the Americas, and the outbreak was not of African origin[28]. Serological investigations on suspected ZIKV patients (n = 1226) recruited from clinics during the epidemic revealed that 226 (18%) were confirmed 'recent infections' by PCR or IgM positive assays, and 311 (25%) samples were IgG or plaque reduction neutralisation test (PRNT) positive[28]. An analysis of *Ae. aegypti* mosquitoes (n = 816) collected across Praia suggested there was low-level Zika virus circulation in mosquitoes (< 0.5%) shortly after the outbreak (August–October 2016)[32].

Here, we present the findings from a serological surveillance study performed in Cabo Verde with convenience samples collected in and around Praia shortly after the conclusion of the ZIKV outbreak in 2016. By combining a panel of arbovirus antigens in an enzyme-linked immunosorbent assay (ELISA) and equivalent commercial solutions, with the analysis of extensive metadata representing a cross-section of the population in Praia, we infer rates of arbovirus seroconversion following the ZIKV outbreak on Cabo Verde, and identify risk-factors for ZIKV seropositivity.

Materials and methods

Study site and sample collection. This study was carried out in the city of Praia located on the island of Santiago, the region with the highest prevalence of vector-borne diseases in Cabo Verde. The municipality of Praia has an area of 102 km^2, has a population of 142,009 and is divided into 88 localities[33]. To compare transmission across high and low prevalence regions, two locations, Plateau (1019 inhabitants) and Tira Chapéu (5785 inhabitants) were selected for this study. These locations reported the fewest and the most suspected Zika cases, according to the data from the Praia Health Delegacy. Residents from the islands of Fogo, Santiago, Boa Vista, Sal, and Brava were also included in this study but were all sampled at the collection centres in Praia.

Blood (5 ml) samples were collected for 7 weeks (August 24, 2016, to October 12, 2016), after the final ZIKV case was reported in Plateau and Tira Chapéu. These samples were obtained at the Health Center in Tira Chapéu and at the Health Department in Plateau. Blood samples were collected by venipuncture and transported same day to the UniPiaget laboratory for separation into serum and plasma. Before initiating sample collection, an awareness campaign was carried out in the two localities, with door-to-door home visits and distribution of information material. All individuals who participated in this study gave informed consent for sample and data collection. The data collection form contained three sections: personal characteristics (e.g., age, occupation), infectious disease history, and Zika symptoms. Ethical approval for this study was given by the Cabo Verde National Ethics Committee in research for health (ref. CNEPS 28/2016). All methods were performed in accordance with the regulations associated with the ethics approval. The study design is summarised in Fig. S1.

Arbovirus indirect ELISA. Participant sera samples were tested for IgG antibodies against ZIKV (Native Antigen Company (NAC): ZIKVSU-NS1), DENV 1–4 (NAC: DENVX4-NS1-100), YFV (NAC: YFV-NS1-100) NS1 and CHIKV E1 (NAC: CHIKV-E1-100) protein antigens. In addition to the standard indirect NS1 ELISA, we employed a commercial DAB assay to increase assay specificity, given the expected high dengue seroprevalence. Samples were defrosted from − 80 °C storage at 7 °C overnight. Storage conditions for each plate were monitored and plates were defrosted at the same time and did not exceed 3 defrost cycles. Each 96 deep-well serum plate was centrifuged at 7 °C at full speed for 10 min. Lipaemic samples were identified, and the lipid layer was aspirated. Following a checkerboard titration for each antigen and a subset of samples, a serum dilution of 1:400 was identified as the optimal analyte concentration and an antigen coating concentration of 1 µg/ml for NS1 proteins and 2 µg/ml for CHIKV E protein was determined to provide optimal assay sensitivity. The antigens were coated on to HBX 96 well ELISA plates (Thermo: 3355) at the aforementioned concentrations in freshly prepared carbonate-bicarbonate buffer pH 10.6 overnight at 7 °C. Plates were blocked in 150 µl phosphate buffered saline + 0.05% tween (PBST) and 1% nonfat-dried milk (NFDM) at room temperature for 3 h and washed 5 times in PBST. Serum was diluted in 1% NFDM and 50 µl was pipetted on to each plate with ZIKV, DENV, YFV and CHIKV positive controls as well as negative controls for the four respective antigens and incubated at room-temperature for three hours and washed. 50 µl of secondary goat anti-Human IgG (H + L) secondary HRP antibody (Thermo: #A18805) was applied at a 1:7000 dilution (PBST) to each well and the plates incubated at room temperature for two hours and washed. Finally, 100 µl of tetramethylbenzidine substrate (TMB) (tebu-bio: TMBW-1000-01) was applied to each well and incubated at room temperature in low-light conditions for 15 min, stopped with 50 µl 1 M H_2SO_4 then plates were read at 450/620 nm.

ZIKV double antigen binding (DAB) assay. HBX 96 well ELISA plates (Thermo: 3355) were coated with 1 µg/ml of ZIKV NS1 (NAC: ZIKVSU-NS1) covered overnight at room temperature and were washed for 5 cycles (PBST 200 µl) with an automated 96 well plate washer. Blocking was performed with 200 µl 3% BSA in

DPBS for one hour at room temperature. Samples were diluted at a 1:100 ratio in to 0.5% BSA in DPBS with 0.2% tween and 100 µl applied to each well using a 12-channel automated multichannel pipette, followed by a 30-min room-temperature incubation on an orbital plate shaker at 800 rpm and washed, as above. 100 µl of biotinylated ZIKV NS1 (NAC) diluted in 0.5% BSA in DPBS with 0.2% tween was added to the plate and incubated for 30 min on an orbital plate shaker at 800 rpm. 100 µl of polystreptavidin HRP conjugate (NAC) in 0.5% BSA in DPBS with 0.2% tween was added and incubated for 30 min on an orbital plate shaker at 800 RPM and washed. Finally, 100 µl of TMB substrate (tebu-bio: TMBW-1000-01) was applied to each well and incubated at room temperature in low-light conditions on an orbital plate shaker at 800 RPM for 15 min, stopped with 100 µl 1 M H_2SO_4 then plates were read at 450/620 nm.

Assay comparison. To further validate the IgG ZIKV and DENV ELISA assays, we tested a random subset of participants (n = 84), using an additional commercial indirect NS1 ELISA for ZIKV (ZG) and DENV (DG) IgG and IgM (EI 2668-9601). Samples were prepared as above, and the kit used as per manufacturer's instructions. Each assay was adjusted using either our own or manufacturer's internal negative controls. The analysis of optical density (OD) correlation was performed and plotted with GGally[34], with Spearman's rank correlation coefficient (r_s).

Statistical models. We applied a Gaussian mixture model to classify samples into serological positive, intermediate, and negative groups. This involved the application of an expectation–maximisation algorithm to estimate the model parameters and to identify the most suitable number of Gaussian components using the Scikit-learn GMM package[35,36]. The optimum number of components were chosen using Akaike's (AIC) and Bayesian (BIC) Information Criteria. Cut-offs between groups were determined by the 0.95 posterior probabilities for each component per model and used to classify titre ODs for further analysis (Fig. S2). Intermediate and negative groups were combined and compared to the positive group for the analysis of risk factors. Logistic regression models were applied to assess the association between ZIKV seropositivity and risk factors (e.g., age, gender, location, symptoms, and other serologically determined or self-reported flavivirus infections), summarised by odds ratios and their 95% confidence intervals.

Results

Patient cohort demographics and ZIKV risk. A total of 732 sera samples were collected in Tira Chapéu (n = 395) and Plateau (n = 337) study centres between August 24, 2016, and November 4, 2016, of which a random subsample (n = 431, 58.9%) were processed (Fig. S1, Table 1). The median age of the 431 participants was 35 years (range 20–72 years), and the majority were female (n = 272, 63.1%). Of the 207 women aged between 20 and 44 years, 74 (35.7%) were pregnant. The participants with self-reported ZIKV infections (suspected infections) (6.7%) and symptoms of Zika virus disease (ZVD; rash, myalgia, nausea, or fever; 6.7%), had minimal overlap (n = 1). For self-reported infections, ZIKV was in lower prevalence than that of historical DENV infection (21.3%) and greater than malaria (1.9%). The median tympanic temperature of participants was 36.1 °C (range 23.3–37.7 °C), most of which were within the established normal range for this method (35.4–37.8 °C)[37]. Most participants reside in Praia (Santiago Island, n = 341, 79.1%) and on São Vicente Island (n = 31, 7.2%), which have high population densities and a high prevalence of ZIKV cases (n = 2480, October 2015 to April 2016; 79% Praia) (Table S1). Geospatial analysis of ZIKV surveillance data (n = 2480 cases) indicated that the majority of the suspected Zika infected population live in or around Tira Chapéu, one of the study sites in Praia (Fig. 1). The other Praia study site, Plateau, had a comparatively lower, but still relatively high ZIKV incidence during the epidemic (Fig. 1).

Assay comparison. On a random subset of samples from our cohort (n = 84/431; 19.5%) we compared 3 commercial assays (ZIKV IgG indirect ELISA (ZG); DENV indirect IgG ELISA (DG); and ZIKV total Ig DAB ELISA) alongside an unmodified in-house ZIKV NS1 IgG indirect IgG ELISA and DENV NS1 indirect IgG ELISA using commercially acquired recombinant antigens. The strongest correlations were between the ZIKV and DENV NS1 IgG ELISA (Spearman's $r_s = 0.80$) and the DG and ZG assays ($r_s = 0.80$), whereas the correlation between ZIKV DAB and DG assays was lower ($r_s = 0.54$). The ZIKV commercial assays were highly correlated with the ZIKV NS1 assay (DAB: $r_s = 0.72$; ZG: $r_s = 0.71$). We observed a reduction in DENV NS1 IgG ELISA positive samples classified as ZIKV positive when the ZIKV DAB assay (n = 3, 33% of DENV positive) was compared to the ZG assay (n = 8, 88% of DENV positive) (Figs. S3, S4). With this, we chose the ZIKV DAB as our primary ZIKV assay for the classification of the larger dataset (n = 431).

In addition to the IgG assays, we screened the subset for ZIKV and DENV IgM using additional ZG and DG IgM assay variants. The ZIKV DAB assay, which detects all isotypes, can infer the presence of IgM (or IgA) by an observable differential with the ZIKV NS1/ZG IgG assay. We found no samples that exhibited elevated anti-ZIKV IgM or IgM/IgA using any of the respective assays.

ZIKV and other arbovirus serology results. Analysis of the 431 participant sera samples revealed that ZIKV, DENV 1–4, YFV and CHIKV seropositivity was 10.9%, 12.5%, 0.2% and 2.6%, respectively (Table 1). Almost a quarter (23.1%; 19/82) of ZIKV or DENV seropositive participants were found to be reactive to both DENV-NS1 (1–4) and the ZIKV-DAB assay. Of those, 63.1% (12/19) self-reported as having had either ZIKV or DENV infections, 5.2% (1/19) reported both, and 31.6% (6/19) no infections. The majority of ZIKV seropositive samples (63.8%) were collected in Tira Chapéu, consistent with the high ZIKV transmission reported in the geospatial analysis (Fig. 1). The distribution of DENV seropositive participants was spread across study centres (Tira Chapéu 46.2%, Plateau 38.8%, outside Praia 12.9%). Low levels of CHIKV seropositivity were observed

Characteristic	N (median)	% (range)
Age (years)	35	20–72
Female	272	63.1
Location		
Praia	341	78.9
Sao Vicente	31	7.2
Other	59	13.9
Self-reported Zika	29	6.7
Self-reported malaria	8	1.9
Self-reported dengue	92	21.3
Body temperature (°C)	36.1	32.3–37.7
Any Zika symptoms*	29	6.7
Zika serology**		
Negative	193	44.7
Intermediate	191	44.3
Positive	47	10.9
Dengue serology**		
Negative	224	52.0
Intermediate	153	35.5
Positive	54	12.5
Yellow fever serology**		
Negative	255	59.2
Intermediate	175	40.6
Positive	1	0.2
Chikungunya serology**		
Negative	363	84.2
Intermediate	57	13.2
Positive	11	2.6

Table 1. Study characteristics. *A rash, myalgia, nausea, or fever; **based on a 3 components mixture model.

(2.6%) with only a single participant YFV IgG positive. The latter findings are in line with the Ministério da Saúde e da Segurança Social epidemiological report, in which there were no reported YFV cases from 2015 to 2019[38]. The median age of all ZIKV, DENV and CHIKV seropositive participants (37 years; range 23–65 years) was higher than those seronegative (34 years, range 20–72 years) (Wilcoxon test $P = 0.009$). There was an association between self-reported ZIKV infection and seropositivity (Chi-square test, $P < 0.001$), as well as between DENV self-reporting and its seropositivity (Chi-square $P = 1 \times 10^{-7}$), but no evidence of association between self-reported ZIKV and DENV ($P = 0.746$).

There were high correlations between ELISA assay ODs (Fig. S5), especially between the two ZIKV assays (in-house ZIKV NS1 and ZIKV DAB, $r_s = 0.74$), in-house ZIKV NS1 and DENV NS1 (1–4) ($r_s = 0.70$) and DENV and YFV ($r_s = 0.51$). The correlation between ZIKV DAB and DENV NS1 ($r_s = 0.34$) was lower than for the NS1 assay, reflecting the greater specificity of ZIKV DAB methodology. Other high correlations ($abs(r_s) > 0.4$) were between CHIKV and YFV infections, and there were low correlations between ELISA assay ODs and age (max. $|r_s| < 0.2$) (Fig. S5).

Risk factors and correlates of ZIKV seropositivity. Univariate analysis revealed several risk factors potentially associated with ZIKV seropositivity (positive 47 vs. non-positive 384), including DENV seropositivity (odds ratio (OR) 6.766, $P < 0.001$) and self-reported ZIKV (OR 6.213, $P < 0.001$) (Table 2). There was marginal evidence of ZIKV seropositivity associations with (higher) age (OR 1.025, 95% CI 0.998–1.051, $P = 0.07$), self-reported DENV infection (OR 1.303, 95% CI 0.998–1.051, $P = 0.046$) and CHKV seropositivity (OR 3.205, 95% CI 0.998–1.051, $P = 0.094$). There were no associations with sex, location or ZVD associated symptoms ($P > 0.287$) (Table 2). These risk factor results were robust to their inclusion in a multivariate model, except self-reported CHKV ($P = 0.574$) and DENV infection ($P = 0.349$), where the latter is correlated with DENV seropositivity.

Discussion

Cabo Verde has been the location of several infectious disease outbreaks in the past 12 years, involving DENV, ZIKV and malaria infections. Within this historically high transmission setting, our study focused on a 3-month period after the recent Zika outbreak in 2016 and applied ELISA assays to perform a serological based assessment of the prevalence of ZIKV and other arbovirus diseases. We selected the DAB methodology as our primary

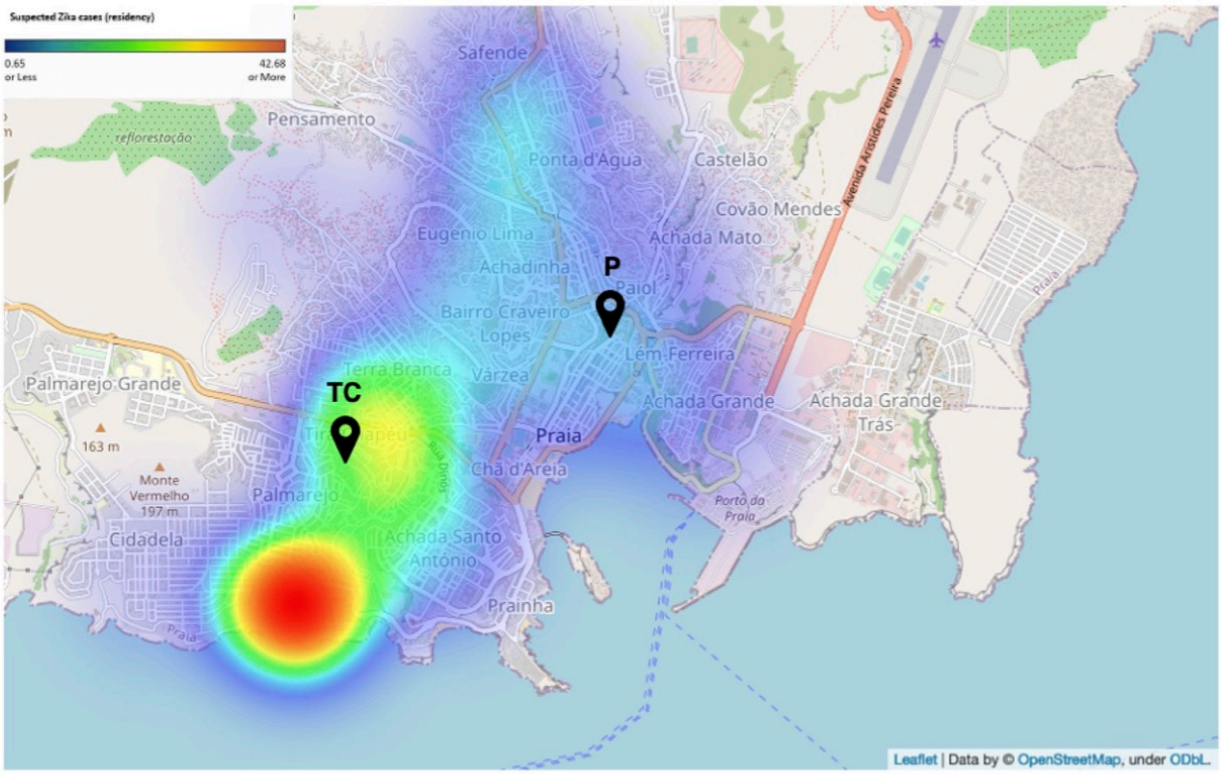


Figure 1. Geospatial distribution of suspected Zika infections in Praia, Cabo Verde. Data extracted and geocoded from the government's surveillance programme. Cases recorded between October 2015 to April 2016. Collection centres in Praia indicated by TC (Tira Chapéu) and P (Praia). The map was generated by OpenStreetMap under ODbL and Folium (https://python-visualization.github.io/folium/).

Risk factor	Zika non +ve N (median)	Zika non +ve % (range)	Zika +ve N (median)	Zika +ve % (range)	OR	95% CI	P-value	AOR	95% CI	P-value
Age (years)	34	20–72	39	23–63	1.025	0.998–1.051	0.068	1.029	1.000–1.060	0.048
Male vs. female	239	62.2	33	70.2	0.699	0.352–1.325	0.287	0.654	0.314–1.361	0.256
Praia vs. other	301	80.2	39	83.0	1.199	0.564–2.860	0.658	–	–	–
Any Zika symptoms*	27	7.0	2	4.3	0.588	0.093–2.053	0.478	–	–	–
Dengue +ve	35	9.1	19	40.4	6.766	3.409–13.326	3.3×10^{-8}	6.783	3.128–15.101	1.6×10^{-6}
Yellow fever +ve	1	0.3	0	0	–	–	–	–	–	–
Chikungunya +ve	8	2.1	3	6.4	3.205	0.998–1.051	0.094	–	–	–
Self-reported Zika	18	4.7	11	23.4	6.213	2.660–14.043	1.4×10^{-5}	4.889	1.913–12.493	9.1×10^{-4}
Self-reported dengue	80	20.8	12	25.5	1.303	0.998–1.051	0.046	0.671	0.292–1.546	0.349

Table 2. Odds ratios (ORs) for risk factors for Zika positivity (n = 47) versus non-positivity (n = 384). *AOR* adjusted odds ratios estimated by a multivariate model that includes all listed risk factors, *CI* confidence intervals; *a rash, myalgia, nausea, or fever.

ZIKV assay as it exhibited a reduction in the classification of DENV NS1 positive samples as ZIKV positive, when compared to the other indirect ELISA methodologies with known ZIKV-DENV cross reactivity.

Analysis of 431 participant sera samples revealed ZIKV and DENV (1–4) seropositivity in excess of 10%. Our study consisted of only adults, and predominantly women, with a high incidence of pregnancy (37.3%). We hypothesise that women are also more likely to visit medical clinics than men, with differences exacerbated due to concerns arising from the causal link between ZIKV infection and congenital Zika syndrome (similar to[39]). Our logistic regression modelling approach adjusted for gender and identified a significant link between greater risk of ZIKV seropositivity and increasing age. For the duration of the study, poster advertisements were placed encouraging the public to engage with the study. While we believe we have captured a balanced demographic in our cohort, a bias toward participants who suspect they may have been exposed to ZIKV may exist.

Our cohort consisted of primarily residents in the municipality of Praia (Santiago Island, 79.1%), the urban capital region with the greatest reported transmission of the archipelago, and São Vicente (7.2%) another island 266 km away from Praia. Additionally, participants reported as residing on Fogo, Boa Vista, Sal, and Brava islands (6.3%). The ZIKV vector, *Ae. Aegypti* thrives in urbanised environments, and therefore, is well positioned for arbovirus transmission in such areas[40,41]. This supposition is reflected in our study, in that the municipalities of Praia and São Vicente had the greatest number of ZIKV and DENV cases and were the two most densely populated municipalities in Cabo Verde.

In our cohort, 6.7% of participants self-reported with a ZIKV infection, of which 37.9% were ZIKV seropositive, which is consistent with another study in Cabo Verde that estimated 43.8% of self-reported suspected cases were ZIKV positive from PCR, IgM or IgG assay assessments[28]. Similarly, 29.3% of our participants who self-reported dengue infection were DENV seropositive, consistent with a report in Brazil (n = 20,880, 26.0–32.5%)[42]. Interestingly, there was a weak overlap between reporting of ZIKV symptoms and self-reported suspected ZIKV infection. It may be that participants recall of multiple symptoms is weaker than that of a single diagnosis. Further, the ambiguous presentation of ZVD in the context of a pandemic may act to confound epidemiological studies of this nature and should be interpreted only with robust supporting data.

CHIKV transmission has been reported to both enhance transmission and pathogenicity of ZIKV infections[15,16]. CHIKV is not included in the 'Priority conditions and diseases under epidemiological surveillance' programme[38]. We have observed low seropositivity of CHIKV in this cohort compared to that reported in Brazil[16]. While there have been no confirmed reports of autochthonous cases of CHIKV in Cabo Verde, this result emphasises the possibility of an introduction event occurring, especially given the number of seropositive (exposed) participants. This possibility is further compounded by the demonstrated competence of the local *Ae. Aegypti* vector population for the transmission of arbovirus pathogens.

Despite the more recent ZIKV outbreak, DENV seroprevalence was marginally greater. The DENV outbreak occurred in 2009 with 25,088 suspected cases, more than three times that of the ZIKV burden[43]. Since then, there have been low levels (36 cases, 2015–2019) of autochthonous DENV2 and DENV4 transmission[38,44]. In addition, there have been reports of secondary ZIKV infections activating memory anti-DENV B and T-cell effector responses through antigenic 'original sin' mechanisms[45,46]. These mechanisms potentially explain the high anti-DENV optical densities, and the significant concordance between the non-specific NS1 indirect ZIKV and DENV assays. Despite the optimisation of the commercial DAB and EI ZIKV NS1 assays, the possibility of ZIKV/DENV false-positives due to cross-reactive antibody responses is a study limitation. The application of PRNT assays would mitigate these confounding effects, however, they are not always possible and applicable in epidemiological contexts.

Overall, our findings provide new insights into the dynamics of a ZIKV outbreak in Africa, introduced from Brazil. Cabo Verde's proximity to mainland Africa and close links with Brazil make it an involuntary trans-Atlantic hub for the introduction of infectious diseases to new continents. Our survey has provided a snapshot of arbovirus seropositivity across Cabo Verde, detailing demographics, and testing assays, while emphasising the requirement for sustained epidemiological surveillance to reduce the burden of future outbreaks, and potential pandemics.

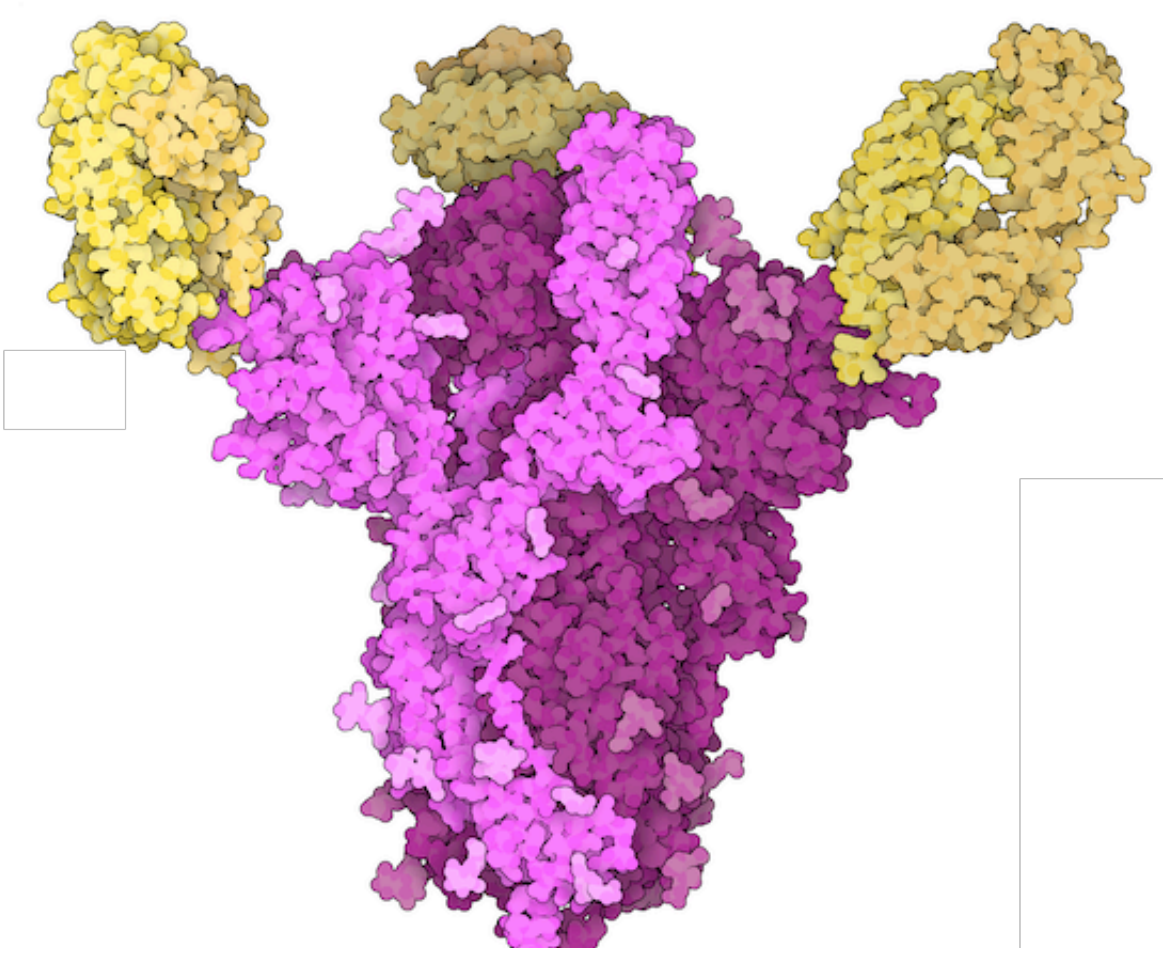
Received: 28 February 2022; Accepted: 5 July 2022
Published online: 09 July 2022

References

1. Rasmussen, S. A., Jamieson, D. J., Honein, M. A. & Petersen, L. R. Zika virus and birth defects—Reviewing the evidence for causality. *N. Engl. J. Med.* **374**, 1981–1987 (2016).
2. Ximenes, R. *et al.* Health outcomes associated with Zika virus infection in humans: A systematic review of systematic reviews. *BMJ Open* **9**, e032275 (2019).
3. dos Santos, T. *et al.* Zika virus and the Guillain–Barré syndrome—Case series from seven countries. *N. Engl. J. Med.* **375**, 1598–1601 (2016).
4. Cao-Lormeau, V.-M. *et al.* Guillain–Barré syndrome outbreak associated with Zika virus infection in French Polynesia: A case–control study. *Lancet* **387**, 1531–1539 (2016).
5. Metsky, H. C. *et al.* Zika virus evolution and spread in the Americas. *Nature* **546**, 411–415 (2017).
6. Zhang, Q. *et al.* Spread of Zika virus in the Americas. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4334–E4343 (2017).
7. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406–410 (2017).
8. Duffy, M. R. *et al.* Zika virus outbreak on Yap Island, Federated States of Micronesia. *N. Engl. J. Med.* **360**, 2536–2543 (2009).
9. Musso, D. *et al.* Zika virus in French Polynesia 2013–14: Anatomy of a completed outbreak. *Lancet Infect. Dis.* **18**, e172–e182 (2018).
10. Dupont-Rouzeyrol, M. *et al.* Co-infection with Zika and dengue viruses in 2 patients, New Caledonia, 2014. *Emerg. Infect. Dis.* **21**, 381–382 (2015).
11. Faria, N. R. *et al.* Zika virus in the Americas: Early epidemiological and genetic findings. *Science* **352**, 345–349 (2016).
12. WHO. *Countries and territories with current or previous Zika virus transmission*. https://www.who.int/emergencies/diseases/zika/countries-with-zika-and-vectors-table.pdf (2019).
13. WHO. Zika cumulative cases. *PAHO-WHO* http://www.paho.org/hq/index.php?option=com_content&view=article&id=12390&Itemid=42090&lang=en (2018).
14. Vogels, C. B. F. *et al.* Arbovirus coinfection and co-transmission: A neglected public health concern?. *PLoS Biol.* **17**, e3000130 (2019).
15. Magalhaes, T. *et al.* Sequential infection of *Aedes aegypti* mosquitoes with chikungunya virus and Zika virus enhances early Zika virus transmission. *Insects* **9**, 177 (2018).
16. Campos, M. C. *et al.* Zika might not be acting alone: Using an ecological study approach to investigate potential co-acting risk factors for an unusual pattern of microcephaly in Brazil. *PLoS One* **13**, e0201452 (2018).
17. de Vasconcelos, Z. F. M. *et al.* Challenges for molecular and serological ZIKV infection confirmation. *Child's Nerv. Syst.* **34**, 79–84 (2018).

Chapter Five

Guiding the improvement of serological diagnostics with *in-silico* immunoanalytic analyses



Neutralising antibody Fab fragments in complex with SARS-CoV-2 spike protein. The Protein Data Bank's molecule of the month (April 2021). The 4A8 antibody depicted here binds away from the receptor binding domain (RBD) but still neutralises the function of the spike protein. **Yellow:** 4A8 antibodies. **Pink:** Sars-CoV-2 spike trimer. (Chi, Xiangyang, *et al*., 2020)

5.1 Introduction

Dissecting the degree to which an antibody response may or may not cross-react with discrete viral antigens is a key step in the development of serological assays. This involves using techniques from several scientific disciplines, some of which are investigated in the following sections. In this chapter, methodologies aimed at improving immunoassay design for the surveillance of emerging infectious diseases are explored. Through this process, bioinformatic analyses are integrated with publicly available *in-vitro* data to design and test novel ZIKV antigens.

5.1.1 Antigen Conservation across Flavivirus species

The influence of cross-reactivity across *Flavivirus* species on adaptive immune responses is not fully understood. But its origins can be explained through understanding their common ancestral history. Each species of the *Flavivirus* genus carries with them, orthologous genes, bound together evolutionarily by their conserved function in viral parasitism, the specific host-range and transmission vectors [1]. This conservation culminates in a complicated immunological landscape, particularly regarding antibody responses, where serological assays are confounded by non-specific antigenic determinants. This conservation is illustrated in the comparison of common human infecting *Flavivirus* species **(Figure 1)**. Here, the polyproteins of four common *Flavivirus* species (and the four DENV serotypes) were aligned and compared with ZIKV, in an analysis that assesses the physicochemical properties of residues and scores them based on their similarity. The significant level of conservation shown in the amino acid sequence is the root cause of the immune cross reactivity observed in immune responses to *Flavivirus* species.

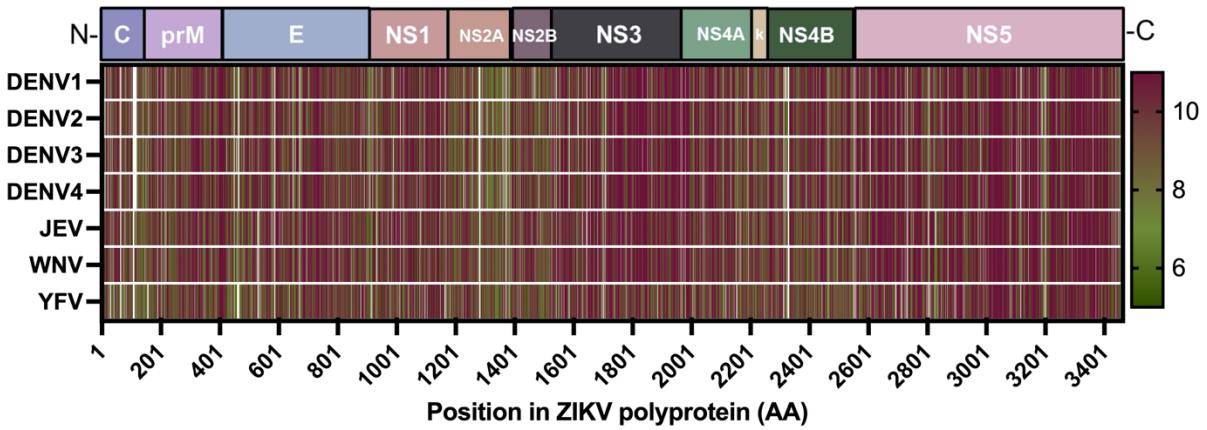


Figure 1. Flavivirus amino acid sequence conservation heatmap. Identical residues score 11, a substitution of aliphatic residues such as alanine and leucine scores 10, whereas a substitution of the basic residue arginine to alanine would result in a score of 4. DENV1-4 Dengue virus 1-4; JEV Japanese encephalitis virus; WNV West Nile Virus; YFV Yellow Fever Virus.

On visual inspection the heatmap reveals two dense regions of high amino acid conservation, NS3 (non-structural protein 3) and NS5 (non-structural protein 5). The data from the plot is dissected in **Table 1**, where the above observations are summarised empirically. The NS3 and NS5 proteins are the most conserved, with their functions evolutionally rooted in the fundamental processes of *Flavivirus* replication; proteolytic cleavage of the virus polyprotein (NS3) and RNA replication (NS5). This conservation is of little concern in the context of antigen design, as these intracellular proteins are not secreted, neither are they surface exposed and would therefore not *dominate* the anti-*Flavivirus* antibody response. The third and fourth-most conserved proteins, the virus envelope (E) and the secreted non-structural protein 1 (NS1), on the other hand, present a complex situation. The E protein is subject to an immunodominant neutralising humoral response, with its role as the primary structural protein and function in viral fusion [2]. While the E protein is glycosylated, the role of these carbohydrates is confined to fusion, replication and virulence, and not immune (epitope) evasion, unlike other viral surface protein glycans [3,4]. The E protein, and its three domains, have remained the focus of research, predominantly in the field of vaccinology and diagnostics. Particularly, the third

domain of the E protein (EDIII) has been found to be both serotype and virus specific, with numerous studies proposing it as a vaccine candidate and effective diagnostic antigen [2,5–7]. NS1, on the other hand, is a multifaceted protein with numerous putative functions. In its intracellular dimeric form, it plays numerous roles in viral replication, after which, it is displayed on the host cell surface membrane and later, secreted as a hexamer. This secreted protein species modulates the immune system, driving inflammatory responses and enhancing propagation, while inhibiting complement attack [8]. Given its intracellular *and* extracellular locale, NS1 is a key target of humoral immune responses, and has played a central role in *Flavivirus* diagnostics, both in antigen capture and antibody detection formats [9–11].

Table 1. The amino acid conservation across human-infecting *Flavivirus* spp.
The data are separate by each protein of the translated polyprotein. Scores are based on pairwise comparisons with the same ZIKV gene. Global sequence identity is a metric comparing the raw sequence alignment quality across the entire polyprotein. DENV1-4 Dengue virus 1-4; JEV Japanese encephalitis virus; WNV West Nile Virus; YFV Yellow Fever Virus.

Species	C	PrM	E	NS1	NS2A	NS2B	NS3	NS4A	2K	NS4B	NS5	Global sequence identity (%)
DENV1	8.37	8.80	9.29	9.21	7.99	8.71	9.68	9.10	9.09	9.09	9.63	**41**
DENV2	8.26	8.78	9.23	9.24	8.04	8.98	9.71	9.31	9.00	9.22	9.64	**42**
DENV3	8.39	8.75	9.27	9.28	8.07	8.94	9.72	8.97	9.35	9.13	9.64	**43**
DENV4	8.26	8.86	9.27	9.18	7.99	9.02	9.72	9.13	9.43	9.04	9.70	**43**
YFV	7.52	8.47	8.72	8.94	8.34	8.73	9.12	8.95	9.22	8.53	9.39	**24**
WNV	8.56	8.98	9.37	9.26	8.64	9.31	9.81	9.21	9.17	8.88	9.80	**42**
JEV	8.44	8.96	9.38	9.35	8.75	9.38	9.70	9.09	9.35	9.05	9.75	**44**
Average	8.26	8.80	9.22	9.21	8.26	9.01	9.64	9.11	9.23	8.99	9.65	**40**

This preliminary analysis illustrates, both, the limited target range for unmodified immunogenic diagnostic antigens and the potential challenges involved in ensuring specificity in the context of multi-*Flavivirus* endemic contexts.

5.1.2 The design and expression of a concatenated ZIKV NS1 GST fusion-peptide

For the first exploration in to generating specific antigens for *Flavivirus* immunoassays, I used a combination of rudimentary informatic analysis to probe the NS1 protein for ZIKV-specific antigenic regions. Three short contiguous peptides based on our analyses were selected for further study.

The process for selecting peptides was designed first and foremost, to increase *Flavivirus* diagnostic specificity by eliminating cross reactivity between ZIKV and DENV epitopes. The initial stage consisted of finding regions on the ZIKV AA polyprotein which were specific, on the sequence-level, when compared to DENV. To achieve this, ZIKV and DENV *Flavivirus* NS1 amino acid sequences were aligned, visually inspected, and selected regions with low identity. To further enhance the probability of finding immunologically significant regions, several *in-silico* epitope prediction software packages were employed (ABCpred [12], DRREP [13] and ElliPro [14]). Between them they employ an array of strategies to make their predictions. DRREP and ABCpred both use a deep neural network trained on extensive datasets of known epitopes, while Ellipro uses 3D structure and the biochemical properties of a given protein to make predictions. With a list of residue positions with reduced homology with DENV, and three lists of predicted epitopes, the collated data and with overlapping epitope 'hits' were mapped onto the 3D crystal structure of ZIKV NS1 (PDB-5GS6). Surface exposed contiguous peptides where then selected resulting in 3 unique ZIKV NS1 based peptides, 23, 32 and 50 residues in length. The coding-sequences for the selected peptides were concatenated in a synthetic construct and codon-optimised for expression in *E. coli*, the peptide was expressed using the pGEX-4T-1 (GST fusion) vector. The resulting purified protein, purified by GST affinity chromatography was termed *'Concat'*.

