THE MOLECULAR EPIDEMIOLOGY OF HCV AND RELATED VIRUSES IN AFRICA

James C. Iles

St. Catherine's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2014

Abstract

The Molecular Epidemiology of HCV and Related Viruses in Africa

James C. Iles, St. Catherine's College, University of Oxford

Submitted for the degree of Doctor of Philosophy in Zoology, Trinity 2014

Hepatitis C virus (HCV) causes severe illness in millions of people worldwide, but the epidemic strains responsible for most infections arose within the past hundred years and represent only a small part of total HCV diversity. In this thesis I combine laboratory and computational methods to study HCV in Africa. I aim to characterize its current genetic diversity and its historical transmission prior to the global HCV epidemic.

In Chapter 2 I begin by screening samples from the Democratic Republic of the Congo (DRC) for HCV and the related human pegivirus. I find high HCV sequence diversity, including a putative new subtype, and find significantly higher HCV prevalence in those born before 1950.

Chapter 3 continues this screening, and combines the sequences obtained with those from online databases. Using molcular clock methods I estimate that genotype 4 originated in central Africa around 1733, and that multiple lineages, including subtype 4a which dominates the HCV epidemic in Egypt, have moved to north Africa since ~1850.

In Chapter 4 I analyse sequences sampled from an elderly population in Kinshasa to estimate HCV's transmission history there during the 20th century. The results indicate a rapid increase in HCV transmission between 1950 and 1970 in multiple independent lineages. Possible causes of this increase are discussed. This study population also exhibits high HCV genetic diversity, including the second genotype 7 sample discovered to date.

Finally, Chapter 5 uses a range of sequencing techniques, including RNAseq, to characterise two putative HCV recombinants from Cameroon. I confirm that both sequences are recombinants, and generate a full genome sequence for one. I also develop new tools to distinguish between dual infection and recombination in next-generation sequencing data, and discuss how recombination might affect HCV diversity and treatment.

Acknowledgements

I would like to thank the following people for their help during my DPhil work:

Oliver Pybus and Paul Klenerman for the enormous level of supervisory help and support they have provided.

Oliver Pybus has been approachable, patient, and supportive, and he has provided the push I needed to make sure I get the most out of this thesis. I can thank his tireless efforts for the number of publications this DPhil has yielded and the number of skills I have been able to learn. Oliver also made sure I had enough financial support to make it through my last year, meaning I could focus entirely on finishing my DPhil.

Paul Klenerman has always been ready to provide advice on how to proceed in the laboratory, and put me in touch with experts able to help me with whichever problem I had run into that time. Additionally, he and his research group have improved my presentation skills immensely by being the audience and asking questions on dozens of presentations over the past four years.

Abby Harrison was invaluable in getting me used to laboratory work for the first two years of the DPhil, and has always been willing to give me advice on how to plan and carry out laboratory research.

Oliver Pybus, Paul Klenerman, Jayna Raghwani, Peter Markov, Nuno Faria, Tommy Lam, Sinead Lyons, Peter Simmonds, Jacques Pépin, Gkikas Magiorkinis, Adrian Smith, Abby Harrison and Narayan Ramamurthy have all been helpful with advice, feedback, practical ideas, and a generally inspiring academic environment.

Emma Culver, Johnny Halliday and David Bonsall provided a friendly and welcoming office environment, and were a great sounding board for my ideas.

Global Viral, Metabiota, Nathan Wolfe, the Department of Zoology and St. Catherine's College have all helped with funding, equipment or by providing resources.

Heather Green, Lynne Richardson, Lorraine Hogg, Simon Ellis, Jason Hogg, Kathryn Sankey, Janet Wood, Tommaso Pizzari, Kath Wondrak and Lizzie Andrews have provided managerial or administrative assistance.

My wife Liz, my parents John and Linda, my sister Jenny, and my family-in-law Matthew, Blair and Kathy have provided boundless love and support through this thesis, and I couldn't have done it without them.

Ellie Williams, James Grover, Harry Heaton, Helen Burt, Chris Longhurst, Cecily Pearson, Peter Morgan, and the many members of OURPGsoc and the Church of St. Mary the Virgin have been great friends, wonderful company, and a pleasant community to be a part of.

Many others were helpful in unending ways, too many to list here.

Thank you all!

Table of Contents

1 INTRODUCTION

The hepatitis C virus (HCV) is a major human pathogen, and causes substantial morbidity and mortality. Globally, 130-170 million people are chronically infected with HCV, and there are 3-4 million new infections each year. Infection with the virus is generally asymptomatic or unspecific in the initial stages, but can lead to chronic hepatitis, liver cirrhosis and fibrosis in the long term (Lauer and Walker, 2001).

In the 25 years since its discovery (Choo *et al.*, 1989), most research has focused on the development of drug, vaccine or preventative strategies needed to halt the spread of HCV worldwide. This drive has largely eliminated transmission of the virus from such sources as blood transfusion and infected blood products, and now the majority of transmission comes from injecting drug users (IDU) using unsterile needles, with a small number of infections being sexually transmitted and an unknown number being caused by unsafe and unsterile injections in developing countries.

The historical and evolutionary background of the virus has been the focus of a steadily-growing body of research, but this has faced difficulties - as the virus has very general symptoms and was only discovered recently, historical records are not useful in tracking its history, and so phylogenetic methods have come to the fore in understanding the past diversity and epidemiological behaviour of HCV (e.g. Pybus *et al.,* 2003, or Markov *et al.,* 2009). The virus has its roots in Africa and Asia, but there is little known about precisely where and when it originated (Simmonds, 2013).

In this Introduction I will first describe the basic biology of HCV, and its global distribution. Next, I will examine the epidemic behaviour of the virus over the past century globally and in Africa, which is best characterised as a series of man-made, accidental epidemics that greatly amplified HCV's prevalence across a country. Third, I address the research surrounding the history of HCV prior to 1900 and its endemic

transmission, and fourth discuss the use of molecular clocks, maximum likelihood frameworks and coalescent theory in the estimation of phylogenies and epidemic histories. Finally, I give an outline of the chapters of this thesis and describe their basic findings.

1.1 THE BIOLOGY OF HCV

1.1.1 Natural History of the Virus

Although my thesis focuses on the evolutionary dynamics and spread of HCV across large populations, it is important that this work is grounded in an understanding of the different functions of the virus' genes and its behaviour inside in the host.

Following infection, the virus starts by invading its target cells, hepatocytes, and replicating extremely rapidly. Within 2-12 weeks of exposure, the disease enters a period of acute infection with very few clinical characteristics; only in rare cases are such symptoms as jaundice, malaise and nausea reported. Between 6 to 8 weeks after acute infection starts, seroconversion (the production of anti-HCV antibodies and their detection in the patient's serum) occurs as the host immune response begins. This can be detected via an enzyme immunoassay, normally containing Core protein as well as non-structural proteins 3, 4 and 5 (Ghany *et al.*, 2009).

This stage is a pivotal point in the HCV infection; the patient either experiences spontaneous clearance where the immune response completely removes the HCV infection, or progresses to chronic infection, characterised as a prolonged period in which there are no symptoms. Which outcome occurs depends on many factors, but primarily is determined by the symptoms experienced in the acute stage. In patients who experienced symptomatic hepatitis (10-15%), 25-52% clear the infection while 48-75% progress to chronic infection. In those who experience asymptomatic infection

(85-90%), 85-90% progress to chronic infection while only 10-15% clear (Maheshwari *et al.*, 2008).

The symptomless period of infection, characterized by low levels of ALT and HCV RNA in serum, has a highly variable length; one third of patients progress to serious and chronic liver disease within 20 years of infection, while another third have no progression after 30 years or longer (Poynard *et al.*, 1997).

Once in the chronic hepatitis stage, the patient has a roughly 20% chance of developing cirrhosis and, following that, a 1-4% risk of carcinoma per year (Lauer and Walker, 2001). The source of this liver damage is poorly understood; suggested causes include the cleavage of mitochondrial proteins by the viral core protein causing build-up of reactive oxygen species and oxidative stress, production of destructive proteins by hepatic stellate cells induced by viral proteins, or cellular apoptosis induced by the action of viral proteins (Li *et al.*, 2007; McCartney *et al.*, 2008; van der Poorten and George, 2008). This liver damage can be exacerbated by co-factors; alcohol abuse is the most significant of these, and can interfere with dendritic cells used in the immune response and increase the oxidative stress caused by HCV (Szabo *et al.*, 2010).

The chronic stages are the ones that cause the major burden of disease in HCV infection. Treatment regimes of ribavirin and interferon-alpha can be effective in curing the chronic infection, clearing the virus in 40-80% of cases, but this treatment is expensive and comes with many side effects (Ilyas and Vierling, 2011). This treatment regime has been augmented recently by the development of the direct-acting antiviral agents such as boceprevir and telaprevir which directly inhibit viral proteins such as the NS3/4A protease or the NS5B RNA-dependent RNA polymerase (Kiser and Flexner, 2013). These drugs have dramatically increased the proportion of patients

cleared of HCV when used in conjunction with ribavirin and/or interferon alpha, and in some cases have reported no detectable HCV RNA in 100% of patients twelve weeks after cessation of treatment (Lok *et al.,* 2012). In addition, some direct-acting antivirals such as simeprevir and sofosbuvir can be administered orally and reduce side effects by eliminating the need for interferon, making them ideal for use worldwide (Au, Destache and Vivekanandan, 2015). As these drugs directly target viral proteins, however, they may be more susceptible to escape mutations than ribavirin and interferon alpha (Gaudieri *et al.,* 2009, Sarrazin and Zeuzem, 2009).

In patients that have progressed to cirrhosis and hepatocellular carcinoma, the only treatment available is liver transplantation, although reinfection of the transplanted liver is nearly inevitable (Feray *et al.*, 1999).