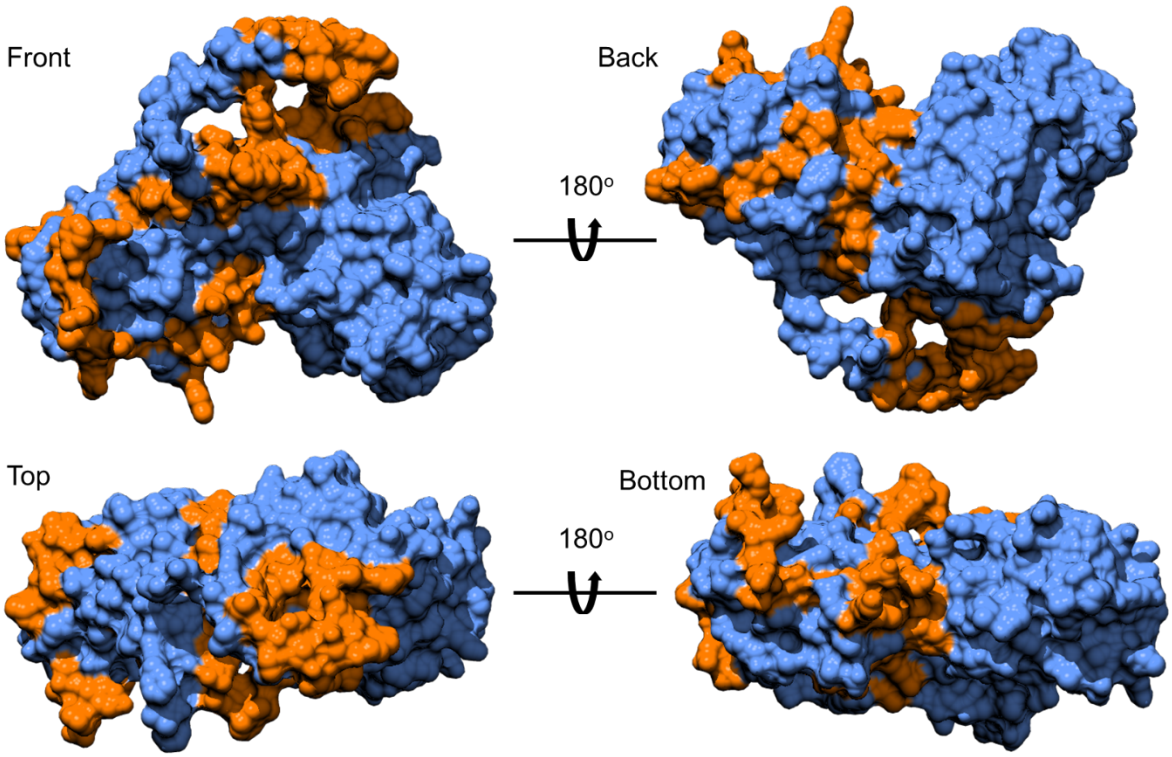


Figure 2. 3D surface model of the Zika NS1 protein monomer. Orange highlighted residues represent the regions selected by the diagnostic peptide pipeline. Image was rendered using Chimera [26].

The process detailed above outlines the first effort at utilising *in-silico* techniques to guide the design of synthetic antigens for diagnostic use. In the results section, I report the testing of this peptide, *in-vitro*, against control ZIKV and DENV immune samples and a subsection of 86 human serological samples, which were featured in the previous chapter. I use a custom Luminex microsphere assay and compare their performance against other existing NS1 immunoassays and antigens. As a follow-up, I explore alternative techniques and suggest improved methods for *in-silico* antigen design, for application in ZIKV diagnostics. In the final section I present a publication detailing the refinement and application of these combined methods in the context of another emerging infectious disease, SARS-CoV-2.

5.2 Methods

Zika 'Concat' protein expression

Three amino acid sequences were identified by the epitope prediction step. These were concatenated into a single 105 AA contig and codon optimised for expression in an *E. coli* system. A PGEX-4T-1 expression vector containing the concatenated peptide-coding sequence was transfected by heat shock at 42^{O}C for 45 seconds and transferred to LB media with ampicillin. Transformed colonies were identified by PCR and Sanger sequencing and selected for culture. The pellet was lysed using a high-pressure call press homogeniser, and purified using Glutathione Agarose (Thermo: 16100). The product was then dialysed and quantified using Bradford reagent (Bio-Rad: 5000001), as per manufacturers instruction.

Commercial Zika/dengue ELISA panel

Eighty-six serum samples were selected for screening using Euroimmun Zika IgG + IgM (EI 2668-9601), and dengue IgG + IgM kits (EQ 266a-9601). Samples were diluted to 1:100 and the ELISA protocol was performed according to manufacturer's specification. Plate reading was performed at 450 nm.

Zika antigen ELISA

The Zika ELISA was performed using NS1 recombinant protein (Native Antigen Company: ZIKVSU-NS1-100), Concat purified protein and a recombinant GST protein. Plates were coated with NS1, Concat and GST individually diluted in coating buffer (Na_2CO_3 + $NaHCO_3$) and incubated overnight at 7^{O}C. The coating concentrations of Concat and GST proteins were adapted to allow for the subtraction of relative GST signal in present the GST-Concat fusion protein. Serum was added at a concentration of 1:500 with 1% milk powder block in a 1% PST-

Tween solution followed by an incubation step overnight at 7^{O}C. Plates were washed 3 times in PBS-Tween and anti-human-IgG HRP antibody at 1:15000 in 1% PBS-Tween was added and incubated for 3 hours at RT. Following another 3 wash steps, 100 µL of TMB One Component HRP Microwell Substrate (TMBW-1000-01) was added, followed by the addition of 0.2 M H2SO4 after 15 minutes and the plates were read at 450 nm.

Zika NS1 and Concat Luminex assay

Recombinant antigens were coupled to MagPlex© COOH-microspheres (Luminex Corp., Austin TX) following the protocol described by Luminex Corparation. The optimal concentrations for each antigen were used for large volume bead coupling (>11 x 10^6 beads, ~900 µL), with coupled beads re-suspended in 1mL of storage buffer (1xPBS, 0.05% Tween, 0.5% bovine serum albumin (BSA), 0.02% sodium azide, 0.02% Pefabloc (Sigma) and stored at 4ºC until further use. An initial mixture containing 8 µl of each antigen-coupled microsphere set and 5 ml of Buffer A (1xPBS, 0.05% Tween, 0.5% BSA, 0.02% sodium azide) was prepared, yielding approximately 1,000 beads per region per well. Next, 50 µl of this combined microsphere mixture was added to a 96-well flat bottom plate (BioPlex Pro™, Bio-Rad Laboratories, UK) and washed once with 100 µl of PBS-TBN (1xPBS, 0.05% Tween-20, 0.5% BSA and 0.02% sodium azide) on a BioPlex Magnetic Hand Washer. 50 µl of samples and controls were added to the plate and incubated in the dark at room temperature (RT) on a microplate shaker at 500 rpm for 90 minutes. Following three washes, 50 µl of fluorescent secondary antibody (Jackson Immuno 109-116-098: Goat Anti-human Fcy-fragment specific IgG conjugated to R-Phycoerythrin (R-PE)), diluted to a 1:200 dilution with Buffer A, was added to all wells and incubated for 90 minutes in the dark at RT at 500 rpmon an orbital plate shaker. After a further three washes, the plate was incubated in 50 µl of Buffer A for 30 minutes. Plates were washed one additional time, after a final addition of 100 µl 1xPBS, were read using

the Luminex MAGPIX© analyser. At least 50 beads per analyte were acquired per sample and median fluorescent intensity (MFI) data were used for analysis. Glutathione S-transferase (GST) coupled beads were included as a control to test for GST-specific immunoglobulin (IgG) response against GST-tagged fusion proteins.

In-silico specificity analysis using k-mer mapping

I used the 'canonical' reference amino acid sequences (from UniProt) for each Arbovirus screened in this analysis. The *seqkit* software package [15] 'sliding' function was implemented to generate 15-mers using a sliding window, iteratively moving 1 residue at a time, resulting in a 15-mer pool of 34666. This pool, with the ZIKV k-mers subtracted, was mapped on to the ZIKV polyprotein, using blastp software [16]. A cutoff of 80% identity was set for a 'successful hit'. A custom output format of the blastp analysis yielded the peptide length and the mapping position on the ZIKV proteome. An in-house script was used to map each of the hits on to a linear axis, which represented each residue of the ZIKV polypeptide. For added functionality, not utilised here, I added a traceability function, which reports the location of each k-mer in the original polyprotein; for each hit, it is possible infer which region in the original virus' proteome the hit originated from. For this analysis, I accounted for only the DENV sequence hits, as this is the only virus with reported transmission in Cabo Verde. The sequence hits for each residue were added to the main 'data pool'.

Alignment conservation analysis

Jalview software [17] was used to assess sequence conservation, the same metric used to generate **(Figure 1)**. In this software, I assessed a multiple sequence alignment consisting of 566 ZIKV whole-polyprotein sequences, and extracted the 'conservation' metric. This metric is generated by measuring the number of conserved physio-chemical properties for each

column of the alignment [18]. For the entire polyprotein, these data were extracted and added to the main ‘data pool’.

Immune epitope database data mapping

From the IEDB database search tool, I extracted 1232 linear human ZIKV confirmed ‘reactive’ epitopes [19]. Like the k-mer mapping process, I used blastp and an inhouse script to map each epitope onto the polyprotein, with the identifying epitope number intact. The metrics for epitope mapping were collated and added to the final data pool.

Implementation of epitope prediction meta-analysis

For each tool, I used respective amino acid sequences for *Flavivirus and Alphavirus* protein as the input. For prediction of potential linear B-cell epitopes using Bepipred software [20], the output was in a similar format to that of the data pool, that is, on a per-residue basis. This was pasted into the pool directly, aligning each protein. For DRREP [13], Lbtope [21], bcepreds [22] and ABCpred [12] software packages, a list of peptide sequence candidates was outputted, which was mapped using blastp, as above. Each column in the data pool corresponds to respective epitope prediction tools, and was minimum-maximum scaled, before aggregation with equal weighting added to each prediction.

5.3 Results

5.3.1 Expression of GST-fusion ZIKV peptides

To begin testing the expressed recombinant Concat antigen, samples were selected which tested positive for ZIKV or DENV as according to the commercial ‘ZG’ or ‘DG’ ELISA kit, featured in **Chapter 3**. Five samples of ZIKV positive and DENV positive serum were pooled and diluted to a 1/400 concentration. A titration of the Concat antigen coating concentration, as

well as a full-size recombinant NS1 protein (Native Antigen Company) was performed, to find the optimal coating conditions for the assay. Because the Concat protein was fused with a GST protein tag, a GST control plate was ran in parallel, which was coated with the relative amount present on the GST-fusion-Concat plates, enabling the removal of any background reactivity to the GST fusion.

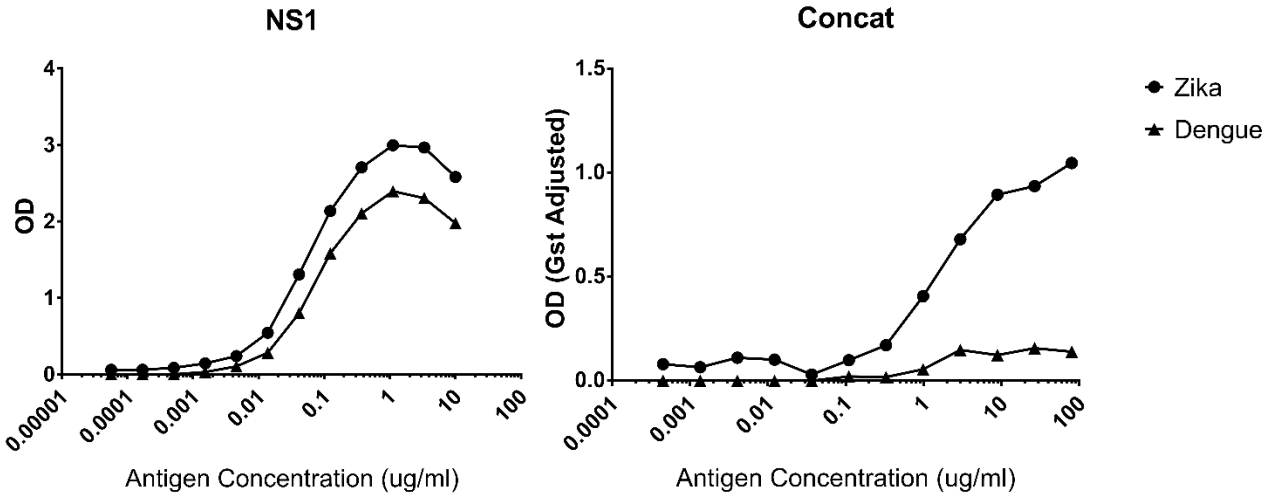


Figure 3. Antigen titration ELISA. Plates were coated with Concat and NS1 at 80 µg/ml and 10 µg/ml respectively and diluted threefold 11 times. GST signal was subtracted from the Concat raw OD. Blank well background was subtracted on all plates. Pooled ZIKV+ or DENV+ serum was diluted at 1/500. Each analysis was run in duplicate, and the mean OD calculated.

The extent to which DENV immune sera cross-reacts with the ZIKV NS1 antigen is shown (**Figure 3)**. Assessing the six highest antigen concentrations (range: 0.1 µg – 10 µg), DENV positive serum resulted in 74% of the reactivity to ZIKV NS1 when compared to ZIKV positive serum. Under the same comparison, the Concat antigen saw 15% DENV cross reactivity when compared to the ZIKV control, showing an apparent reduction in non-specific signal from our DENV immune polyclonal control. Notably, however, the lack of a sigmoidal curve in the Concat analysis indicates that the assay did not reach saturation. The maximum optical density

(OD), when adjusted for background reached 1 optical density unit (ODU), compared to the maximum of 3 ODU with the full ZIKV NS1 protein. In a repeat experiment, a plate with an increased concentration of the Concat antigen was coated, which resulted in a similar limit. In **Figure 3**, the coating concentration was 8 times higher than that of the full NS1 protein. This demonstrates a reduced sensitivity of the antigen, most likely due to the shorter peptide length, which in-turn, reduces the availability of epitopes to be bound by immune sera. Nonetheless, having established that the Concat antigen is serologically reactive and that the perceived cross-reactivity with DENV immune sera in this sample set appears to be diminished compared to the full-length NS1 protein, the assay was subsequently tested on a panel of 256 Cabo Verdean serological samples using the Luminex platform. As before, full-size NS1, Concat and GST antigens were run in parallel to enable the removal of GST background signal.

Figure 4 (left) illustrates the strong correlation between the reactivity to a GST antigen alone, analysed in parallel with the GST-fused Concat protein ($r = 0.71$, $p = < 0.001$). The reactivity to the GST antigen was subtracted from the signal from the GST-Concat fusion peptide signal **Figure 4 (right)**. Little correlation was observed between the reactivity of GST-Concat, which was adjusted for GST background reactivity, and the full ZIKV NS1 protein ($r < 0.001$, $p = 0.001$). The red colouration of the data points in the GST adjusted assay (**Figure 4 – right)** illustrates no link between the increased GST-Concat reactivity and the GST signal exhibited prior to adjustment (**Figure 4 – left).**

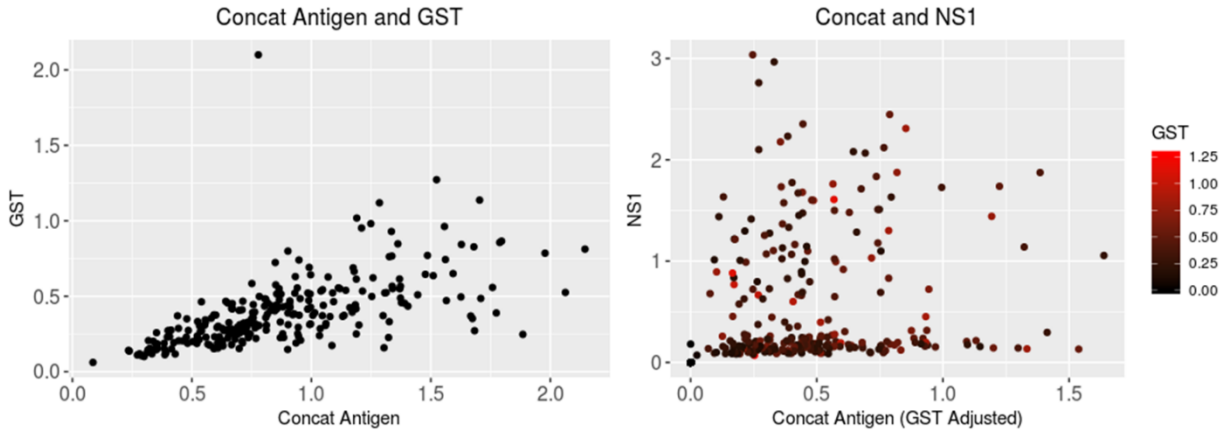


Figure 4. NS1, Concat and GST ELISA assay results. **(Left)** Raw Concat optical density (OD) compared with GST OD run in parallel. **(Right)** NS1 OD compared with GST adjusted Concat data. Colouration scale refers to the pre-adjustment GST signal. All samples were diluted at 1:500. Each analysis was run in duplicate, and the mean OD calculated.

The GST adjusted Concat signal was analysed alongside the ZG and DG commercial assays based on the subset of 86 samples (from **Chapter 3**). **Figure 5 (left)** illustrates the comparison of the full ZIKV NS1 protein assay and the commercial ZG assay, which uses a proprietary NS1 antigen. These assays were well correlated ($r = 0.65$, $p < 0.001$). As reported in the previous chapter, the commercial ZIKV 'ZG' and DENV 'DG' assays were also highly correlated ($r = 0.80$, $p < 0.001$), even more so than the full ZIKV NS1 antigen and the ZIKV commercial assay, which may imply a lack of assay specificity. **Figure 5 (right)** is the comparison of the GST-Concat signal (adjusted) with the ZG assay. No correlation was observed between these two assays ($r < 0.001$, $p < 0.001$), as well as between GST-Concat and the DENV 'DG' assays ($r = 0.03$, $p = < 0.782$).

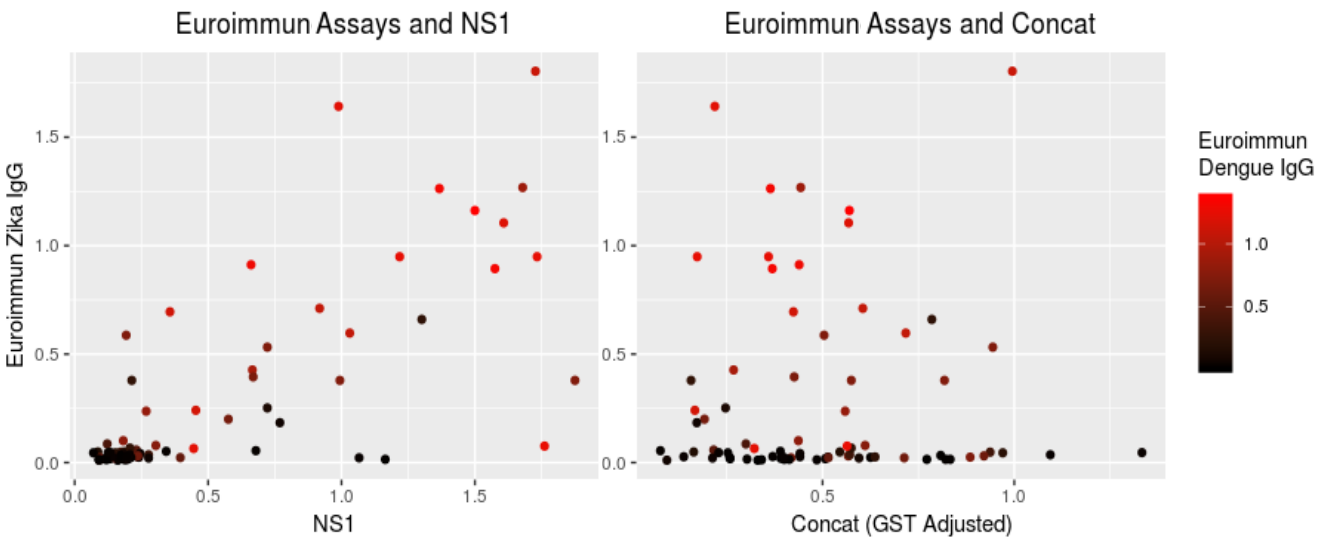


Figure 5. NS1, Concat and GST ELISA assay results. **(Left)** Raw Concat optical density (OD) compared with GST OD run in parallel. **(Right)** NS1 OD compared with GST adjusted Concat data. Colouration scale refers to the pre-adjustment GST signal. All samples were diluted at 1:500. Each analysis was run in duplicate, and the mean OD calculated.

Overall, a strong correlation between the reactivity of GST-Concat and the GST protein reactivity was observed. Considering the length of the truncated peptides, and the reduced signal observed when compared to larger, more immunogenic proteins, an alternative methodology of expressing peptides alone, or with a smaller fusion partner, may be beneficial. While the exclusion of the GST signal appeared to reduce the background reactivity to the fusion partner, leaving acceptable levels of signal for this analysis, the apparent baseline reactivity of GST-fusion proteins to the sample set makes this expression system a poor choice. The reactivity of the Concat peptide appeared to hold no resemblance to either ZIKV NS1 assays, which may indicate this construct is of a sub-optimal design.

5.2.2 Refining in-silico techniques for specific ZIKV antigen discovery

Given the sub-optimal results demonstrated above, I sought to refine two of the key aspects of the antigen design process, within the wider aim to improve ZIKV immunoassay performance. The first objective was improving the *in-silico* analyses used to guide the selection of antigen peptides, which will be addressed in this section. The second, improving the methodologies used to express targets, addressing the fusion-protein issue demonstrated above, is explored in the si chapter.

In this analysis, a dissection of each discrete step of the second iteration of the *in-silico* pipeline is demonstrated. This is later applied in a further example, in the final section of this chapter, demonstrating the methodology's utility in the context of SARS-CoV-2 research. Here, each protein of the ZIKV genome was analysed and included in the final candidate panel, but for illustrative purposes, the search is reduced to four genes: E and NS1 antigens, and the conserved serine protease (NS3) and RNA polymerase (NS5) proteins.

Firstly, the pipeline was improved through analysis automation. The method described previously was time-consuming to perform, and included an element of visual interpretation, which may add error and subjectivity to the design process. In this example, the pipeline was built around a central data-pool, presented in a per-residue format, which would contain the combined analyses designed to guide the selection of antigens (**Table S1**). Each individual analysis was parsed and pooled by a scripted process, which was augmented by the calculation of summary statistics to inform decision making when choosing candidates.

As in the previous analysis, one of our central objectives was to predict specificity, particularly in the context of DENV background. While a sequence alignment, such as that featured in

Figure 2, would be suitable for comparing proteins in a linear fashion, a different approach was used, termed ‘k-mer mapping’. This methodology allows the comparison of each protein at single-epitope resolution. One further benefit of this methodology is its utility in assessing homology in lesser related virus species. With similar endemicities for *Flavivirus* species, the *Alphavirus* species Chikungunya and Mayaro virus, can be added to the analysis, which would be prohibited using standard alignment techniques due to the significant differences in genome structure.

For this process, a function that splits each *Flavivirus* protein in to overlapping 15-mer stretches of AA sequence, using a sliding window which covers the entire proteome was implemented. These sequences are pooled and mapped on a linear axis, which in this case was the ZIKV polyprotein. With a pool containing 34666 15-mers from common human *Flavivirus* and *Alphavirus* species, the identification of homologous sequence fragments that may result in cross reactive epitopes is made possible. The graphical representation of this analysis is shown (**Figure 6**)**.** Here, the three aforementioned ZIKV proteins are expressed on a linear axis, with the occurrence of non-ZIKV k-mers with identity shared at that position in the target ZIKV protein denoted by solid colour bars.

Firstly, this analysis supported the previous findings guiding the design of the first ZIKV antigen, Concat. The peptide positions highlighted (**Figure 6;** grey) indicate the regions used in the Concat construct. On visual inspection, the three components represent the largest contiguous regions of the ZIKV NS1 protein with no DENV identity indicated. Importantly however, the analysis did identify regions with greater sequence identity shared with YFV, WNV and JEV, which should be factored in when considering the use-case of potential candidates. Assessing the mapping of k-mers to the E protein, 75.0% of the residues exhibit identity shared with DENV, compared to 60.6 % in NS1. The increased identity exhibited across all four *Flavivirus* species with the NS3 and NS5 proteins further illustrates the conservation of these proteins. With no orthologous genes, the *Alphavirus* species exhibit an overall reduced amino acid sequence identity when compared with other *Flavivirus*

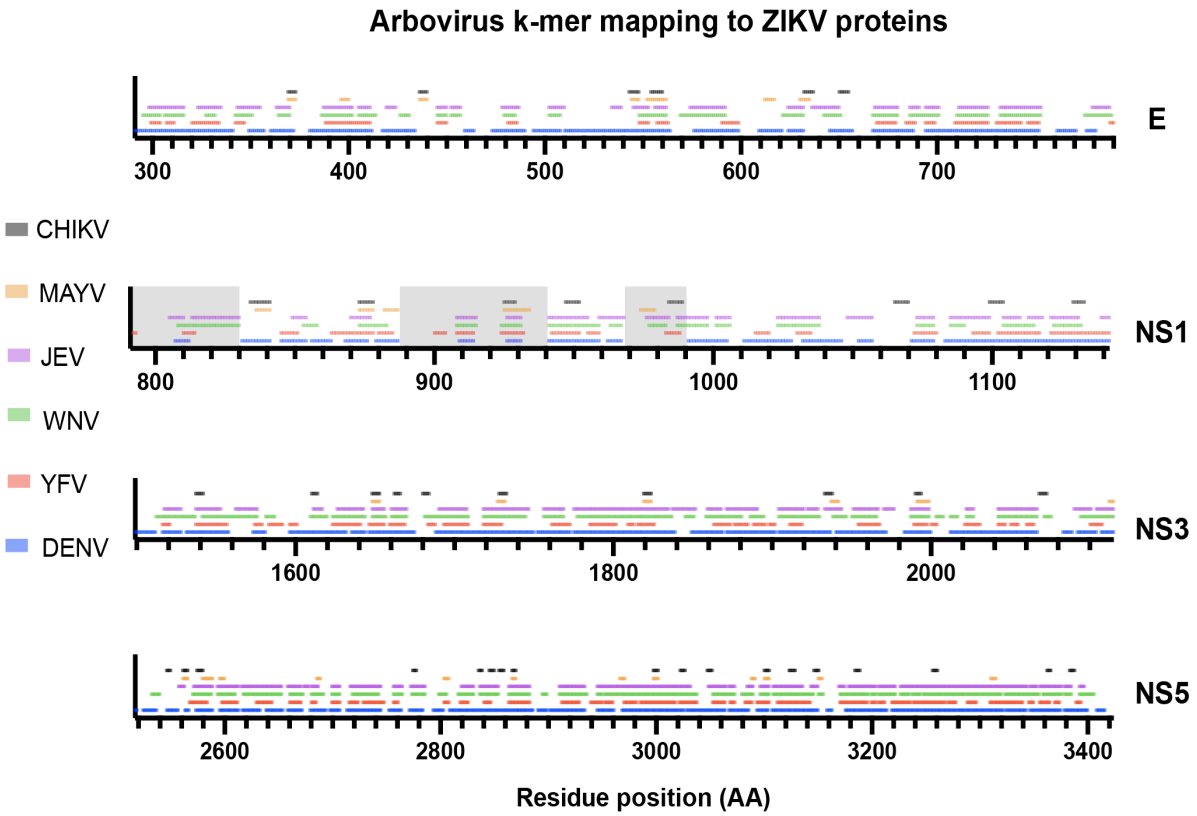


Figure 6. The library of 34666 15-mer amino acid sequence fragments from the *Flavivirus* species DENV (1-4), YFV, WNV, JEV and the *Alphavirus* species CHIKV and MAYV, mapped to the canonical ZIKV amino acid sequences of E, NS1, NS3 and NS5 proteins. The mapping process was performed using blastp. The NS1 regions highlighted denote the three 'Concat' peptides discussed in the previous section.

arboviruses. The longest *Alphavirus* peptide with ZIKV identity was 10 residues long, which was mapped to the NS1 protein.

The next analysis is similar to that shown in **Figure 1**. It assesses the amino acid conservation at each position in an alignment of 566 ZIKV sequences sampled between years 2014 and 2019 (**Figure 7**). The least conserved protein was E (Average score = 10.59). NS3 and NS5 were equally conserved (10.70) and NS1 was comparable to the E protein (10.60). The least conserved positions assessed in these four proteins were found in E, corresponding to a section within domain II (highlighted in red). This domain has a putative role in the early

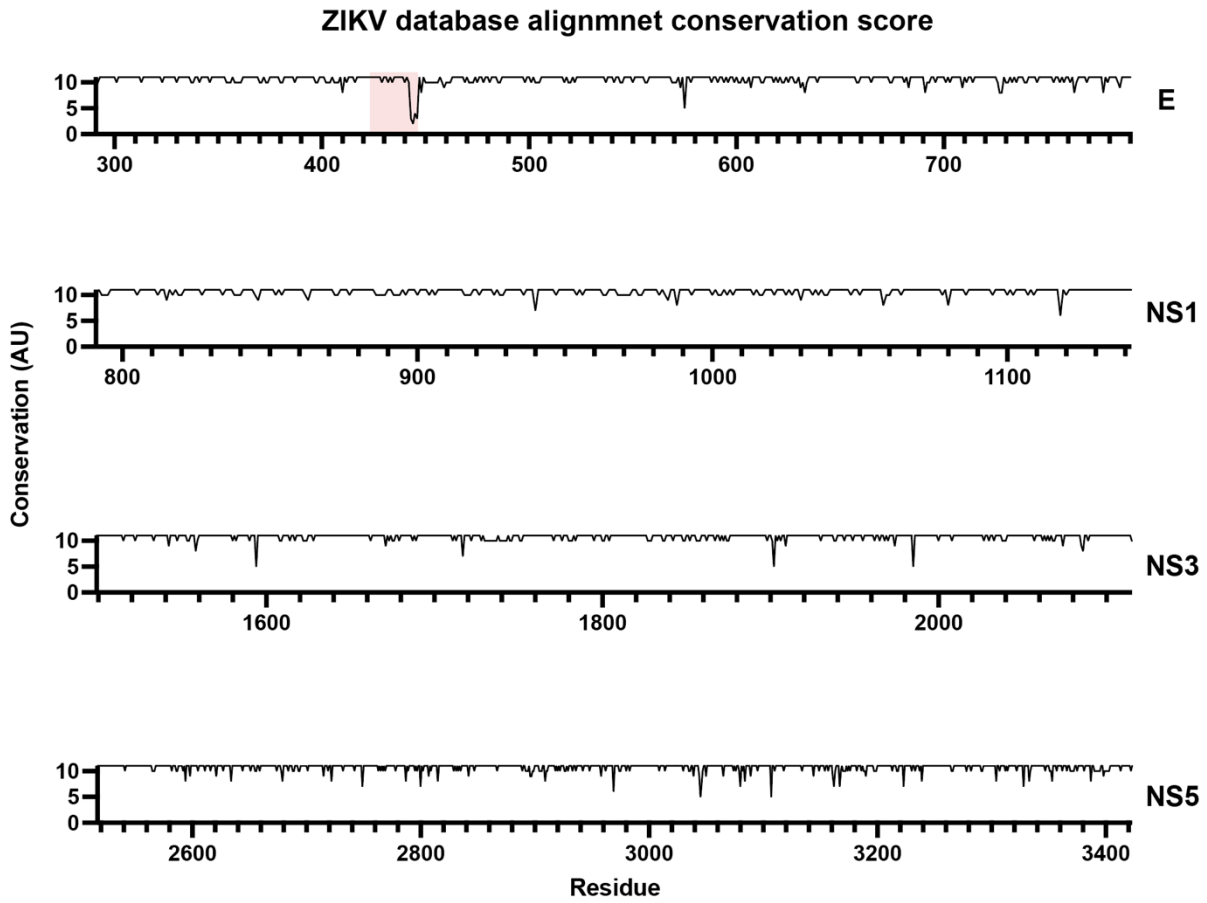


Figure 7. Alignment conservation score of the E, NS1, NS3 and NS5 proteins, based on a dataset of 566 ZIKV full-proteome sequences, from the Asian strain of ZIKV. This metric is measured as a numerical value reflecting the conservation of physico-chemical properties in the alignment: Identities score highest, and the next most conserved group contain substitutions to amino acids lying in the same physico-chemical class.

infection process in numerous *Flavivirus* species, positions which are likely under selective immune pressure [23]. Polymorphism in diagnostic antigen proteins may result in variable sensitivity across regions with discrete virus genotypes. While the analysis cannot describe the extent to which amino acid variability will affect epitope conservation, the metrics shown here are incorporated into the final aggregate score, under a binary scoring scheme. Positions with < 10 conservation are penalised, which will affect the avoidance of such regions.

The next analysis focused on B-cell epitopes that have been identified through *in-vitro* techniques. This dataset was sourced from the Immune Epitope Database (IEDB). The search was limited to Human linear B-cell epitopes only, which were identified in the literature using immunoassays or blots, crystallography, electron microscopy, flow cytometry, nuclear magnetic resonance, or surface plasmon resonance. All epitopes defined with 'positive reactivity' were mapped to the ZIKV polyprotein. The resulting analysis of 1232 epitopes, ranging from 11 to 30 residues in length, is illustrated in **Figure 8,** which shows the frequency of mapped IEDB linear peptides at a given residue position in the four example proteins. A high frequency at a given position indicates that, firstly, this section has been studied in *in-vitro* epitope mapping analyses. It is important to note that in some cases, the IEDB frequency may only be a function of its interest in the literature. Secondly, with the epitopes having been filtered for positive reactivity, a high frequency indicates detectable immunogenicity in assays with human sera.

The mean frequency of the global (entire polyprotein) epitope mapping has been plotted (**Figure 8 - red line = 5.3**) on each protein. The mean frequency for each protein is also shown (**Figure 8 – box**). The protein with the greatest mean IEDB epitope mapping frequency was NS1 (7.3), which was to be expected, given its central role in immunoassay development, and its reported high-level immunogenicity. Interestingly, the second protein with the greatest mean frequency was the NS5 RNA polymerase. This protein has been shown to possess immunodominant B-cell epitopes, despite its intracellular locale, which emphasises the requirement to assess all proteins in these analyses, regardless of their assumed function in viral humoral responses [24]. Conversely, NS3 (serine protease) exhibited the lowest mean frequency of mapped IEDB epitopes (4.1).

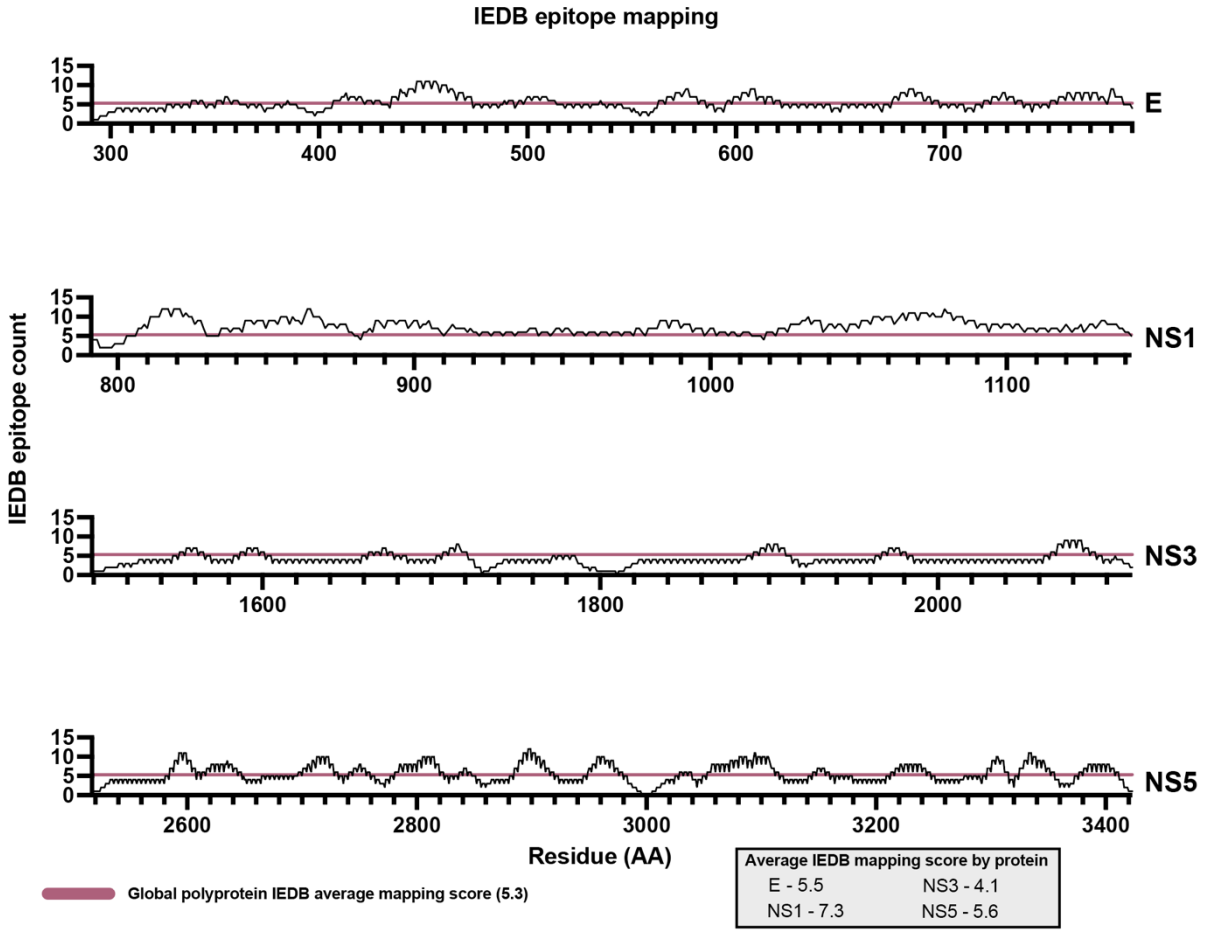


Figure 8. Mapping frequency of *in-vitro* IEDB epitopes to ZIKV proteins. 1232 epitopes were mapped across the ZIKV polyprotein. **Red line –** Mean IEDB mapping frequency across ZIKV polyprotein.

In the final analysis, five epitope prediction tools are combined in a consensus approach to guide the selection of peptides for the final panel. BepiPred, Bcepreds DRREP, LBtope and ABCpred outputs were chosen for this analysis. **Table 2** outlines details of the methodologies underpinning each predictive tool, all of which are trained on a range of datasets from public databases of known epitopes.

Table 2. Epitope prediction servers, the date of publication, the primary machine-learning methodology used, and the public dataset used to train it.

Name	Published	Method	Training
BepiPred-2.0	2017	Ensemble learning	IEDB Linear Epitope Dataset
DRREP	2017	Deep neural network	BC-Pred homology reduced
lbtope	2013	Support vector machine	IEDB Linear Epitope Dataset
Bcepreds	2008	Subsequence kernel	Bcipep database + non-epitopes
ABCPred	2006	Recurrent neural network	Bcipep database + non-epitopes

These predictive tools, except for BepiPred produce a set of candidate peptides in the form of an amino acid sequence. BepiPred produces a linear scale of the epitope probability, on a per-residue basis. For the latter, like the IEDB and k-mer mapping methodology, blastp was used to map the epitopes to the 'continuous' residue number, to enable the comparison of each predictive tool on a single axis. The consensus approach utilises '*min-max'* scaling to balance epitope scores prior to aggregation. With this, an equal weighting was applied to the predictive output of each tool. The association between the tools was assessed across predictions covering the whole ZIKV polyprotein. **Table 3** is a matrix containing the Phi coefficient of each predictive tool. Weak positive relationships were found between the predictions of DRREP and Bcepreds (φ=0.26), DRREP and BepiPred (φ=0.24), and Bcepreds and Bepipred (φ=0.20). ABCped appeared to carry no association with the predictions of the other tools (φ=-0.07 – 0.03). Interestingly, neither of the tools using the same training datasets were found to be associated in their predictions.

Table 3. Phi coefficient matrix comparing the results of epitope prediction software, collated across the entire ZIKV polyprotein. The tools were run with default parameters.

Tool	BepiPred	ABCpred	Bcepreds	DRREP
ABCpred	-0.01			
Bcepreds	0.20	0.02		
DRREP	0.24	-0.07	0.26	
Lbtope	0.13	-0.03	0.03	0.09

Figure 9 illustrates the output of each predictive tool mapped to the AA sequence of the ZIKV NS1 protein. Each coloured line represents each tool's prediction on a 0 to 1 probability scale. The grey line is the aggregate of all tools, scaled between 0 to 5. In this analysis, numerous occasions are observed where three or more tools converge in their predictions, resulting in the peaks seen in the aggregate. There was positive correlation between the increased IEDB epitope mapping score and the aggregate epitope prediction scores, both with the individual NS1 protein ($r = 0.32$, $p < 0.001$) and the entire ZIKV polyprotein ($r = 0.19$, $p < 0.001$).

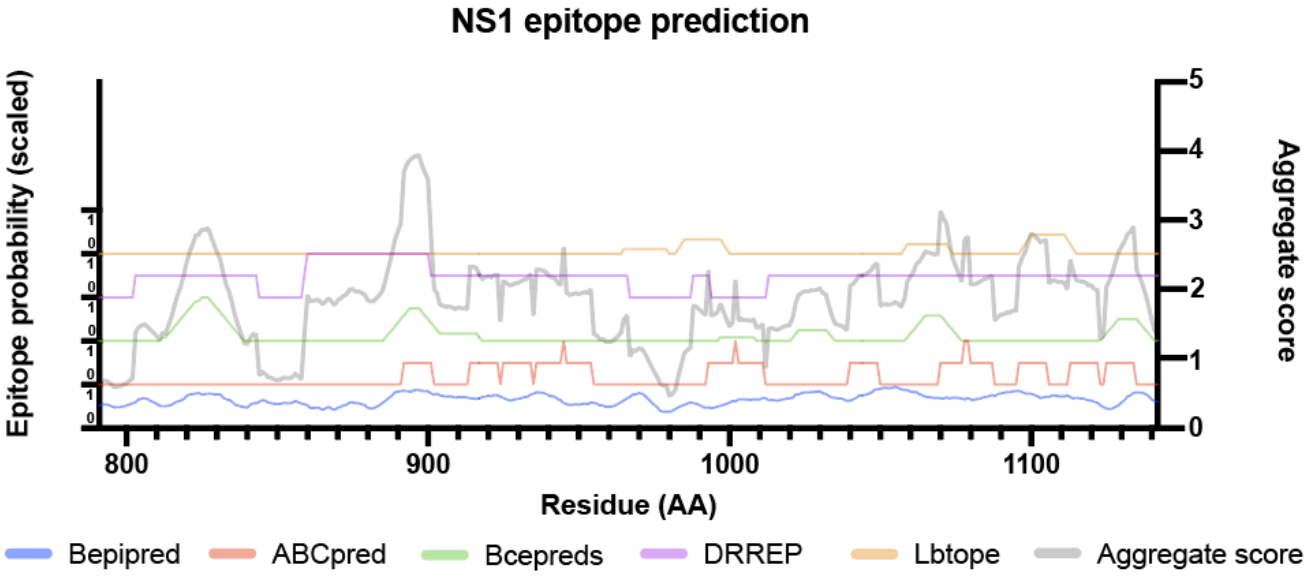


Figure 9. Parallel comparison of epitope predictions across the five tools chosen, focusing on the NS1 protein only, with the aggregate prediction score superimposed (**Right Y axis**). All results from each tool were scaled between 1 and 0.

To integrate these analyses into a final metric to inform the design of our candidate antigen panel, a 'metascore' was constructed. Customised weighting was added to each of the aggregate scores of the above analyses, tailored to an understanding of the significance of each metric. All the scores from these analyses, spanning the entire ZIKV polyprotein were scaled proportionately between 0 and 1. The specificity score from the DENV k-mer mapping, the aggregate epitope prediction score and the IEDB contributed to 84% of the final metascore, in equal parts. Positions with < 10 units of amino acid conservation were penalised with a reduction of 16% to the total score. The mean metascore for the entire polyprotein gene was used to set a threshold for peptide selection. The final filter was a length cut-off that resulted in only the selection of peptides > 15 AA long. Any region > 15 AA that was +1 standard deviation (SD) above the mean metascore was selected for an antigen panel. The metascore for the NS1 protein, and the final candidate peptides selected are shown (**Figure 10).** A summary of peptides selected based on the above analyses, spanning the entire ZIKV polyprotein, is shown in **Table 4.** Also added to this panel, were the results of a literature search, targets from both *in-vitro* and *in-silico* studies, which were tested in **Chapter 6**.

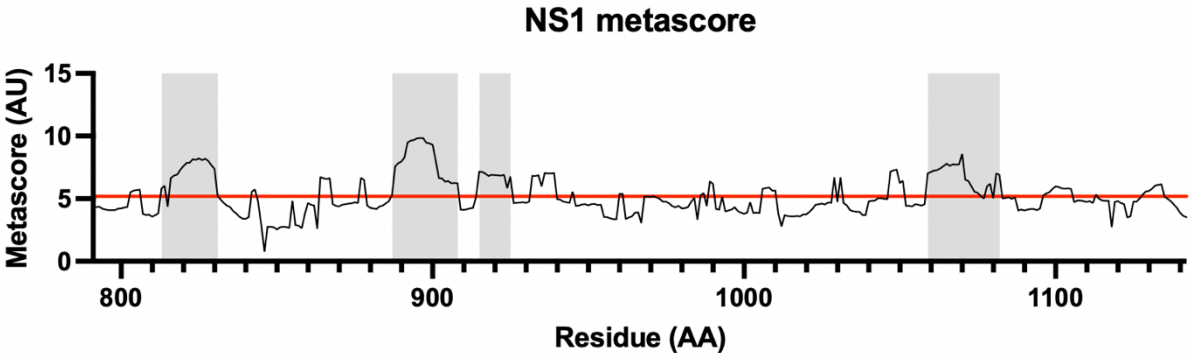


Figure 10. The final metascore, consisting of the aggregated conservation, k-mer mapping, IEDB mepping and epitope prediction scores. The regions highlighted are candidate peptide, selected due to their >15 AA length and 1 SD increase over the mean polyprotein metascore.

Table 3. Candidate peptide list from the peptide epitope prediction pipeline, incorporated with those obtained five other publications (See table S1)

Internal Code	Publication	Pathogen	Gene	Sequence	Notes
PRED_ZIVK_1	NA	ZIKV	prM	MSYECPMLDEGVEPDDVDCW	Epitope prediction workflow
PRED_ZIVK_2	NA	ZIKV	prM	KKGEARRSRRAVTLPSHSTRKLQTRSQTWLESREYTKH	Epitope prediction workflow
PRED_ZIVK_3	NA	ZIKV	E	GSQHSGMIVNDTGHETDENRAKVEITPNSPRAEATLGG	Epitope prediction workflow
PRED_ZIVK_4	NA	ZIKV	NS1	VFVYNDVEAWRDRYKYHP	Epitope prediction workflow
PRED_ZIVK_5	NA	ZIKV	NS1	KNPMWRGPQRLPVPVNELPHG	Epitope prediction workflow
PRED_ZIVK_6	NA	ZIKV	NS1	SYFVRAAKTNNSFVVDGDTLKECPLKHRAWN	Epitope prediction workflow
PRED_ZIVK_7	NA	ZIKV	NS1	KGPWHSEELEIRFEECPGTKVHVE	Epitope prediction workflow
PRED_ZIVK_8	NA	ZIKV	NS3	SYCGPWKLDAAWDGHS	Epitope prediction workflow
PRED_ZIVK_9	NA	ZIKV	NS3	RNPNKPGDEYLYGGGCAET	Epitope prediction workflow
PRED_ZIVK_10	NA	ZIKV	NS3	CFDGTTNNTIMEDSVPAEV	Epitope prediction workflow
ZIKV1	1	ZIKV	prM	HMCDATMSYECPMLDEG	Epitope prediction workflow
ZIKV2	1	ZIKV	prM	HKKGEARRSRRAVTLPSH	
ZIKV3	1	ZIKV	E	TVSNMAEVRSYCYEASIS	
ZIKV4	1	ZIKV	E	ISDMASDSRCPTQGEAYL	
ZIKV_ETP_1	1	ZIKV	NS1	TGVFVYNDVEAWRDRYKY	"Common early marker for ZIKV infections, especially in ZIKV-infected pregnant women"
ZIKV5	1	ZIKV	NS1	REGYRTQMKGPWHSEELE	
ZIKV7	2	ZIKV	NS2	DITWEKDAEVTGNSPRLDVA	
ZIKV8	3	ZIKV	prM	SFPTTLGMNKCYIQIMDL	
DENV1	3	DENV	prM	TSTWVTYGTCNQAG	
DENV2	3	DENV	E	YENLKYTVIITVHTGDQH	
ZIKV9	3	ZIKV	E	GALEAEMDGAKGRLSSGH	
ZIKV10	3	ZIKV	E	FKSLFGGMSWFSQILIGT	
DENV3	3	DENV	NS1	RPGYHTQTAGPWHLGKLE	
DENV4	3	DENV	DENV	LDFNYCEGTTVVITENCG	
ZIKV11	4	ZIKV	ZIKV	SDLIIPKSLAGPLSHHNTREGYRTQMKGPWHSEEL	High performing peptide from screening
ZIKV12	4	ZIKV	NS1	SDLIIPKSLAGPLSHHNTREGYRTQVKGPWHSEEL	Same as above, but with a mutation only found in SEA/AFR
ZIKV13	4	ZIKV	NS2B	VDMYIERAGDITWEKDAEVTGNSPRLDVALDESGDFSLVEDDGPPMREI	
ZIKV14	4	ZIKV	NS2B	VDMYIERAGDITWEKDAEVTGNSPRLDVALDESGDFSLVEEDGPPMREI	Same as above, but with a mutation only found in SEA/AFR
ZIKV15	4	ZIKV	NS3	TRVEMGEAAAIFMTATPPGTRDAFPDSNSPIMDTEVEVPERAWSSGFDWVTDHSGKTVWFVPSVRNGNE	
ZIKV16	4	ZIKV	NS4B	VVTDIDTMTIDPQVEKKMGQVLLIAVAVSSAILSRTAWGWGEAG	
ZIKV17	4	ZIKV	NS4B	VVTDIDTMTIDPQVEKKMGQVLLIAVAISSAVLLRTAWGWGEAG	Same as above, but with a mutation only found in SEA/AFR
WNV1	4		WNV	E (domain 3) GSQLKGTTYGVCSKAFKFLGTPADTGHGTVVLELQYTGTDGPCKVPISSVASLNDLTPVGRLVTVNPFVSVATANAKVLI	
WNV_2	4	WNV	EDIII	ELEPPFGDSYIVVGRGEQQINHHWHKSG	
ZIKV18	5	ZIKV	(preM)	AIAWLLGSSTSQKV	Peptide array
ZIKV19	5	ZIKV	(E)	DRGWGNGCGLFGKGS	Peptide array
ZIKV20	5	ZIKV	(E)	EALVEFKDAHAKRQT	Peptide array
ZIKV21	5	ZIKV	(E)	GGALNSLGKGIHQIF	Peptide array
ZIKV22	5	ZIKV	(NS1)	SKKETRCGTGVFVYN	Peptide array
ZIKV23	5	ZIKV	(NS1)	VNELPHGWKAWGKSH	Peptide array
ZIKV24	5	ZIKV	(NS1)	KECPLKHRAWNSFLV	Peptide array
ZIKV25	5	ZIKV	(NS1)	LECDPAVIGTAVKGK	Peptide array
ZIKV26	5	ZIKV	(NS2)	STSMAVLVAMILGGF	Peptide array
ZIKV27	5	ZIKV	(NS3)	VAAEMEEALRGLPVR	Peptide array
ZIKV28	5	ZIKV	(NS3)	YIMDEAHFTDPSSIA	Peptide array
ZIKV29	5	ZIKV	(NS3)	IAACLTKAGKRVIQL	Peptide array
ZIKV30	5	ZIKV	(NS4)	TFVELMKRGDLPVWL	Peptide array
ZIKV31	5	ZIKV	(NS4)	DGTTNNTIMEDSVPAEVWTRH	Peptide array
ZIKV32	5	ZIKV	(NS4)	LGASAWLMWLSEIEP	Peptide array
ZIKV33	5	ZIKV	(NS4)	SYNNYSLMAMATQAGVLFGMGKGMPFYAWDFGV	Peptide array
ZIKV34	5	ZIKV	(NS4)	VEKKMGQVLLIAVAI	Peptide array
ZIKV35	5	ZIKV	(NS4)	SSAVLLRTAWGWGEA	Peptide array
ZIKV36	5	ZIKV	(NS5)	SSPEVEEARTLRVLS	Peptide array
ZIKV37	5	ZIKV	(NS5)	DRFAHALRFLNDMGK	Peptide array
ZIKV38	5	ZIKV	(NS5)	HRRDLRLMANAICSSV	Peptide array
ZIKV39	5	ZIKV	(NS5)	CSSVPVDWVPTGRTTWSIHGKGEW	Peptide array
FLAVMIX1	5	ZIKV	(E)	GSVGGVFNSLGKGIH	Reactive to all flavivirus
FLAVMIX2	5	ZIKV	(NS1)	DGVEESDLIIPKSLA	Reactive to all flavivirus
FLAVMIX3	5	ZIKV	(NS3)	DGDIGAVALDYPAGT	Reactive to all flavivirus
FLAVMIX4	5	ZIKV	(NS5)	SLVNGVVRLLSKPWD	Reactive to all flavivirus

5.3 Discussion

In this chapter, a range of techniques that can be used to guide antigen design have been explored, with its primary application in increasing the likelihood of selecting both reactive and specific candidates. The first application, expressing the Concat peptide, yielded sub-optimal results. It is apparent that the large protein fusion partner that was chosen, dwarfed the signal obtained from the Concat peptide. GST fusions have been used previously in immunoassays for malaria surveillance [25], however, in this application, I believe the 26 kDa GST fusion is too reactive, when considering the mass of Concat is less than half (11.5 kDa). A further limitation in this methodology lies in the concept of joining discrete peptides in a single contiguous chain, as Concat was. It is possible that the synthetic concatenation of these peptide may create new peptide sequences that mimic other linear epitopes, besides those present on the NS1 protein, which may increase unwanted background signal. It was observed that the peptide yielded a significantly weaker signal, when compared to other, larger antigens. While this may not necessarily be a major limitation, providing there is a differential between immune and non-immune signals detected, it would be prudent to explore ways to increase the overall signal of small peptide antigens.

The summary analyses comprising the final metascore, may require tuning, given the complexities associated with integrating such datasets. For example, while the literature for each epitope prediction tool reports the performance of classification by means of area under curve (AUC) metric, the heterogeneous training and testing conditions associated with each tool make assessing their validity difficult, a problem discussed in the final chapter. It is for this reason a consensus approach may be effective. Depending on the importance of each characteristic to the main objectives of the antigen panel design, weightings can be adjusted accordingly.

Data pools were generated for all the arboviruses featured in **Figure 6,** including YFV, WNV, DENV, JEV, CHIKV and MAYV. To expand on our analysis here, I'd like to combine these datasets, cross-examining them and comparing the resulting metascores, to determine if there are features that are common between the species. Additionally, another analysis was omitted, which is featured in the SARS-CoV-2 example. This is based on secondary structural features inferred by predicative/analytical tools as well as data obtained from the UniProt website. An example plot of these data is illustrated in **Figure S1**. While it is generally understood that 'turns' or disordered loops between helices or sheets are more likely to be a linear epitope, it was felt that more information was needed before including it in the final metascore calculations.

The k-mer mapping analysis incorporated a function I called 'tracing'. For each kmer in the pool, is was assigned a unique code, allowing the identification of the origin of the mapped k-mer. I believe that with this information, there is are a wealth of analyses to be performed, not only in the context of *Flavivirus* species, but in our wider understanding of antibody cross reactivity, particularly background signal. Again, this is expanded on further in the final chapter.

5.4 References

1. Pettersson JH-O, Fiz-Palacios O. Dating the origin of the genus Flavivirus in the light of Beringian biogeography. Journal of General Virology. Microbiology Society; 2014;95:1969–82.

2. Chávez JH, Silva JR, Amarilla AA, Figueiredo LTM. Domain III peptides from flavivirus envelope protein are useful antigens for serologic diagnosis and targets for immunization. Biologicals. Elsevier; 2010;38:613–8.

3. Carbaugh DL, Lazear HM. Flavivirus envelope protein glycosylation: impacts on viral infection and pathogenesis. Journal of virology. Am Soc Microbiol; 2020;94:e00104-20.

4. Grant OC, Montgomery D, Ito K, Woods RJ. Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. Scientific reports. Nature Publishing Group; 2020;10:1–11.

5. Shukla R, Ramasamy V, Shanmugam RK, Ahuja R, Khanna N. Antibody-dependent enhancement: a challenge for developing a safe dengue vaccine. Frontiers in Cellular and Infection Microbiology. Frontiers; 2020;597.

6. Denis J, Attoumani S, Gravier P, Tenebray B, Garnier A, Briolant S, *et al.* High specificity and sensitivity of Zika EDIII-based ELISA diagnosis highlighted by a large human reference panel. Rodriguez-Barraquer I, editor. PLOS Neglected Tropical Diseases. Public Library of Science; 2019;13:e0007747.

7. Tripathi NK, Shrivastava A. Recent Developments in Recombinant Protein-Based Dengue Vaccines. Frontiers in immunology. NLM (Medline); 2018. p. 1919.

8. Rastogi M, Sharma N, Singh SK. Flavivirus NS1: a multifaceted enigmatic viral protein. Virology Journal. Virology Journal; 2016;13:131.

9. Mora-Cárdenas E, Aloise C, Faoro V, Gašper NK, Korva M, Caracciolo I, *et al.* Comparative specificity and sensitivity of NS1-based serological assays for the detection of flavivirus immune response. PLOS Neglected Tropical Diseases. Public Library of Science; 2020;14:e0008039.

10. Bosch I, de Puig H, Hiley M, Carré-Camps M, Perdomo-Celis F, Narváez CF, *et al.* Rapid antigen tests for dengue virus serotypes and Zika virus in patient serum. Science translational medicine. American Association for the Advancement of Science; 2017;9:eaan1589.

11. Chaterji S, Allen Jr JC, Chow A, Leo Y-S, Ooi E-E. Evaluation of the NS1 rapid test and the WHO dengue classification schemes for use as bedside diagnosis of acute dengue fever in adults. The American journal of tropical medicine and hygiene. The American Society of Tropical Medicine and Hygiene; 2011;84:224.

12. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins: Structure, Function, and Bioinformatics. 2006;65:40–8.

13. Sher G, Zhi D, Zhang S. DRREP: deep ridge regressed epitope predictor. BMC Genomics. BioMed Central; 2017;18:676.

14. Ponomarenko J, Bui H-H, Li W, Fusseder N, Bourne PE, Sette A, *et al.* ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics. BioMed Central; 2008;9:514.

15. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PloS one. Public Library of Science San Francisco, CA USA; 2016;11:e0163962.

16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215:403–10.

17. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. Narnia; 2009;25:1189–91.

18. Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Bioinformatics. Oxford University Press; 1993;9:745–56.

19. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, *et al.* The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Research. Oxford University Press; 2019;47:D339–43.

20. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic acids research. 2017;45:W24–9.

21. Singh H, Ansari HR, Raghava GPS. Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence. Schönbach C, editor. PLoS ONE. Public Library of Science; 2013;8:e62216.

22. Davydov YI, Tonevitsky AG. Prediction of linear B-cell epitopes. Molecular Biology. Springer; 2009;43:150–8.

23. Hu T, Wu Z, Wu S, Chen S, Cheng A. The key amino acids of E protein involved in early flavivirus infection: viral entry. Virology Journal. 2021;18:136.

24. Fumagalli MJ, Figueiredo LTM, Aquino VH. Linear and Continuous Flavivirus Epitopes From Naturally Infected Humans. Frontiers in cellular and infection microbiology. Frontiers; 2021;705.

25. Tetteh KKA, Wu L, Hall T, Ssewanyana I, Oulton T, Patterson C, *et al.* Optimisation and standardisation of a multiplex immunoassay of diverse Plasmodium falciparum antigens to assess changes in malaria transmission using sero-epidemiology. Wellcome Open Research. F1000 Research Ltd; 2020;4:26.

5.5 Supplementary figures

Figure S1. Structural metanalysis of ZIKV NS1 protein. The predictions were compiled using BepiPred software. The UniProt structural information was extracted from an XML formatted file downloaded from UniProt website on 08/02/2022.

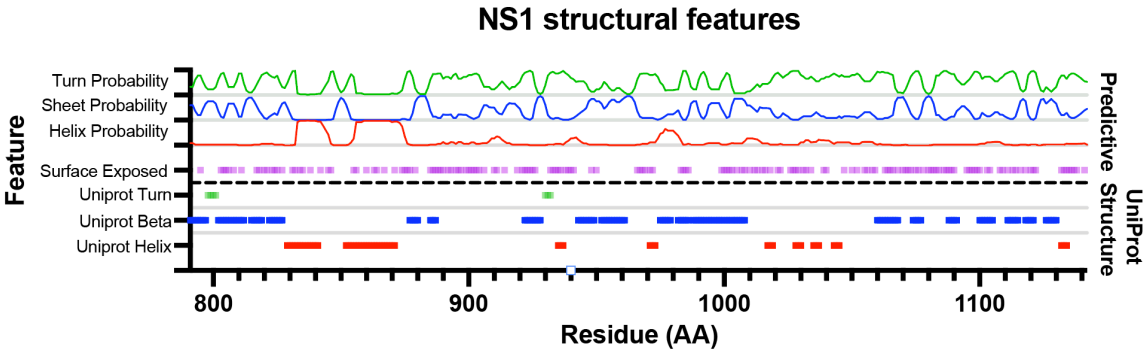


Table S1. Sources for ZIKV peptide metanalysis

Paper Title	Citation
ZIKV-Specific NS1 Epitopes as Serological Markers of Acute Zika Virus Infection	27
Identification of diagnostic peptide regions that distinguish Zika virus from related mosquito-borne Flaviviruses	28
Diagnosis of Zika Virus Infection by Peptide Array and Enzyme-Linked Immunosorbent Assay	29
Differential human antibody repertoires following Zika infection and the implications for serodiagnostics and disease outcome	30
Diagnosing Zika virus infection against a background of other flaviviruses: Studies in high resolution serological analysis	31

5.6 RESEARCH PAPER COVER SHEET

SECTION A – Student Details

Student ID Number	**1603403**	**Title**	Mr
First Name(s)	Daniel		
Surname/Family Name	Ward		
Thesis Title	Zika virus surveillance in human and mosquito populations in Cabo Verde – exploring molecular and serological tools for the surveillance of emerging infectious diseases		
Primary Supervisor	Prof. Taane Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	**Genome Medicine**		
When was the work published?	**January 2021**		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	**N/A**
Please list the paper's authors in the intended authorship order:	N/A
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	**DW: Assembly of data, design of analyses, tools and website. DW wrote the manuscript.** **MH and JP contributed to backend code for the mutation tracking function**

SECTION E

Student Signature 

Date 20/04/2022

Supervisor Signature 

Date April 20, 2022

Ward *et al. Genome Medicine* (2021) 13:4
https://doi.org/10.1186/s13073-020-00822-6

Genome Medicine

DATABASE Open Access

An integrated in silico immuno-genetic analytical platform provides insights into COVID-19 serological and vaccine targets

Daniel Ward[1*], Matthew Higgins[1], Jody E. Phelan[1], Martin L. Hibberd[1], Susana Campino[1] and Taane G. Clark[1,2*]

Abstract

During COVID-19, diagnostic serological tools and vaccines have been developed. To inform control activities in a post-vaccine surveillance setting, we have developed an online "immuno-analytics" resource that combines epitope, sequence, protein and SARS-CoV-2 mutation analysis. SARS-CoV-2 spike and nucleocapsid proteins are both vaccine and serological diagnostic targets. Using the tool, the nucleocapsid protein appears to be a sub-optimal target for use in serological platforms. Spike D614G (and nsp12 L314P) mutations were most frequent (> 86%), whilst spike A222V/L18F have recently increased. Also, Orf3a proteins may be a suitable target for serology. The tool can accessed from: http://genomics.lshtm.ac.uk/immuno (online); https://github.com/dan-ward-bio/COVID-immunoanalytics (source code).

Keywords: SARS-CoV-2, COVID, Human-coronavirus, Immuno-informatics, Mutation, Epitopes, Cross-reactivity, Surveillance

Background

COVID-19, the disease caused by the SARS-CoV-2 virus, was first characterised in the city of Wuhan, Hubei, and has now spread to 190 countries. With over 60 million confirmed cases worldwide and more than 1.26 million deaths, the COVID-19 pandemic has placed a high burden on the world's healthcare infrastructure and economies, with projected final costs of 28 trillion or 31% of the global gross domestic product [1, 2]. The majority of infections are either asymptomatic or result in mild flu-like symptoms, with severe cases of viral pneumonia affecting between 1.0% (≥ 20 years) and 18.4% (≥ 80 years) of diagnosed patients [3]. Its variable infection outcome, mode of transmission and incubation period together have enhanced the ability of the pathogen to spread efficiently worldwide. As a result, there has been an urgent push for the development of diagnostics, therapeutics and vaccines to aid control efforts.

Current front-line diagnostic strategies apply a quantitative reverse transcription PCR (RT-qPCR) assay on patient nasopharyngeal swabs, using primer/probe sets targeting the *nsp10*, *RdRp*, *nsp14*, envelope and nucleocapsid genes; tests endorsed by a number of agencies and health systems [4, 5]. Patients hospitalised with severe respiratory disease who are RT-qPCR negative may be diagnosed radiographically (chest x-ray or computerised tomography scan), but in limited resource or high infection rate settings, these methods maybe unviable. Considering the inherent limitations in the sample collection process and transient viral load, RNA detection-based diagnostics can vary in their sensitivity. The demand for serological diagnostics is high, particularly because these tests are capable of detecting SARS-CoV-2 antibodies, which are biomarkers indicative of current infection that remains present after viral clearance [6]. These tools are essential to address crucial sero-epidemiological questions, like understanding viral

* Correspondence: Daniel.ward1@lshtm.ac.uk; Taane.clark@lshtm.ac.uk
[1]Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
Full list of author information is available at the end of the article

prevalence across a population, the percentage of asymptomatic patients and longevity of antibody responses post infection.

Numerous lateral flow rapid diagnostic tests (RDTs) and enzyme-linked immunosorbent assays (ELISA) tests have been developed, including an approved IgM/IgG RDT which uses the nucleocapsid protein as a target for the detection of seroconverted individuals [7]. Other assays use the spike protein as an antigen, with some using the receptor-binding domain (RBD) as a target, a region with a high level of diversity between alphacoronavirus species [8]. Unlike RNA detection methods, these platforms can identify convalescent patients, which further informs outbreak control efforts.

Long-term control strategies will involve vaccine roll-out. As of November 2020, there were more than 50 vaccines at development phase 1 or greater, with at least 10 vaccines in phase 3 [9, 10]. Vaccines at the forefront include those based on a non-replicating adenovirus vector base (ChAdOx-nCoV-19 and Ad5-nCov), an LNP-encapsulated mRNA (BNT162 and mRNA-1273), protein subunit (NVX-CoV2373) or inactivated virus (BBIBP-CorV and CoronaVac) [11]. The discovery, development and management of efficacious vaccines, as well as sensitive and specific serological diagnostics, are both dependant on the availability of up-to-date information on viral evolution and immune-informatic analyses. The identification of variable or conserved regions in the proteome of SARS-CoV-2 can inform the rational selection of reverse-design targets in both vaccinology and diagnostic fields, as well as indicate immunologically relevant regions of interest for further studies to characterise SARS-CoV-2 immune responses. Whilst the availability of biological data for SARS-CoV-2 in the public domain has increased, insights are most likely to come from its integration informatically in an open and accessible format.

Construction and content

Rationale

Here, we present an online integrated immuno-analytic resource for the visualisation and extraction of SARS-CoV-2 meta-analysis data [12]. This website was built around an automated pipeline for the formation of a whole genome sequence based variant database for SARS-CoV-2 isolates worldwide (as of November 2020, $n = 150{,}090$). We have integrated this dataset with a suite of B cell epitope prediction platform meta analyses, HLA-I and HLA-II peptide prediction, an 'epitope mapping' analysis of available experimental in vitro confirmed epitope data from the Immune Epitope Database (IEDB) [13] and a protein orthologue sequence analysis of six relevant coronavirus species (SARS, MERS, OC43, HKU1, NL63 and 229E), with all data updated and annotated regularly with information from the UniProt database. Additional functionality enables users to visualise external analytical datasets presented in the literature (e.g. [14]). Moreover, we have added functionally to spatio-temporally track non-synonymous mutations of interest through the dataset, allowing up-to-date surveillance of mutations that may be of immunological relevance. With this resource, users can browse the annotated SARS-CoV-2 proteome and extract meta data to inform further research and analyses.

Whole genome sequence data analysis

SARS-CoV-2 nucleotide sequences were downloaded from NCBI (https://www.ncbi.nlm.nih.gov) and GISAID (https://www.gisaid.org). As part of an automated in-house pipeline, sequences were aligned using MAFFT software (v7.2) [15] and trimmed to the beginning of the first reading frame (orf1ab-nsp1). Sequences with > 20% missing were excluded from the dataset. Using data available from the NCBI COVID-19 resource, a modified annotation (GFF) file was generated and open reading frames (ORFs) for each respective viral protein were extracted (taking in to account ribosomal slippage) using the bedtools 'getfasta' function [16]. Each ORF was translated using EMBOSS transeq software [17], and the variants for each protein sequence were identified using an in-house script [18]. As a part of our analysis pipeline, we generated consensus sequences for each SARS-CoV-2 protein from the nucleotide database using the EMBOSS Cons CLI tool [17]. These canonical sequences were used as a reference for prediction, specificity and epitope mapping analyses.

B cell epitope prediction meta-analysis

Six epitope prediction software platforms were chosen for this analysis (Bepipred [19], AAPpred [20] DRREP [21], ABCpred [22], LBtope [23] and BCEpreds [24]). For each tool, we used the settings and quality cut-offs as recommended by their respective authors. The scores across the predictive platforms were then normalised (minimum-maximum scaled) to ensure that no single tool skewed the aggregate 'consensus' score, and combined to provide a single consensus B cell epitope prediction score. Within the 'raw data table' (accessed from the tool's landing page), users can dissect each score depending on their preference of methodology.

HLA-I and HLA-II peptide prediction

We have incorporated an HLA-I peptide prediction analysis within the tool to aid in the scrutiny and development of vaccine candidates. $CD8^+$ effector immunity has been reported to play a central role in the response to SARS-CoV infection, as well as infection mediated immunopathology [25–27]. We used a database of 2915

HLA-A, HLA-B and HLA-C alleles to make HLA-I peptide binding predictions using the netMHCpan server (v4.1) [28], with peptide lengths of 8 to 14 amino acids across the entire SARS-CoV-2 proteome. We chose to use the netMHCpan server for our HLA-I peptide prediction analysis, due to its high overall performance and its extensive HLA-I allele database [28]. We ran predictions for a total of 2915 alleles (HLA-A 886, HLA-B 1412 and HLA-C 617) across all peptide lengths (8–14 amino acids). The analysis generated 1.1 billion candidates. After quality control, we selected a total of 736,073 peptides based on strong binding affinity across the allele database. We selected strong binding affinity peptides based on the tools internal binding scoring metrics. Only 'strong binding' alleles were selected for further analysis. For each position with a ligand with high binding affinity, we analysed the percentage representation of the respective HLA-I type across the allele database. For predicting HLA-II peptides we used the MARIA online tool [29]. We pre-processed the SARS-CoV-2 canonical protein sequences using a 15 amino acid sliding window. Predictions were made for all available HLA-II alleles. A 95% cut off was chosen for a positive HLA-II presentation. All data for each 15-mer is displayed on the tool.

Epitope mapping

B cell epitopes for coronavirus species were sourced from the Immune Epitope Database (IEDB) resource (https://www.iedb.org, updated: October 2020) [13]. Using BLASTp [30], we mapped short amino acid epitope sequences onto the canonical sequence of SARS-CoV-2 proteins. A BLASTp bitscore of 25 with a minimum length of 8 residues was selected as a quality cut-off for mapped epitopes. The frequency of mapped epitopes was logged for each position in the protein and parsed for graphical representation.

Coronavirus homology analysis

Reference proteomes for SARS, MERS, OC43, 229E, HKU1 and NL63 α and β coronavirus (-CoV) species were sourced from UniProt database. These sequences were processed into 10-mers using the *pyfasta* platform and mapped on to the canonical sequences of SARS-CoV-2 proteins using the aforementioned 'epitope mapping' process. The k-mer mapping technique applied a matching threshold of at least 10 residues in orthologous viral proteins, which is of sufficient length to cover HLA-bound peptides and/or whole or part of a B cell epitope, something that is challenging using only pairwise multiple sequence alignments. Homologous peptide sequences with a BLAST bitscore indicating 10 or more residues mapped to the target sequence were recorded and parsed for display on the graph.

Online SARS-CoV-2 "Immuno-analytics" resource and analysis software

We developed an online immuno-analytics resource with an interactive plot that integrates up-to-date SARS-CoV-2 genetic variation analysis, T and B cell epitope prediction and mapping, human coronavirus homology mapping, literature meta-analysis and an accessible database for extracting data for further study. This tool is available online from http://genomics.lshtm.ac.uk/immuno. The source code for the website and up-to-date raw data files are available at https://github.com/dan-ward-bio/COVID-immunoanalytics [18] (see Additional file 1: Fig. S1 for screenshots). The BioCircos.js library [12] was used to generate the interactive plot and *Datatables.net* libraries for the table. The underlying web-tool software and in-house pipelines for data analysis are available at https://github.com/dan-ward-bio/COVID-immunoanalytics [18].

Metadata consisting of collection date and source (geographical) location for each GISAID sequence are analysed. Temporal and geographic data on individual mutations can be found on the 'Mutation Tracker' page, accessed via the tool's home page. For the spatio-temporal mutation plots, we partitioned the whole genome sequencing dataset by week and continent and plotted non-synonymous allele frequencies using Google Charts JavaScript libraries. To improve sustainability of the tool, all functions and data of the website are generated and updated using automated data scripts developed in-house.

Utility and discussion

To demonstrate the functionality of the immuno-analytics tool, we present an analysis of the SARS-CoV-2 spike, nucleocapsid and orf3a proteins, which are vaccine and serological targets. Analysis of 150,090 SARS-CoV-2 sequences identified 911,324 non-synonymous mutations across 16,951 sites in protein-coding regions; 0.71% of these mutations are singleton events and 0.03% (46) of these mutations have a frequency above 1%, occurring in > 1500 samples. The most frequent mutations were the spike protein D614G (87.3%) and nsp12 L314P (87.5%), which were common across all the geographical regions (all > 86%) (Table 1), in keeping with their deep ancestral nature in the SARS-CoV-2 phylogenetic tree [24]. In particular, nsp12 L314P has been used to genotype the putative S and L strains of SARS-CoV-2, which have now been clustered into further groups [31]. Spike D614G lies 73 residues downstream from the spike RBD, a region of interest as it is a primary target of protective humoral responses and bears immunodominant epitopes that play a possible role in antibody dependant enhancement [32–35]. We have observed a strong correlation between the spatiotemporal accumulation of both spike D614G and nsp12 L314P (Fig. 1, Table 1),

Table 1 Most frequent non-synonymous mutations found in the 150,090 global SARS-CoV-2 whole genome sequences

Protein*	Pos.	Ref. Allele	Alt. Allele	Alt. Allele Freq.	Dominant Alt. Allele Freq. (150090)	Europe (88266)	Asia (7818)	NAm (38203)	SAm (1702)	AFR (1211)	OCE (12837)	UK (68017)**
S	614	D	V:G:S:N	1:131490:2:10	0.877	0.975	0.940	0.866	0.890	0.866	0.868	0.883
nsp12	314	N	H:F:L:S	2:114:131053:1	0.876	0.973	0.938	0.863	0.887	0.876	0.865	0.881
N	203	R	G:K:M:I:S	9:57773:73:1:37	0.388	0.426	0.394	0.374	0.316	0.393	0.440	0.380
N	204	G	R:V:L:Q:T	57,327:5:319:4:1	0.385	0.424	0.393	0.373	0.312	0.392	0.439	0.379
orf3a	57	Q	K:Y:R:H:L	1:6:6:35047:2	0.234	0.267	0.227	0.236	0.210	0.193	0.199	0.248
nsp2	85	T	V:I	1:25913	0.173	0.195	0.156	0.180	0.160	0.090	0.153	0.180
nsp6	37	L	F	11,992	0.080	0.087	0.112	0.079	0.061	0.082	0.073	0.079
nsp2	120	I	V:F:M	7:11996:1	0.080	0.094	0.061	0.082	0.059	0.054	0.059	0.083
S	222	A	T:V:P:I:S:F	4:11819:2:1:12:1	0.079	0.099	0.148	0.057	0.177	0.135	0.026	0.085
orf10	30	V	A:I:L	2:2:11619	0.078	0.097	0.146	0.056	0.176	0.132	0.025	0.084
N	220	A	V:T	11,555:5	0.077	0.097	0.146	0.056	0.176	0.131	0.025	0.083
S	477	S	T:R:G:I:N:K	1:19:2:59:9811:1	0.067	0.077	0.044	0.070	0.048	0.045	0.043	0.068
N	194	S	A:L:P:T	30:6441:2:1	0.043	0.052	0.043	0.035	0.031	0.053	0.038	0.049
orf8	84	L	F:C:S:V	1:1:6338:1	0.042	0.046	0.053	0.043	0.049	0.040	0.048	0.042
orf8	24	S	L	6151	0.041	0.053	0.033	0.036	0.034	0.009	0.022	0.048
S	18	L	F:I	5888:1	0.040	0.048	0.071	0.030	0.106	0.059	0.017	0.041
nsp5	15	G	S:D	5433:5	0.036	0.039	0.035	0.036	0.022	0.038	0.044	0.036
orf3a	251	G	V:S:D:C	5252:31:4:6	0.035	0.035	0.063	0.037	0.031	0.040	0.034	0.030
nsp13	541	Y	C	2606	0.017	0.019	0.032	0.017	0.018	0.021	0.019	0.017
nsp13	504	P	L:H:S	2535:1:111	0.017	0.020	0.032	0.017	0.018	0.022	0.021	0.017

Pos. position, *Freq.* frequency, *NAm* North America, *SAm* South America, *AFR* Africa, *OCE* Oceania, *REF* reference, *ALT* alternative, **S* spike, *M* membrane, *N* nucleocapsid, ** included in Europe

due to either a common origin and subsequently linked accumulation by a founder effect or a more complex biological interaction, including positive selection driven in part by increased transmissibility, as suggested by a recent study [36]. Specifically, the spike D614G and nsp12 N314L both appear to have a near-identical frequency with a consistent increase across all geographic regions (negating weeks with poor data collection). In contrast, the frequency of orf3a Q57H appears to fluctuate, increasing and decreasing significantly from the time it was first observed in February 2020 (week 8) to November 2020, week 43) (Fig. 1; Table 1). Using the immuno-analytical tool, spike A222V, S477N and L18F variants were observed to have increased significantly in frequency between May and November 2020 (weeks 23–40). Spike mutations A222V and L18F appear to have become entrenched in Europe reaching a total frequency of 70.6% and 31.6%, respectively (Additional file 1: Table S1, Additional file 1: Fig. S2). Moreover, A222V appears to be increasing in Asia and Oceania from week 41, and S477N has increased to > 95% frequency across Oceania ($N = 8321$) with a peak of 9.3% in Europe (Additional file 1: Table S1, Additional file 1: Fig. S2), consistent with a recent report [37].