1.1.2 Viral Genetics and Life Cycle

Since their discovery in the 1970s, it was clear that Hepatitis A and Hepatitis B viruses were not responsible for the entirety of hepatitis cases arising from blood transfusion. After a decade of work, the virus now known as HCV was finally formally identified in 1989, using random primers to create a cDNA library of the suspected non-A non-B hepatitis agent (Choo *et al.*, 1989; Kuo *et al.*, 1989; Lauer and Walker, 2001).

Very soon after its discovery, HCV was characterised as an enveloped positive-sense virus with an RNA genome approximately 9400bp in length. It encodes a single open reading frame (ORF), flanked by untranslated regions (UTR) at the 5' and 3' ends. Study of HCV's life cycle was complicated by the inability to find working cell culture systems or animal models, but information has been gathered using heterologous expression systems, functional cDNA clones and HCV pseudoparticles (Moradpour *et al.*, 2007). The functions of the virus' genome has been characterised over the stages of

cell infection, translation and polyprotein processing, RNA replication and virion assembly, and I will go through them in order.

Virus particles circulating outside cells associate with low-density lipoproteins, and begin invading a cell by associating with a number of cell receptors in order, starting with the low density lipoprotein receptor and glycosaminoglycans and progressing to the scavenger receptor class B type 1 and the CD81 receptor (Cocquerel *et al.*, 2006). Finally, the tight junction component claudin-1 functions as a HCV co-receptor (Evans *et al.*, 2007). These receptors channel the virus particle to a clathrin-coated pit where it is taken into the cell by endocytosis. The endosome is acidified, inducing the fusion of the HCV particle with the endosome membrane and its uncoating (Moradpour *et al.*, 2007). Following the fusion of viral and cellular membranes, the single-stranded RNA genome is released into the cell's cytoplasm.

Upon infection of a host cell the virus' positive-sense nature allows it to be translated directly, with its internal ribosome entry site (IRES) in the 5' UTR facilitating cap-independent binding with the 40S ribosome subunit (Zhang *et al.*, 1999). Once the ORF has been translated, it is proteolytically cleaved into 10 viral proteins. The first third of the ORF is cleaved by host signal peptidases and signal peptide peptidases to produce three structural proteins (the basic Core protein and glycoproteins E1 and E2) and the membrane protein p7 which seems to function as an ion channel (Pavlović *et al.*, 2003). The rest of the genome is processed by two viral enzymes (the NS2 autoprotease and the NS3-4a serine protease) to produce the non-structural proteins NS2, NS3, NS4A, NS4B, NS5A and NS5b which together manage the virus' intracellular life cycle (Lindenbach and Rice, 2005).

Once cleaved, the different proteins associate with the membrane of the perinuclear endoplasmic reticulum, with NS4B altering its structure via GTPase activity to form a membranous web (Einav *et al.*, 2004; Gosert *et al.*, 2003). Once the web has been formed the proteins form a replication complex as shown in Figure 1.1, and non-structural proteins work together to create copies of the HCV genome.

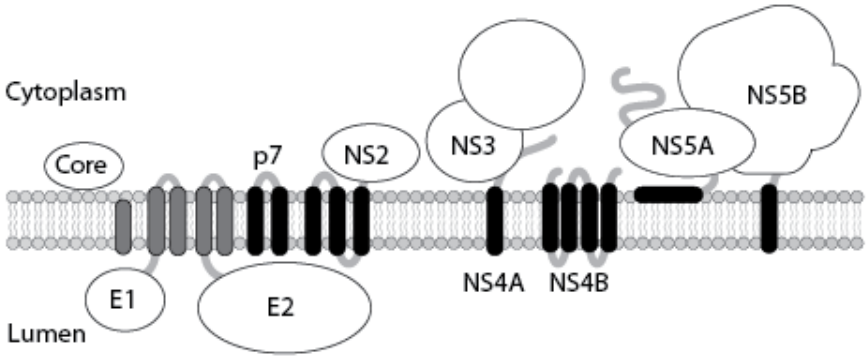


Figure 1.1. *The topology of HCV proteins with respect to the membrane of the endoplasmic reticulum (from Lindenbach and Rice, 2005).*

The exact roles each protein plays in HCV replication is uncertain, but analysis of protein structure has provided insights. NS3 protein has both a serine protease domain (required in ORF cleavage) and an RNA helicase domain that may play a role in initiating RNA synthesis on the HCV genome RNA, dissociating new RNA strands from the template during synthesis, or displacing proteins or trans-acting factors from the genome (Lindenbach and Rice, 2005). NS4A is a small protein that anchors NS3 to the membrane and enables NS3 protease function. NS5A is a large, multi-domain protein whose function is uncertain, but it may involve forming a lattice with other NS5A proteins and using its highly basic groove and acidic 'arms' to provide a transportation network for HCV RNA and protect it from host RNAses (Moradpour *et*

al., 2007). Finally, NS5b is the key enzyme for HCV replication, forming the RNA-dependant RNA polymerase (RdRp). This enzyme creates a negative-strand template from the HCV RNA genome, and then uses that negative-strand template to synthesise many copies of the positive-strand RNA genome.

Once the HCV RNA genome has been synthesised, it is translated by the ribosomes bound to the replication complex's membrane and then transported together with its associated proteins through intracellular vesicles to the Golgi body and the lipid droplet for packaging and assembly of virus particles (Liu *et al.*, 2012). Once packaged the virus particles leave the cell, potentially through the secretory pathways (Serafino *et al.*, 2003), a conclusion supported by the detection of Golgi-specific complex N-linked glycans on the surface of HCV particles found in sera (Suzuki *et al.*, 2007).

1.1.3 Viral Diversity

The process of HCV infection and replication within a host is rapid, producing an estimated 10 million or more virion particles per day (Neumann *et al.*, 1998). This replication lacks a proofreading function, resulting in the generation and evolution of a genetically-diverse population within an infected host.

HCV is highly diverse; as seen in Figure 1.2 it is classified into six distinct genotypes that differ by as much as 33% over the whole viral nucleotide sequence (with a seventh genotype currently proposed), and there are multiple subtypes within each genotype (Simmonds *et al.*, 1993). As shown in Figure 1.3, different parts of the HCV genome vary greatly in how much they are conserved; the Core region has very little variability between genotypes, whereas NS5A is highly diverse. The genome of HCV is highly ordered, forming several complex RNA structures throughout its length, and this is thought to impose a significant restriction on the potential for sequence change in some

regions (Simmonds *et al.,* 2004). This means that even synonymous mutations are very likely to have an impact on viral fitness by causing RNA structures to misfold, thus reducing the number of neutrally evolving sites in the genome compared to what may be expected in viruses without this degree of structure. Simmonds *et al.,* 2004, suggested that this genome-scale ordered RNA structure (GORS) may play a role in the modulation of innate intracellular defence mechanisms, as the existence of GORS in a viral genus correlated strongly with the ability of that genus to persist in its natural host. As discussed in Belshaw *et al.*, 2008, this may be why the substitution rate of HCV over short time scales is so different to the fixation rate in long time scales – the mutation rate remains constant, but the amount of possible change is limited and so the long term fixation rate is slowed.

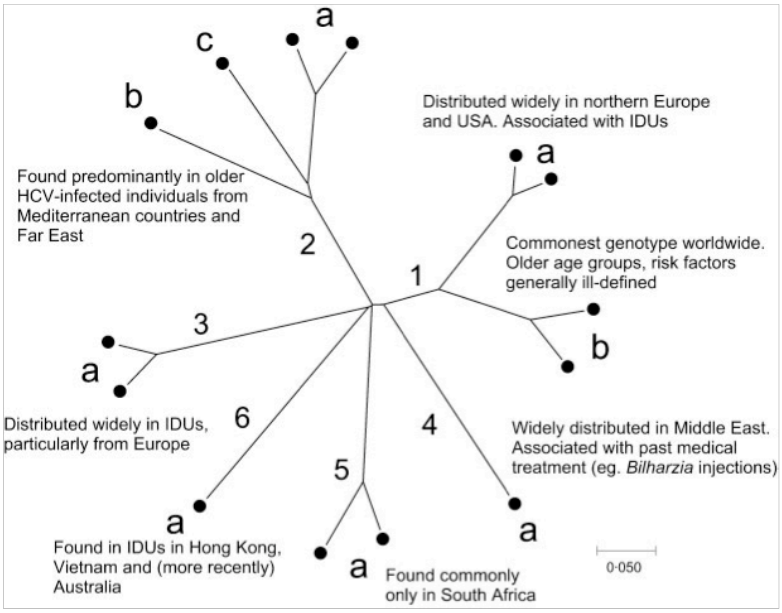


Figure 1.2. *Neighbour-Joining tree and main epidemiological associations of the principal HCV Genotypes (from Simmonds 2004).*