The proximity of the D614G mutation to one of the functional domains of the spike protein has raised concerns, but whether it confers any gain in pathogenicity, transmissibility or immune evasion is still unclear [32]. Other high frequency mutations occur on the nucleocapsid gene (R203K, 38.8.0%; G204R 38.4%; across all geographical regions > 31%; Table 1), which has been the target antigen for several serological RDTs currently in use or in production. Both of these mutations share a near-identical spatio-temporal profile. We have identified 363 non-synonymous variant sites across the nucleocapsid gene with mutations occurring 173,955 times in this dataset. Using the SARS-CoV-2 immuno-analytics platform, we further queried these polymorphic regions for immunological relevance. The 20 residues surrounding the spike mutation D614G (S604-624)

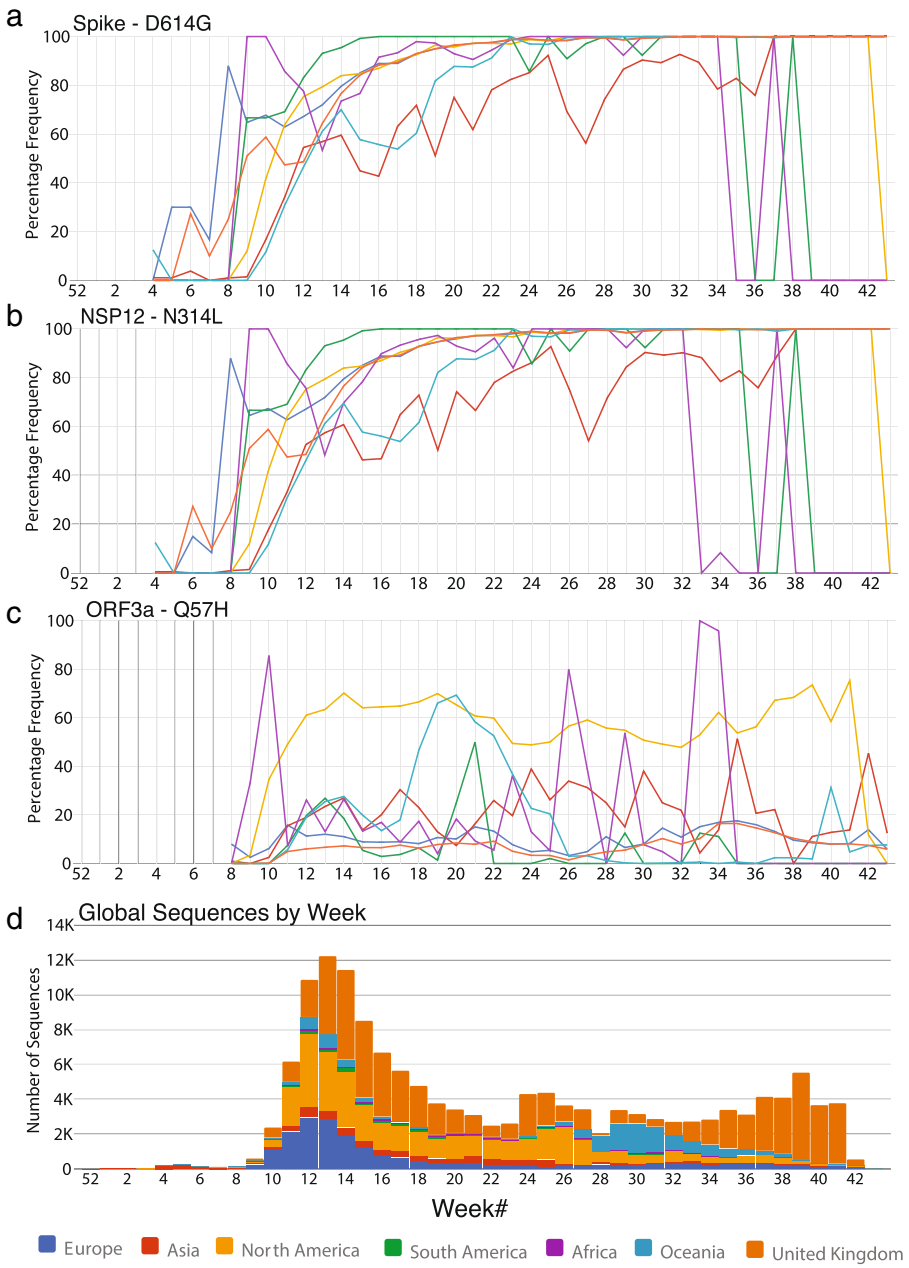


Fig. 1 a–c High frequency non-synonymous spike D614G, NSP12 N314L and orf3A Q57H mutations found in SARS-CoV-2 plotted weekly by continent. This functionality is available on the website. Users can select any mutation from the main plot and visualise it temporally and geographically. UK sequences are not included in Europe due to high frequency. **d** A stacked bar chart representing total sequences published by each continent by week. This chart is included to assist users understand how allele frequencies may be affected by poor sampling. (see http://genomics.lshtm.ac.uk/immuno/mutation_tracker)

(Fig. 2) have a high epitope prediction meta-score (34% increase on the global median) with 204 IEDB epitope positions mapping to the surrounding residues, suggesting that this region is of high interest and may elicit a strong immune response. On top of the high level of SARS-CoV sequence homology reported, we have identified multiple clusters in the S2 domain of the spike protein, with homology to MERS, OC43, 229E, HKU1 and NL63 human coronaviruses, which may elicit a cross-reactive immune response in immune sera. Human coronavirus sequence homology is greatly reduced in the S1 domain, with only two small 10-residue pockets of OC43 and HKU1 identity (see Fig. 2). We observed a 17% increase over the median epitope meta-

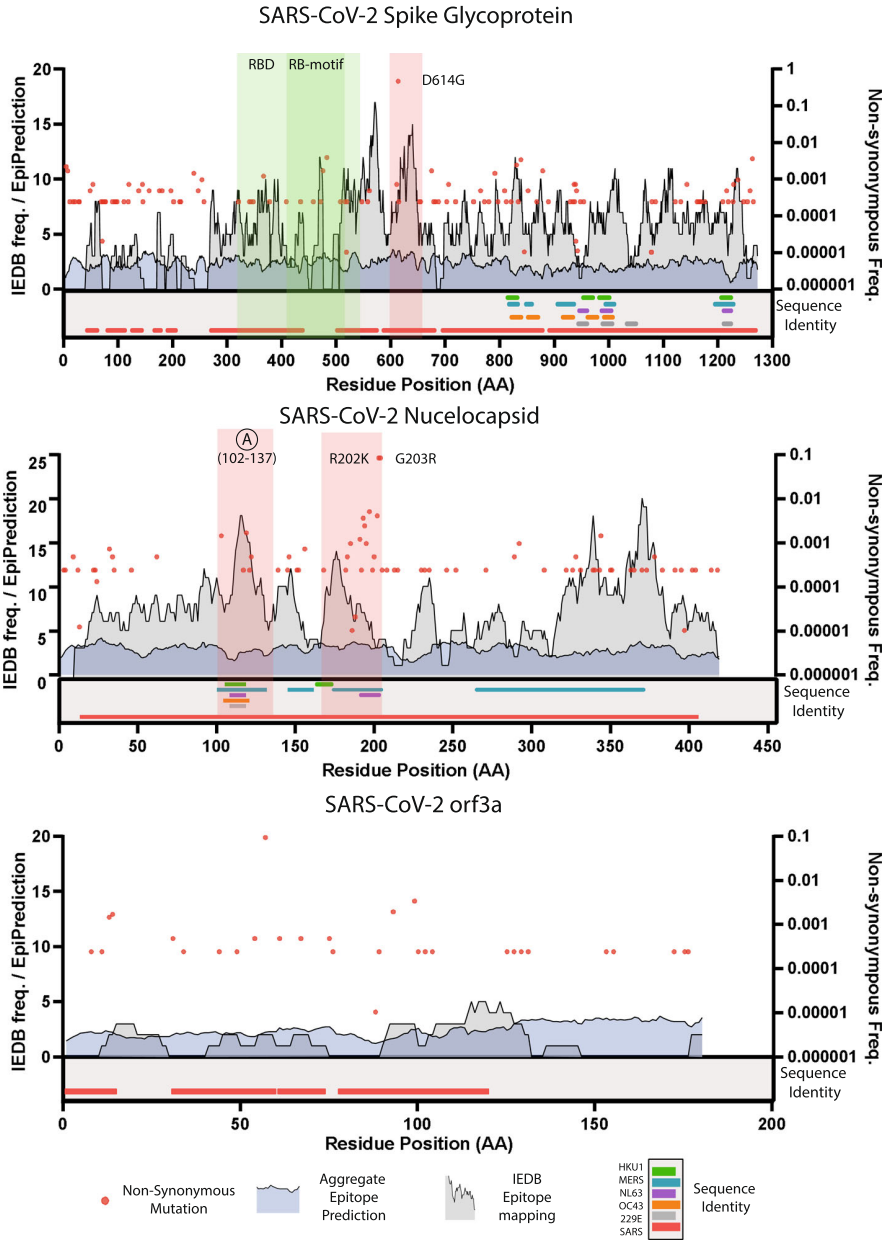


Fig. 2 Linearised extracts from the SARS-CoV-2 immuno-analytics resource database; spike, nucleocapsid and orf3a proteins. Non-synonymous mutations, B cell epitope prediction meta-score, IEDB epitope mapping and sequence identity analyses were plotted (see key). Left axis denotes the scale for epitope prediction and IEDB mapping. Right axis denotes the relative allele frequency found in the genome dataset; AA, amino acid

score within the receptor-binding motif (AA437-508), a region implicated in the direct ACE2 (angiotensin-converting enzyme 2) interaction. HLA-II peptide binding prediction (see Fig. 3) yielded several epitopes within the receptor-binding domain with high HLA-II ligand affinity, as well as strong B cell epitope prediction scores (28% above the global median). Metadata obtained from the UniProt database reveals 3 clusters of glycosylated residues across the spike protein, a characteristic highlighted by this tool that should be considered when choosing expression systems for producing protein/peptides based on these regions.

For two high-frequency non-synonymous nucleocapsid protein mutations (R203K and G204R; co-fixed), all but three of the 30 (N173-234) flanking residues have non-synonymous variants, with 5 sites reporting an

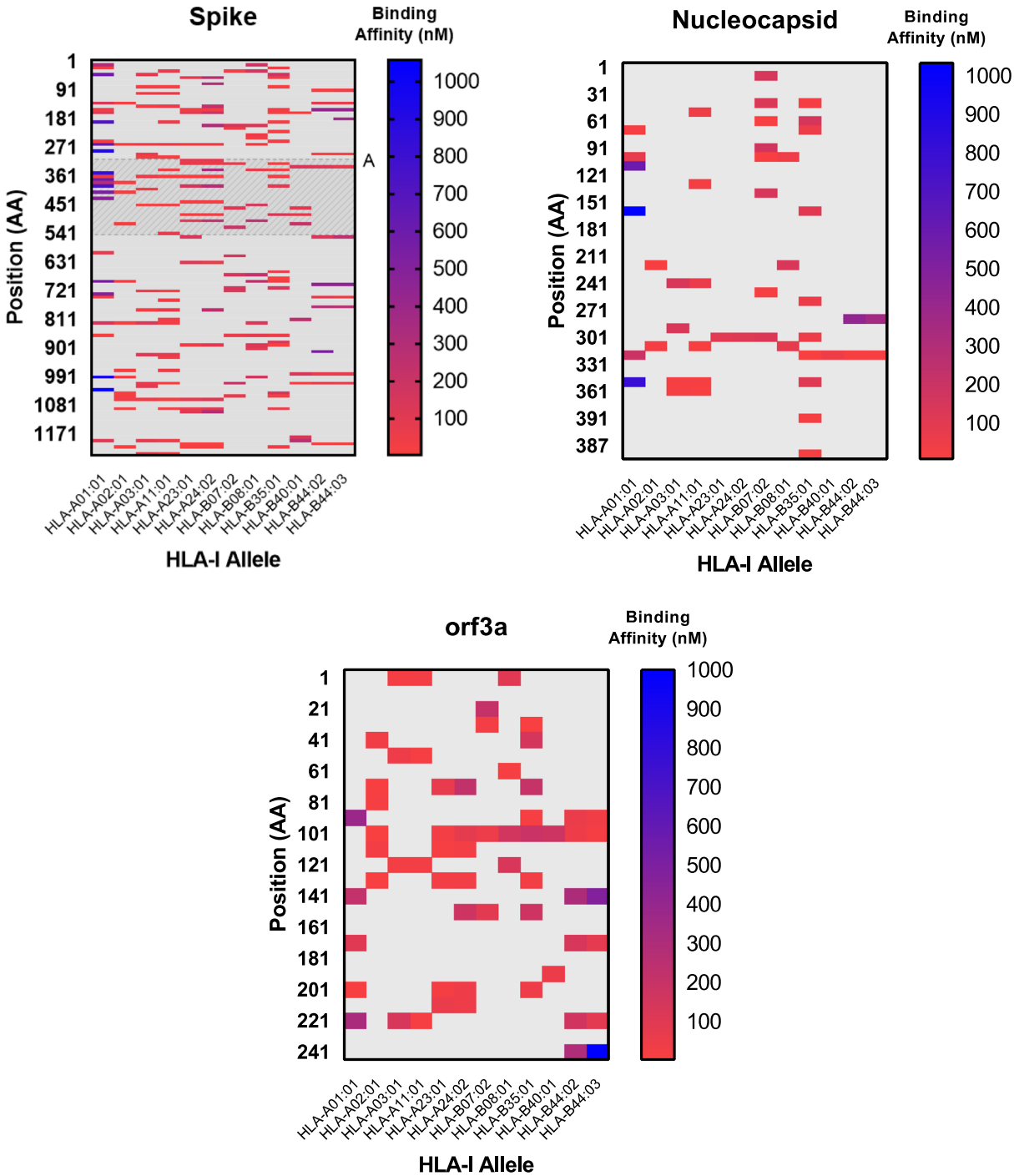


Fig. 3 Mapping of HLA-I A and B epitopes to the spike, nucleocapsid and orf3a protein sequences. Epitopes were inferred using NetMHCpan 4.1. Only strong binding epitopes were selected for analysis, based on the tool's internal classification system. A representative set of HLA-I alleles were used as described in Grifoni et al. [14]. Binding affinity is based on predicted IC50 values (nM) (a lower concentration equates to a stronger affinity). Shaded region (denoted by A) indicates the spike receptor-binding domain (RBD)

alternative allele frequency greater than 1% (A220V, S194L, S197L, M234I and P199L:S). The average epitope meta-score for these variant sites is 30% above the global median prediction score, with the two aforementioned high frequency mutant residues scoring 35% above the global median epitope predictive score. The sequence homology analysis of the nucleocapsid protein revealed a high level of shared identity between SARS-CoV (90%) and MERS-CoV (45%) on a per-residue basis. The nucleocapsid protein analysis revealed two clusters of shared human coronavirus orthologue identity (Fig. 2), one of which was found to cross the aforementioned N173-234 region with identity to HKU1, NL63, MERS and SARS detected. Moreover, these clusters were found to have an increased IEDB epitope mapping frequency, high polymorphism frequency and B cell epitope meta-scores (23% above the global median), indicative of potential B cell immunogenicity. We focused on two nucleocapsid protein specific regions of interest (amino acids 102 to 137 and 167 to 206; Fig. 2). Within the first 35-residue region (amino acids 102 to 137), we have detected NL63, SARS, OC43, 229E, MERS and HKU1 human coronavirus homology. Further, we observed an increase in mapped IEDB epitopes, including mapped linear peptidic B and T cell epitopes from avian gammacoronavirus, murine betacoronavirus, feline and canine alphacoronavirus-1 providing in vitro confirmation that peptides within this region may indeed serve as immunogenic cross-reactive epitopes. The second region (amino acids 167 to 206) contains the R203K and G204R mutations along with a cluster of high frequency variants. We detected homology with HKU1, NL63 and MERS human coronavirus species along with a high frequency of SARS and murine coronavirus mapped IEDB epitopes, with a 34% increase on the median B-cell epitope prediction meta-score.

Previous studies of adaptive cellular effector immune responses to SARS-CoV infection have emphasised the importance of spike peptide presentation in the progression and severity of disease; regions of particular interest include the following: S436–443, S525–532, S366–374, S978,and S1202 [25, 26, 38]. We analysed these regions for their performance as HLA-I ligands in silico and found that all of the regions of interest had a high binding affinity scores associated with that position. Moreover, these peptides were widely represented in the predictions made across the 2915 HLA-A, -B and -C alleles used in this analysis. Taking into account all available HLA-A, B and C alleles, we found the spike peptides had an average allele coverage of 21%, 18% and 34% respectively. We performed an analysis to include 12 alleles with the highest frequency observed across the human population, as reported recently [14]. We found that peptides in the S366-374 and S1202 regions had high representation across the subset of 12 high frequency HLA-I alleles (Fig. 3). These findings imply that the peptides as HLA-I ligands may have a putative role in initiating a protective cellular response in SARS-CoV-2 infections across a significant proportion of the HLA-I population worldwide. We have identified another region of interest that scores highly in the HLA-I peptide binding analysis. The S690-700 region has a high frequency of peptides with a high binding affinity with significant representation across all HLA-I alleles (HLA-A 40%, -B 23%, -C 60%). Furthermore, we have observed no mutations present in this region based on our SARS-CoV-2 variant analysis, implying this peptide appears to remain conserved making it a prime candidate for further study. The spike D614G mutation does not appear to have significantly elevated HLA-I epitope prediction scores (Fig. 3), a finding supported by recent work [39]. The biological importance of spike D614G, particularly its immunological relevance and impact on transmission and disease, are still unclear [40, 41].

Protein 3a (orf3a) has been reported to play a role in host immune modulation by decreasing interferon alpha-receptor expression in SARS-CoV-infected cells and activating the NLRP3 inflammasome [42, 43], a response that may boost inflammation mediated COVID-19 pathology. Orf3a has been a target for SARS-CoV vaccinology studies, with reports of it eliciting potentially protective responses in both protein and DNA forms [27, 44]. These immunogenic properties appear conserved in the SARS-CoV-2 orthologue, with consistently strong antibody responses reported in COVID-19 patients [45]. Looking across the SARS-CoV-2 proteome, of the 50 residues with the highest B cell epitope prediction meta-score, orf3a occupies 16%, despite only constituting 2.5% of the total SARS-CoV-2 protein sequence. Moreover, there are numerous high affinity HLA-II epitopes, which may serve to elicit strong antibody responses. Although protein orf3a shares a high level of identity with its SARS-CoV orthologue, we detected no amino acid sequence homology with OC43, NL63, HKU1 and 229E human coronavirus species or any non-SARS-CoV IEDB epitopes.

Our analysis of the 150,090 SAR-CoV-2 whole genome sequences detected 267 variant sites within orf3a, with non-synonymous mutations occurring 68,473 times. A minority of these variant sites are singletons (8.2%) and five (1.8%) have a frequency higher than 1% (> 1500 isolates), with a non-synonymous mutation density 40% lower than that of the nucleocapsid. The variant sites identified in the orf3a gene have a mean epitope predictive meta-score of 2.3, which is equal to the median global score, indicating that these sites may not form a part of a B cell epitope. Comparing the predictive meta-scores of the nucleocapsid protein variant sites, we

observed an increase of 26% over the global median, indicating that nucleocapsid protein non-synonymous mutations may impact epitope variability more than those found in orf3a. $CD8^+$ effector responses to protein 3a have been characterised in SARS-CoV patients and appear to play a significant role in immunity [26, 46, 47]. Notably, alongside two within the spike protein, a peptide in orf3a (orf3a 36-50) has been found to form a part of the public (conserved) T cell epitope repertoire across SARS-CoV patients [47]. This region scores highly in the HLA-II predictions with numerous HLA-A and HLA-B high affinity peptides covered (HLA-A 19%, -B 41%, -C 48%) and is relatively conserved with few low frequency non-synonymous mutations (maximum mutant allele frequency of 0.00219 (65 times)). We have identified one further region in orf3a (101-121) that scores highly with HLA-I epitope prediction across frequent HLA-I alleles (Fig. 3) and therefore may be of interest to those studying HLA-I ligands. For HLA-I prediction, we observed that orf3a performs significantly better than the nucleocapsid. Despite the nucleocapsid protein sequence being 52% larger than that of orf3a, there are 34% more high affinity HLA-I epitopes across our subset of 12 frequent HLA-I alleles (Fig. 3), which may indicate that orf3a has a more immunodominant role in cellular responses following intracellular processing when compared to the nucleocapsid protein.

Overall, we have developed an immuno-analytical tool that combines in silico prediction data with in vitro epitope mapping, SARS-CoV-2 genome variation and a k-mer-based human coronavirus sequence homology with curated functional annotation data. Furthermore, we have added functionality enabling users to track mutations geographically across time. An additional framework exists to annotate positions with relevant findings from the literature to further guide users' research. The integration and co-visualisation of these data support the rational selection of diagnostics, vaccine targets with reverse immunology, and highlight regions for further immunological studies. We demonstrate the utility of the tool through the analysis of three proteins and their mutant positions, which are of relevance to current SARS-CoV-2 research.

Understanding the magnitude of transmission and patterns of infection will lead to insights for post-isolation strategies. The rapid emergence of the SARS-CoV-2 virus called for an expediated process to deploy serological RDTs for the detection of SARS-CoV-2 IgG/IgM antibody responses. There were reports early in the outbreak of lateral flow SARS-CoV-2 Ig RDTs not reaching sufficiently high levels of sensitivity and specificity [43]. While many assays use the spike protein as its sole antigen for antibody detection, others employ a combination of the spike and nucleocapsid proteins; other assays have been based solely on the nucleocapsid protein [7]. Our analyses suggest that, in its native form, the nucleocapsid protein may prove a sub-optimal target for use in serological diagnostic platforms. It possesses the greatest number of residues across all SARS-CoV-2 genes with high-frequency non-synonymous mutations, the majority of which have a high predictive epitope and IEDB epitope mapping scores when compared to variant positions of other genes. This implies that there may be an inherent variability in dominant antibody responses to different nucleocapsid protein isoforms, which may work to confound testing. We have located three regions of homology with other highly prevalent human coronavirus species, which could serve as non-specific SARS-CoV-2 epitopes if used in serological assays. Moreover, we have emphasised the high level of SARS-CoV identity across the SARS-CoV-2 proteome (except in orf8 and orf10), which may have implications for diagnostic deployment in countries that have had outbreaks involving SARS-CoV.

The spike protein has remained a focus of both vaccine and diagnostic research. Its functional role in viral entry imparts this antigen with immunodominant and neutralising antibody responses [29, 44]. This role is reinforced in our analyses, with several clusters of high epitope meta-scores in functional regions and high IEDB epitope mapping counts. The S1 domain has been the focus of a number of studies looking for specific antigens, not least because of its apparent lack of sequence homology with other human coronavirus species when compared to regions in the S2 domain, as well as its strong functional and immunogenic role in SARS-CoV-2 infection [7, 40, 44, 45]. However, as vaccination programmes begin, most of which will target the spike protein, it will become challenging to differentiate vaccination responses from those elicited by SARS-CoV-2 infection. Therefore, alternative viable targets for serological screening may be needed.

The broad nature of the analyses performed by our tool may assist in the understanding of vaccine targets, during both design and testing phases. The prediction of HLA-I ligands is relevant not only to the study of structural viral targets, but the full range of potentially immunologically relevant endogenous proteins that may be presented following intracellular processing, some of which may have less coverage in the literature. Our broad approach to HLA-I ligand prediction ensures that researchers can assess the applicability of in silico informed vaccine targets across different populations. Further, ensuring that targets are both specific and devoid of polymorphism is essential to ensuring the longevity of vaccine responses and diagnostic capabilities, the analysis of which is achieved easily with our tool. The humoral and cellular immune responses as well as the

affects of human coronavirus protein homology to SARS-CoV-2 proteins have yet to be fully characterised. With the significant levels of amino-acid sequence identity between SARS-CoV-2 and other human coronavirus species detected in our analysis, researchers should be wary of the potentially deleterious effects of both non-specific humoral and cellular responses in enhancing infection; a phenomenon observed in a number of other viral pathology models. While the tracing and monitoring of non-synonymous mutations and their spatio-temporal analysis provides an initial indication of their importance, potentially the impact of evolutionary pressures on loci of interest, further analyses on signals of selection may provide additional insights. Computer intensive genome-wide analyses of positive selection are becoming available (e g. http://covid19.datamonkey.org/) and may be used to complement insights from our immuno-analytical tool.

In summary, using the SARS-CoV-2 immuno-analytics platform, we were able to identify shortcomings in current targets for diagnostics and suggest orf3a as another target for further study. This protein has proven in vitro immunogenicity in COVID-19 patients, and promising functional aspects were supported by our integrated data analysis using the tool. The database underpinning the online tool is updated automatically using data parsing scripts that require minimal human curation. The monitoring of the temporal changes in the frequencies of mutations or their presence in multiple clades in a SARS-CoV-2 phylogenetic tree could provide insights for infection control, including post-vaccine introduction. Importantly, our open-access platform and tool enables the acquisition of all of the aforementioned data associated with the SARS-CoV-2 proteome, assisting further important research on COVID-19 control tools.

Conclusions

The SARS-CoV-2 immuno-analytics platform enables the visualisation of multidimensional data to inform target selection in vaccine, diagnostic and immunological research. By integrating genomic and whole-proteome analyses with in silico epitope predictions, we have highlighted important advantages and shortcomings of two proteins at the foci of COVID-19 research (spike and nucleocapsid), while suggesting another candidate for further study (orf3a). Both spike and nucleocapsid proteins have regions of high identity shared with other endemic human coronavirus species. Moreover, several high frequency mutations found in our dataset lie within putative T and B cell epitopes, something that should be taken into consideration when designing vaccines and diagnostics. Further, our work is likely to become more important as the roll-out of vaccines will introduce new selection pressures that will need to be monitored for escape variations. The immuno-analytics tool can be accessed online (http://genomics.lshtm.ac.uk/immuno), and the source code is available on GitHub (https://github.com/dan-ward-bio/COVID-immunoanalytics) [18].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-020-00822-6.

Additional file 1: Table S1. Extracted data from the Immuno-analytics tool. Mutations listed here have increased significantly over the 20-week period (weeks 20 to 40, year 2020) (see **Fig. S2.** for spike mutations A222V, S477N and L18F). **Fig. S1**. Screenshots from the Immuno-analytics webpage (http://genomics.lshtm.ac.uk/immuno). **Fig. S2.** Screen capture from 'Mutation Tracker' page tracing spike mutations accumulating in Europe, North America and Oceania since the last week of December 2019 (week 52) into 2020 (week 1 onwards). Mutations can be traced across continents by week on the 'Mutation Tracker' page. Mutations shown here are in the Spike (A222V, S477N and L18F).

Abbreviations

COVID-19: Coronavirus disease (2019); SARS-CoV-2: Severe acute respiratory syndrome–coronavirus (2); RT-qPCR: Reverse transcriptase-quantitative polymerase chain reaction; RNA: Ribonucleic acid; RDT: Rapid diagnostic test; ELISA: Enzyme-linked immunosorbent assay; RBD: Receptor-binding domain; HLA-I and HLA-II: Human leukocyte antigen; IEDB: Immune epitope database; SARS: Severe acute respiratory syndrome; MERS: Middle eastern respiratory syndrome; GFF: General feature format; ORF: Open reading frame; CLI: Command line interface; CD8: Cluster of differentiation (8); ACE2: Angiotensin-converting enzyme 2; NLRP3: NLR family pyrin domain containing 3; IC50: 50% inhibitory concentration

Acknowledgements

We gratefully acknowledge the laboratories who submitted the data to the IEDB, NCBI and GISAID public databases on which this research is based. We also thank IEDB, NCBI and GISAID for developing and curating their databases. We gratefully acknowledge the availability of the Medical Research Council UK funded eMedLab (HDR UK) computing resource.

Authors' contributions

DW, SC and TGC conceived and directed the project. MH and JEP provided software and informatic support. DW and JEP performed bioinformatic and statistical analyses under the supervision of TGC. DW, MLH, SC and TGC interpreted results. DW wrote the first draft of the manuscript. All authors commented and edited on various versions of the draft manuscript. DW, TGC and SC compiled the final manuscript. All authors read and approved the final manuscript.

Funding

DW is funded by a Bloomsbury Research PhD studentship. SC is funded by Bloomsbury SET, Medical Research Council UK (MR/M01360X/1, MR/R025576/1 and MR/R020973/1) and BBSRC UK (Grant no. BB/R013063/1) grants. TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1 and MR/R020973/1) and BBSRC UK (Grant no. BB/R013063/1).

Availability of data and materials

The sequencing data analysed during the current study are available from GISAID (https://www.gisaid.org) and NCBI (https://www.ncbi.nlm.nih.gov). Full analysis datasets can be downloaded from http://genomics.lshtm.ac.uk/immuno or https://github.com/dan-ward-bio/COVID-immunoanalytics [18]. The source code for the website can be accessed through https://github.com/dan-ward-bio/COVID-immunoanalytics [18].

Ethics approval and consent to participate

Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Received: 12 May 2020 Accepted: 14 December 2020
Published online: 07 January 2021

References

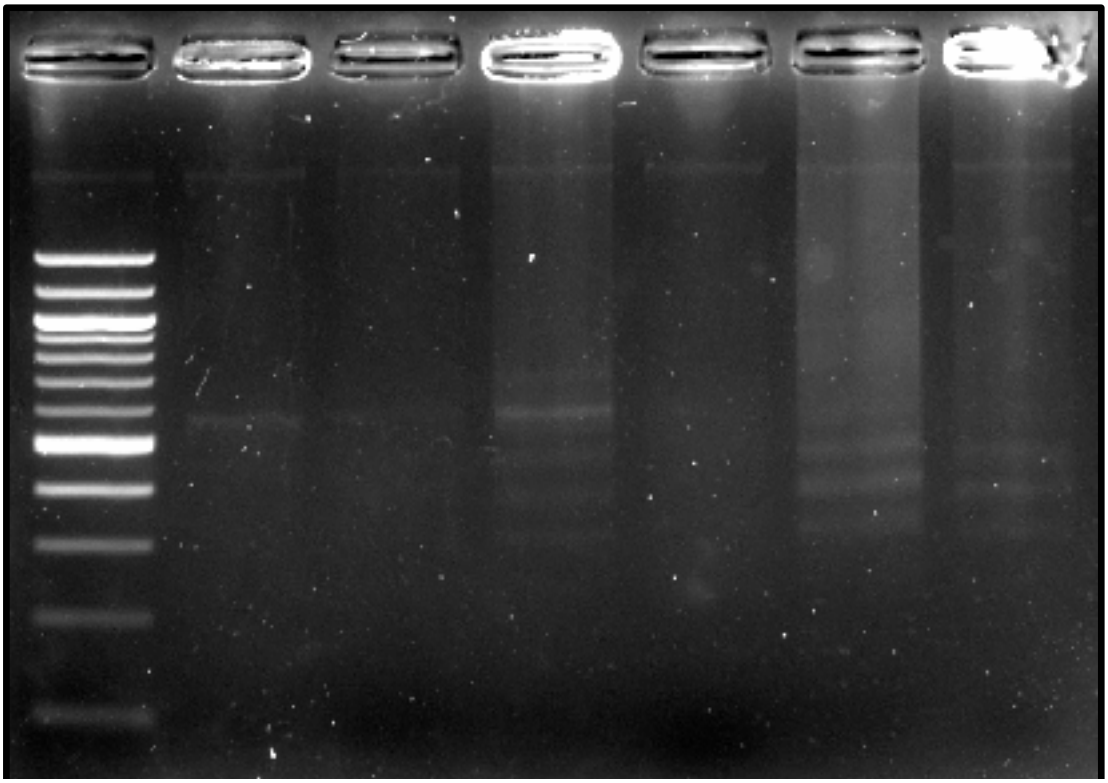
1. IMF. World Economic Outlook Update, June 2020: A Crisis Like No Other, An Uncertain Recovery [Internet]. IMF. 2020. [cited 2020 Nov 10]. Available from: https://www.imf.org/en/Publications/WEO/Issues/2020/06/24/WEOUpdateJune2020.
2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. Elsevier; 2020. p. 533–4. https://doi.org/10.1016/S1473-3099(20)30120-1.
3. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019 : a model-based analysis. Lancet Infect Dis. 2020;3099:1–9.
4. Vogels CBF, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC, et al. Analytical sensitivity and efficiency comparisons of SARS-COV-2 qRT-PCR assays. medRxiv; 2020;2020.03.30.20048108.
5. CDC. Processing of sputum specimens for nucleic acid extraction. 2020.
6. Long Q-X, Liu B-Z, Deng H-J, Wu G-C, Deng K, Chen Y-K, et al. Antibody responses to SARS-CoV-2 in patients with COVID-19. Nat Med. 2020;26:845–8.
7. JHU Centre for Health Security: Global Progress on COVID-19 Serology-Based Testing. http://www.centerforhealthsecurity.org/resources/COVID-19/Serology-based-tests-for-COVID-19.html#sec1. [Accessed 3 Apr 2020].
8. GeurtsvanKessel CH, Okba NMA, Igloi Z, Bogers S, Embregts CWE, Laksono BM, et al. An evaluation of COVID-19 serological assays informs future diagnostics and exposure assessment. Nat Commun. 2020;11:3436. Available from: https://doi.org/10.1038/s41467-020-17317-y.
9. World Health Organisation. Landscape of COVID-19 candidate vaccines [Internet]. https://www.who.int/blueprint/priority-diseases/key-action/novel-coronavirus-landscape-ncov.pdf?ua=1. [Accessed 3 Apr 2020].
10. Thanh Le T, Andreadakis Z, Kumar A, Gómez Román R, Tollefsen S, Saville M, et al. The COVID-19 vaccine development landscape. Nat Rev Drug Discov. https://doi.org/1038/d41573-020-00073-5.
11. Parker EPK, Shrotri M, Kampmann B. Keeping track of the SARS-CoV-2 vaccine pipeline. Nat Rev Immunol. 2020;20:650.
12. Cui Y, Chen X, Luo H, Fan Z, Luo J, He S, et al. BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. Bioinformatics. 2016;32:1740–2.
13. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019;47:D339–43.
14. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe. 2020;27:671–80 e2.
15. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
16. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
17. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. Elsevier Ltd; 2000. p. 276–7. https://doi.org/10.1016/s0168-9525(00)02024-2.
18. COVID Immunoanalytics GitHub Page. https://github.com/dan-ward-bio/COVID-immunoanalytics. Accessed 1 Dec 2020.
19. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 2017;45:W24–9.
20. Davydov YI, Tonevitsky AG. Prediction of linear B-cell epitopes. Mol Biol Springer. 2009;43:150–8.
21. Sher G, Zhi D, Zhang S. DRREP: deep ridge regressed epitope predictor. BMC Genomics. 2017;18:676.
22. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins Struct Funct Bioinforma. 2006;65:40–8.
23. Singh H, Ansari HR, Raghava GPS. Improved method for linear B-cell epitope prediction using antigen's primary sequence. Schönbach C, editor. PLoS One; 2013;8:e62216.
24. Saha S, Raghava GPS. ICARIS 2004, LNCS 3239; 2004.
25. Zhi Y, Kobinger GP, Jordan H, Suchma K, Weiss SR, Shen H, et al. Identification of murine CD8 T cell epitopes in codon-optimized SARS-associated coronavirus spike protein. Virology. 2005;335:34–45.
26. Channappanavar R, Fett C, Zhao J, Meyerholz DK, Perlman S. Virus-specific memory CD8 T cells provide substantial protection from lethal severe acute respiratory syndrome coronavirus infection. J Virol. 2014;88:11034–44.
27. Lu B, Tao L, Wang T, Zheng Z, Li B, Chen Z, et al. Humoral and cellular immune responses induced by 3a DNA vaccines against severe acute respiratory syndrome (SARS) or SARS-like coronavirus in mice. Clin Vaccine Immunol. 2009;16:73–7.
28. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J Immunol. 2017;199:3360–8.
29. Chen B, Khodadoust MS, Olsson N, Wagar LE, Fast E, Liu CL, et al. Predicting HLA class II antigen presentation through integrated deep learning. Nat Biotechnol. 2019;37:1332–43.
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
31. Phelan J, Deelder W, Ward D, Campino S, Hibberd ML, Clark TG. Controlling the SARS-CoV-2 outbreak, insights from large scale whole genome sequences generated across the world. bioRxiv; 2020;2020.04.28.066977.
32. Sui J, Deming M, Rockx B, Liddington RC, Zhu QK, Baric RS, et al. Effects of human anti-spike protein receptor binding domain antibodies on severe acute respiratory syndrome coronavirus neutralization escape and fitness. J Virol. 2014;88:13769–80.
33. Wang SF, Tseng SP, Yen CH, Yang JY, Tsao CH, Shen CW, et al. Antibody-dependent SARS coronavirus infection is mediated by antibodies against spike proteins. Biochem Biophys Res Commun. 2014;451:208–14.
34. Korber B, Fischer W, Gnanakaran SG, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv; 2020;2020.04.29.069054.
35. Huang AT, Garcia-Carreras B, Hitchings MDT, Yang B, Katzelnick LC, Rattigan SM, et al. A systematic review of antibody mediated immunity to coronaviruses: kinetics, correlates of protection, and association with severity. Nat Commun. 2020;11:4704.
36. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell. 2020;182:812–27 e19.
37. Hodcroft EB, Zuber M, Nadeau S, Comas I, González Candelas F, Consortium S-S, et al. Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. medRxiv; 2020;2020.10.25.20219063.
38. Chen H, Hou J, Jiang X, Ma S, Meng M, Wang B, et al. Response of memory CD8 + T cells to severe acute respiratory syndrome (SARS) coronavirus in recovered SARS patients and healthy individuals. J Immunol. 2005;175:591–8.
39. Kiyotani K, Toyoshima Y, Nemoto K, Nakamura Y. Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. J Hum Genet. 2020;65:569–75.
40. Grubaugh ND, Hanage WP, Rasmussen AL. Leading edge making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear; 2020.
41. Volz EM, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole A, et al. Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. medRxiv. 2020;2020.07.31.20166082.
42. Minakshi R, Padhan K, Rani M, Khan N, Ahmad F, Jameel S. The SARS coronavirus 3a protein causes endoplasmic reticulum stress and induces ligand-independent downregulation of the type 1 interferon receptor. PLoS One. 2009;4(12):e8342. https://doi.org/10.1371/journal.pone.0008342.
43. Siu KL, Yuen KS, Castano-Rodriguez C, Ye ZW, Yeung ML, Fung SY, et al. Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. FASEB J. 2019;33:8865–77.

44. Zhong X, Guo Z, Yang H, Peng L, Xie Y, Wong TY, et al. Amino terminus of the SARS coronavirus protein 3a elicits strong, potentially protective humoral responses in infected patients. J Gen Virol. 2006;87:369–74.
45. Wang H, Hou X, Wu X, Liang T, Zhang X, Wang D, et al. SARS-CoV-2 proteome microarray for mapping COVID-19 antibody interactions at amino acid resolution. bioRxiv. 2020;2020(03):26.994756.
46. Oh H-LJ, Chia A, Chang CXL, Leong HN, Ling KL, Grotenbreg GM, et al. Engineering T cells specific for a dominant severe acute respiratory syndrome coronavirus CD8 T cell epitope. J Virol. 2011;85:10464–71.
47. Li CK, Wu H, Yan H, Ma S, Wang L, Zhang M, et al. T cell responses to whole SARS coronavirus in humans. J Immunol. 2008;181:5490–500.

Publisher's Note

Chapter Six

Multi'Mer: A streamlined pipeline for constructing and expressing short, specific tandem repeat peptide antigens for use in diagnostic immunosorbent assays.



Assembling repeated DNA elements using PCR. 2% agarose DNA gel electrophoresis of the first successful MultiMer PCR reaction.ç

MultiMer: A streamlined pipeline for constructing and expressing short, specific tandem repeat peptide antigens for use in diagnostic immunosorbent assays.

In this final Chapter, I describe my work on developing a novel methodology to translate *in-silico* antigen design workflows into high-yield recombinant peptides, for application in immunoassays for disease surveillance. I present a short preliminary analysis of the technique's utility in ZIKV antibody response technology, and outline further experiments required to progress the methodology and work.

6.1 Introduction

Effective antibody detection assays require antigen targets that present a restricted epitope landscape, within which a meaningful and specific cognate antibody-antigen interaction can be measured. However, distinct pathogens can present orthologous and/or conserved protein antigens to the immune system throughout the course of infection, resulting in a non-pathogen-specific or cross-reactive signal that confound serologic assays.

The challenge of identifying unique antigen targets present within a given pathogen's proteome has been approached in numerous innovative ways, some of which have been covered in the previous chapters. Methodologies of testing candidate peptides inferred through *in-silico* or *in-vitro* screening have been applied in the field of *Flavivirology* among others, and usually involves the use of truncated viral antigens[1–5]. These peptides vary in size from 14 residues to entire protein subunits. Most consist of short contiguous sections of viral proteins that have been identified as being unique on the sequence level, have been indicated by epitope

prediction, or found to be favourably represented in epitope mapping studies, using techniques such a peptide microarrays or phage display.

As a means of testing antigen targets, I sought to develop a methodology that would enable the expression of these peptides in the widely available and scalable *E. coli* expression system. One of the primary reasons for this was economy. While peptide synthesis would be a viable means of screening antigen panels for use in diagnostic antibody detection assays, the cost of synthesising such a library would be > £300 per peptide, which proved prohibitive for the scope of this project. Moreover, the context in which these assays would be deployed, medium- to low-income *Flavivirus* endemic settings, makes the inclusion of such costly reagents prohibitively expensive. The cost of the assays featured in **Chapter 3** was > £600 per 96-well plate. And while the NAC DAB and EUROIMMUN assays featured there provided an acceptable level of specificity for the scope of the analysis, I sought an assay format which guaranteed specificity, suitable for contexts with multiple endemic *Flavivirus* species, with an economy suitable for broad-scale implementation.

The expression of peptides in *E. coli* systems can be challenging. While the option of employing vectors with fusion-partners was available, given the sub-optimal results shown in the previous sections which employed GST fusions, I sought an alternative approach. A challenge associated with this requisite, however, is a common limitation found when expressing small peptides in *E. coli.* The most frequent length of peptide in our panel is 15 AA residues long, the coding-sequence for which is 45 nt in length. Both the cloning and expression a panel of 45 nt fragments is invariably awkward. The challenges associated with visualising, purifying and storing such short fragments of short peptides is compounded by the aversion of bacterial expression systems to the heterologous expression of small peptides, which are prone

to proteolytic degradation [6,7]. The knowledge of this, combined with my own unsuccessful preliminary efforts at doing so, suggest that such a synthesis strategy would not be a practical line of enquiry.

In the study of anti-microbial peptides (AMPs), researchers are faced with a similar set of challenges. The lack of scalability in peptide synthesis techniques in the context of AMP research makes *E. coli* systems an attractive option for mass-production and research, but as discussed above, the challenges in its implementation are very much akin to my own. Additionally, the toxic nature of the small bioactive AMPs makes them a difficult target to express in the very microbial systems they are designed to attack. Numerous groups have converged on a methodology that appears to alleviate these issues, which involves the multimerisation of peptides in to a single contiguous tandem-repeated construct [8–11]. Through expressing multimers, the risk of proteasomal degradation is diminished, and the detection and purification of larger proteins is made significantly easier. However, these approaches rely either on 'serial cloning', which is repeated processes of insertion and ligation, or on commercially produced synthetic tandem-repeat constructs, which are time challenging to produce and costly to acquire. While enzymatic multi-fragment assembly technologies are available, such as Gibson or Golden Gate assembly, I have found in my own investigations, these methodologies to be ineffective and inconsistent.

In the following sections, I outline my development and validation of the novel methodology, called *MultiMer*. The primary objective in developing this technique was to produce tandem-repeat peptide constructs, which in this case, would be short candidate ZIKV peptide antigens, for use in antibody detection immunoassays. The MultiMer technique is capable of generating the desired constructs in a single PCR reaction that can be tuned, producing approximate copy

numbers at a desired length. These features may prove to make it a cost-effective and powerful approach by which it is possible to test the performance of a library of ZIKV peptide antigens for use in ZIKV immunoassays, and then scale up its production for wider use in immunosurveillance applications. The technique covers the entire process of antigen validation, from *in-silico* primer analysis design, cloning into a custom destination vector, expression, purification, and validation **(Figure 1)**.

With this methodology, I propose the use of short peptide antigen tandem spaced repeats expressed in *E. coli* using a medium to high throughput construction and cloning protocol, streamlining the economical production of specific immunoassay antigen targets. Generation of the MultiMer antigens requires a single set of primers per peptide and only readily available protein expression reagents and equipment. The process is guided by bespoke, openly available online tools, and yields a high concentration of pure peptide antigen, facilitating the development and sustained usage of in-house immunoassays for application in longitudinal surveillance study contexts.

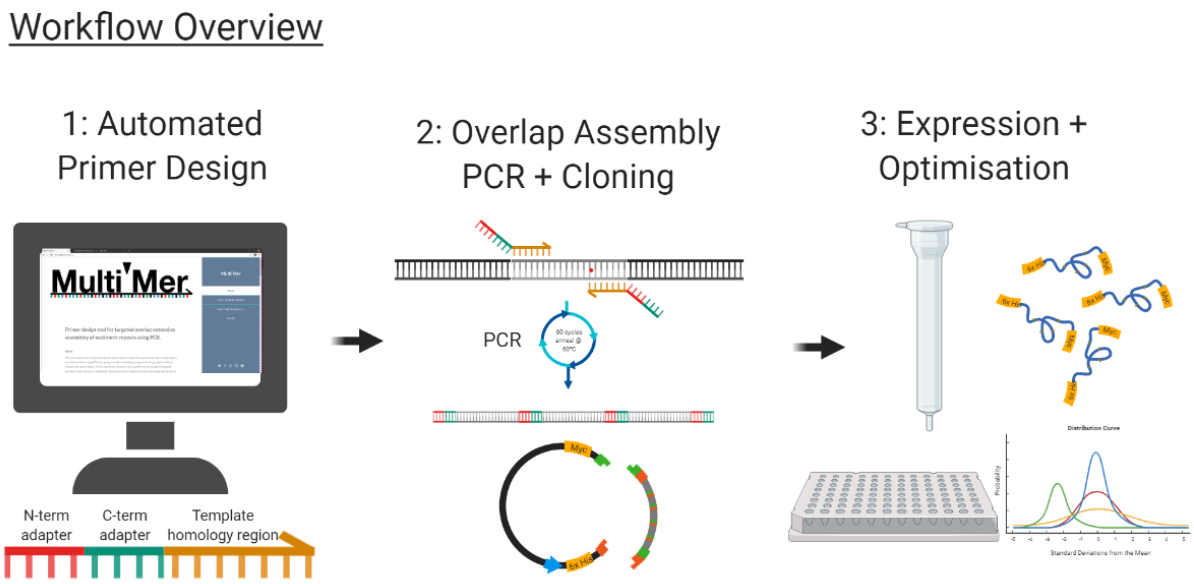


Figure 1. A graphical abstract of the current iteration of the MultiMer methodology. **1.** *in-silico* primer design using the web-based automated toolkit. **2.** PCR overlap assembly and cloning of MultiMer amplicons in to bespoke destination expression vector. **3.** Optimised expression**,** purification, and analysis of peptide antigens.

At its core, I utilise a novel PCR overlap assembly technique to generate tandem-repeats. A single set of primers is required to facilitate this process, which are designed by a bespoke automated *in-silico* pipeline. In this PCR reaction, complementary 5' and 3' adapters, added to a target peptide coding sequence through overlap extension PCR, anneal, create a daisy-chain effect in subsequent PCR cycles, which is capable of generating > 20 X copies of a 48-nucleotide fragment in a single-tube reaction. The adapters are complementary to the cloning site of a modified pER28-a destination expression vector, which enables the efficient expression of the multimeric construct, in a straight-forward way. In downstream steps, I describe an economic methodology whereby these peptides are efficiently purified and applied in immunoassay techniques.

6.2 Development of the MultiMer protocol

6.2.1 Designing multi-purpose adapters for the expression of tandem repeats

Unlike the AMP approach, where peptides were proteolytically cleaved into monomers prior to downstream applications, using un-cleaved tandem repeats as antigens in our application required, firstly, the resolution of an intrinsic defect. Synthetically joining antigens together in a single contiguous peptide may give rise to the formation of new *off-target* mimotopes, at the juncture between the N and C termini of the conjoined peptides. To address this issue, I added linker peptides of ten alternately repeated glycine and alanine residues which would serve the primary purpose of sterically distancing the two repeated antigen peptides from one another. In addition, the spacer would add flexibility to the chain due to the small size of G and A R groups, reducing the formation of secondary structures which might lead do a degree of steric hinderence, a technique which is used commonly in fusion-protein methodologies [12]. I chose

glycine and alanine, also, because of their relatively low impact on overall protein charge, a key property governing epitope specificity. These linker peptides were appended to each end of the coding sequence of the antigen peptide, in-frame, through overlap extension PCR.

An additional function to the linkers was their role as nucleotide adapters for the PCR overlap assembly process. This adapter sequence remains conserved across each target peptide construct, facilitating multimerisation and the cloning of the insert, in a uniform, antigen peptide sequence independent manner. **Figure 2** is an example schematic of one of the epitope MultiMer constructs.

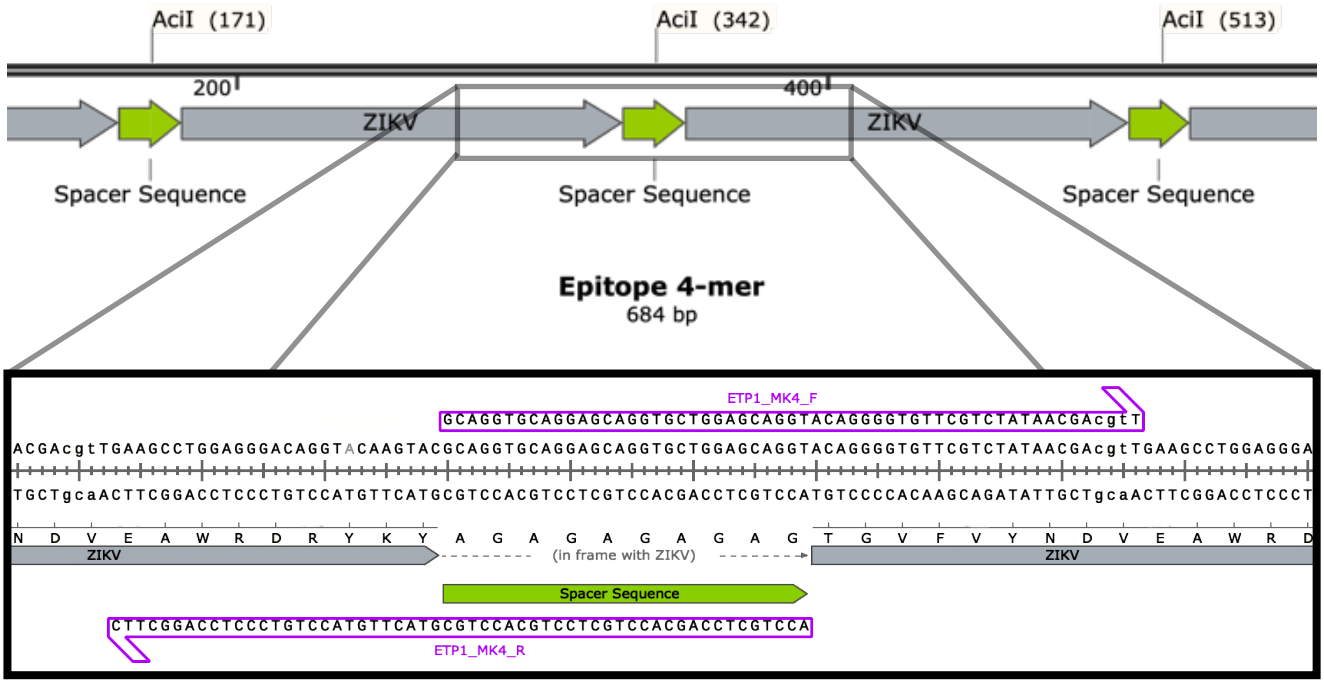


Figure 2. (Top) A map of the coding region of part of the ETP1 (ZIKV1) MultiMer. The antigen peptide sequence (ZIKV) is coloured grey and the conjoined adapter peptide sequences are coloured green. The **AciI** restriction endonuclease site is also indicated. **(Bottom)** The coding sequence of the peptide, which has been sequence verified by sanger sequencing, shown with the MultiMer adapter-primers annealed.

The schematic of the overlap assembly PCR reaction used to generate the MultiMer constructs is shown (**Figure 3).** PCR amplicon repeats are formed following an initial phase of PCR cycles, resulting in the synthesis of monomeric amplicons, with the addition of extra adapter

residues added on the 5’ and 3’ ends of each amplicon, by way of the overlap extension of adapter-primer pairs. In subsequent cycles, the primer concentration decreases, as the adapter-primers are incorporated into amplicons. As this occurs, monomeric strands with complementary 3’ and 5’ adapters begin to prime each other, with the second, complementary strand, synthesised by Q5 polymerase, which complete the duplex MultiMer DNA amplicons. Dimers prime dimers and trimers primer trimers until larger contigs are formed in successive PCR cycles.

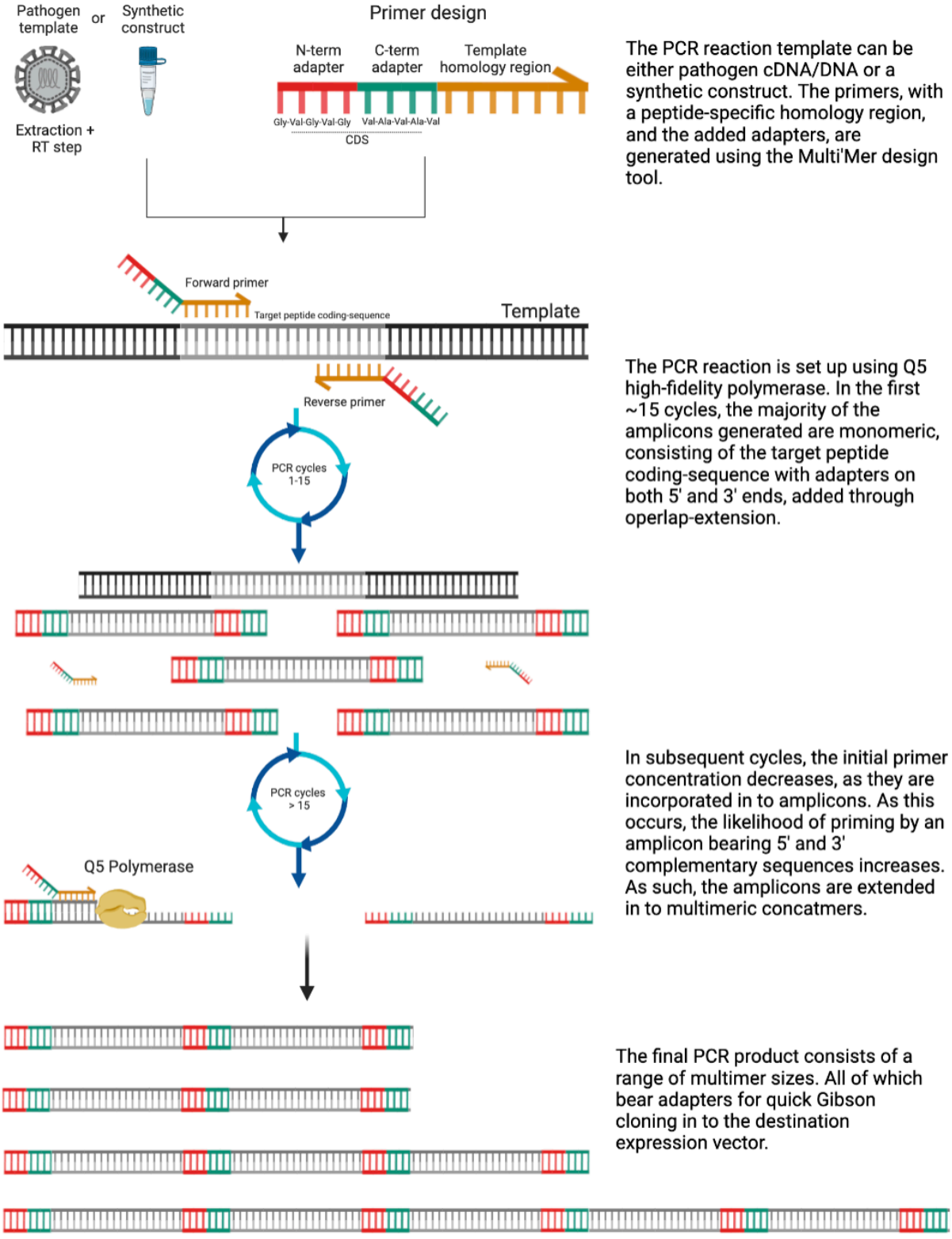


Figure 3. A schematic of our initial understanding of how the PCR overlap assembly reaction functions.

An example of the resulting PCR product is shown **(Figure 4),** with the gel-electrophoresis of two discrete MultiMer amplicons, ETP1 and ETP2. In the far lanes, I analyse the digestion of the AciI restriction endonuclease site, present in each adapter sequence to show the cleavage of multimeric fragments, back in to their monomeric forms.

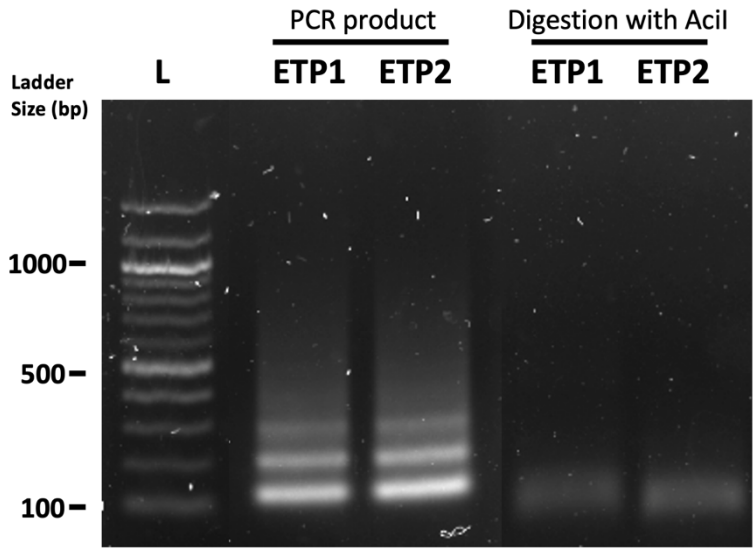


Figure 4. Gel electrophoresis of crude overlap assembly PCR reaction product (lanes 2 and 3), and the subsequent disassembly of MultiMer fragments by AciI digestion. The PCR product was generated as per the protocol outlined in the methods section and run on a 1.5% agarose gel. The amplicon was cleaned using column purification prior to digestion, which was completed in 30 minutes, as per manufacturer's instruction.

6.2.2 *In-silico* modelling of the overlap assembly PCR reaction

A model was constructed to further an understanding of the PCR reaction underpinning the overlap assembly. The model included incorporated the number of PCR cycles, units of Q5 DNA polymerase, the moles of primer and the initial concentration of template, with the output being the size and number of (MultiMer) amplicons. The polymerase unit count, and the template concentration were assumed as being in excess throughout the reaction. In the first implementation of the model, I sought to understand the relationship between cycle number and amplicon priming. **Figure 5** illustrates the first application of this model, with a typical

overlap assembly reaction scenario. As PCR cycles increase as the reaction progresses, the adapter-primers in the initial PCR mix are consumed as they are incorporated in to amplicons, and in subsequent cycles, template priming occurs exclusively through MultiMer-adapter sequence annealing.

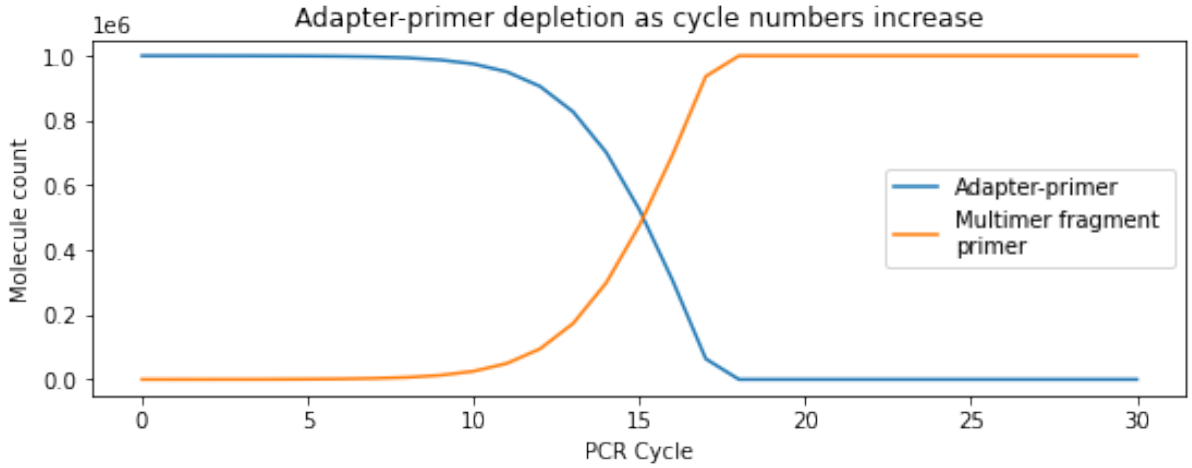


Figure 5. A model illustrating the overlap assembly PCR reaction, visualised as the count of adapter-primers (oligonucleotides added to the PCR mix before initiating) and MultiMer fragment primers (cross-priming amplicons).

A second investigation using the model was undertaken to understand the rate of amplicon size growth as PCR cycles increase (**Figure 6)**. The model incorporated previously established understanding of primer depletion leading to amplicon co-priming. Based on this assumption, I theorise that as the cycle number increases past 15 cycles, the amplicon concentration gradually supersedes that of the adapter-primers. The probability of a 1-mer amplicon priming another 1-mer amplicon increases, over that of an adapter-primer priming the template, which results in the synthesis and accumulation of 2-mer amplicons. As the concentration of 1-mers decreases (as they themselves are incorporated in to 2-mers), the probability of a 3-mer (or a 4-mer) forming increases, and *so on.* This continues indefinitely, until the polymerase efficiency drops due to successive melting steps. While this model is rudimentary and not

exhaustive, it illustrates well the current understanding of how the overlap assembly process functions.

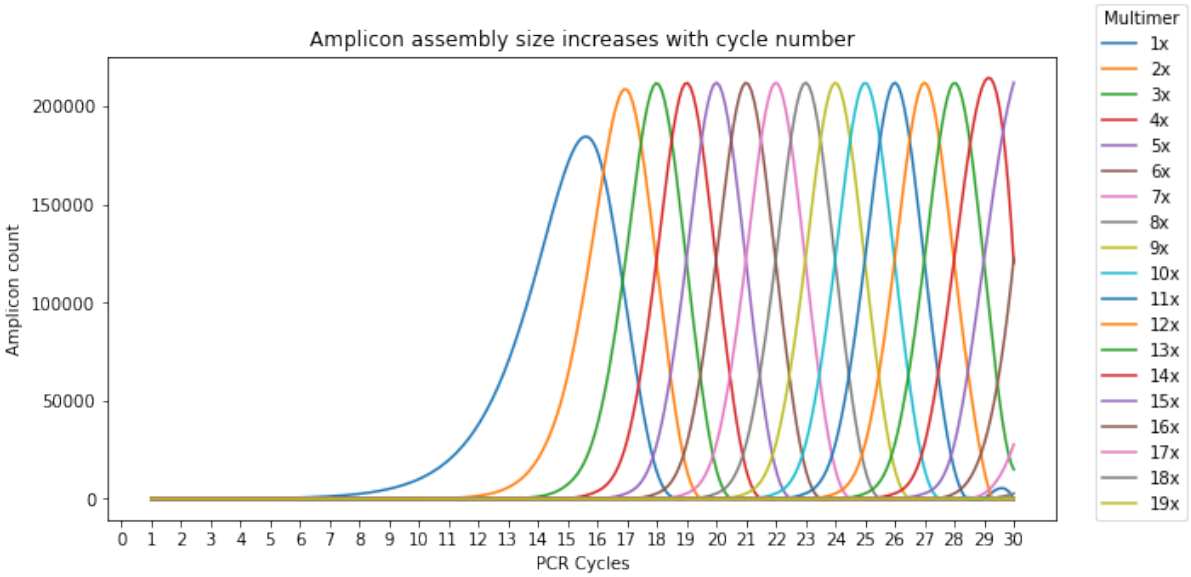


Figure 6. The frequency of increasing size MultiMer amplicons increases with cycle number, as larger fragments prime and assemble, depleting the previous size while incorporating in the next.

6.2.3 In-vitro modelling of the overlap assembly PCR reaction

The findings in the previous section demonstrated that there were two key variables that dictate the frequencies of fragment-sizes produced. The initial concentration of adapter-primer in the reaction dictates the speed at which amplicon synthesis switches from priming with the adapter-primers to priming with amplicons, which in-turn drives forward the assembly of multimers. The second factor was the number of PCR cycles. As the PCR cycle count increases, the depletion of successive *n-mers* occurs, again, promoting increasingly larger MultiMer formation. I theorised that these two variables could be balanced to produce amplicons within a controlled size-parameter.

To validate these assumptions, a series of experiments were performed (see **Figure 7 (a,b)** and **Figure 8)**. I tested the effect that variation in initial adapter-primer concentration, cycle number, extension time and template concentration have on the length of synthesised MultiMer

amplicons (**Figure 7**). In this experiment, I used a primer concentration of 10 μM (the standard concentration for use with Q5 polymerase), 5 μM and 3 μM. The reaction was sampled at cycle increments of 10, from 30 to 60. I tested differences in the initial template concentration and the effect of varying 72ºC (standard Q5 polymerase extension temperature) extension times between 1 and 5 seconds. The findings appeared to validate the fundamental conclusions of our *in-silico* model. While a slight increase in MultiMer size can be seen when changing the template concentration, the bulk of the effect is observed in relation to the other primer concentration and cycle number variables. The increase in extension time appears to be detrimental to the integrity of both reactions, with the 48 nt reaction (ZIKV21) showing a marked decrease in the clarity of the visualised bands. The *smeared* appearance (**Figure 7)** for the shorter length repeats in the 5 second elongation condition, indicated the possible incomplete/truncated extension of fragments. With a longer extension time, this would seem counter intuitive. However, given the length and melting temperature of the large MultiMer fragments in later stages of the reaction, I theorise that the longer elongation time results in an increased probability of additional fragments annealing to the single-stranded template during synthesis, which elongates the fragment without complete synthesis, creating incomplete amplicons and a *smeared* appearance. In subsequent experiments (data not shown), I tested different lengths of adapter sequence, their melting temperature and PCR primer annealing times, which yielded no insights of particular interest.

Overall, an increase in elongation time above 1 second appears to be detrimental to the reaction, and that the key variables effecting MultiMer length appear to be primer concentration and PCR cycle count.

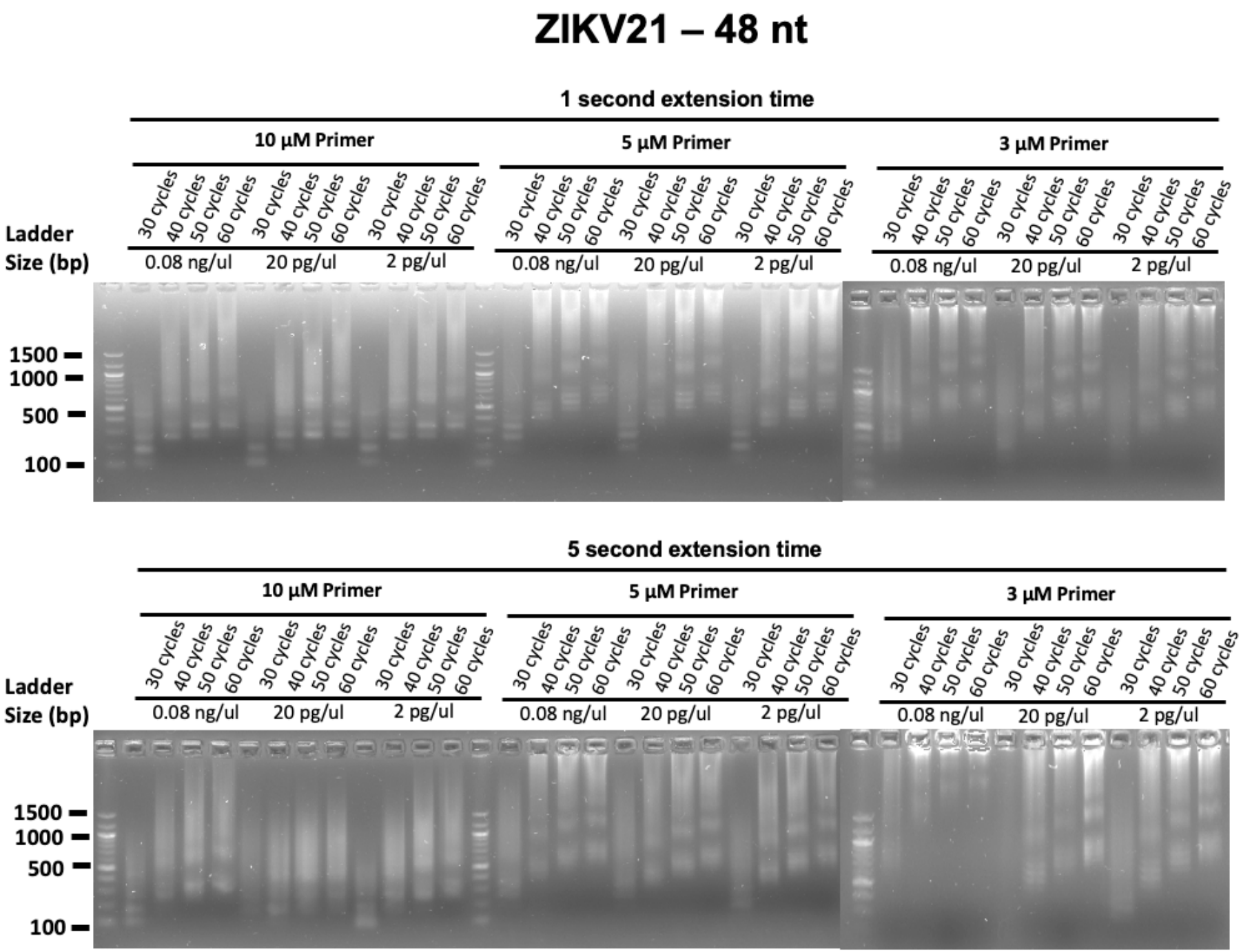


Figure 7a. Gel electrophoresis of MultiMer reactions on a 2% agarose gel. 2 µL of each DNA template was added to the PCR reactions in the defined concentrations. The reactions were performed using the same thermal cycler, with the steepest ramp times available. Each gel is shown with a 100 bp ladder.

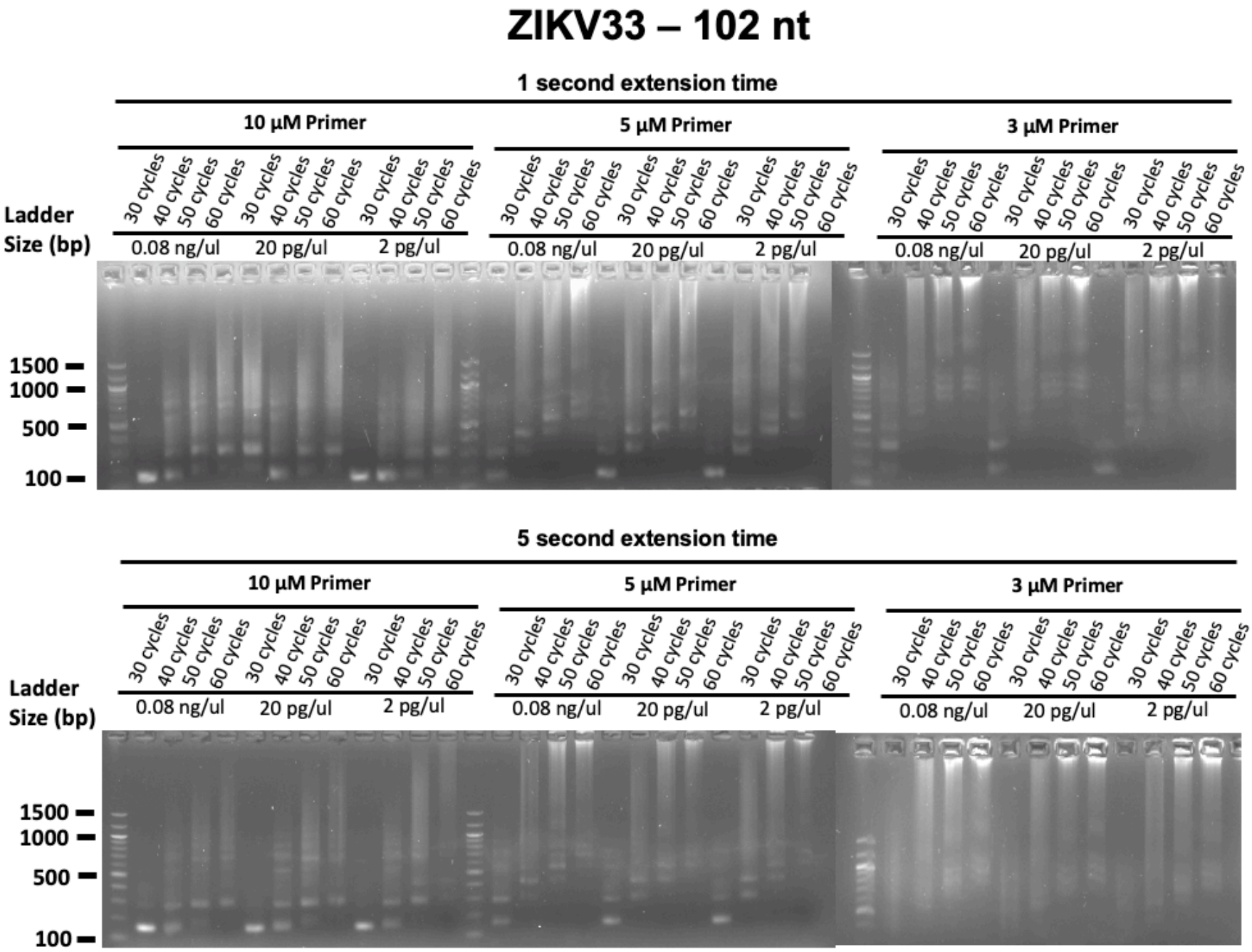


Figure 7b. As above. In this experiment, a 102 nt peptide construct was multimerised almost double the size of that shown in **7a.**

Building on our previous assumptions that primer depletion drives MultiMer formation, I tested the effect of reducing primer concentration on the rate of increase in MultiMer size (**Figure 8**). In this PCR reaction, PCR conditions remain the same for each reaction is limited to 30 cycles. I decrease the concentration of adapter-primer in the initial PCR mix, ranging from 4.5 µM to 1.92 µM.