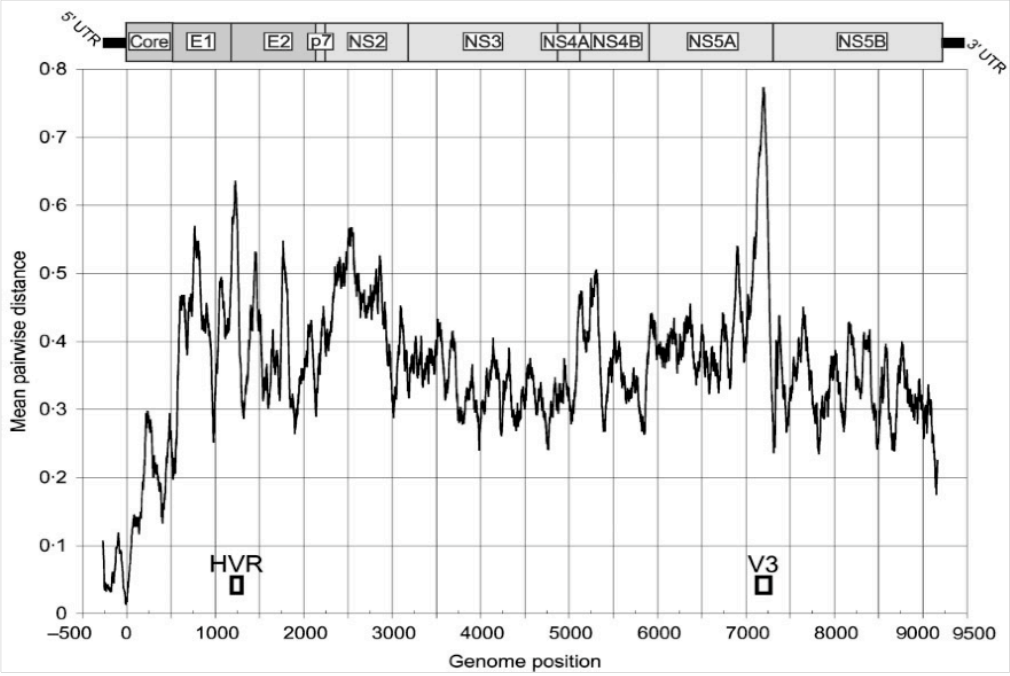


Figure 1.3. *Map of the HCV genome (top) showing positions of the structural and non-structural proteins. Plot of sequence diversity (bottom) among HCV genotypes, calculated in windows of 150nt across the genome. The 5'UTR and Core genes are highly conserved, while the envelope genes and NS5A show high variability (from Simmonds 2004).*

1.1.4 Related Viruses

While anti-HCV antibodies have been detected in a number of primate species in captivity (Korzaya *et al.,* 2002), epidemiological studies of non-human primates in geographic locations where HCV is highly endemic (such as Gabon) have so far failed to discover HCV infection despite finding high amounts of other hepatitis viruses such as HBV (Makuwa *et al.*, 2003,2006). In the absence of an established non-human reservoir, it is difficult to draw conclusions about how HCV emerged in human hosts, but investigating related viruses may provide clues. HCV is a member of the genus *Hepacivirus* in the family *Flaviviridae*, and is unique within that genus in infecting humans. For a long period of time the closest-related viruses to HCV that we knew

about were the GB-viruses, first discovered as part of the search for the non-A non-B hepatitis agent in 1967. When tamarins were experimentally inoculated with serum from an acute hepatitis sufferer, patient GB (Deinhardt *et al.*, 1967), one of the tamarins developed hepatitis, and over the next thirty years passage of the 'GB agent' through tamarins was studied and profiled. In 1995, two viruses from a tamarin in the 11th passage were identified and classified as GBV-A and GBV-B (Simons *et al.*, 1995). In the same year, two groups discovered the same virus independently; one classified it as GBV-C, and the other as HGV, but later sequence comparison showed the two samples were likely to be variants of the same species (Simons, Leary, et al. 1995; Linnen et al. 1996).

These viruses (GBV-A, B and GBV-C) all showed significant sequence similarity to HCV; GBV-B was closely related to HCV, while GBV-A and GBV-C formed a separate cluster (Kim and Fry, 1997; Leary *et al.*, 2005; Muerhoff *et al.*, 1995). As GBV-B is the virus most likely to have been the 'GB agent' but clusters separately from the rest of the GB viruses, it has recently been suggested that the GB viruses be renamed (Stapleton *et al.*, 2011). This proposal recommends that GBV-B be simply termed GBV to maintain the 'GB agent' link, and be placed within the genus *Hepacivirus*. The other GB viruses would be renamed according to their host species, and placed in a new genus, *Pegivirus*. Thus GBV-A would become Simian Pegivirus (SPgV), the human clade of GBV-C would become Human Pegivirus (HPgV), and the chimpanzee clade of GBV-C would become chimpanzee Simian Pegivirus ($SPgV_{cpz}$). The new classification system will be used in this thesis, as it provides more clarity on the different host ranges of each virus.

The tree of the Hepaciviruses has grown rapidly recently in the wake of a new species discovered in New York kennels and initially termed Canine Hepacivirus or CHV

(Kapoor *et al.*, 2011). Following further investigation, much greater diversity of this virus was found in horses, leading to the virus being renamed Non-primate Hepacivirus (NPHV) (Burbelo *et al.*, 2012). Since then, many more Hepaciviruses have been discovered in bats, horses and rodents, as have many new members of the closely-related Pegivirus genus (Epstein *et al.,* 2010; Chandriani *et al.,* 2013; Drexler *et al.,* 2013; Kapoor *et al.,* 2013a,b; Lauck *et al.,* 2013; Quan *et al.,* 2013). Figure 1.4 shows a phylogeny of these new viruses in relation to HCV, coloured according to their host species.

While HCV has a worldwide prevalence of roughly 3%, the prevalence of the GB viruses is less certain. HPgV is the most studied, since it is a human virus. Studies suggest that 1-4% of healthy blood donors in developed countries are viraemic, and another 13% have anti-HPgV antibodies (Blair *et al.*, 1998; Gutierrez *et al.*, 1997; Pilot-Matias *et al.*, 1996; Tacke *et al.*, 1997). This proportion is much higher in risk groups for sexually transmitted infections; in one study of HIV-infected homosexual men, 39.6% were viraemic and 46% had anti-HPgV antibodies (Williams *et al.*, 2004). This suggests that sexual transmission of HPgV is much more efficient than in HCV, and together with the earlier reports suggests that perhaps a quarter or more of the world's population may have been infected with HPgV at some point.

The GB viruses show significant diversity in regards to their host range. SPgV has been found in New World Primates (NWP) in the wild, and the same group of primates have been experimentally infected with GBV (Bukh and Apgar, 1997; Bukh *et al.*, 2001; Muerhoff *et al.*, 1995; Simons *et al.*, 1995).

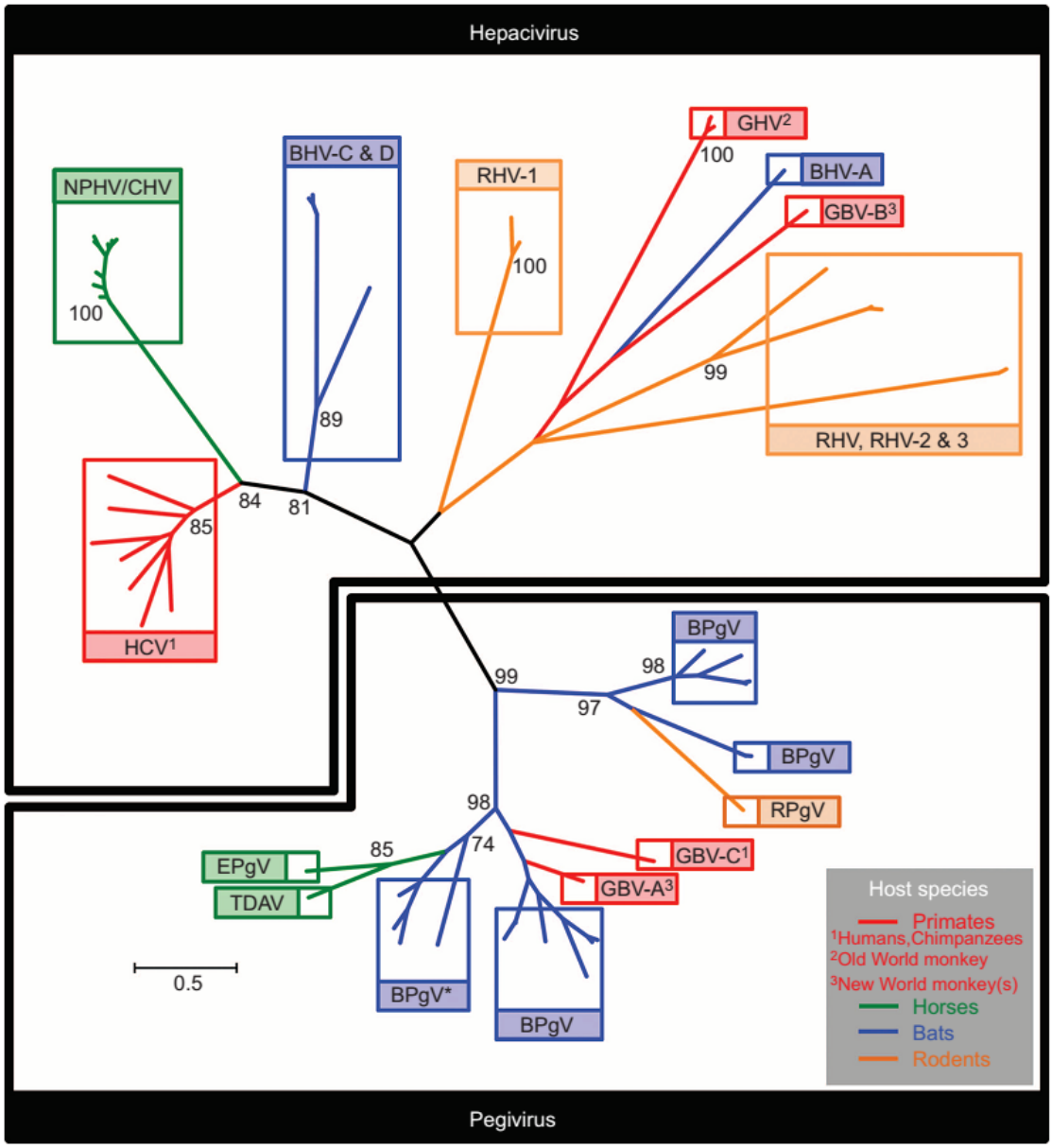


Figure 1.4. *Maximum likelihood tree of the NS3 protease domain in selected members of the hepaciviruses and pegiviruses. Nodes are labelled with bootstrap support, with only values >70% shown. Viral clades are colour-coded according to the host species discovered. Branch lengths are proportional to the scale bar and represent nucleotide substitutions per site. BHV – bat hepacivirus; BPgV – bat pegivirus (clade BPgV* contains the virus previously designated GBV-D; RCV – rodent hepacivirus; RPgV – rodent pegivirus; EPgV – equine pegivirus; TDAV – Theiler's disease-associated virus; GHV – guereza hepacivirus. Figure from Pfaender* et al.*, 2014.*

On the other hand HPgV infects humans and chimpanzees but not NWP and the isolates obtained from chimpanzees form a separate phylogenetic group to human HPgV, hence the division between HPgV and $SPgV_{cpz}$ (Adams *et al.*, 1998). BPgV has so far been found only in bats (Epstein *et al.*, 2010). In comparison, HCV naturally is found only in humans, although chimpanzees have been experimentally infected (Bukh, 2004). Most of these newly-discovered viruses have not been shown to be pathogenic, although rodents carrying RHV showed some sign of liver inflammation (Drexler *et al.,* 2013), and TDAV has been indicated as a causative agent for an outbreak of acute hepatic disease on a horse farm (Chandriani *et al.,* 2013).

The genome organisation of the GB viruses, shown in Figure 1.5, is similar to that of HCV. The HCV genome is cleaved by cellular signal peptidases and the NS3/NS4a autoprotease complex; amino acids required for these enzymes to function in HCV are highly conserved in the GB viruses, making it likely that they are cleaved in the same way. The genomes do vary in certain ways, however. SPgV and HPgV do not appear to have a Core region; they have cleavage site only 17 to 21 amino acids downstream of the putative initiation site. Despite this, biophysical characterisation of HPgV virus particles appears to show they have the nucleocapsid that is encoded by the Core protein in GBV and HCV. Several hypothesis have been proposed to explain this; the nucleocapsid could form from the very small peptide at the start of the genome, it could be translated from alternative reading frames of the HPgV genome, or it could be a repurposed host protein (Theodore and Lemon, 1998; Xiang *et al.*, 1998).

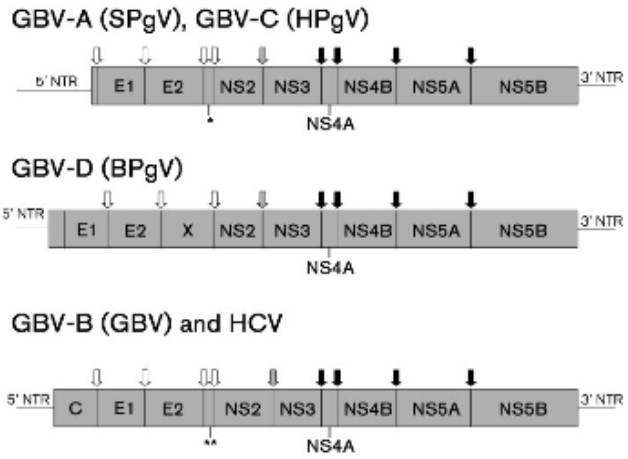


Figure 1.5. *Genome organisation of the GB Viruses and HCV. Open arrows show cleavage points for cellular signal peptidases, shaded arrows show cleavage points for the NS2-NS3 autoprotease, and block arrows show cleavage points for the NS3-NS4A protease complex (from Stapleton* et al., *2011).*

BPgV has a large protein after E2 and before NS2, provisionally termed the 'X' protein, that is not found in the other GB viruses; its function and structural role is so far unknown (Epstein *et al.*, 2010). Finally, while the 3' untranslated region of HCV and GBV contain a poly-U tail, a structure important for RNA stability and translation, the other GB viruses do not (Epstein *et al.*, 2010; Kim and Fry, 1997; Leary *et al.*, 2005; Muerhoff *et al.*, 1997). The loss of multiple features of the HCV and GBV genomes in the other GB viruses is likely a result of them belonging to different families, but it is interesting that many of the lost features are associated with human pathogenesis. The newly-discovered Hepaciviruses, on the other hand, have a largely identical genome structure to HCV, although some isolated rodent hepaciviruses may be lacking an NS4A region and the length of each gene varies between species (Pfaender *et al.,* 2014).

1.2 THE MOLECULAR EPIDEMIOLOGY OF HEPATITIS C

1.2.1 The Global State of the Virus

When analysing the global spread of the virus, I will be drawing a distinction between its behaviour in three different states: as an endemic strain, as a local epidemic strain, and as a global epidemic strain (Stumpf and Pybus, 2002). While endemic and local epidemic strains are discussed later, an overview of HCV must focus on the global epidemic strains. These strains, a small selection of subtypes (1a, 1b, 2a, 2b and 3a), are globally epidemic and are responsible for the majority of HCV infections worldwide (Simmonds, 2004). These subtypes arose comparatively recently - molecular clock methods estimate them to be roughly 100 years old (Simmonds, 2004). Epidemic subtypes show little in the way of genetic variability, but have a wide geographic dispersal. This, combined with their recent origin, implies that these strains managed to take advantage of the novel transmission routes created over the 20th century, such as intravenous drug use, blood transfusion, use of blood products etc., to greatly increase their spread worldwide (Magiorkinis *et al.*, 2009).

Geographic distribution and risk factors vary greatly among subtypes. Subtype 1a is distributed widely across northern Europe, USA and the other western countries, and is generally associated with IDUs. Subtype 1b is usually found in older age groups and more strongly associated with blood transfusion. Subtypes 2a and 2b are found in older individuals, particularly in the Mediterranean and the Far East. Subtype 3a is found globally among IDUs. Subtype 4a is found in the Middle East, particularly in Egypt, and is and associated with past medical treatment (Simmonds *et al.*, 2005).

1.2.2 Risk Factors

The Hepatitis C virus is limited in its transmission routes, requiring percutaneous transmission (i.e. from the blood of an infected individual to the blood system of the person being infected). Thus its main routes of transmission require methods that transfer blood; the unhygienic sharing of needles by injecting drug users is the major method of spread today, responsible for 68% of current infections in the US (Alter, 2002).

Blood transfusion and the use of other infected blood products obtained from infected blood are a second major causes of infection, although since the virus' discovery screening programs in the developed world have stopped infection via these routes almost entirely. In developing countries the use of unscreened blood is still a problem, however, with the WHO estimating that 43% of blood used in the developing world is not adequately screened for HCV or other blood-borne viruses (Shepard *et al.*, 2005).

Another problem in the third world is infection from unsterile medical equipment, as informal healthcare providers often lack the training or the resources to ensure proper sterilisation. It is estimated that unsafe injections are responsible for 2 million new cases of HCV per year, making this an important factor in the spread of the disease (Hauri *et al.*, 2004).

The virus can be transmitted perinatally from an infected mother to her children in an estimated 5% of cases (Shepard *et al.*, 2005), although this rate can be higher in cases of HIV/HCV co-infection (Thomas *et al.*, 1998). The virus is generally not spread sexually; over the course of a long-term monogamous partnership the virus is only transmitted in 0-0.6% of case (Terrault, 2002), although risky and promiscuous behaviour, and co-infection with ulcerative STDs, can raise the chances of infection.

1.2.3 The Past Century

Although HCV currently has a global distribution and is a major cause of disease worldwide, at the start of the 20th century it was probably much less prominent. The major routes of HCV infection in modern times, (e.g. the unhygienic use of needles, transfusion of infected blood or blood products, IDUs) have certainly increased greatly in the last century. This can be seen in the US population, where a well-established cohort of 'baby bommers', people born between 1945 and 1965, are significantly more likely to be HCV-infected – while they represent ~27% of the population, they represent ~75% of HCV infections and 73% of HCV-associated mortality in the US (Smith *et al.,* 2012). In Africa, which is the focus of the work presented in this thesis, it has been repeatedly indicated that a major cause of HCV transmission is iatrogenic transmission during mass public health campaigns (Pybus *et al.,* 2003; Njouom *et al.,* 2009; Pépin *et al.,* 2010; Njouom *et al.,* 2012). One of these campaigns has been profiled in great detail: the antischistosomal treatment campaign in Egypt during the twentieth century.

Egypt currently has one of the highest rates of prevalence of HCV in the world, with around 20% of blood donors seropositive by ELISA for HCV antibodies (Arthur *et al.*, 1997). Roughly 90% of Egyptian HCV infections belong to the 4a subtype (Ray *et al.,* 2000). Put together, the high prevalence and dominance of one subtype imply that HCV in Egypt is epidemic rather than endemic. Such an epidemic would require a highly efficient transmission method for the blood-borne virus, and one that wasn't shared with other countries. A suitable route is the extensive use of parenteral antischistosomal therapy (PAT) in a mass treatment setting, practiced between the 1920s and the 1980s and previously implicated as a risk factor for HCV antibody positivity (Darwish *et al.*, 1993). This campaign could easily have spread HCV in great

amounts; at its height, doctors were injecting about 500 people with tartar emetic in the space of three hours, taking just five seconds on average to inject each patient. The needles were sterilised by washing them through and boiling for one or two minutes before being refilled and reused, easily allowing HCV to remain and be spread to more individuals. Additionally, patients kept coming back to the clinic for multiple injections or the clinic travelled to them, further raising the chance of infection (Strickland, 2010). As support for this hypothesis, cross-sectional epidemiological analyses have shown a correlation between the level of PAT and HCV prevalence in different age groups and regions (Frank *et al.*, 2000).