Illustrated clearly here, in an example (ZIKV21 – 48 nt) MultiMer reaction, is the effect that adapter-primer concentration alone has on the assembly of multimers. The greatest concentration of adapter-primer results in a minimum assembled fragment size of ~200 bp (3-mer (accounting for adapter sequences)), while the lowest primer concentration produces a minimum fragment size of ~500 bp. The minimum length resolvable band in each condition increases as the primer concentration decreases.

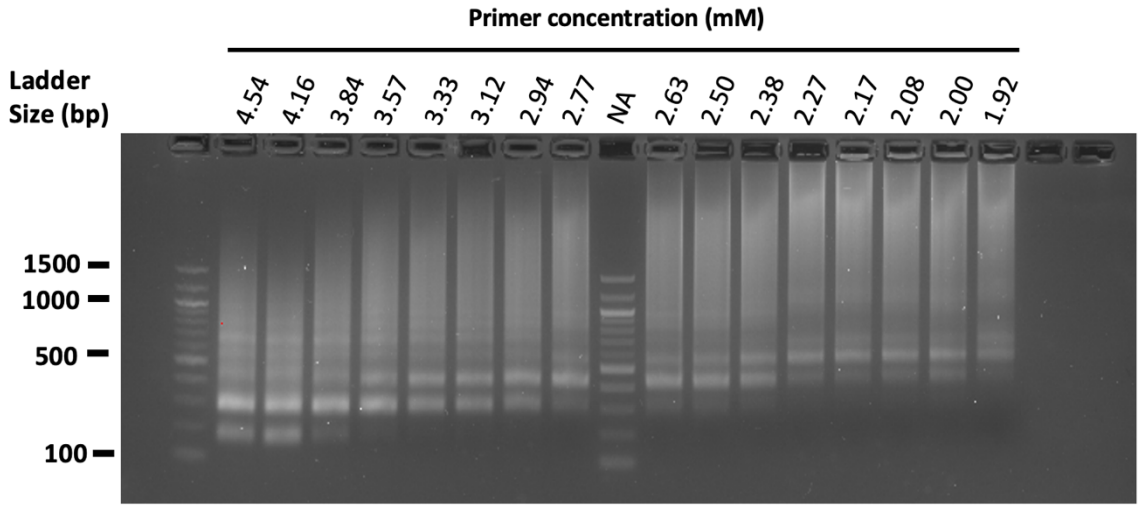


Figure 8. Decreasing primer concentration in a 48 nt MultiMer reaction of 30 cycles, increases in a consistent and repeatable manner, the minimum size MultiMer amplicon. 10 µL of crude PCR reaction for each condition was loaded on to a 2% agarose gel and run at 100 v for 25 minutes. A 100 bp ladder was used.

Overall, it may be possible to control the size-range of the amplicons generated in the overlap assembly PCR reaction. This would enable the consistent and repeatable assembly of tandem-repeat constructs of a pre-defined size in an economic and simple methodology.

6.2.4 Designing a bespoke pET vector for MultiMer expression

For the expression of MultiMer peptides, I opted for a BL21-DE3 *E. coli* expression system, which is highly scalable, accessible in most synthetic biology laboratories, and is known to produce high yields with relative ease, when compared to other expression/synthesis systems. To express the MultiMer amplicons in this system, I chose the ubiquitous pET-28a vector, compatible with the DE3 lysogen, as a template for our recipient vector. The template vector was modified in two successive steps, using primer overlap extension PCR, to better equip the vector to receive MultiMer amplicon inserts. A map of the vector is shown (**Figure 9),** with the cloning site highlighted**.** The first modification involved the removal of the 6X poly-histidine tag on the C-terminus of the multiple cloning site (MCS) protein coding region, and the simultaneous addition of a Myc epitope tag. This change was performed in a single overlap extension PCR reaction using the pET28a-Myc_mod_REV and pET28a-Myc_Nterm_adapter primers (see **Figure 9**). The Myc tag was added to enable the visualisation of peptides in subsequent SDS-PAGE Western blotting steps. The key function of this tag, besides general epitope-tagging/visulisation, lied in its positioning at the C-terminus of the protein coding sequence. This position was chosen to enable the verification of complete translation of the entire cloned amplicon coding sequence. If any frame-shifting mutations, or truncation of the peptide during cloning or translation occur, the amino-acid sequence of the Myc tag would not be translated, which would be clearly indicated in subsequent immunoblotting analyses. This would enable the efficient screening of *E. coli* colonies for positive expression of the heterologous peptide sequence. The second overlap-extension PCR reaction was performed

using the pET-28a-Myc-Cterm_adapter and pET28a-Myc_Nterm_adapter primers, which add the adapter complementary sequences to each end of the plasmid cloning site **(Figure 9 (green) – "N Terminus adapter" and "C Terminus adapter")**. These sequences are complementary to the conserved 5' and 3' ends of the adapter sequences, present on each amplified MultiMer construct. Through the addition of these sequences to the plasmid backbone cloning site, the efficient direction-dependant 1-

step Gibson cloning of MultiMer PCR products in to the MCS is made possible. The final PCR modification generated a linearised 5372 bp amplicon, which would serve as the

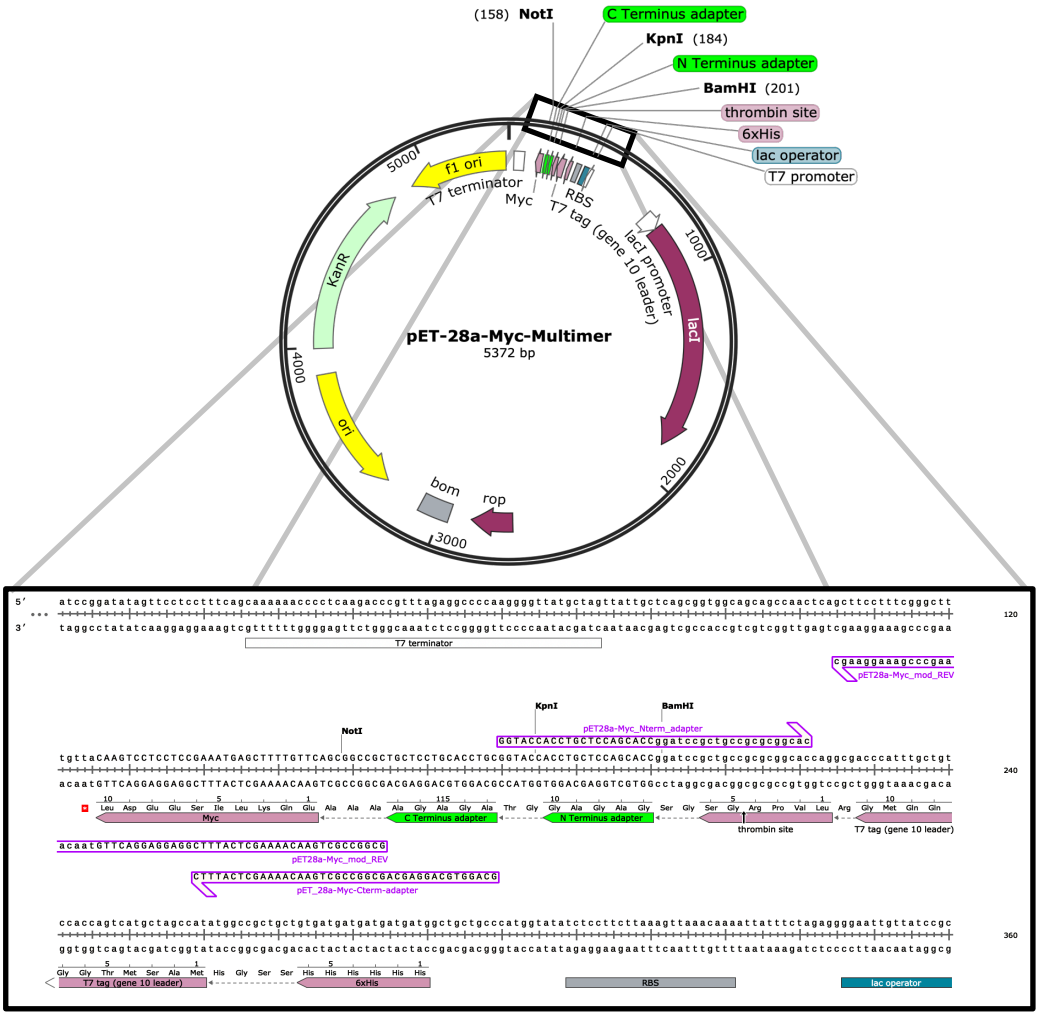


Figure 9. Plasmid map **(upper)** of the pET28a-Myc-Multimer vector backbone. Flanking NotI and BamHI restriction sites can be used for the excision of inserts, for the transferral of the adapter sequences to another backbone. **(lower)** The multiple cloning site (MCS) of the backbone vector. The plasmid is linearised using the primers shown in the diagram, producing amplicons for insert cloning.

recipient plasmid DNA, compatible with the Gibson cloning reaction without PCR clean-up. Importantly, the Q5 high fidelity polymerase was used for both of these modifications, which has both a high processivity and low-error rate. The first PCR was limited to 15 cycles, and the second to 25. While this decreases the total yield of amplicon, it reduces the probability of mutations occurring during the amplification of long, mutation sensitive constructs. I capillary sequenced the cloning site of the plasmid to confirm the correct addition of the aforementioned features.

6.2.4 Gibson cloning of *MultiMer* amplicons into the *pET-28a-Myc-Multimer* plasmid

In a single reaction without any clean-up steps, the PCR product of the MultiMer reaction can be cloned into the recipient *pET-28a-Myc-Multimer* vector. The 15 bp adapter sequences, present on both vector and amplicon insert provide an adequate annealing sequence for specific cloning of fragments in the correct reading frame and orientation. For the propagation of the cloned vector, I used the NEB Stable *E. coli* strain, which is a recA1 and endA1 deficient line, a genotype which reduces the chances of recombination/excision of highly repetitive sequences. Despite this countermeasure, when scaling up the cloning operation to insert several peptide constructs, I noticed a high level of insert recombination in some cloning conditions, illustrated in a diagnostic PCR of cloning sites, amplified using MCS flanking T7 promoter and T7 terminator primers (T7 sites; **Figure 9**), as shown (**Figure 10a).** To understand whether the band pattern is indeed the amplification of discrete recombined-species or a PCR artifact, I selected a repeated the cloning experiment, isolated a subset of clones exhibiting recombination in the colony selection process and generated plasmid preparations. The insert was excised using the flanking BamHI and NotI restriction sites (**Figure 9**) and the product analysed using DNA gel electrophoresis (**Figure 10b/c)**. Based on this analysis, I found that the recombination

occurred seemingly at random in a peptide sequence independent manner. While the recombination is caused by the tandem-repeated sequences, the range or intensity of the bands that are present in T7 PCR screen of colonies appear to bear no indication of the final length of the insert in the final plasmid preparation. Reassuringly, I found that colonies with no apparent recombination indicated during T7 PCR, or colonies with low-level recombination with a *dominant* band at a desired length, remained stable, even when propagated at 37°C. From this, I theorise that the risk of recombination is present only in the initial stages of transformant plasmid replication, most likely due to nicks present in the annealed insert or other structural damage inflicted during assembly. While aberrant recombination in cloned fragments remains one of the central flaws in the current MultiMer workflow, I found that performing the

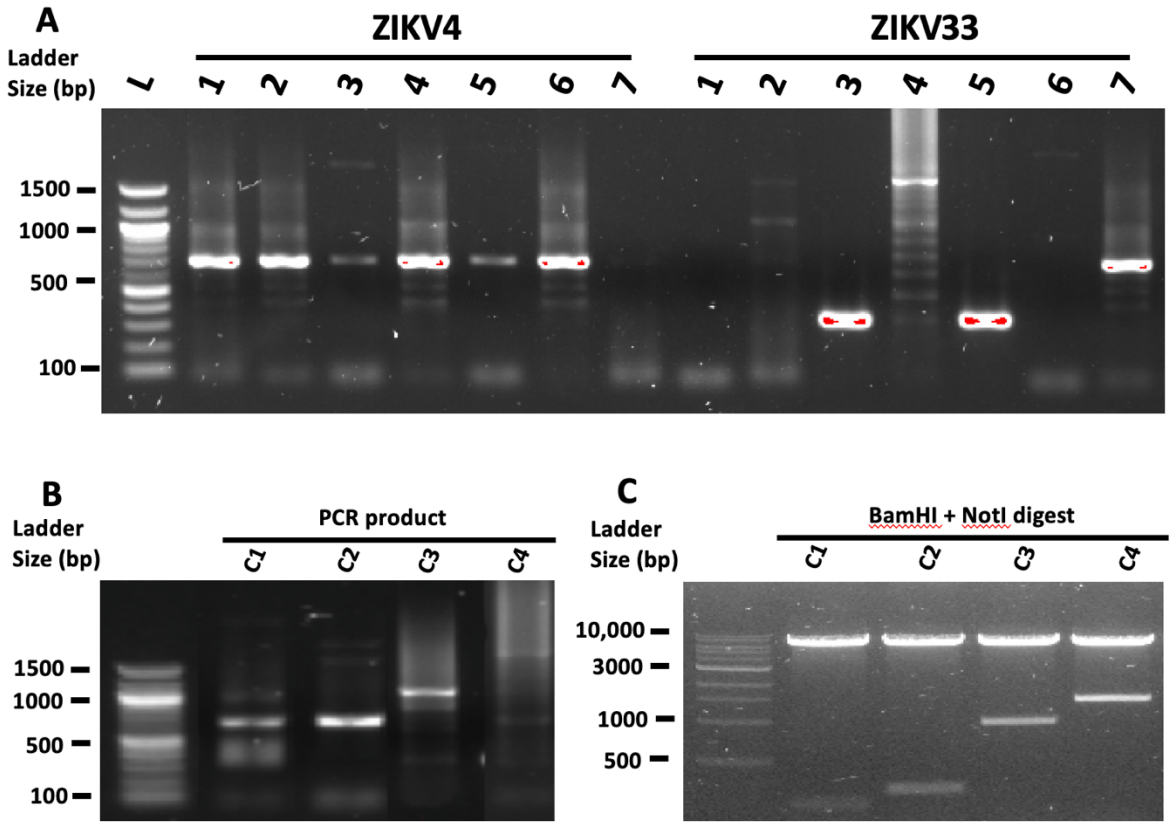


Figure 10. (**A**) T7 diagnostic PCR (2%) gel electrophoresis of single colonies, assayed after cloning and growth at 37°C for 24 hrs. Recombination events in the MCS are indicated by the presence of >1 band, which represent polyclonality of plasmids in the screened colony. Bands at ~300 bp represent an empty MCS. **(B)** T7 diagnostic PCR, as above. **(C)** Diagnostic restriction digest using BamHI and NotI sites flanking MCS. 1 µg of column-cleaned amplicons were digested for 30 minutes as per manufacturer's instruction.

transformation, outgrowth and agar plate colony growth stages at room temperature resulted in an efficiency of 12.5-25%, where approximately 1-2 in 8 clones were stable and retained an insert of adequate size.

6.2.5 Protein expression screening and purification

The expression of the cloned MultiMer constructs was optimised in such a way that up to 24 constructs could be expressed and screened in a two-day protocol. I opted to use the autoinduction technique [13], instead of the typical laboursome IPTG induction methodology. In short, this methodology operates on the principle that T7 repression, controlled by the *lac* operon will *automatically* induce expression of a protein transcript as the glucose in the ZYP-5052 media is consumed, and the ß-galactosidase product (allolactose) concentration increases. This methodology results in high-density cultures, which require no monitoring or addition of reagents throughout the expression process. With this methodology, the high-throughput expression of numerous constructs can be achieved, with significantly reduced *hands-on* time per-peptide.

Purifying the MultiMer peptides from the *E. coli* lysate has been optimised for the initial screening of multiple constructs, and can be scaled up to produce > 10 mg of peptide per batch. It was found that sonication was the most effective means of lysing > 10 candidates in the screening stage, while for increased batch sizes, a French press was used for lysates of a > 20 mL volume. For the purification of poly-histidine tagged peptides, I opted for a high-throughput microcentrifuge purification technique. This technique allowed the small-scale purification of 24 peptides in a single day, ready for ELISA screening. For greater throughput, I used Ni-resin beads, in a modified centrifugation technique, which produced > 5 mL of > 1 mg/mL protein eluate. The screening of peptides can be performed using either a dot-blot

technique or an SDS-PAGE Western, however, it was found that the dot-blot is more economic to screen several clones of each transformed cultured line.

6.2.6 Development of the online MultiMer primer design tool

To streamline the design of MultiMer adapter-primers for the peptide dataset, I generated an automated pipeline, with a user-friendly web interface. The workflow makes the design of adapter-primers simple, with only two user inputs required. The first is a template FASTA formatted file. This should be the exact sequence of the template material (DNA or cDNA) to be added to the PCR reaction, possessing the peptide-coding sequence to be multimerised. In the second input, the user adds a list of FASTA formatted amino-acid peptide sequences; the peptide targets to be multimerised. A diagram outlining the backend pipeline is shown (**Figure 11).** In brief, the script uses the reverse-translating ability of tblastn [14], which searches the input template DNA sequence for the exact, or an approximation of the target peptide coding amino acid sequence. The sequence is then extracted, where another module searches for compatible forward and reverse primer pairs. With integrated quality control (QC) to highlight mismatches in the peptide sequence, the script then outputs a CSV formatted table with all of the information required to order adapter-primers for MultiMer reactions (**Table 1)**. The frontend uses a flask web-container with a HTML5/ JavaScript interface, with included diagrams, and schematics of the backbone vector explaining the MultiMer process (**see screenshot** – **Figure S1)**. This pipeline was used to generate a library of oligonucleotides for the generation of our own example MultiMer peptides.

The final output consists of a table (**Table 1),** detailing the backend generation of adapter-primers. This table highlight peptides with a mismatch in the uploaded template sequence, or peptides that may be truncated. It also outputs the target genomic DNA sequence and the melting temperature of the primer pairs.

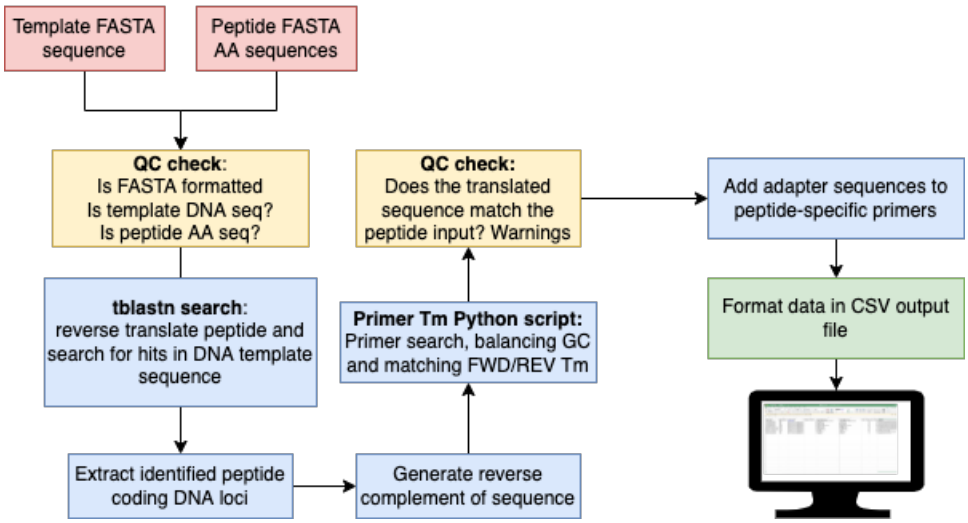


Figure 11. A diagram of the *backend* of the MultiMer design tool. The flowchart represents a vague overview of the program's function. **Red –** user input **Yellow –** QC stage **Blue –** data analysis **Green –** results output.

Indentity Report	Identifier	Genomic Sequence	Percentage Peptide Match	Query Peptide Sequence	Target Peptide Sequence	Genom. Start Pos	Genom. End Pos	Genomic Seq	Genomic Seq (rev complement)	Possible Primer Target (FWD)	Possible Primer Target (REV)	ID	Forward Primer Sequence	Forward Primer Melting Temp	Reverse Primer Sequence	Reverse Primer Melting Temp
100% Peptide identity match	ZIKV33	MF073357|Homo-sapiens|Brazil@2016/02/01	100	SYNNYSLMAMATQAGVLFGMGKGMPFYAWDFGV	SYNNYSLMAMATQAGVLFGMGKGMPFYAWDFGV	7095	7193	TCATACAACAACTACTCCTTAATGGCGATGGCCAAAATCCCATGCGTAGAATGGCATCGGATTTTGGAGTC	GACTCCAAAATCCCATGCGTAGAATGTCATACAACAAGGAGTAGTTGTTGTATGA	AATGGCGATGTCTAGTCCTGAAGTGGAAG	GACTCCAAAATCCCAATGCGAT	27	gcagtagcagtagcagttggtgttggtgttTCATACAACAAT	60.08375899	aacaccaacaccaactgctacTgctactgcGACTCCAAAATCCCATG	58.46170425
100% Peptide identity match	ZIKV36	MF073357|Homo-sapiens|Brazil@2016/02/01	100	SSPEVEEARTLRVLS	SSPEVEEARTLRVLS	8121	8165	TCTAGTCCTGAAGTGGAAGAAGCACGGACGCTCAGAGTCCTCTCC	GGAGAGGACTCTGAGCGTCCGTGCTTCTTCCACTTCAGGACTAGA	GGAGAGGACTCTGAGCGTCCGTGCTTCAAGCACGGACG	CGTGCTTCTTCCG	33	gcagtagcagtagcagttggtgttTCTAGTCCTGAAGTGGAAG	60.08981243	aacaccaacaccaactgctacTgctactgcGGAGAGGACTCTGAGC	61.49950795
100% Peptide identity match	ZIKV37	MF073357|Homo-sapiens|Brazil@2016/02/01	100	DRFAHALRFLNDMGK	DRFAHALRFLNDMGK	9687	9731	GATAGGTTTGCACATGCCCTCAGGTTCTTGAATGATATGGGAAAA	TTTTCCCATATCATTCAAGAACCTGAGGGCATGTGCAAACCTATC	GATAGGTTTGCAGGTTCTTGCACAGAAGG	TCATTCAAGAACCACAGATGA	34	gcagtagcagtagcagttggtgttTATAGGTTTGCACATGCC	59.68982341	aacaccaacaccaactgctacTgctactgcTTTTCCCACCACATGC	59.82601246
100% Peptide identity match	ZIKV38	MF073357|Homo-sapiens|Brazil@2016/02/01	100	HRRDLRLMANAICSSV	HRRDLRLMANAICSSV	9975	10022	CACAGAAGGGACCTCCGACTGATGGCCAATGCCATTTGTTCATCTGTG	CACAGATGAACAAATGGCATTGGCCATCAGTCGGAGGTCCCTTCTGTG	GACCTCCGACTGATGGCCAA	ACAAATGGCATTGGCCATCAG	35	gcagtagcagtagcagttggtgttCACAGAAGTGGACCTC	58.88634008	aacaccaacaccaactgctacTgctactgcCACAGATGAACAAATGGC	60.33540939
100% Peptide identity match	ZIKV39	MF073357|Homo-sapiens|Brazil@2016/02/01	100	CSSVPVDWVPTGRTTWSIHGKGEW	CSSVPVDWVPTGRTTWSIHGKGEW	10011	10082	TGTTCATCTGTGCCAGTTGACTGGGTTCCAACTGGGAGAACTACCTGGTCAATCCATGGAAAGGGAGAATGG	CCATTCTCCCTTTCCATGGATTGACCAGGTACCACAGATGAACA	TGTTCATCTGTGGCCAGTTGACTGGGTT	CCATTCTCCCTTGCCATGGTACCAGTC	36	gcagtagcagtagcagttggtgttTGTTCATCTGTGCCAG	60.13338692	aacaccaacaccaactgctacTgctactgcCCATTCTCCCTTTCCAT	59.48345206
100% Peptide identity match	FLAVMIX3	MF073357|Homo-sapiens|Brazil@2016/02/01	100	DGDIGAVALDYPAGT	DGDIGAVALDYPAGT	4971	5015	GATGGGGACATTGGAGCGGTTGCGCTGGATTACCCAGCAGGAACT	AGTTCCTGCTGGGTAATCCAGCGCAACCGCTCCAATGTCCCCATC	TTGCGCTGGATGACAAAGGA	AGCGCAACCGCTGGGTAA	41	gcagtagcagtagcagttggtgttGATGGGGACATTGGAGCGGTTG	59.94793249	aacaccaacaccaactgctacTgctactgcAGTTCCTGCTGGGTAATCCAGCGCAAC	59.95341052
WARNING! Non-exact peptide match	ZIKV19	MF073357|Homo-sapiens|Brazil@2016/02/01	100	DRGWGNGCGLFGKGS	DKGWGHWSGCAGLPS	4964	5008	GACAAAGGATGGGGACATTGGAGCGGTTGCGAACCAATGTCCCCATCCTTTGTC	GCTGGGTAATCCAGCGCAACCGCTCCGGAGCGGTTGC	TGGGGACATTGGAGCGGTTGC	TCCAGCGCAACCGCTCCAATG	2	gcagtagcagtagcagttggtgttGACAAAGGATGGGGAC	60.22521328	aacaccaacaccaactgctacTgctactgcGCTGGGTAATCCAGC	60.35159909
WARNING! Non-exact peptide match	ZIKV23	MF073357|Homo-sapiens|Brazil@2016/02/01	100	VNELPHGWKAWGKSH	VNELPHGWKAWGKSY	2811	2855	GTGAACGAGCTGCCCCACGGCTGGAAGGCTTGGGGGAAATCGTAC	GTACGATTTCCCCCAAGCCTTCCAGCCGTGGGGCAGCTCGTTCAC	GTGAACGAGCTGCCCCACGG	GTACGATTTCCCCCAAGCCTTCCAGCCCT	8	gcagtagcagtagcagttggtgttGTGAACGAGCTGCCCCACG	64.30348946	aacaccaacaccaactgctacTgctactgcGTACGATTTCCCCCAA	59.38793703
WARNING! Non-exact peptide match	FLAVMIX5	MF073357|Homo-sapiens|Brazil@2016/02/01	100	EDMLMVWNRVWIEEN	EDMLVVWNRVWIEEN	10092	10136	GAAGACATGCTTGTGGTGTGGAACAGAGTGTGGATTGAGGAGAAC	GTTCTCCTCAATCCACACTCTGTTCCACACCACAAGCATGTCTTC	CTTGTGGTGTGGAACAGAGTG	GTTCTCCTCAATCCACACTC	43	gcagtagcagtagcagttggtgttGAAGACATGCTTGTGGTGTGGAACAG	60.88877583	aacaccaacaccaactgctacTgctactgcGTTCTCCTCAATCCACA	59.41102277
WARNING! truncated Peptide	ZIKV19	MF073357|Homo-sapiens|Brazil@2016/02/01	53	GWGNGCGL	AWGKLCSL	10420	10443	GCTTGGGGAAAGCTGTGCAGCCTG	CAGGCTGCACAGCTTTCCCCAAGC	GCTTGGGGAAAGCTGTGCAGCCTG	CAGGCTGCACAGCTTTCCCCAAGC	3	gcagtagcagtagcagttggtgttGCTTGGGGAAAGCTGTGCAGCCTG	61.42044909	aacaccaacaccaactgctacTgctactgcCAGGCTGCACAGCTT	63.8310241
WARNING! truncated Peptide	ZIKV19	MF073357|Homo-sapiens|Brazil@2016/02/01	80	WGNGCGLFGKGS	WKTSCGDQ*KGS	2677	2712	TGGAAAACATCATGTGGAGATCAGTAGAAGGGGAGC	GCTCCCCTTCTACTGATCTCCACATGATGTTTTCCA	TGGAAAACATCATGTGGAGATCAGTAGAAG	GCTCCCCTTCTACTGATCTCCACATGATGT	4	gcagtagcagtagcagttggtgttTGGAAAACATCATGTGGAGA	60.09044749	aacaccaacaccaactgctacTgctactgcGCTCCCCTTCTACTGA	60.60005346
WARNING! truncated Peptide	ZIKV23	MF073357|Homo-sapiens|Brazil@2016/02/01	60	ELPHGWKAW	EVPREWQSW	2224	2250	GAGGTGCCAAGAGAATGGCAGTCTTGG	CCAAGACTGCCATTCTCTTGGCACCTC	GAGGTGCCAAGAGAATGGCAGTCTTGG	CCAAGACTGCCATTCTCTTGGCACCTC	9	gcagtagcagtagcagttggtgttGAGGTGCCAAGAGAAT	59.38860272	aacaccaacaccaactgctacTgctactgcCCAAGACTGCCATTCTCT	59.76115872

Table 1. CSV formatted output of the MultiMer pipeline. The output includes an 'identity report' which notifies the user of any discrepancies in the sequence search process. The right-most columns contain the final adapter-primer sequences and their melting temperatures.

6.3 MultiMer Protocol

Primer Design

The MultiMer online tool has been designed to generate specific primers pairs for amplifying a target protein coding sequence from a specified template while adding adapter sequences for the multimerisation process and subsequent cloning into the *E. coli* expression vector. This pipeline uses NCBI tblastn to reverse translate target peptide sequences and locate them within the template sequence. An in-house script cross-checks reverse-translated matches and generates Tm balanced primers flanking the protein coding sequence.

This process has three steps:

1. Users specify a template by uploading a FASTA formatted file to the 'Template Upload' form.
2. Generation of a list of FASTA formatted target peptide sequences, known to translate from the given template and upload to the 'Peptide Upload' form.
3. Download the resultant CSV file for inspection.

The output details the primer analysis for each targeted peptide. The first column contains a report of the respective peptide primer search. Rows with '100% peptide match' indicate a successful match. The primers in the final columns may be used in the following amplification steps. Primers are usually between 45 and 50 nt long and have a Tm of 60^{o}C ± 2^{o}C.

Important Note

It is important to check that the query and subject peptides match. Mismatches or truncated forms of the target peptide are common. Analyse the sequences to find the perfect match. If

none exist, consider the implications of the mismatch, and continue or use an alternative template.

1. Overlap extension PCR for the generation of pET-vFusion plasmid

Any N-terminus His-tagged tagged pET (or comparable) vector is suitable for this reaction providing it is compatible with BL21 (DE3) *E. coli* protein expression lines. The primers for this step are described in the '**Primer Design - Part 1'** stage in the introduction.

a) Prepare the PCR overlap extension reaction in a 0.2 mL PCR tube. This will add the cloning adapters to the recipient vector to allow for the cloning of the MultiMer insert.

Component	**Volume (µl)**	**Final Concentration**
Q5 Buffer	10	1X
dNTP	1	200 µM
Q5 Polymerase	0.5	0.02 U/µl
Fwd Primer (10 µM)	2.5	0.5 µM
Rev Primer (10 µM)	2.5	0.5 µM
Template	Variable	
Nuclease free water	To final volume	
	Total: 50 µl	

Step	**Temperature (ºC)**	**Duration**
Initial denaturation	98	30 seconds
Cycle 25 times		
Denaturation	98	20 seconds
Annealing	72	30 seconds
Extension	72	4 minutes 30 seconds
End Cycling		
Final extension	72	2 minutes

Add 1 µl of DpnI enzyme mix to the completed PCR reaction and leave for >1 hr at 37ºC.

- If high background is indicated after PCR colony screening in **Step 4**, consider using a lower template concentration in **Step 1**, or increasing DpnI digestion duration.

b) Quantify the vector using a UV-Vis spectrophotometer (PCR clean-up required) or a fluorometric quantification assay.

2. MultiMer insert overlap assembly PCR reaction

For this reaction, the template bearing the protein-coding sequence targeted in the MultiMer website stage of the primer design process must be present at a sufficient concentration for PCR amplification.

a) Mix the two primers for each pair generated in the '**Primer Design – Part 2'** step together equally to a working concentration of 10 µM per pair (5 µM + 5 µM); half the concentration of a conventional primer working stock and set up the following PCR reaction:

- Lyophilised primers are usually produced to be diluted to a stock concentration of 100 µM per tube. Perform this dilution as per the documentation. Take 5 µl from the two 100 µM stocks for a single primer pair and dilute this in to 90 µl of PCR clean water to generate the 5 µM + 5 µM stock required for the MultiMer reaction.

Component	**Volume (µl)**	**Final Concentration**
Q5 Buffer	5	1X
dNTP	0.5	200 µM
Q5 Polymerase	0.25	0.02 U/µl
MultiMer Primer Mix 5 µM + 5 µM	2.5	0.5 µM + 0.5 µM
Template	Variable	
Nuclease free water	To final volume	
	Total: 20 µl	

Step	Temperature (°C)	Duration
Initial denaturation	98	30 seconds
Cycle 60 times		
Denaturation	98	10 seconds
Annealing	60	15 seconds
Extension	72	2 seconds
End Cycling		
Final extension	72	2 minutes

b) Once the PCR reaction has completed, take 5 µl of each reaction and run it on a **2% agarose gel** with a 100 bp ladder. Visualise the gel using an imager compatible with the gel stain used. Successful reactions will be indicated by a laddering effect on the gel.

- To increase the average molecular weight of each MultiMer to the desired range, replace the PCR reaction into the thermal cycler and run the above protocol for a further 10 cycles. The upper size for a MultiMer limit is approximately 1 kb using this method.
- An insert with a specific number of repeats can usually be selected in **Step 4** (colony selection), however if the size of the insert is not the dominant band indicated in the previous gel electrophoresis visualisation step **(Step 2b)**, then the band can be excised from the gel using a gel purification kit (QIAGEN - 28706X4). However, this will increase greatly the bench-time as well as decrease cloning efficiency. If a larger/smaller average size of insert is required, optimise **step 2a** and adjust the PCR cycle count accordingly.

3. NEBuilder® HiFi DNA assembly reaction

The vector generated in **Step 1** and the MultiMer PCR reaction generated in **Step 2** can now be assembled using the NEBuilder HiFi DNA Assembly Reaction Protocol. The MultiMer PCR reaction will have generated a range of molecules around the size of 500 bp, the most

concentrated of which will be preferentially cloned into the recipient vector using the 30 bp adapter sequences added to the both the vector and the insert in the respective PCR reactions.

a) Prepare the NEBuilder HiFi DNA Assembly Reaction using the table below as a guide. The size of the reaction was scaled down by 3 times to conserve the NEBuilder HiFi DNA Assembly Reaction master mix.

    - Keep all the reagents both before and after the reaction on ice.

Component	**Volume (µl)**	**Final Concentration**
NEBuilder HiFi DNA Assembly Master Mix	1.25	1X
PCR amplified vector **(step 1)**	0.2	12.5 ng total
MultiMer amplicons **(step 2)**	0.3	Variable
PCR clean water	**To total: 2.5 µl**	

b) Incubate the prepared reaction mixture for 15 minutes at 50 ºC and place on ice.

4. NEB Stable E. coli transformation

Transform the NEB Stable *E. coli* cell line using the assembly reaction from the previous step. The NEB Stable line was chosen to reduce recombination during propagation of the plasmid. Other lines will work with varying efficiency, providing each 37 ºC growth step is reduced to room temperature. Stocks of the Stable cell line were prepared using the Hanahan Competent Cell Protocol [15], aliquoting the resulting competent cells into 100 µl shots.

a) Transform the competent cells as per the following modified NEB protocol.

    i. Thaw an aliquot of cells on ice for 10 minutes.

    ii. Add 1 µl Assembly reaction to the competent cells.

    iii. Mix cells and DNA gently by flicking the tub. Do not vortex.

    iv. Place the mixture on ice for 30 minutes.

    v. Heat shock at exactly 42°C for exactly 30 seconds. Do not mix. Place on ice for 5 minutes. Pipette 500 µl of room temperature LB medium into the mixture.

Incubate the cells with shaking at RT for 3 hours. Shake the tube horizontally at 250rpm or rotate.

vi. Leave antibiotic selection plates at RT for 30 minutes.

vii. Mix the cells thoroughly by flicking, centrifuge the tube at 800 RCF for 5 minutes, discard the supernatant and resuspend the pellet in 100 µl of LB media by gently pipetting, then streak 100 µl of cells or diluted cells onto a selection plate and incubate at RT for 24 to 36 hours.

5. Colony selection and plasmid preparation

Following the 24 to 36 hours of RT incubation time, small colonies should have grown on the antibiotic LB agar plate. Analyse at least 10 clones by colony PCR using a conventional Taq polymerase with T7/T7-Term primers or equivalent *'sequencing'* primers with annealing sites on the backbone of the recipient plasmid flanking the insert cloning site.

- Do not use primers complementary to the inserted sequence, as this will produce a similar *laddering* effect seen in **Step 2** and will not reflect the true size of the insert.

For high throughput colony screening, scraping single colonies with a sterile pipette tip I suggested, and subsequently depositing them each into a single well on a 96 well PCR plate with 10 µl of LB media (without antibiotic) in each well. Leave this plate on a shaker for 5 minutes to effectively inoculate the media, gently shaking so the tips move in a controlled circular motion, then add 0.5 µl of the inoculated LB to the screening PCR reaction.

Component	Volume (µl)	Final Concentration
10 µM Forward Flanking T7 FWD Primer	0.5 µl	0.2 µM
10 µM Reverse Flanking T7-Term Primer	0.5 µl	0.2 µM
Inoculated LB media	variable	variable
One*Taq* 2X Master Mix with Standard Buffer	12.5 µl	1X
Nuclease-free water	to 25 µl	

a) Set up the following PCR reaction to screen the colonies for the desired MultiMer insert size. For this reaction OneTaq® 2X Master Mix with Standard Buffer (NEB - M0482S) was used.

Step	Temperature (°C)	Duration
Initial Denaturation	94°C	30 seconds
30 Cycles	94°C	15 seconds
	55°C	30 seconds
	68°C	2 minutes
Final Extension	68°C	5 minutes
Hold	4-10°C	

Run the colony screening PCR reactions on a 1% agarose gel with a 100 bp ladder. Analyse the results and identify colonies with the correct inset size insert. Select 2 colonies corresponding to a suitable insert size for plasmid preparation.

- The pET-28a-IVfusion plasmid has an empty cloning site amplicon of ~300 bp. Amplicons exceeding this size indicate the respective copies of the MultiMer.
- Lanes with multiple bands (a band a 300 bp and another faint band at 800 bp) indicate clones that have undergone recombination and insert excision. These should not be selected. Only colonies with a single band >300 bp should be selected for plasmid preparation.

b) Prepare 5 ml LB media cultures with kanamycin (50µg/ml) in 17 x 100 mm culture tubes aiming to prepare 2 different clones per MultiMer construct. Inoculate a 5 ml

LB culture with a 10 µl screening culture corresponding to the positive-screened colony. Leave overnight to grow at 30°C.

c) Following overnight growth, using the QIAprep Spin Miniprep Kit (QIAGEN – 27104) or for higher throughput applications, QuickLyse Miniprep Kit (QIAGEN – 27405). Prepare purified plasmid preparations for each construct and quantify the final eluate.
    - Plasmid concentration may be lower if using Stable lines. If a higher concentration is required for sequencing applications, grow the 5 ml Stable culture for 24 hours.

6. Protein expression and Western blotting

Expression of the MultiMer polypeptides is possible in a range of conventional *E. coli* cell lines and methodologies. I have chosen the BL21 (DE3) line paired with an autoinduction methodology [13]. As before, I prepared our own shots of BL21 (DE3) chemically competent cells using the Hanahan methodology [15]. The expression process is scalable depending on which stage of development the user is in peptide development. For the initial discovery/screening stage, I recommend using 17 x 100 mm culture tubes with 500 µl of ZYP-5052 media + kanamycin. For greater yields, 1 L baffled Erlenmeyer flasks with 200 ml ZYP-5052 media + kanamycin.

a) Transform the BL21 (DE3) *E. coli* using the plasmid preparation generated in the previous step.
    i. Thaw a tube of BL21(DE3) Competent E. coli cells on ice for 10 minutes
    ii. Add 1 µl of plasmid DNA to the cell mixture
    iii. Carefully flick the tube 4–5 times to mix cells and DNA. Do not vortex

iv. Place the mixture on ice for 30 minutes. Do not mix

v. Heat shock at exactly 42°C for exactly 30 seconds. Do not mix

vi. Place on ice for 5 minutes. Do not mix

vii. Pipette 950 µl of room temperature LB broth into the mixture

viii. Place at 37°C for 60 minutes. Shake vigorously (250 rpm) or rotate

ix. Warm selection plates to 37°C

x. Mix the cells thoroughly by flicking the tube and inverting

xi. Streak 50–100 µl onto a selection plate and incubate overnight at 37°C

b) Pick two single colonies for each construct from the BL21 (DE3) selection plate and culture them in a 17 x 100 mm culture tube with 500 µl of ZYP-5052 autoinducing media + kanamycin. This culture will serve as both an analyte for the Western blotting process and as a starter culture for large-scale culture. Place the tubes in a 37°C shaker-incubator for >20 hours.

c) Following successful overnight growth, pipette 200 µl from each culture tube in to a 1.5 ml microcentrifuge tube and place the 17 x 100 mm culture tubes in the refrigerator. Centrifuge the 1.5 ml microcentrifuge tubes for 3 minutes at full speed.

- If the time between completing the Western blot **(Step 6d)** and initiating large-scale cultures **(Step 7a)** will be greater than 24 hours, add 1 ml of ZY broth to each 17 x 100 mm culture tube for longer term storage.

d) To each microcentrifuge tube, add 20 µl of 1x Laemmli buffer + 350 mM DTT and pipette vigorously to mix. Boil the samples using a heating block at 100°C for 30 minutes. After boiling, the sample is at all viscous, add a further 10 µl of 1x Laemmli buffer + 350 mM DTT and repeat the boiling step.

e) Prepare an SDS-PAGE apparatus capable of resolving 15-80 kDa protein analytes. The Mini-PROTEAN electrophoresis system (Bio Rad - 1658005EDU) was used with 4–15% Mini-PROTEAN™ TGX Stain-Free™ Protein Gels, 15 well, 15 µl (Bio-Rad 4568086). Load 10 µl of each sample from the boiled microcentrifuge tube, with a ladder on to the gel and run according to manufacturer's instruction.

- Bio-Rad stain-free gels (and comparable similar technologies) allow visualisation of total protein and transfer on to blotting membrane from a single gel. These technologies are recommended for high-throughput applications of this protocol.
- Due to the nature of stain-free visualisation chemistry, some peptides that are deficient or rich in tryptophan may be over or underrepresented on the gel. In these cases, a second gel with Coomassie staining may be appropriate. Although total protein visualisation is not always essential when Western blotting is performed.

f) Transfer the gel on to a PVDF or nitrocellulose membrane using the transfer apparatus of choice. Blot using a rat-α-Myc-tag antibody or, additionally a mouse-α-6xHis antibody and an antibody against the protein from which the MultiMer peptide originates. Polyclonal Human-anti-ZIKV was used in this case.

- With this blotting configuration, 3 fluorescent secondary antibodies α-mouse-488, α-rat-546 and α-human-800 can be used on a single blot to identify each MultiMer's N-terminus 6xHis tag, the C-terminus Myc tag and serve as a preliminary analysis to screen Multi'Mers against immune serum.

i. Block membrane using PBS/T + 5% milk for 1 hour at RT on a shaker. Discard the block solution.
ii. Mix primary antibodies at the appropriate concentration with PBS/T + 5% milk and incubate with the blot at RT for 1 hour. Remove antibody-block solution and freeze at -20°C for future blots.
iii. Wash three times for 5 minutes on a shaker in PBS/T.
iv. Add secondary antibodies at the appropriate concentration into a PBS/T solution and incubate in an opaque container for 1 hour at RT on a shaker and discard the solution.
v. Wash three times for 5 minutes on a shaker in PBS/T.
vi. Visualise the membrane using an appropriate florescent imager.

Using the Western blot image, select clones that have successfully expressed the MultiMer peptides. This is indicated by the presence of rat-α-Myc-tag antibody signal. The Myc tag is located on the C-terminus of the peptide, which implies that the full-length peptide has been expressed.

7. High volume culture, cell lysis and protein purification

At this stage, successful MultiMer expressing BL21 (DE3) clones can be achieved in a larger volume. Following this, lysis and protein purification will yield highly pure > 0.5 mg/ml MultiMer peptide preparations for downstream immunoassay applications. Purification of the peptides involves the re-solubilisation of inclusion-bodies from a washed cell pellet. I have validated that buffer exchange from out of the denaturing solution is not required for ELISA or Luminex applications. For downstream applications that require the MultiMer peptides to

be in a different buffer, Amicon® Ultra-15 Centrifugal Filter Units (Merck - UFC900308) utilised with a high-volume dilution and centrifugal concentration protocol, as per product supporting documentation are a suitable methodology for exchange. Modulate the pH of the recipient buffer to increase polypeptide solubility during centrifugal buffer exchange.

a) Using the Western blot in the previous step as a guide, from the refrigerated liquid culture stocks made in **Step 6c**, inoculate a 200 ml ZYP-5052 antibiotic culture in a 1 L baffled Erlenmeyer flask with the entire contents of the 17 x 100 mm culture tube. Leave overnight shaking at 200 RPM in a 37°C incubator.
b) Following incubation, decant the bacteria into a container for centrifugation at full speed for 15 minutes and discard the now translucent supernatant.
    - A single 50 ml falcon tube for each culture was used, repeating centrifugation for the entire 200 ml volume.
c) To the 50 ml falcon tube with the entire bacterial pellet, add the lysis buffer of choice up to the 40 ml mark. Each buffer should be suited for the method of cell lysis:
        i. For probe sonication use PBS (water-bath sonication is not sufficient).
        ii. For French (cell) press, use PBS.
        iii. For enzymatic lysis use 0.5 mg/ml lysozyme from egg white supplemented with 800 U/ml DNase and 24 U/ml RNase.
        iv. Freeze thaw cycles, use PBS.

d) Following thorough cell lysis, transfer 2 ml of the lysate to a 2 ml microcentrifuge tube and centrifuge at full speed >15,000 RCF for 10 minutes.
    - I have chosen to use small volumes for the subsequent purification steps, as each lysis preparation usually produces >2 mg per run.

e) Discard the supernatant and resuspend the pellet in pre-solubilisation buffer up to a volume of 1.5 ml **(see materials for recipe).** Using a probe sonicator to resuspend the lysate.

f) Centrifuge at full speed >15,000 RCF for 10 minutes and repeat **Step 7e.**

g) Centrifuge at full speed >15,000 RCF, add solubilisation buffer to a final volume of 1.5 ml **(see materials for recipe),** and resuspend by sonication. Centrifuge once more at full speed >15,000 RCF for 10 minutes and transfer the supernatant to a fresh 1.5 ml microcentrifuge tube.

h) Perform the NEBExpress® Ni Spin Column Reaction Protocol (NEB #S1427) as per manufacturer's instruction, using all buffers supplemented with 8 M urea. Collect the flow-through from the 'Lysate Binding' step and re use in subsequent purification repeats.

i) Repeat protocol until the desired volume of eluate is obtained.

j) Quantify the peptide using the Bradford assay with a BSA standard curve. Do not use BCA quantification as the contents of the buffer are not compatible.

    - In the output from the MultiMer website primer generation step, a column is included with the extinction coefficient of the specific peptide. This can be used to increase the accuracy of spectrophotometric protein quantification methods using the Beer-Lambert Law.

k) Store the final eluate, crude lysate, solubilised protein at -80°C.

8. ELISA – identification of reactive MultiMer candidates

This final step serves as a preliminary analysis on whether the expressed MultiMer construct functions effectively as an antigen or not. For this analysis I use a conventional *checkerboard* ELISA approach, titrating both antigen and serum concentration. I have used a polyclonal human-anti-ZIKV (internal control) with a rabbit-anti-human-HRP (thermo: A18805) secondary antibody.

As noted in **Step 7**, the effect of urea and imidazole present in the antigen buffer on the ELISA and Luminex assays has been found to be negligible. It is therefore not essential to exchange buffer prior to ELISA analysis. I have used a carbonate-bicarbonate coating buffer (pH 9.6) supplemented with 2 M urea and 125 mM imidazole to ensure peptide solubility. Coating buffer with 8 M urea and 500 mM imidazole will work, but has been found to be unnecessary, and high concentrations urea may precipitate in 7^{o}C incubation steps.

6.3.1 Materials - Reagents

Generation of pET-vFusion plasmid

- Appropriate pET or compatible T7/lacUV5 *E. coli* protein expression plasmid.
- Q5 Hot Start Polymerase, or any comparable high-fidelity polymerase. (NEB-M0491S)
- dNTPs (NEB-N0447S)
- Forward and reverse adapter sequence primers (see primer design section).
- DpnI restriction endonuclease (NEB-R0176S)
- *(Optional)* QIAquick PCR Purification Kit (QIAGEN -28104)

Overlap assembly PCR reaction

- DNA Template genomic material.
    - For ZIKV, three sets of primers were used to amplify 3 kb lengths spanning the genome from cDNA (See appendix).

- Forward and reverse MultiMer DNA primers, each at a final concentration of 5 μM (see primer design section).
- Q5 Hot Start Polymerase, or any comparable high-fidelity polymerase. (NEB-M0491S)
- dNTPs (NEB-N0447S)

DNA Gel electrophoresis

- Gel Loading Dye, Purple (6×) (NEB-B7024)
- 100-bp DNA Ladder (NEB-N3231)
- SeaKem LE Agarose (Lonza-5000)
- 10× TAE
- SYBR Safe DNA Gel Stain (Thermo Fisher Scientific-S33102)

Cloning of MultiMer fragment in pET-vFusion to vector

- MultiMer PCR reaction (confirmed on gel)
    - Purification of the DNA fragment from the previous step is not necessary.
    - Optionally, for increased specificity of the size of insert, gel excision can be used.
- NEBuilder® HiFi DNA Assembly Cloning Kit (NEB-E5520S).
- 50 μg/mL Kanamycin LB agar plates (2 plates per construct).
- 50 μg/mL Kanamycin LB broth (20 ml per construct).
- 'Economy' Taq Polymerase (NEB- M0480S).
- QIAprep Spin Miniprep Kit (QIAGEN- 27104).
- Competent cells
    - Optional: For longer repeats, to increase insert stability >500bp we recommend using NEB® Stable Competent E. coli. Our own chemically competent Stable cells were produced using the Hanahan method. We found acceptable results with XL10-G or DH5α grown at RT.

Expression of Cloned MultiMer Peptides

- MultiMer plasmid prep

- BL21 (DE3) (NEB-C2527H) or BLR (DE3) (Merck-69053-3) competent cells.
- 50 µg/mL Kanamycin LB agar plates

Reagents for autoinduction using the Studier autoinduction methodology:

- ZYP 5052 media
    1. 1 L = 10g bacto tryptone, 5 g yeast , 925 water
    2. 20x NPS + 50X 5052 + MgSO4 (see Studier methodology)
    3. Kanamycin to final conc. of 100 µg/mL (or antibiotic of choice)

Lysis and purification:

- Phosphate buffered saline
- Pre solubilisation buffer – 2% Triton X-100, 2 M urea, 20 mM sodium phosphate, 300 mM NaCl, pH 7.4
- Solubilisation Buffer – 8 M 8rea, 20 mM sodium phosphate, 300 mM NaCl, pH 7.4
- NEBExpress Ni Spin Wash Buffer – 8 M urea, 5 mM imidazole 20 mM sodium phosphate, 300 mM NaCl, pH 7.4
- NEBExpress Ni Spin Elution Buffer – 8 M urea, 500 mM imidazole, 20 mM sodium phosphate, 300 mM NaCl, pH 7.4
- NEBExpress® Ni Spin Columns (25 Pack) (NEB- S1427L)

SDS-PAGE and Western Blotting Purified MultiMer Peptides

- Mini-PROTEAN TGX Stain-Free Protein Gels, 15 well, 15 µl (Bio-Rad – 4568096).
- PageRuler™ Prestained Protein Ladder, 10 to 180 kDa (Thermo - 26617).
- Laemmli buffer.
- Transfer buffer.
- Anti-Myc primary antibody (Abcam: ab10910).
- Goat anti-Rat IgG (H+L) Cross-Adsorbed Secondary Antibody, HRP (Thermo - A10549).
    - Fluorescent conjugated antibodies can also be used (Thermo - A-21096). This eliminates the need for HRP substrate.
- Pierce™ ECL Western Blotting Substrate (Thermo - 32109).

MultiMer Peptide ELISA

- Quantified antigen (MultiMer Peptides)
- Controls:
    1. Polyclonal immune positive to target organism.
    2. Confirmed negative sera.
- TMB one component HRP microwell substrate (Tebu-bio laboratories - TMBW-1000-01).
- Tween 20.
- Skimmed milk powder.
- 0.2M sulphuric acid

6.3.2 Materials - Equipment

Amplification

- PCR tubes. (STARLAB-I1402-4300)
- Gel electrophoresis tank
- UV gel imaging device
- Thermal cycler

Cloning

- 17 x 100 mm culture tubes
- Rotary shaker
- Water bath/heating block

Protein Expression

- 1 L baffled Erlenmeyer flasks
    - Baffling is essential. Flask-culture volume ratio is central to efficient expression. See Studier documentation.
- Sonicator with probe
    - A cell press or enzymatic lysis can be used to similar effect.
- 2 mL Eppendorf tubes (Eppendorf – 0030120094)
- High speed microcentrifuge (>15,000 RCF)

SDS-PAGE and Western blot

- *Mini*-PROTEAN electrophoresis system (Bio-Rad - 1658005EDU)

- 4–15% Mini-PROTEAN™ TGX Stain-Free™ Protein Gels, 15 well, 15 µl (Bio-Rad 4568086)
- *Western blotting transfer apparatus.*
- *Chemiluminescent or florescent Western blot imager.*

Peptide ELISA

- Immulon 4 HBX 16-well ELISA plates (Thermo – 3855)
- Plate reader
- Refrigerator

6.4 Results

6.4.1 High throughput cloning of MultiMer peptide libraries

With the MultiMer PCR reaction well characterised, and the Gibson cloning strategy operating at an acceptable level, the production of MultiMer constructs was scaled up to cover 12 of the peptide candidates. For the initial MultiMer PCR reaction, 2.5 µM of each primer was used

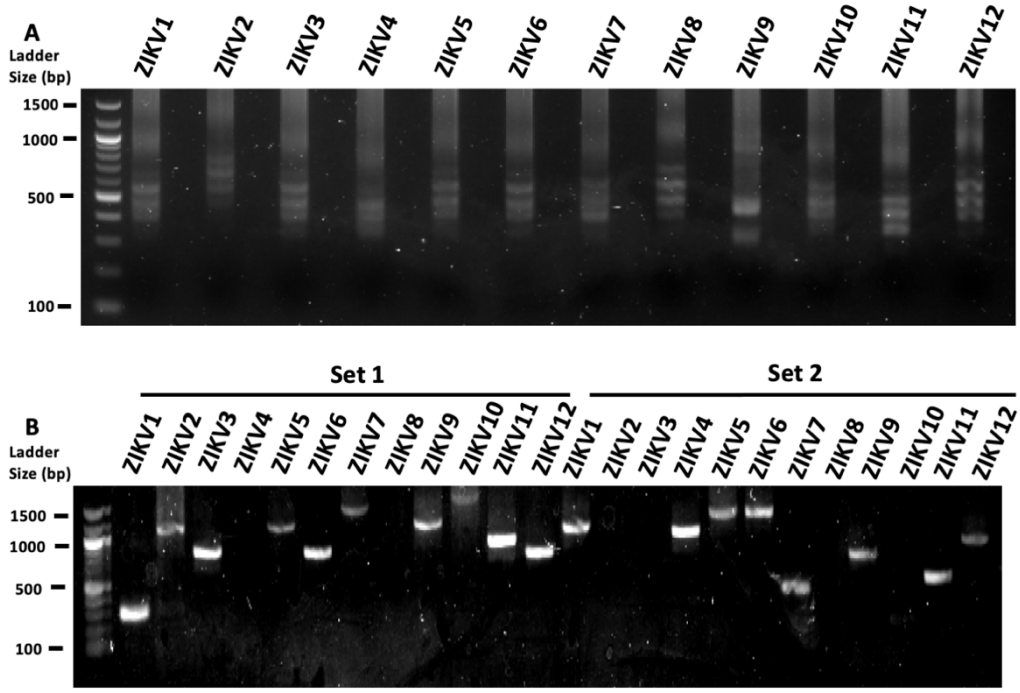


Figure 12. (a) Multimer PCR reaction of 12 ZIKV peptides coding sequences of varying length. The crude PCR product was run on a 2% agarose gel, with a 100 bp marker. **(b)** T7 PCR of final plasmid-prepped cloned constructs. The band size represents the size of the insert (minus the 300 bp N and C term tags. The gen was run, as above. Two plasmid preps were generated for each single construct (**Set 1, Set 2).**

with 40 cycles, which based on the modelling outlined in the previous section, would produce Multimers of approximately 500 bp in length. The gel electrophoresis of the fragments is shown (**Figure 12a).** The range of fragment sizes yielded from the reaction were between ~350 – 800 bp. I added 0.2 µL of each fragment to a Gibson cloning reaction with the pET28a-Myc-Multimer plasmid and cloned (at room temperature) and selected 8 colonies from each agar plate. I found at least one stable colony (out of 8 tested) with a cloned insert of ~500 bp. Using the T7 primers, the MCS was amplified, the resulting gel of which is shown in **Figure 12a.** Interestingly, most stable cloned fragments with the 350 bp N and C termini tags subtracted, lay between 600 and 1200 bp, which is significantly greater than the ~350 – 800 bp visualised in **Figure 12b.** A subset of the resulting plasmids was sequence verified using flanking T7 promotor and terminator primers. No frame-shifting mutations were observed in any of the constructs (**Figure S2**)

6.4.2 High-throughput expression of MultiMer constructs

Following transformation of the BL21-DE3 *E. coli* system with each respective pET28a-Myc-Multimer-insert plasmid, cultures were grown using the autoinduction methodology. Grown at 37°C, all the peptides were expressed and enriched from inclusion bodies using sonication and successive treatments with a chaotropic agent. As an initial screen for the expression of

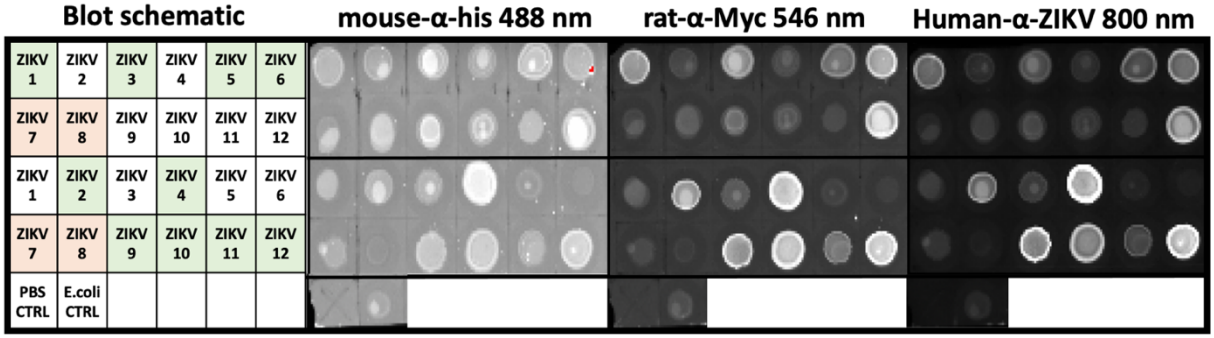


Figure 13. A single Immunostained dot-blot of concentration normalised clarified induced *E. coli* lysates. 10 µL of the lysate was added to a nitrocellulose membrane, which was washed 3 times, blocked, and blotted with the aforementioned antibody pairs. The blot was visualised at excitation wavelengths of 488, 546 and 800 in tandem on a Bio-Rad gel-doc system.

peptides, a dot-blot was preformed using the clarified *E. coli* lysate, blotted with three primary and secondary antibody pairs. Mouse-anti-6X-his tag + goat-anti-mouse-488, rat-anti-Myc + anti-rat-546 and Human-anti-ZIKV + anti-Human-800 **(Figure 13)**. These antibodies were chosen to identify the 6X histidine tag and Myc epitope tags, present on the N and C termini of each peptide, respectively, and a Human ZIKV immune polyclonal antibody, which serves as a preliminary indicator of ZIKV reactivity. The anti-6X-his tag antibody exhibited non-specific activity, due to the ambiguous nature of the histidine tag epitope. The Myc and ZIKV antibodies appeared to bind specifically, with little to no reactivity with the *E. coli* control **(Figure 13 – CTRL)**. Out of the 24 lysates screened (12 peptides, two replicates), 11 of them expressed levels of Myc tag detectable in this assay.

The expression of the Myc tag on the blot indicated that the construct had been expressed in full. Ten *E. coli* lysates were selected (**Figure 13)**, and nickel-bead purification was performed using a streamlined micro-centrifugal column protocol. The columns yielded a highly purified protein product which was eluted using a 500 mM imidazole and 2 M urea buffer. The purified proteins were diluted to a consistent concentration and analysed using an SDS-PAGE and Western blot technique (**Figure 14)**. The total protein was visualised using the incorporated dye in Bio-Rad 'Stain Free' gels, and the transferred gel was blotted with rat-anti-Myc + anti-rat-546 and Human-anti-ZIKV + anti-Human-800 antibodies. In this iteration, apparent truncation of the peptides occurs, as indicated by the laddered effect seen on the total protein gel. However, the primary signal form both the Myc tag and the ZIKV immune sera, indicates that most of the protein was expressed in full. On top of the 10 MultiMer peptides, the full-length ZIKV and DENV NS1 proteins were expressed (lanes 1 and 2) as a positive control for immunoassay comparison.

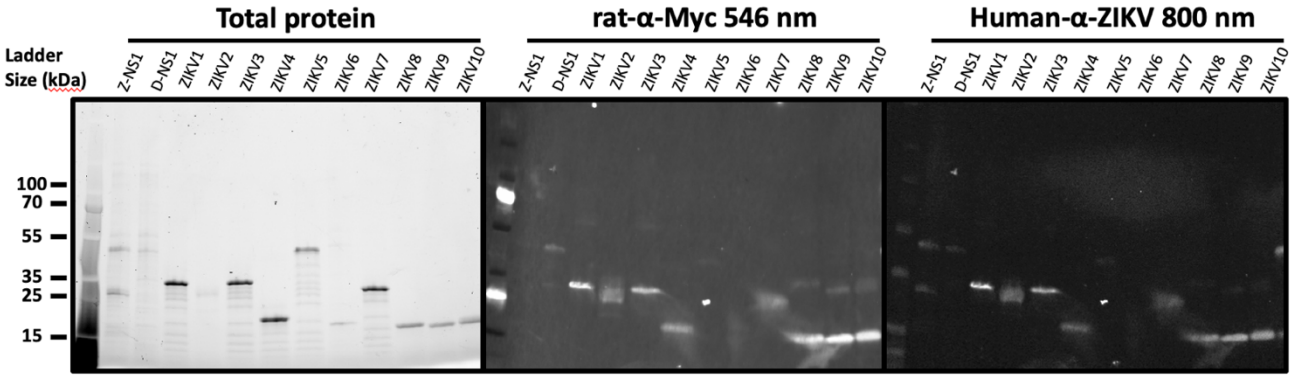


Figure 14. SDS-PAGE 15% Tris Glycine gel, with a dual antibody (Myc and ZIKV) western blot. A Page Ruler 10 kDa to 180 kDa pre-stained ladder was loaded in the left-most lane. Protein analytes were loaded in equal concentration.

6.4.4 Preliminary analysis of peptide reactivity to ZIKV immune sera

To ascertain whether the peptides were reactive to ZIKV positive polyclonal sera, ten MultiMer peptides were screened in a titration against an immune ZIKV positive sera and a serologically confirmed negative control. Firstly, a single MultiMer antigen, ZIKV1, was screened with ZIKV-NS1 with 3 different dilutions of human ZIKV immune sera, 1:50, 1:100 and 1:400 (**Figure 15).** For the full-length ZIKV-NS1 protein, both 1:50 and 1:100 sera dilutions exhibited similar reactivity to both the immune and non-immune sera. The typical sigmoidal curve seen during ELISA titration experiments indicated that the assay reached saturation at approximately 100 µg/mL. Unlike ZIKV-NS1, the ZIKV1 MultiMer peptide did not reach saturation, but there does appear to be a small difference between the 1:50 and 1:100 dilutions. In subsequent repeats on other MultiMer peptides, a similar lack in this differential was observed, and a lack of saturation, even with coating concentrations of 300 µg. Given these results, for the preliminary screening of the ten MultiMer peptides, a 1:50 serum dilution was chosen.

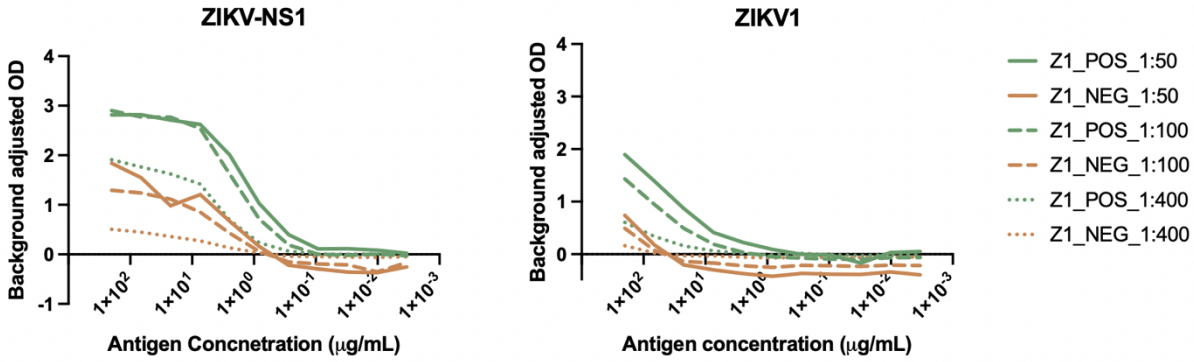


Figure 15. Titration of antigen concentration against analyte ZIKV +ve and ZIKV -ve polyclonal sera. Coating antigen concentration ranges from 300 µg to 0.005 µg.

With the titration of a further nine peptides at a serum concentration of 1:50 (**Figure 16**), only two of the peptides, ZIKV1 and ZIKV5, reached an OD of greater then 0.6, 50% of the signal compared to ZIKV-NS1. Despite this drop, both peptides exhibited 3-fold greater reactivity when compared to the negative samples. The other eight peptides exhibited with low overall reactivity, or a low ratio when comparing the reactivity of ZIKV positive sera, to negative. No correlation was observed between the number of repeats in the peptides and the overall reactivity on ELISA.

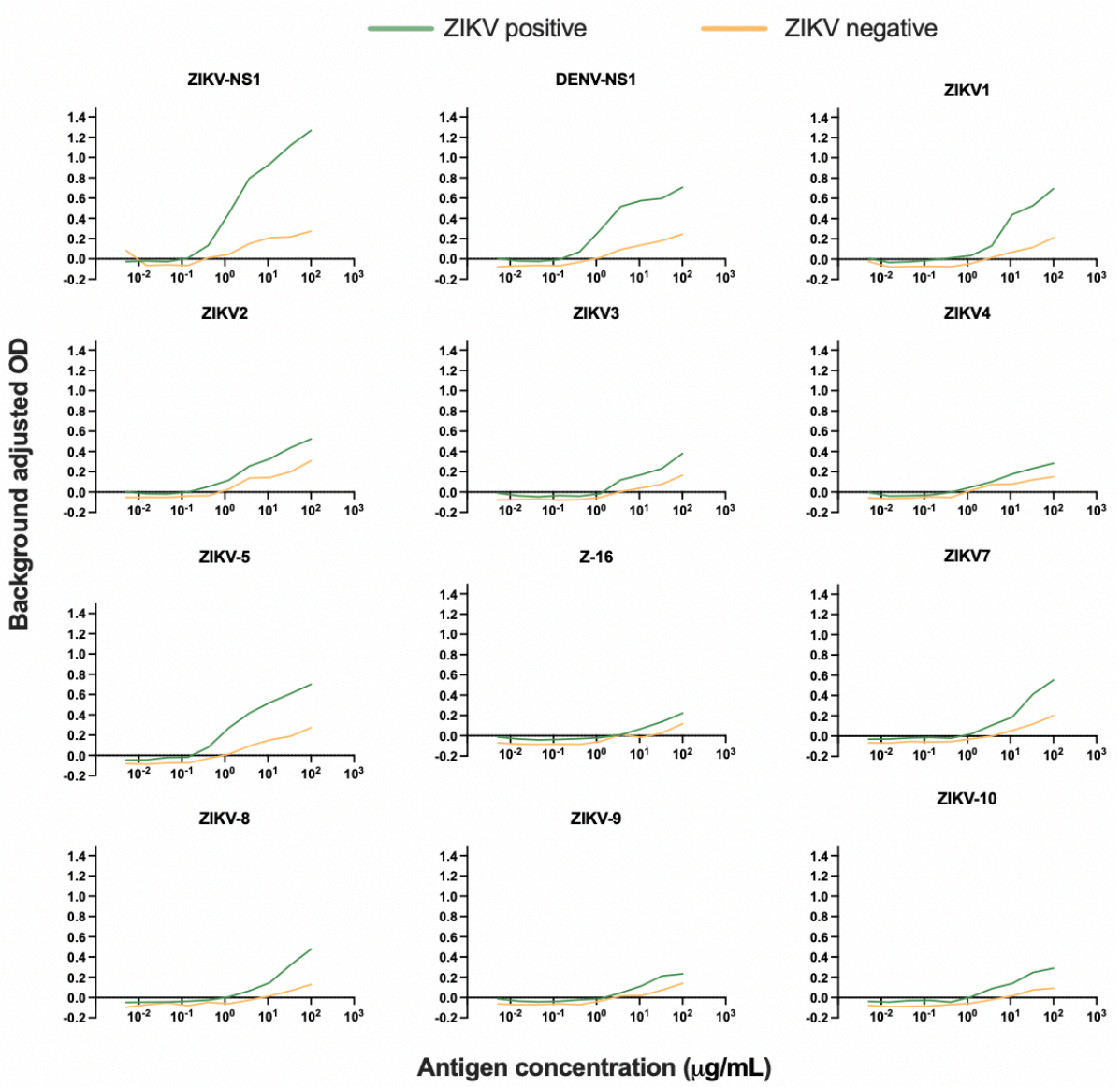


Figure 16. Titration of antigen concentration against analyte ZIKV +ve and ZIKV -ve polyclonal sera. 12 of the expressed peptides were coated on to ELISA plates in the presence of 2M urea. Coating antigen concentration ranges from 300 µg to 0.005 µg.

6.5 Discussion

In this section, the development and validation of the central molecular mechanics of the MultiMer methodology has been described. The modelling of the MultiMer PCR reaction, both *in-silico* and *in-vitro,* has elucidated the key variables that drive the assembly reaction. I have demonstrated that the modulation of adapter-primer concentration facilitates the predictable

assembly of MultiMer repeats, in a simple one-step reaction. I believe that this methodology can not only be applied in the context shown here, but also other synthetic biology applications that require the generation of such constructs. This methodology may also benefit the field of antimicrobial-peptide study, eliminating the need to time-consuming serial cloning techniques, synthesis or enzymatic assembly.

While the technique has limitations, I have described a novel methodology and laid the foundation for its continued improvement and development. The low cloning efficiency (**Figure 10)** can be rectified with modification of the adapter sequences. I believe the limited diversity in the nucleotide content inherent of the Gly-Ala codon repeats facilitates hairpin formation, and backbone self-annealing, which can be corrected with further adjustment. In future work, it would be possible to understand the effect of modifying adapters, both at sequence and the amino-acid level, on cloning efficiency and expression levels. The adapter peptides present further opportunities for exploring additional functionality in this system. The addition of a protease cleavage site would allow for exploring the possibilities of cleaving MultiMer peptides back into their monomeric form, which would yield peptides not too dissimilar to those produced in chemical synthesis.

The preliminary analysis of the MultiMer peptides demonstrated that high yields of MultiMer peptides can indeed be expressed using the technique described. The laddering effect (seen in **Figure 14)** may be corrected by moving the his-tag to the C terminus of the plasmid backbone MCS. This moving would ensure that only the full-length product is purified eliminating the truncated species during purification; thereby allowing the production of precise repeat numbers, crucial to the validation of this technology.

The preliminary ELISA screening of the 10 MultiMer peptides produced in the example shown here was insufficient to gain a robust understanding of the antigenic qualities of the target peptides. Reassuringly, both immunoblots (**Figure 13; Figure 14)** appear to demonstrate reactivity to ZIKV immune sera. In numerous iterations of ELISA experiments, I have attempted to increase the signal obtained from each peptide titration, while broadening the panel of peptides. Modifying the buffer in which the peptide is stored, the coating concentration and the plate medium, appeared to yield a no more typical binding curve than that shown here. Performing titrations using only the attached Myc epitope tag and its cognate antibody yields a similar result, which indicates that this result is not peptide-specific and may be an intrinsic flaw in the proposed ELISA methodology, most likely regarding the association of the peptide with the solid-phase. Moving forward, it would be possible to test small MultiMer fusion partners, which may increase the binding efficiency of the peptides.

The final caveat lies in the difficulty associated with the visualisation and quantification of synthetic peptides. When visualising peptides using the SDS-PAGE 'Stain Free' technology or other similar products, the incorporated stain covalently binds with the aromatic tryptophan residues to visualise protein bands following UV activation. The ZIKV2 peptides (amongst others), along with the linkers and affinity/epitope tags contain no tryptophan residues, which can result in underrepresentation in such analyses. In other iterations, I have used Coomassie blue stain to visualise tryptophan devoid peptides (**Figure S3).** This alternative is not elegant and highlights a neglected issue. Most visualisation or quantification techniques operate on the principle of a *reaction* with residue functional group(s). As such, the outcome of the assay is entirely sequence dependent. Coomassie, Bradford and bicinchoninic acid assays react with aromatic groups, silver stain with carboxylic acid groups, Lowry assay with cysteine, and 280 nm spectrophotometric analysis with aromatic rings.

While in most scenarios, quantifying *natural* proteins, these residues are usually well balanced, here, a peptide may consist almost entirely of the same 17 repeated residues, which presents a unique issue. In exploring this challenge, I have attempted to capitalise on the 205 nm absorbance of peptide bonds, compensating with a pre-calculated extinction co-efficient of peptides. However, absorbance at this wavelength is shared with several common compounds, including salts present in most protein storage buffers. Despite these limitations, the proposed approach offers a new technique which may yield insights into repeat construct synthesis, with applications across synthetic biology.

6.6 References

1. Hansen S, Hotop S-K, Faye O, Ndiaye O, Böhlken-Fascher S, Pessôa R, *et al.* Diagnosing Zika virus infection against a background of other flaviviruses: Studies in high resolution serological analysis. Sci Rep [Internet]. 2019;9:3648. Available from: https://doi.org/10.1038/s41598-019-40224-2

2. Mishra N, Caciula A, Price A, Thakkar R, Ng J, Chauhan L V, *et al.* Diagnosis of Zika Virus Infection by Peptide Array and Enzyme-Linked Immunosorbent Assay. MBio [Internet]. American Society for Microbiology; 2018 [cited 2019 Feb 14];9:e00095-18. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29511073

3. Kam Y-W, Leite JA, Amrun SN, Lum F-M, Yee W-X, Bakar FA, *et al.* ZIKV-Specific NS1 Epitopes as Serological Markers of Acute Zika Virus Infection. J Infect Dis [Internet]. Oxford University Press; 2019 [cited 2021 Feb 10];220:203–12. Available from: https://academic.oup.com/jid/article/220/2/203/5371191

4. Auerswald H, Klepsch L, Schreiber S, Hülsemann J, Franzke K, Kann S, *et al.* The Dengue ED3 Dot Assay, a Novel Serological Test for the Detection of Denguevirus Type-Specific Antibodies and Its Application in a Retrospective Seroprevalence Study. Viruses [Internet]. MDPI AG; 2019 [cited 2021 Feb 10];11:304. Available from: https://www.mdpi.com/1999-4915/11/4/304

5. Lee AJ, Bhattacharya R, Scheuermann RH, Pickett BE. Identification of diagnostic peptide regions that distinguish Zika virus from related mosquito-borne Flaviviruses. Wang T, editor. PLoS One [Internet]. Public Library of Science; 2017 [cited 2021 Feb 10];12:e0178199. Available from: https://dx.plos.org/10.1371/journal.pone.0178199

6. Li Y. Carrier proteins for fusion expression of antimicrobial peptides in Escherichia coli. Biotechnol Appl Biochem. Wiley Online Library; 2009;54:1–9.

7. Jin F, Xu X, Zhang W, Gu D. Expression and characterization of a housefly cecropin gene in the methylotrophic yeast, Pichia pastoris. Protein Expr Purif. Elsevier; 2006;49:39–46.

8. Tian Z, Dong T, Yang Y, Teng D, Wang J. Expression of antimicrobial peptide LH multimers in Escherichia coli C43 (DE3). Appl Microbiol Biotechnol. Springer; 2009;83:143–9.

9. Matthyssen T, Li W, Holden JA, Lenzo JC, Hadjigol S, O'Brien-Simpson NM. The Potential of Modified and Multimeric Antimicrobial Peptide Materials as Superbug Killers. Front Chem. Frontiers; 2022;1108.

10. Nagaraj S, Reddy PN, Ramlal S, Paul S, Peddayelachagiri B, Parida DM. A novel tandem repeat cloning technique for creation of multiple short peptide repeats to differentiate closely related antigens. J Immunol Methods [Internet]. 2019;469:11–7. Available from: https://www.sciencedirect.com/science/article/pii/S0022175918303430

11. Lee JH, Minn IL, Park CB, Kim SC. Acidic peptide-mediated expression of the antimicrobial peptide buforin II as tandem repeats inEscherichia coli. Protein Expr Purif. Elsevier; 1998;12:53–60.