Using phylodynamic approaches, which employ the coalescent framework to estimate population size history (see Section 1.4.1), previous work has reconstructed the epidemic and reached a similar result. This work found that HCV in Egypt switched from a low level of endemic infection to rapid exponential growth in the 1930s, lasting until the 1950s and resulting in a 50-fold increase in infections (Pybus *et al.*, 2003). Furthermore, the speed of this growth was far greater than that seen in subtypes spread by IDUs, such as 1a and 1b, making PAT the most probable explanation for the epidemic. A similar cohort effect was seen in Cameroon, with a HCV prevalence dramatically higher in those born before 1945 (Nerrienet *et al*., 2005). Further studies have implicated the treatment campaigns carried out between the mid-1930s and 1950s against yaws (a tropical infection of the skin, bones and joints caused by the bacteria *Treponema pallidum pertenue*) and the use of intravenous treatment for malaria as the most likely causes of this cohort effect (Pépin *et al.*, 2010a; Pépin, 2011).

1.2.4 Hepatitis C in Africa

Table 1.1 compares the seroprevalence of HCV in various different countries throughout Africa. There are some clear trends visible; firstly, wherever age

differences are noted, the older group has a higher seroprevalence. This cohort effect is likely due to treatment campaigns in the past century, as discussed in the previous section. Secondly, countries in Central Africa (DRC, Republic of the Congo, Cameroon, and Gabon) seem to have on average higher seroprevalences, compared to the much lower levels in southern Africa (South Africa, Mozambique, Tanzania). In most cases HCV RNA was not screened for and so it is uncertain how accurate these seroprevalence rates are – as seen in Njouom *et al.,* 2012, HCV RNA may be detectable in only half of the seropositive samples, raising questions about the false positivity rate of the serological testing methods. The cause of this high false positive rate deserves investigation. Egypt has very high levels of seroprevalence due to specific epidemiological factors, as discussed earlier.

Table 1.1: *HCV seroprevalence in different African countries.*

Country	Source	Anti-HCV%	Population Screened (Age)
Burkina Faso	Nagalo *et al.,* 2012	16.3%	Blood donors
Cameroon	Noubiap *et al.,* 2013	4.8%/6.9%	New blood donors (>18/>50)
	Pépin *et al.*, 2010a	55.9%	Elderly (>60 years old)
	Nerrienet *et al.*, 2005	4.5%/16.7%	SE Cameroon (>15/>50)
		4.3%/9%	West Cameroon (>15/>50)
		24.6%/47.2%	SW Cameroon(>15/>50)
		29.7%/49.5%	South Cameroon (>15/>50)
		11.6%/32.8%	Yaounde (>15/>50)
Central African	Njouom *et al.,* 2009	8.5%/12.8%/ 13.8%	Rural villagers (55-64/65-74/>75)

Republic			
Côte d'Ivoire	Combe *et al.*, 2001	3.3%	Gynecology clinics
Democratic Republic of Congo	Laurent *et al.*, 2001 Tibbs *et al.,* 1991	6.6% 4.3% 6.4%	Commercial Sex Workers Pregnant Women Blood donors
Egypt	Guerra *et al.,* 2012	14.7%/46.3%	Nationwide survey (15-59/50-59)
	Frank *et al.*, 2000	21.9% 50-55% 40-50% 35-40%	General population (<50) Lower Egypt (40-50) Middle Egypt (40-50) Upper Egypt (40-50)
Equatorial Guinea	Basaras *et al.*, 1999	1.7%/4.6%	General population (all/>40)
Gabon	Ndong-Atome *et al.*, 2008 Njouom *et al.,* 2012	1.9%/ 4.1%/ 6% 1.6%/12.4%/ 20.5%	Pregnant women (26-30/31-35/>35) Randomly selected villagers (<25/46-55/>55)
Ghana	Candotti *et al.*, 2003 Acquaye *et al.*, 2000	1.3% 5.2%	Blood Donors Blood Donors
Mozambique	Cunha *et al.,* 2007	1.1%/2.2%	Blood donors (all donors/>49)
Nigeria	Balogun *et al.,* 2012	2.1%	Blood donors.
Republic of the Congo	Sousa *et al.*, 2010	12% 3.6%	Ngala subgroup Teke subgroup

		5.6%	Kongo subgroup
		3.8%	Pygmies
South Africa	Ellis *et al.*, 1990	1.2%	Urban black blood donors
		0.8%	Urban Asian blood donors
		0.6%	Urban white blood donors
Tanzania	Tess *et al.*, 2000	1.2%	General Population

1.3 HEPATITIS C IN HISTORY

1.3.1 Endemic Spread

While subtypes 1a, 1b, 2a, 2b and 3a are responsible for the majority of epidemic HCV, and thus the majority of morbidity and mortality caused by HCV over the past century, the rest of the genotypes and subtypes of HCV show more endemic behaviour. Some endemic subtypes are restricted to a certain area, with a great deal of local genetic variation, long-term persistence and low levels of transmission, and are typically found in the tropics (Pybus *et al.*, 2007). Others show wider regional distribution; genotypes 1 and 2 are endemic in west Africa, genotype 4 in central Africa and the Middle East and genotype 6 in south-east Asia (Candotti *et al.*, 2003; Jeannel *et al.*, 1998; Mellor *et al.*, 1995; Ndjomou, 2003; Pybus *et al.*, 2008).

These distributions have been summarized by Messina *et al.,* 2014, who reviewed 1,217 studies of HCV published across the world between 1989 and 2013 to create a comprehensive picture of the distribution of different HCV genotypes across 117 countries. Their main results are shown in Figure 1.6, and highlight the difference between genotypes with globally epidemic subtypes (genotypes 1, 2 and 3) and those more restricted to certain regions (genotype 4, 5 and 6). Interestingly, the more restricted genotypes still can dominate locally – for example genotype 4 in the Middle

East and genotype 5 in South Africa. These results imply that social, behavioral and demographic factors are more important than viral genetic variation in determining the global prevalence of different genotypes, and thus that many currently-endemic strains could become globally prevalent if amplified by human factors such as injecting drug use or unsafe injections during public health campaigns.

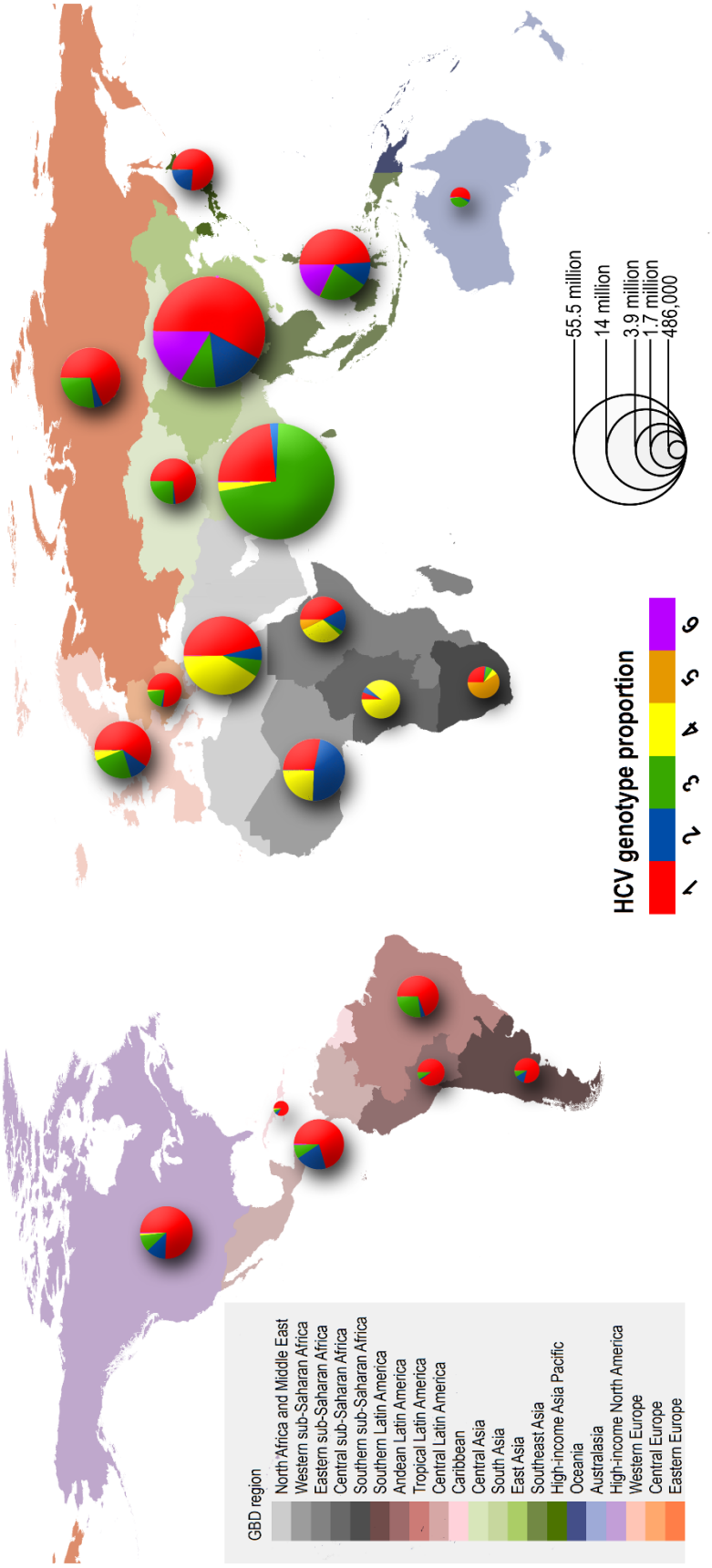

55.5 million
14 million
3.9 million
1.7 million
486,000
HCV genotype proportion
1
2
3
4
5
6
GBD region
North Africa and Middle East
Western sub-Saharan Africa
Eastern sub-Saharan Africa
Central sub-Saharan Africa
Southern sub-Saharan Africa
Southern Latin America
Andean Latin America
Tropical Latin America
Central Latin America
Caribbean
Central Asia
South Asia
East Asia
Southeast Asia
High-income Asia Pacific
Oceania
Australasia
High-income North America
Western Europe
Central Europe
Eastern Europe

Figure 1.6 (previous page). *Relative prevalence of each HCV genotype by GBD region (see legend). Size of pie chart corresponds to the number of seroprevalent cases as estimated by Mohd Hanafiah et al., 2013. Figure from Messina et al.,2014.*

Bayesian coalescent and molecular clock approaches (see Section 1.4) can be used to estimate how the virus has spread endemically over the past thousand years; for example, Markov *et al.*, (2009) analysed all known genotype 2 sequences spanning a certain region of the NS5B gene, along with 56 new sequences from Guinea Bissau. When a maximum likelihood phylogeny was estimated, it demonstrated a distinct spatial structure, with samples taken from the same location in Africa clustering together. Sequences from Guinea-Bissau, but also from Senegal, Guinea and the Gambia, were the most ancestral. A second cluster nested within this ancestral group contained sequences from Burkina Faso, Benin and Ghana, and finally a third cluster nested within the second contained sequences from Cameroon and the Central African Republic. Samples from central Africa were almost entirely located within this third group.

Markov *et al* (2009) used a NS5B substitution rate of 0.0005 per site year (Pybus *et al.*, 2001) to estimate the ages of these clusters. The most recent common ancestor (MRCA) of the African G2 strains was dated to 1470 (95% CIs: 1326-1541), this also being the age of the older Guinea-Gambia cluster. The second cluster from Benin-Ghana was dated to 1627 (1556-1680) and the third to 1637 (1611-1743).

Together, these results imply a West African origin of genotype 2, with an eastwards expansion across the continent. Other studies have suggested potential evolutionary origins for genotypes 1 and 4, which share a common ancestor. Ndjomou *et al.,* 2003, place this common ancestor in the region of Africa that is now Cameroon, and

hypothesised that two genotypes diversified there, with genotype 1 heading westwards into west Africa and genotype 4 heading eastward and northward into central Africa and the Middle East, but this has not been tested yet. Additionally, genotypes 3 and 6 seem to have been present in South Asia and South-East Asia respectively for a long time; genotype 6 seems to have been in Asia for more than a thousand years, and is highly diverse (Pybus *et al.*, 2008). There is as yet no conclusive evidence about which continent HCV began on, or indeed whether it had one origin or many.

The high diversity of endemic HCV lineages in west and central Africa suggests that the virus may have originated there, and raises the possibility that a zoonotic reservoir may be found. Although evidence for historical scenarios will always be indirect, as more viral sequences are gathered in Africa and from around the world this hypothesis can be addressed. There are obvious parallels here with HIV-1, which originated in Central Africa from a zoonotic reservoir of SIV in chimpanzees before spreading around the world in the past century (Vidal *et al.*, 2000; Keele *et al.*, 2006; Pépin, 2011). While HIV-1 spread incredibly rapidly across the world in three decades, HCV seems to have spread slowly worldwide following human migrations over thousands of years, presenting very different challenges in tracking its origins (Markov *et al*, 2012).

1.3.2 The Maintenance of the Endemic State

The finding that HCV has likely existed in human populations for millennia is problematic when the main risk factors known today are modern inventions. Parenteral drug use was invented in the past century, and it took the advent of cities and a rise in prostitution that followed to make sexual transmission of HCV at all relevant. As discussed earlier, natural methods of transmission such as from mother-to-child or between partners in a monogamous relationship are very low in frequency, likely too low to maintain the observed patterns of viral transmission. Given that the main cause

of HCV transmission seen today is the transfusion of infected blood and the use of unsterilized needles, it makes sense to consider what cultural practices or environmental factors there may have been to create similar levels of blood transmission.

Cultural dynamics may have played a role in HCV transmission. As well as transmission between sexual partners and from mother to child as discussed earlier, ritualised cultural behaviours such as tattooing, scarification, piercing and circumcision could potentially have spread the disease (Shepard *et al.*, 2005). These have the advantage that they are often preserved as traditions through time, but their variability from location to location means that they cannot have been the main factor for the maintenance of endemic HCV. Some effect on transmission is likely, however; a history of tattooing was a major risk factor for HCV in blood donors in the UK, but not ear-piercing or acupuncture (Neal *et al.*, 1994). Conversely, acupuncture was significantly associated with HCV in a Taiwanese cross-sectional seroprevalence study, but not tattooing (Sun *et al.*, 1999). Unfortunately, this only serves to underline the regional variability in these transmission methods.

Mechanical transmission by vector is another possibility. As the virus does not seem to be adapted to reproduce within an arthropod vector (Woelk and Holmes, 2002), infection would have to come from blood left on the vector's feeding apparatus. This is made plausible by the fact that all other human pathogens in the flaviviruses are spread by vector (although other flaviviruses replicate within their hosts and there is no evidence of this in HCV) and that the majority of sites where HCV has had a long history of endemic transmission are in equatorial regions - hotspots for biting arthropods that would work well as vectors. Additionally, a mathematical model estimated that only one in 16,000 host-to-host transmission events needs to transmits

the virus in order to maintain long-term endemic transmission (Pybus *et al.*, 2007). Though there has been no direct evidence of an arthropod vector involvement in HCV transmission, it is a possible alternate explanation to cultural practices for endemic transmission.

Finally, the recent discovery that bats and rodents can carry hepaciviruses closely related to HCV (see section 1.1.4) has led to speculation about the role that non-primate hosts, especially bats and rodents, might play in the origins and endemic transmission of HCV. Pybus and Gray (2013) suggest three possible scenarios as more viruses are discovered (figure 1.7): first, that all newly discovered rodent hepaciviruses are only distantly related to HCV and so its origins are unresolved; second, that newly discovered species are closer to HCV than NPHV, implying that HCV strains arose from a single ancestral transfer from rodents to humans; and third, that newly discovered species group within the known diversity of HCV, which would imply that many independent transmissions from rodents to humans had occurred.

This third possibility presents a solution to the question of how the long-term endemic state of a diverse HCV population in Africa can be maintained in the absence of an effective route of transmission, as the rodent population could act as a natural reservoir for the virus (similar interactions have been reported in the transmission of coronaviruses from bats to humans – see van Boheemen *et al.,* 2012).

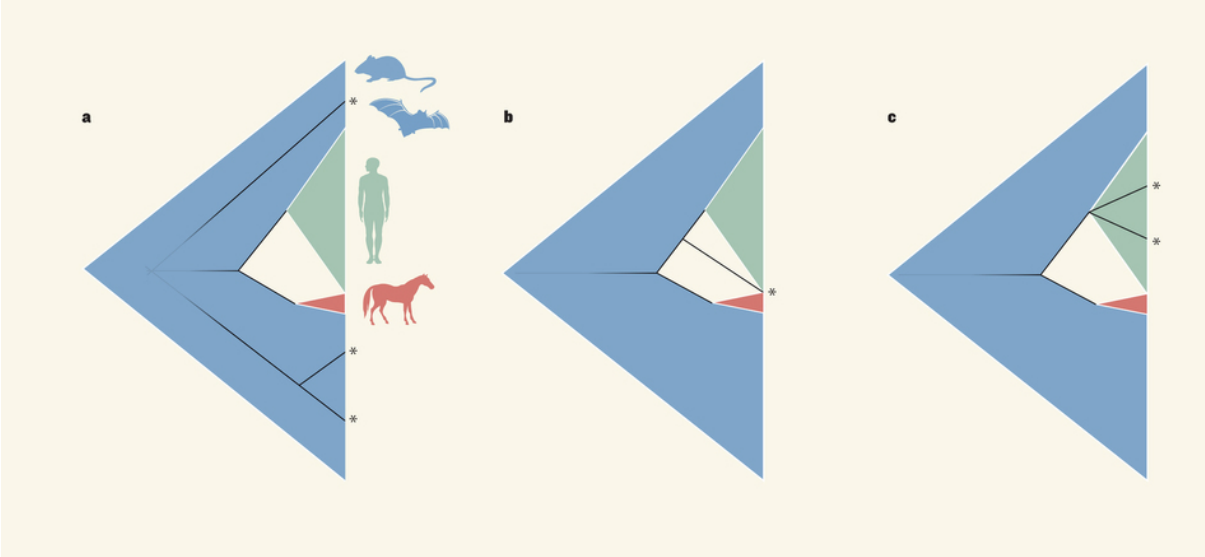


Figure 1.7*. The genetic diversity of the hepaciviruses in bats and rodents (blue), humans (green) and horses (red). Future surveys in bats, rodents and other animals may discover new hepaciviruses (asterisks) in three different locations – (a) within the existing rodent diversity, (b) more similar to HCV than equine hepaciviruses, (c) the new viruses group within the genetic diversity of HCV. Figure from Pybus and Gray, 2013.*

1.4 PHYLOGENETICS AND COALESCENT THEORY

In order to estimate a phylogenetic tree from a set of nucleotide sequences in a statistical framework (e.g. Bayesian inference or maximum-likelihood), three things are needed: a method for proposing a tree, a method for finding the conditional probability of the sequences given the tree and evolutionary model (the phylogenetic likelihood), and a method for improving or sampling the likelihood of the tree until a given threshold is reached. In this section, I will go through each of these methods in turn and explain how they will be used in this thesis.