12. Chen X, Zaro JL, Shen W-C. Fusion protein linkers: property, design and functionality. Adv Drug Deliv Rev. Elsevier; 2013;65:1357–69.

13. Studier FW. Protein production by auto-induction in high density shaking cultures. Protein Expr Purif [Internet]. Protein Expr Purif; 2005 [cited 2021 May 19];41:207–34. Available from: https://pubmed.ncbi.nlm.nih.gov/15915565/

14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

15. Green MR, Sambrook J. The hanahan method for preparation and transformation of

competent Escherichia coli: High-efficiency transformation. Cold Spring Harb Protoc [Internet]. Cold Spring Harbor Laboratory Press; 2018 [cited 2021 May 18];2018:183–90. Available from: https://pubmed.ncbi.nlm.nih.gov/29496820/

Table S1. Multimer protein concentration determined by Bradford assay standard interpolation.

Peptide/Protein	Concentration (mg/mL)
Z-NS1	1.58
D3-NS1	1.61
ZIKV1	1.59
ZIKV2	0.80
ZIKV3	1.26
ZIKV4	1.24
ZIKV5	0.58
ZIKV6	0.83
ZIKV7	1.36
ZIKV8	0.20
ZIKV9	1.23
ZIKV10	1.14

Figure S1. Screenshots of the MultiMer primer design website.

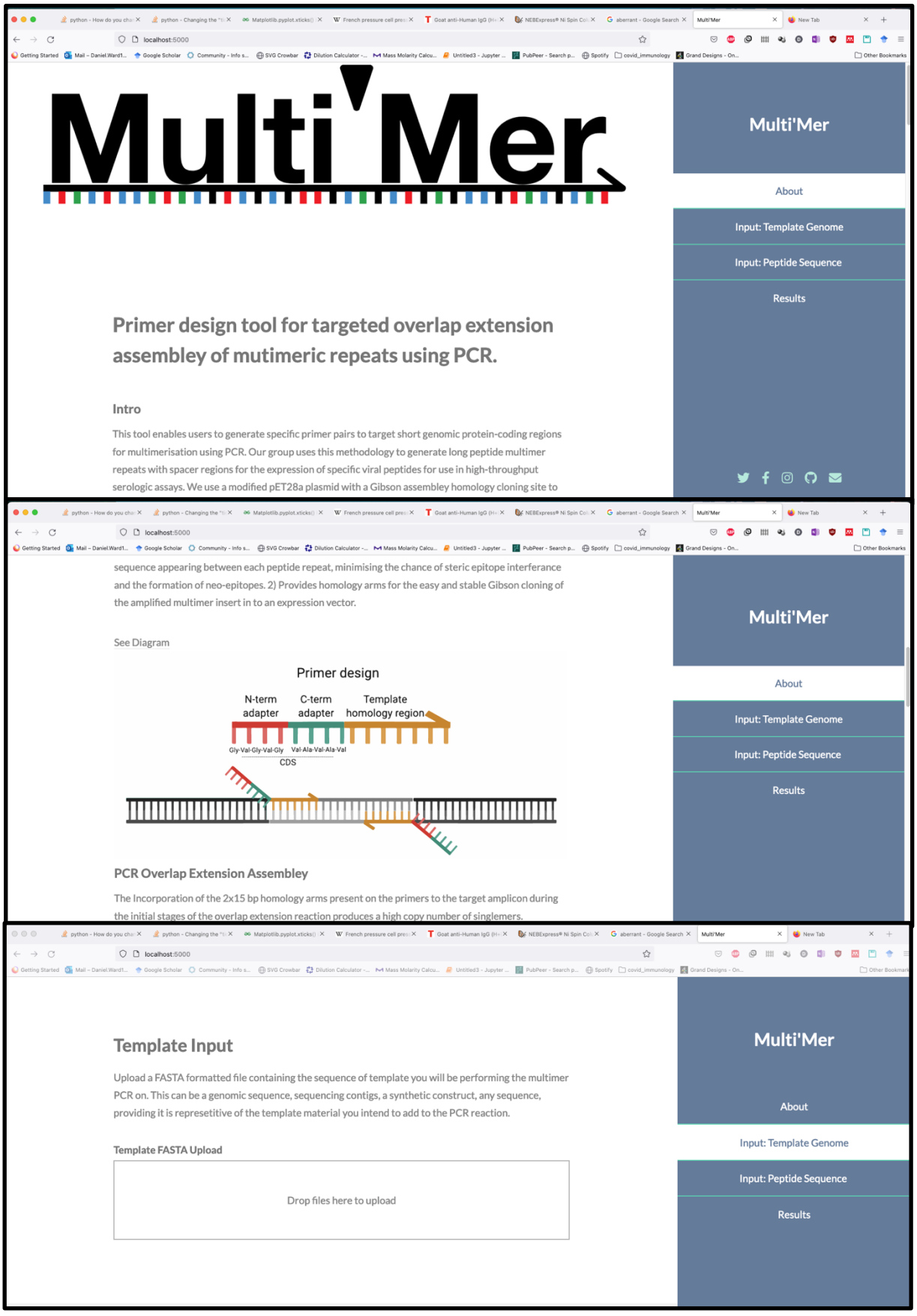


Figure S2. Sanger sequencing verification of Multimer construct. T7 sequencing primer PCR amplicication of the cloned, plasmid prepeared Multimer in the pET28a-Myc-Multimer vector.

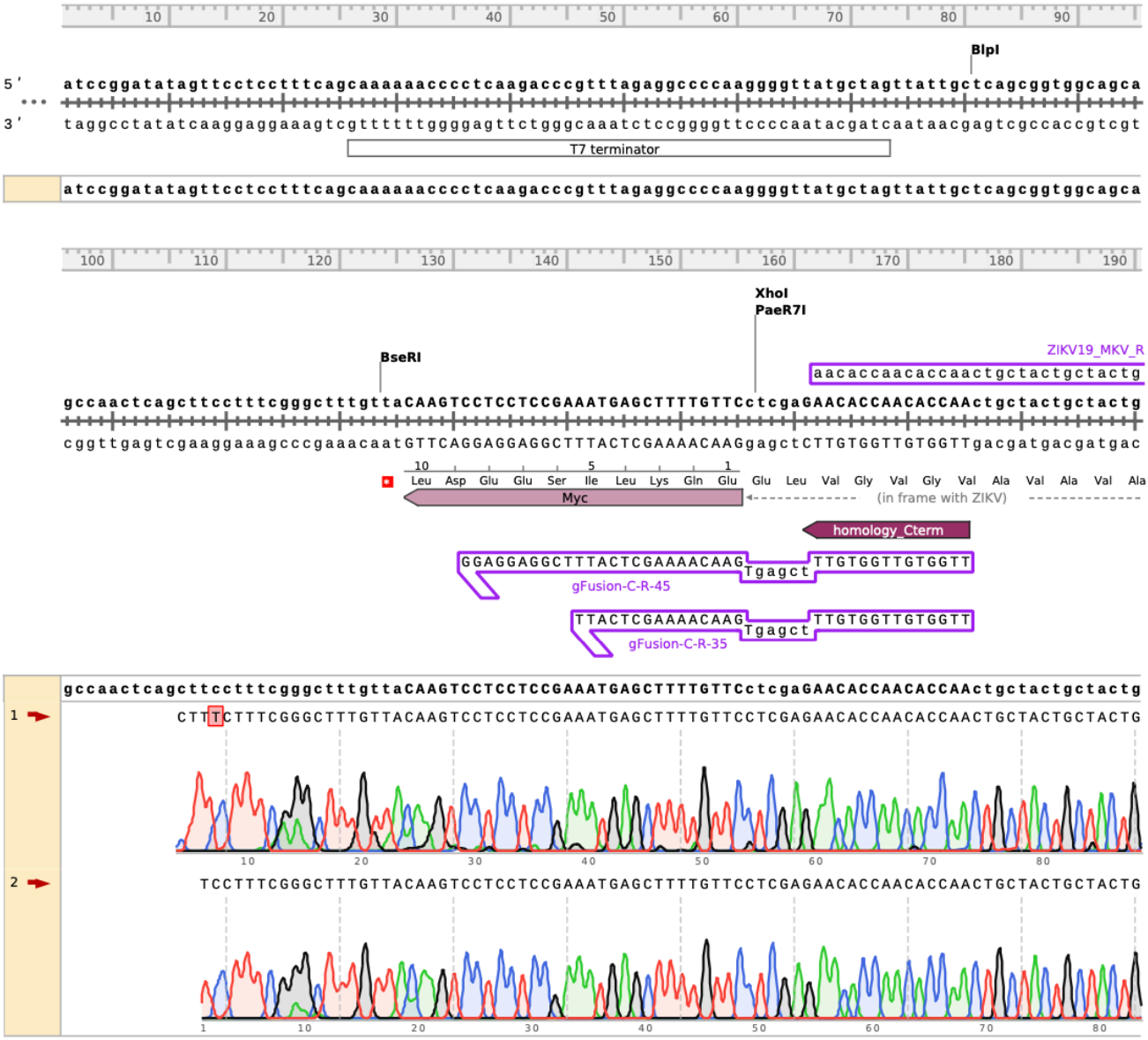


Figure S2.1

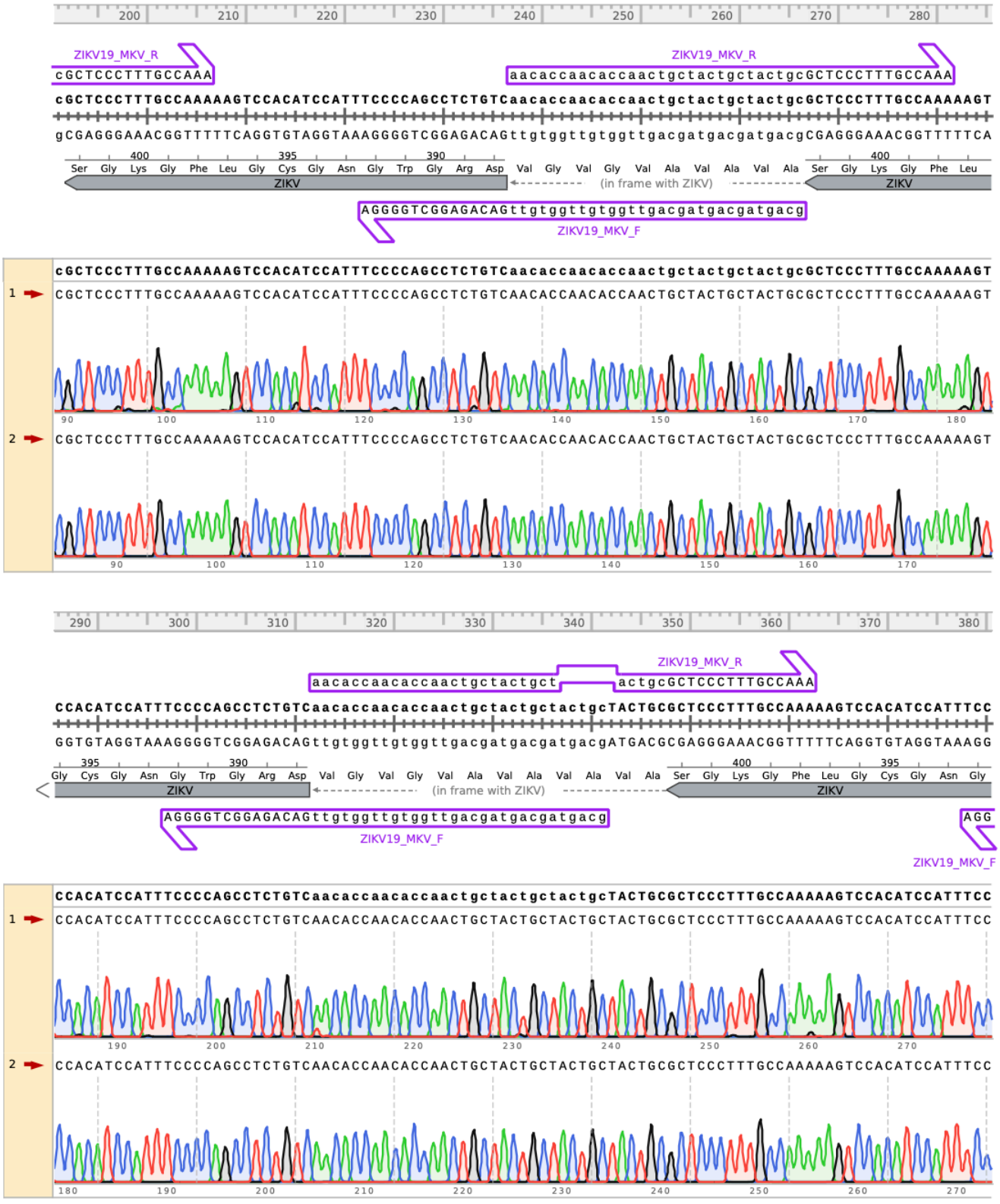


Figure S2.2

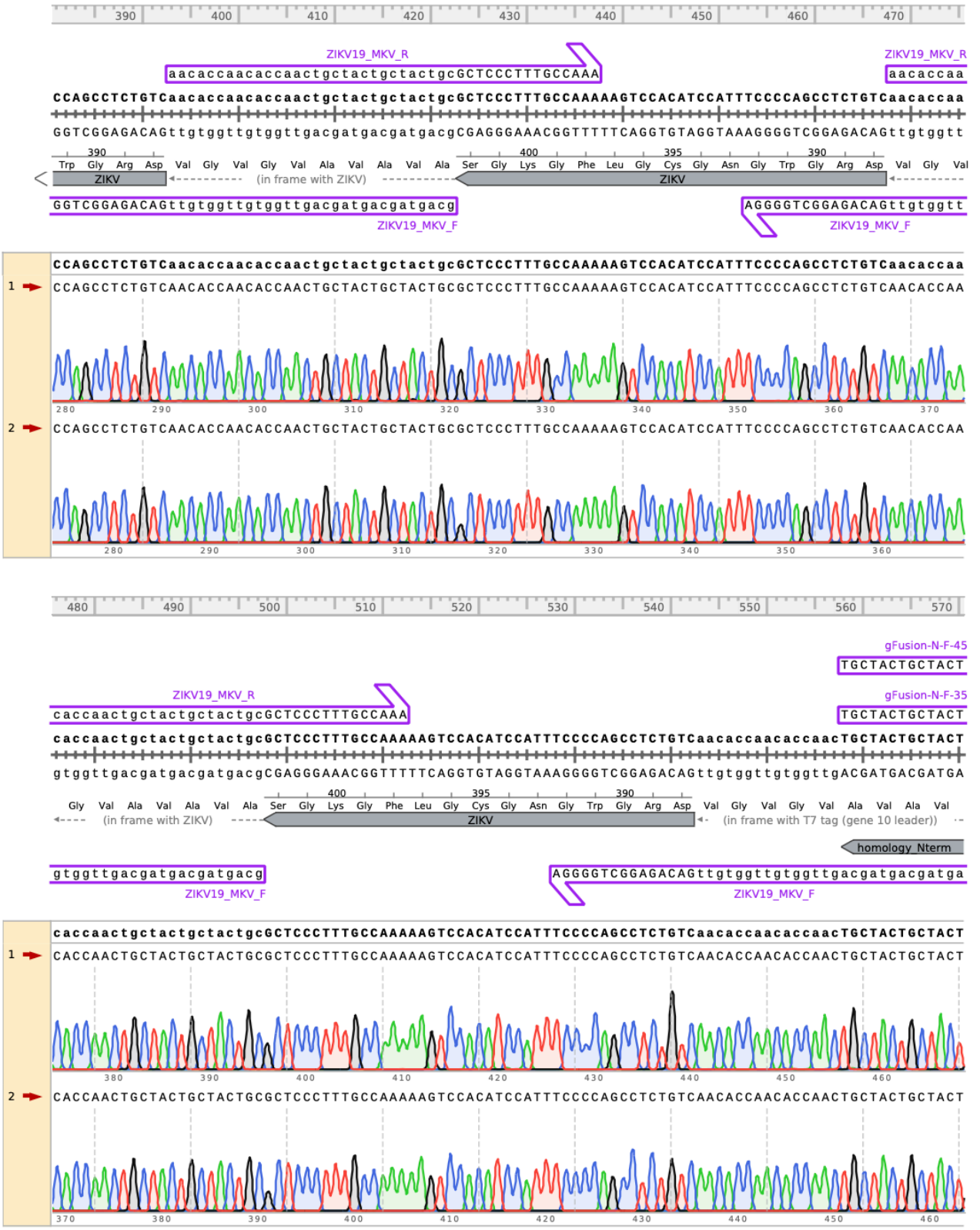

ZIKV19_MKV_R
ZIKV19_MKV_F
ZIKV
(in frame with ZIKV)
gFusion-N-F-45
gFusion-N-F-35
homology_Nterm
(in frame with T7 tag (gene 10 leader))

Figure S2.3

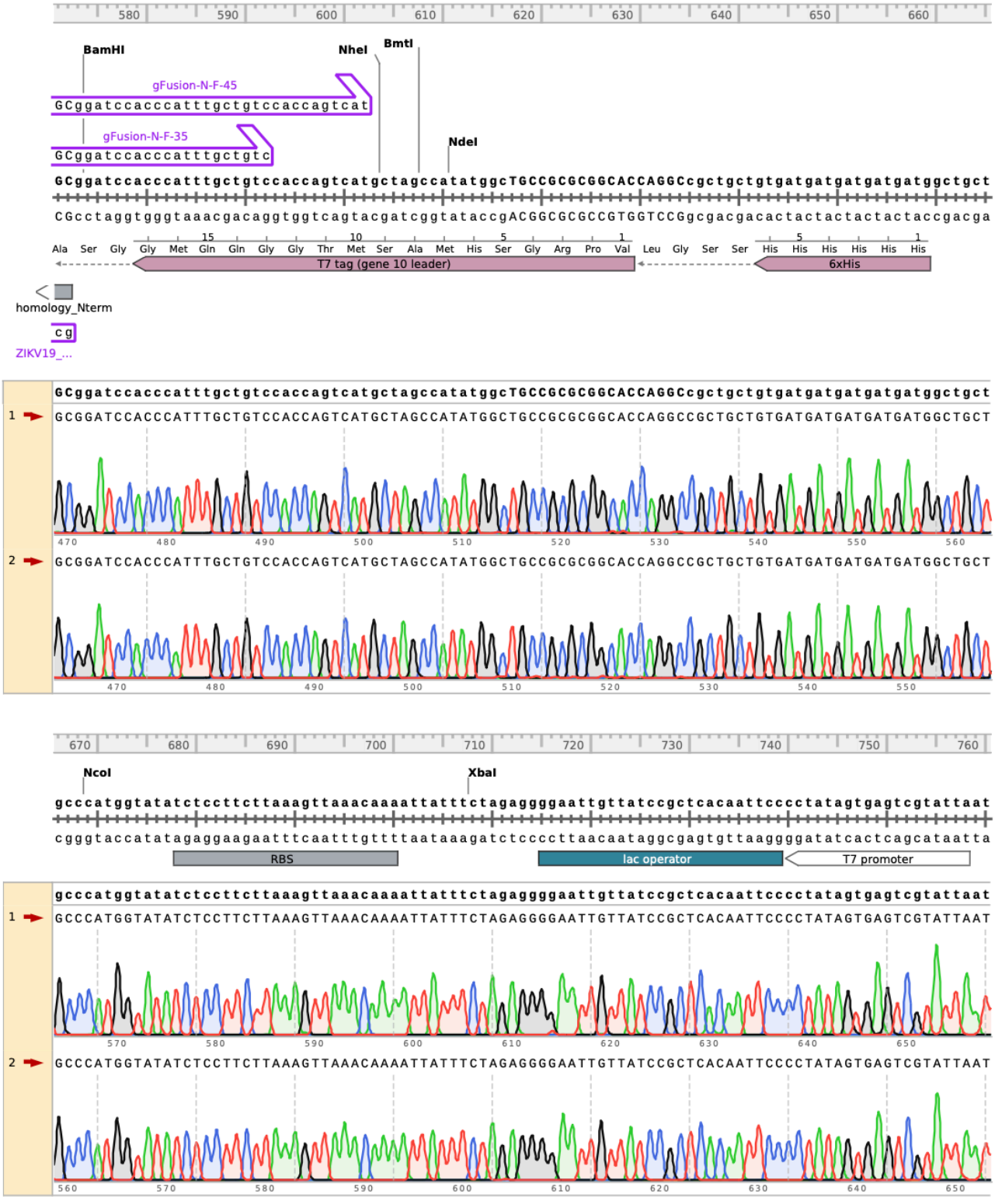
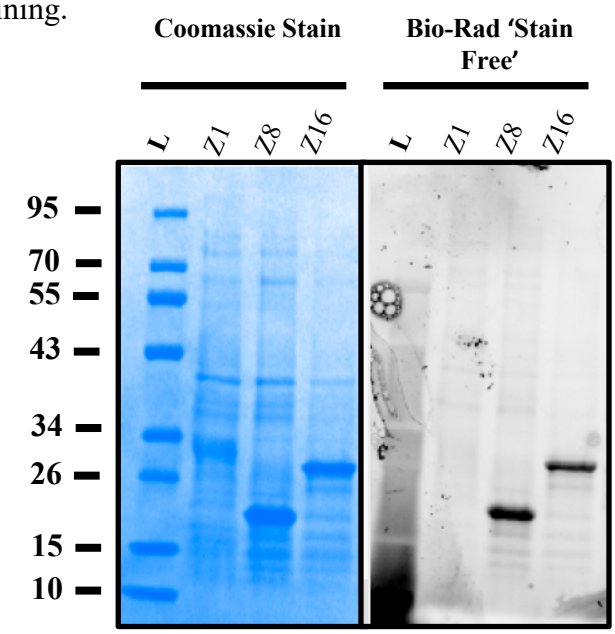

BamHI
NheI
BmtI
NdeI
gFusion-N-F-45
gFusion-N-F-35
T7 tag (gene 10 leader)
6xHis
homology_Nterm
ZIKV19_...
NcoI
XbaI
RBS
lac operator
T7 promoter

Figure S3. Comparison of SDS page total protein visualisation methodologies. The same gel was firstly images with UV (1 min activation) followed by Coomassie staining overnight followed by de-staining.



Chapter Seven

Discussion

“WHO continues to work with regional and national health authorities to enhance health system capacity to respond to the continued threat of ZIKV transmission”. A polluted reservoir in the Apipucos neighborhood of Recife. Researchers say new cases of congenital Zika syndrome are occurring and that the virus is still circulating in Brazil and beyond. Image: NYT (2022) ; WHO Zika epidemiology update (Feb 2022).

Discussion

In this thesis I have studied the ZIKV outbreak on Cabo Verde, in a multi-faceted analysis that addresses different components which describe ZIKV transmission during the 2015/16 epidemic. Together, this provides epidemiological insights that contribute to current research on ZIKV control strategy, a virus that has caused ongoing health consequences for children and is likely to re-emerge with significant morbidity. The inevitable global increase in temperature, due to climate change, will facilitate expansion of *Ae. aegypti* populations, including those with insecticide resistance phenotypes. The suitability of new habitats increasing at a rate of 4.4% per decade to 2050, up from 1.5% in the 20th century [1]. With this exists an inherent risk of increased arbovirus transmission, and current measures of control may be insufficient without additional scientific insights and resourcing.

In **Chapter 2**, the *Ae. aegypti* vector population was surveyed shortly after the ZIKV outbreak on Cabo Verde using xenomonitoring techniques and amplicon sequencing. The application of qPCR in this instance appeared to be sensitive. However, in the future, the pooling of samples and use of automated biological material extraction could be an economic and timely improvement on current practice. A combined RNA/DNA column extraction protocol was implemented, which enabled the analysis of both, RNA, for transcriptomics and virus detection, and DNA, for the study of insecticide resistance and population genetics, in a single extraction protocol. This combined approach is more economic and presents greater scope for impactful research for a single sample set. The amplicon sequencing analyses revealed susceptibility to common pyrethroid insecticides through the analysis of the *VGSC* gene, while describing the population genetics of the vector population using phylogenetic inference and haplotype mapping approaches. In the years since the sampling, there has been the first report in Cabo Verde of insecticide resistance in *Anopheles arabiensis,* a malaria vector [2]. This

report detailed resistance mutations in the *kdr* gene, which like the *VGSC* gene, confers pyrethroid resistance. These findings emphasise the requirement for sustained genomic surveillance of insecticide resistance on Cabo Verde, which should be supplemented with bioassay techniques.

In the third chapter, the molecular characterisation of sampled *Ae. aegypti* mosquitoes was reported with a focus on RNA virus detection. Using transcriptomic sequencing on a third generation nanopore platform, pooled *Ae. aegypti* total RNA isolates were characterised. This led to the assembly of an EIAV virus partial genomic sequence. To build on these findings, I would like to screen more mosquitoes from our dataset, and produce a research note on the presence of EIAV in the Cabo Verdean *Ae. Aegypti* population. The profile of the transcriptome was surprisingly complex. I found that the most frequent taxa assigned in our samples was *Homo sapiens* and *Sus scrofa* (wild boar). Considering that many of the mosquitoes collected were blood-fed, the former is expected. However, no evidence was found of residents keeping *Sus scrofa or Sus scrofa*-like species near to the collection sites, and it is possibly an example of misclassification by the Kraken metagenomic software. The detection of EIAV is more robust, especially as the sequences were compared and validated against the publicly available viral sequence data in the phylogenetic analysis.

The RNA yield from the *Ae. aegypti* RNA isolations was low. A whole-transcriptome amplification technique was employed, which *without bias,* amplified RNA transcripts at a starting concentration of ~10 ng/µL to > 1 µg/µL. In this process, random hexamers were mixed with the ZIKV specific primers, as well as the reverse-transcription mix, which contained oligo-dT primers. This kit was likely a sub-optimal choice. While many RNA virus species' transcripts are polyadenylated at the 3' UTR [3], *Flavivirus spp* are not, and feature a hairpin

loop instead. Because of this, the amplification process may have selected against ZIKV, or any other *Flavivirus* transcripts. Producing the whole-genome ZIKV sequence was therefore complex and challenging. The qPCR positive isolates were subjected to the tiling amplicon approach, but I was unable to produce any coverage above 20% of the genome and exhausted the RNA stock (after whole transcriptome amplification). In a final attempt, 95 samples with metadata indicating that the mosquitoes that were captured in a similar area, trap, or timeframe were pooled together. By changing the reverse transcriptase (NEB ProtoScript II) in the protocol, to a more *feature rich* variant, Thermo SuperScript IV, a proprietary MMLV mutant with increased processivity and greater thermostability, the results reported in **Chapter 3** were obtained. Similar issues with amplification sensitivity were encountered in the 'Zibra project', which sampled low-viraemia human isolates.

Genomic recombination in ZIKV and other *Flavivirus spp*. has been previously reported [4,5]. Phylogenetic reconstructions are built on the assumption that taxa have a single ancestry; recombinant genomes violate this. This topic is seldom discussed in *Flavivirus* genomics, possibly because the reported signals of recombination appear to be isolated to only the E and M genes, and manifest in only minor changes, given the typical clonality of co-circulating sub-lineages. In this thesis, I discuss instances where there *are* multiple, discrete phylogenetic lineages, co-circulating in a single geographic and temporal space, which challenge phylogenetic assumptions. Revisiting **Chapter 3**, a significant differential in the clock-rate of some clades on the maximum likelihood tree for the temporal phylogenetic reconstruction was observed (**Figure 9)**. While this could be a simple case of isolate mislabelling, or an undescribed shift in selective pressure, the differential could also be a signature of recombination, which should be the subject of further investigation.

The sero-epidemiological work yielded a snapshot of *Flavivirus* seroconversion within the diverse demographic setting of the Cabo Verdean population. Interestingly, one of the findings indicated that self-reporting ZIKV infection was not associated with ZIKV seroconversion. Given that, in many cases, *Flavivirus* infections do present in a flu-like manner, it is possible that infections were erroneously reported. A further line of inquiry would be to expand the antigen panel to respiratory diseases, which might answer this question. To fundamentally improve this study, I would like to have implemented PRNT assays, the gold standard in the study of *Flavivirus* humoral responses. This implementation would have enabled a robust comparison of assay specificity but was not possible due to budget constraints. The DAB assay did perform better than any of the ZIKV assays featured. Using a DAB format with DENV antigens would have provided a more robust differentiation between responses. However, given the significant cost of these assays, the work was limited to deducing specificity through the assumption of ZIKV DAB specificity. Despite reports of broader cross reactivity between *Flavivirus spp,* I observed no obvious manifestation of this on the analysis of YFV responses. The CHIKV seroconversion observed was surprising. The Cabo Verdean Ministério da Saúde does not routinely screen for this *Alphavirus* but could consider doing so. None of the participants with apparent seroconversion declared recent travel to CHIKV endemic regions. Given the presence of the *Ae. aegypti* vector, and the swift introduction of ZIKV to Cabo Verde described by the molecular analyses, which occurred shortly after the establishment of autochthonous transmission in Brazil, the potential for a CHIKV outbreak on Cabo Verde is high. This insight is an important implication from my research, and given the recent history of ZIKV, COVID-19 and other pandemics, it is important for infection control stakeholders to have readied alert and prevention systems across a range of infections and their potential vectors.

The results in **Chapter 4** demonstrated that ZIKV assays are easily confounded by cross reactive antibody responses. Accurately determining the etiological agent behind a person's ambiguous febrile presentation is of the utmost importance in ZIKV endemic regions. The morbidity of pre-natal ZIKV infection has been substantial, with long-term commitments to microcephaly cases crucial. Diagnoses facilitate decision making, through which the provision of maternal health services, pre-natal testing, and if appropriate, safe termination, counselling, and post-natal care services is possible [6]. Improving assay specificity was the primary goal of the final two chapters. Despite its partial completion, the methodological exploration of several novel avenues that could be harnessed for the development of cost-effective diagnostics were explored, with potential applications across a range of contexts in infection biology. The implementation of *in-silico* analyses, in an 'omics-like methodology, is a new approach to designing antigen panels, with which, I highlight ten peptides for further study. Applying these methodologies in the context of SARS-CoV-2 diagnostics, in the form of an interactive website, has led to a well-used resource, receiving over 1000 repeat (non-bot) users.

One of the limitations of my research is the focus on linear epitopes. It has been previously reported that of the of the total array of epitopes present on an antigen, only a fraction of them are linear, the rest being conformational or 'discontinuous' epitopes [7]. A lot of reactivity is therefore missed but mimicking the three-dimensional structure such macromolecules such that it resembles the native protein is of great difficulty, especially given the confines of our analytical pipeline. There are tools that consider 3D structure in epitope prediction, such as the ElliPro software. However, coding this, and translating findings into something applicable in the formats I have explored would be a significant undertaking. I demonstrate the rendering of custom annotations in 3D protein models as a *beta* add-on to the 'mutation tracker' of the

Immunoanalytics tool. On this foundation, I will base my future explorations in to incorporating 3D structural analyses and conformational epitopes.

The SARS-CoV-2 pandemic presented an excellent opportunity for conducting meta-analyses, especially with the research community's efforts leading to the generation of numerous datasets for both comparison and repurposing. An opportunity for future work is the integration of my meta-analytical approach with mapping data of linear epitopes from SARS-CoV-2 proteome microarrays, ELISA, and phage-display techniques. These data could provide insights into whether *in-silico* predictions translate. Relatedly, the analysis of the high-dimensional data could involve the implementation of machine learning techniques. For example, it is possible to train neural networks to predict antigen peptides, employing not only current sequence data tools, but also metrics employed in my work, to improve predictions.

The work in validating the Multimer methodology was time-consuming. The research reported in **Chapter 6** is a highly condensed report of the numerous iterations my research explored. My initial attempts at assembling tandem-repeated fragments using conventional methodologies were unsuccessful. My first iteration, which was not reported in this thesis, resembled somewhat, a modified Golden Gate genotyping array technique. Using restriction endonucleases with compatible cohesive recognition sites, sympathetic to spacer codon sequences, I devised a one-step methodology for assembling short, repeated constructs enzymatically. The attempted validation of these constructs through diagnostic PCR assays revealed the potential for the overlap assembly methodology, explored in **Chapter 6**. In characterising the reaction, I believe the MultiMer methodology will prove a significant contribution to those seeking to build tandem-repeat constructs.

While the expression of peptides in inclusion bodies (IBs) is considered in many cases, sub-optimal, here, it presents several benefits. The high yield in which they are expressed, and ease of which extraction and purification is performed from IBs streamlines the Multimer protocol. These peptides are inherently non-conformational. The treatment with chaotropic agent, not only disassembles the aggregates, but its sustained presence in downstream processes ensures no secondary structures form, right up to solid-phase antigen immobilisation. I validated the use of urea in ELISA coating processes and found no significant effect on coating efficiency. Typical polystyrene ELISA plate coating occurs through hydrophobic interactions, which are not hindered by chaotropic agents.

With the sub-optimal results exhibited when testing the Multimer antigens, similar to the lack of saturation exhibited by the Concat antigen, I tested several ways by which antigen coating and the reactivity of peptides could be improved i.e., the sensitivity of the assay. I have explored using MBP fusions, which may improve coating efficiency. Polystyrene binding motifs have been reported to increase the efficiency by which small peptides associate with plastic surfaces [8,9]. For future endeavours in improving the reactivity to Multimer peptides, it would be possible to build fusion constructs containing this motif on the N-terminus. In the eventuality that the MultiMer methodology is completed, there are several applications in downstream analyses to dissect *Flavivirus* antibody responses. Its foremost application would be in a multi-epitope mixture for *Flavivirus* diagnostic assays, with implementation in both high throughput research settings, such as that on Luminex platforms, and low cost RDT-LFA antibody assays. One of the outcomes of developing the Multimer methodology was its application in the research undertaken in **Chapter 4.** Antibody repertoires develop throughout the course of infection, by way of class switching, somatic hypermutation and affinity maturation. Responses can also be marked by the development infection pathology and host-factors, both genetic and

environmental. With a wide panel of specific antigens, I believe the dissection and identification of key responses marking significant stages of infection or host factors is possible. It has been previously demonstrated that an NS1 peptide might be an early marker for ZIKV infection in infected pregnant women [10]. Building on this, with access to paired-timepoint sera from a non-human primate (NHP) challenge study, it is possible to model differentials in reactivity to specific peptides at timepoints of infection and apply this in human epidemiological settings contexts.

Many of the *in-silico* processes applied in this thesis have been implemented using automated scripts. It would be important and useful to package these in a single, robust, public codebase for future immunoanalytic analyses. The k-mer, IEDB, prediction and structural mapping processes described in **Chapter 5** would all benefit from integration within such a single package, especially as there other no tools available with this broader functionality. The inclusion of tools for 3D structural analyses would add significant utility. I have tested how such a suite of tools might be implemented, resulting in a beta-phase tool called *Antigen Profiler* (**Figure S1)**. Many complex analyses of this sort, implemented in publications, are hard to reproduce, especially for those with less developed informatics skills. Building or extending capacity to perform such analysis, especially through openly available tools, is essential to making scientific advances in the field.

The insights reported in this thesis have contributed to the control of ZIKV and SARS-CoV-2 viruses, two greatly important EIDs of the recent decade. The UK Health Research Analysis 2018 shows that non-commercial research investment in cancer biology was significantly greater than that immunology or infection biology [11]. As pathogens emerge and re-emerge, increasingly more so in the 21st century, the development, investment in integration and

application of multi-disciplinary methodologies, such as those explored here, is central to effective control strategy. My thesis has described serological and molecular techniques that have contributed to the significant reduction and elimination of infectious diseases. As seen by the recent collective research and operational efforts during the COVID-19 pandemic, a unified strategy can reduce the global burden of infectious diseases, thereby improving human and animal health equity worldwide.

7.2 References

1. Iwamura T, Guzman-Holst A, Murray KA. Accelerating invasion potential of disease vector Aedes aegypti under climate change. Nat Commun. Nature Publishing Group; 2020;11:1–10.

2. da Cruz DL, Paiva MHS, Guedes DRD, de Souza Gomes EC, Pires SG, Gomez LF, *et al.* First report of the L1014F kdr mutation in wild populations of Anopheles arabiensis in Cabo Verde, West Africa. Parasit Vectors. BioMed Central; 2021;14:1–7.

3. Huo Y, Shen J, Wu H, Zhang C, Guo L, Yang J, *et al.* Widespread 3′-end uridylation in eukaryotic RNA viruses. Sci Rep [Internet]. 2016;6:25454. Available from: https://doi.org/10.1038/srep25454

4. Dermendjieva M, Donald C, Kohl A, Evans D. Recombination between African and Asian lineages of Zika virus in vitro and its consequences for viral phenotype. Access Microbiol. Microbiology Society; 2020;2:790.

5. Worobey M, Rambaut A, Holmes EC. Widespread intra-serotype recombination in natural populations of dengue virus. Proc Natl Acad Sci. National Acad Sciences; 1999;96:7352–7.

6. WHO -PAHO. ZIKA ETHICS CONSULTATION: ETHICS GUIDANCE ON KEY ISSUES RAISED BY THE OUTBREAK. 2016; Available from: http://iris.paho.org/xmlui/bitstream/handle/123456789/28425/PAHOKBR16002_eng.pdf

7. Van Regenmortel MH V. What is a B-cell epitope? Ep Mapp Protoc. Springer; 2009. p. 3–20.

8. Qiang X, Sun K, Xing L, Xu Y, Wang H, Zhou Z, *et al.* Discovery of a polystyrene binding peptide isolated from phage display library and its application in peptide immobilization. Sci Rep [Internet]. Nature Publishing Group; 2017 [cited 2021 Jun 14];7:1–11. Available from: www.nature.com/scientificreports

9. Kogot JM, Sarkes DA, Val-Addo I, Pellegrino PM, Stratis-Cullum DN. Increased affinity and solubility of peptides used for direct peptide ELISA on polystyrene surfaces through fusion with a polystyrene-binding peptide tag. Biotechniques [Internet]. 2012 [cited 2021 Jun 14];52. Available from: www.BioTechniques.com

10. Kam Y-W, Leite JA, Amrun SN, Lum F-M, Yee W-X, Bakar FA, *et al.* ZIKV-Specific NS1 Epitopes as Serological Markers of Acute Zika Virus Infection. J Infect Dis [Internet]. Oxford University Press; 2019 [cited 2021 Feb 10];220:203–12. Available from: https://academic.oup.com/jid/article/220/2/203/5371191

11. HRCS. UK Health Research Analysis 2018 - HRCS Online [Internet]. 2020. Available from: https://hrcsonline.net/reports/analysis-reports/uk-health-research-analysis-2018/

Figure S1. Antigen profiler beta. The website employs a window-based user interface, to make the use of tools, in combination simple. The plot in the top right was an automatically generated epitope prediction metanalyses for NS1. Users will be able to either add their own PDB file or use an existing entry, after which all analyses will be run automatically, producing similar insights to those in **Chapter 5**.