1.4.1 The Coalescent and Genealogies

To investigate the historic behavior of HCV it is important to reconstruct the virus' population dynamics over time, and the research presented in this thesis uses analytical methods based on coalescent theory. The coalescent process, first developed in Kingman, 1982, traces back the ancestry of a group of samples (n) taken from a larger population (of size N) by representing the sample's genealogy as a series of coalescence events, at which two lineages converge, until there is only one lineage left – the MRCA of the sample. The key assumptions of the coalescent model are that n is much smaller than N, and that the shape of the phylogeny is not strongly affected by natural selection, recombination, or population subdivision. This means that the genealogy of the samples can be considered to be independent of the mutational process, and so the genealogy can be treated as a separate mathematical process.

Mathematically deriving the coalescent process starts with determining the probability that two lineages in a particular generation will coalesce in the immediately preceding generation. First, we must make some assumptions: the population is haploid, the effective population size (N_e) is constant and so each sampled lineage has N_e potential ancestors in the previous generation, the sample size n is much smaller than the population size N, and finally, that offspring are associated at random with parents in the previous generation. In these circumstances the probability that a pair of sampled lineages share the same parent, and so coalesce in the parental generation, is $1/N_e$, and inversely the probability that they do not coalesce is $1-(1/N_e)$. At every generation, then, the probability of coalescence in the whole sample is the chance a coalescence occurs among one pair of lineages multiplied by the number of possible pairs of lineages: $\frac{\binom{n}{2}}{Ne}$

Under this model, the waiting times between coalescent events in the sample is geometrically distributed, such that the probability of coalescence *g* generations prior to the present is given by:

$$P(coalescence\ at\ generation\ g) = \left(1 - \frac{1}{Ne}\right)^{g-1} \left(\frac{\binom{n}{2}}{Ne}\right)$$

As N_e becomes very large (in comparison to *n*) this can be approximated with an exponential distribution:

$$P(coalescence\ at\ generation\ g) = \binom{N}{2} \frac{1}{Ne} e^{-\frac{g-1}{Ne}}$$

This distribution has an expected value and standard deviation of N_e, meaning that while on average it will take two samples N_e generations to coalesce this time can vary greatly.

Although these equations apply to a pair of sequences, a genealogy for *n* individuals will contain *n-1* coalescences. Assuming that no coalescence events happen simultaneously, this means that there are a range of time points t_1, t_2, … t_{n-1} at which a coalescence occurs, and time periods T_1, T_2, … T_{n-1} between them. Every coalescence between two lineages is exponentially distributed as described above, meaning that T_n increases as the number of lineages shrinks. See Figure 1.6 for an example coalescent genealogy.

Under the model assumptions, the rate of coalescence is inversely proportional to N_e; in a larger population any given pair of lineages is less likely to coalesce in a set length of time than in a small population. This means that the demographic history of a population can be inferred from the shape of a generated genealogy; in the classic skyline plot (Pybus, Rambaut and Harvey, 2000) this is done by breaking down the

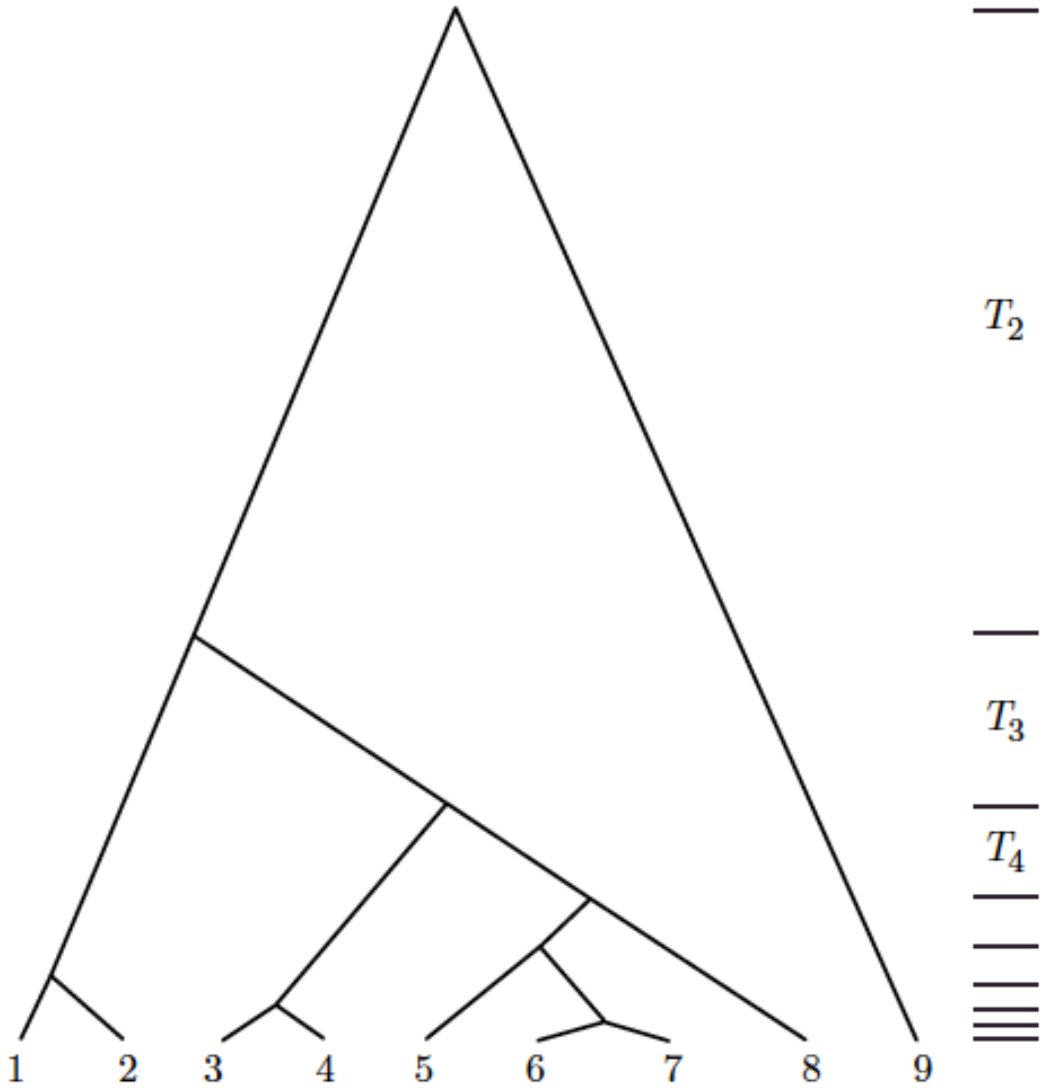


Figure 1.7. *A coalescent genealogy from nine samples. Note that as the number of lineages (i) decreases, the time between coalescent events (T_i) increases (on average). From Wakeley, 2008.*

genealogy into windows that correspond to coalescent intervals and calculating the population size within each window with the equation:

$$N_e \; in \; window = duration \; of \; window \frac{n(n-1)}{2}$$

Where *n* is the number of lineages present in the window. When the estimated N_e at each time point is plotted this produces a graph of effective population size through time, but as the duration of the window can vary greatly this can lead to a very 'noisy' plot. To counteract this, the generalised skyline plot (Strimmer and Pybus, 2001) groups adjacent windows so that only a preset and limited number of windows remain. Within each window, the population size $M_{n,k}$ (where *n* is the number of lineages at the start of the window, and *k* the number of coalescences within the window) can be found using a method of moments estimator as follows:

$$M_{n,k} = duration\ of\ window \frac{n(n-k)}{2k}$$

There are other methods of inferring N_e from a coalescent genealogy, and they are discussed further in chapter 4. With these methods, we have a means of generating a genealogy and the ability to infer useful properties from it. Next, we must assess how reliable the conclusions drawn from the genealogy are.

1.4.2 Likelihood estimation

Beyond simulated data, we do not know for certain the true genealogy of sampled nucleotide sequences. In order to discern among genealogies and choose the one that best describes the ancestral relationships among our sequences, then, we must use statistical models of the process of molecular evolution.

The first step to determine is how one sequence might mutate into another. This involves a range of factors, which are incorporated into a *nucleotide substitution model*: the frequencies of each nucleotide (A, C, T, G) in the sequences; whether there are invariant sites, and how many of them there are; whether rates of nucleotide evolution varies between sites (if it does, it is usually approximated with a gamma distribution); and whether evolutionary changes among different nucleotides are

equally frequent – often transitions (A↔G and C↔T) are more common than transversions (A↔C, A↔T, G↔C, G↔T) , reflecting the greater biochemical change of the latter (Whelan, Liò and Goldman, 2001). Nucleotide substitution models range from the very simple JC model which assumes all base frequencies and substitution rates are equal (Jukes and Cantor, 1969) to the complex GTR model that allows every nucleotide frequency and nucleotide substitution rate to differ (Tavare, 1986). A useful midpoint between these is the HKY model, which assumes that nucleotide frequencies are unequal, that all transitions happen at the same rate, and that all transversions happen at a different rate (Hasegawa, Kishino and Yano, 1985).

The next step is to consider the length of the phylogeny branches, which represent the genetic distances among samples. Coalescent trees are rooted and their branches can be measured in units of time; hence a *molecular clock model* is required to describe the relationship between genetic distance (substitutions per site), time, and evolutionary rate (denoted μ; substitutions per site per unit time). Under the simple strict molecular clock model, μ is constant across the tree. In contrast the relaxed molecular clock model gives each branch a different μ, drawn from an underlying distribution (often the lognormal distribution), with the variance in μ given by a parameter called the coefficient of variation σ (Drummond *et al.,* 2006). When using a clock model the likelihood of the branch lengths in a phylogeny is assessed by calculating the probability of the node heights (and thus divergence times) given μ (or a vector of μ values when a relaxed clock is used).

The final component of the phylogenetic likelihood is the topology of the tree, which may be proposed by the coalescent model detailed in section 1.4.1. Putting these components together, we have sequence data (*D*), a phylogeny relating the sequences

(*T*), the lengths of the phylogeny branches (*B*), and a model of the nucleotide substitutions process (*M*). The likelihood (*L*) of the variables given the data is defined as proportional to the probability of *D* given the hypothesised model (*H*), as discussed in Zwickl, 2006:

$$L \propto P(D|H) = P(D|T, M, B)$$

This value can be calculated using standard algorithms (Felsenstein, 1981). The resulting likelihood provides a metric by which trees can be statistically compared.

1.4.3 Bayesian and Maximum Likelihood inference of phylogenies

Two phylogenetic inference frameworks are used in this thesis, *maximum likelihood* and *Bayesian coalescent* inference. In maximum likelihood (ML) inference, the values of *T*, *M*, *B* that maximize the phylogenetic likelihood defined in section 1.4.2 are identified. This is typically performed by an iterative algorithm that identifies the highest likelihood possible for a given topology. The algorithm works by taking each branch and altering it until the length of the branch is that which gives the highest likelihood. The process is then repeated for all other branches so that the maximum likelihood of *B* is found for the entire tree.

In addition to this calculation, the tree topology with the highest likelihood must also be found. As a set of *n* sequences has $(2n-5)!/[(n-3)!\ 2^{n-3}]$ possible unrooted bifurcating trees, it quickly becomes impractical to exhaustively search through all possible topologies for the one with the highest likelihood. Instead, another iterative algorithm is used to sequentially propose different tree topologies, and for each topology the maximum likelihood arrangement of branch lengths is calculated, as above. The topology yielding the highest likelihood is retained, and the rearrangement process repeats until there is no further likelihood improvement.

While I will use the Maximum-Likelihood (ML) approach to estimate molecular phylogenies in this thesis, I also use Bayesian MCMC inference when molecular clock and coalescent models are involved in analysis. I use the Bayesian approach first developed by Drummond *et al.,* 2002 that is implemented in the software package BEAST, which can jointly estimate rates of evolution, effective population sizes, and the genealogy by using Markov Chain Monte Carlo (MCMC) integration. Briefly, this method randomly samples a high-dimensional posterior distribution using a Metropolis-Hastings algorithm that randomly walks through parameter space (including the space of all possible phylogeny shapes) to determine which parameter values to keep and which to reject (Metropolis *et al.*, 1953; Hastings, 1970; Drummond and Rambaut, 2007). The posterior distribution is a complex function of the sequence data, the prior probability distributions for each evolutionary parameter, and the phylogenetic, nucleotide substitution, coalescent, and molecular clock models as described in 1.4.2. Once the random walk has been repeated on enough states, the distribution of randomly sampled states should match that of the target posterior distribution, allowing the estimation of the median values and probability distributions for each evolutionary parameter. Similarly, the phylogenies sampled during Bayesian MCMC inference represent the phylogenetic information contained in the sequences and can be summarised using a "maximum clade credibility tree" that represents a best-estimate of the phylogenetic topology, its branch lengths, and the temporal position of its nodes (Heled and Bouckaert, 2013). The results obtained using Bayesian MCMC inference can depend on the prior distributions for the evolutionary parameters and therefore it is important to ensure that appropriate priors are used and that the results are robust to the choice of prior (this issue is addressed in more detail in Chapter 4).

1.5 THESIS OUTLINE

Chapter 2 details the screening of a population of 299 plasma samples from soldiers from the DRC for HCV. This work was a pilot study to demonstrate the feasibility of, and assist in the planning of, a large-scale screening effort. This screening successfully provided information on the prevalence of HCV in the population, and also indicated a cohort effect whereby individuals older than 50 were significantly more likely to be HCV positive. Conversely, HPgV did not show any cohort effect, implying that the cohort effect is specific to HCV. Since this study indicated a reasonably high prevalence for HCV in the DRC and the possibility of a cohort effect was worth investigating, I moved ahead with a full screening program using these results to modify the age distribution of samples requested.

In Chapter 3, I conduct the first comprehensive molecular epidemiological analysis of genotype 4 in Africa, with the goal of understanding the past century of HCV's history, during which it grew to global prominence, and the centuries of endemic transmission beforehand. This chapter includes the sequences resulting from the large-scale survey of HCV in our DRC cohort, as well as samples gathered from online databases. The samples from the DRC continued to show an age cohort effect, and genotype-level estimates of epidemic history obtained using coalescent analysis showed that HCV genotype 4 transmission across the whole of Africa rose in the middle of the 20th century. Additionally, the origin of genotype 4 is dated and placed in central Africa, not the Middle East, and the chapter discusses the historical factors that may have lead to the current geographical distribution of the genotype 4 subtypes.

Chapters 2 and 3 indicated a potential age cohort effect in the DRC. In chapter 4, I build upon these findings by performing an in-depth study of the epidemic history of HCV in the DRC using coalescent theory, and I directly test the hypothesis that HCV

infection rates in the DRC rose rapidly halfway through the 20th century. This chapter also compares and contrasts the use of different coalescent methods when analyzing a range of different alignment types, and describes the circumstances when one might be preferred over others.

The screening programs performed in chapters 2, 3 and 4 each showed a high genetic diversity of HCV in Central Africa, including the identification of some potentially recombinant samples that seemed to belong to different genotypes in the core and NS5B genome regions. Chapter 5 describes how I identified and characterised two putative recombinant samples originally isolated in Cameroon in order to investigate the role recombination might play in the evolution and epidemiology of HCV. In the course of this chapter I also investigated the possibilities offered by high-throughput RNA sequencing, and developed tools to understand and interpret next-generation sequencing data. Both samples were confirmed as recombinants and a full genome was obtained for one, making it the first inter-genotypic recombinant involving genotype 4 to be discovered and the first to be fully sequenced.

Finally, in chapter 6 I summarise the conclusions of this thesis and discuss the potential future research that may follow on from this work.

1.6 REFERENCES

Acquaye, J. K. and Tettey-Donkor, D. (2000). Frequency of hepatitis C virus antibodies and elevated serum alanine transaminase levels in Ghanaian blood donors. *West African Journal of Medicine* 19: 239–241.

Adams, N. J., Prescott, L. E., Jarvis, L. M., Lewis, J. C., McClure, M. O., Smith, D. B. and Simmonds, P. (1998). Detection in chimpanzees of a novel flavivirus related to GB virus-C/hepatitis G virus. *The Journal of General Virology* 79: 1871–1877.

Alcantara, L. C. J., Cassol, S., Libin, P., Deforche, K., Pybus, O. G., Van Ranst, M., Galvao-Castro, B., Vandamme, A. M. and de Oliveira, T. (2009). A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Research* 37: W634–W642.

Alter, M. J. (2002). Prevention of spread of hepatitis C. *Hepatology* 36, S93–8.

Alter, M. J., Kuhnert, W. L. and Finelli, L. (2003). Guidelines for laboratory testing and result reporting of antibody to hepatitis C virus. Centers for Disease Control and Prevention. *Morbidity and Mortality Weekly Report Recommendations and Reports* 52: 1–13.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

Arthur, R. R., Hassan, N. F., Abdallah, M. Y., El-Sharkawy, M. S., Saad, M. D., Hackbart, B. G. and Imam, I. Z. (1997). Hepatitis C antibody prevalence in blood donors in different governorates in Egypt. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 91: 271–274.

Au, T. H., Destache, C. J. and Vivekanandan, R. (2015). Hepatitis C therapy: looking toward interferon-sparing regimens. *Journal of the American Pharmacists Association* 2015:e72-e86.

Basaras, M., Santamaría, A., Sarsa, M., Gutiérrez, E., de Olano, Y. and Cisterna, R. (1999). Seroprevalence of hepatitis B and C, and human immunodeficiency type 1 viruses in a rural population from the Republic of Equatorial Guinea. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 93: 250–252.

Batina Agasa, S., Dupont, E., Kayembe, T., Molima, P., Malengela, R., Kabemba, S., Andrien, M., Lambermont, M., Cotton, F., Vertongen, F. and Gulbis B. (2010). Multiple transfusions for sickle cell disease in the Democratic Republic of Congo: the importance of the hepatitis C virus. *Transfusion Clinique et Biologique* 17: 254–259.

Berkes, J. and Cotler, S. J. (2005). Global Epidemiology of HCV Infection. *Current Hepatitis Reports* 4: 125–130.

Belshaw, R., Gardner, A., Rambaut, A. and Pybus, O. G. (2008). Pacing in a small cage: mutation and RNA viruses. *Trends in Ecology and Evolution* 23: 188-193.

Bhattarai, N. and Stapleton, J. T. (2012). GB virus C: the good boy virus? *Trends in Microbiology* 20: 124–130.

Blair, C. S., Davidson, F., Lycett, C., McDonald, D. M., Haydon, G. H., Yap, P. L., Hayes, P. C., Simmonds, P. and Gillon, J. (1998). Prevalence, Incidence, and Clinical Characteristics of Hepatitis G Virus/GB Virus C Infection in Scottish Blood Donors. *The Journal of Infectious Diseases* 178: 1779–1782.

Bukh, J. (2004). A critical role for the chimpanzee model in the study of hepatitis C. *Hepatology* 39: 1469–1475.

Bukh, J. and Apgar, C. L. (1997). Five new or recently discovered (GBV-A) virus species are indigenous to New World monkeys and may constitute a separate genus of the Flaviviridae. *Virology* 229: 429–436.

Bukh, J., Apgar, C. L., Govindarajan, S. and Purcell, R. H. (2001). Host range studies of GB virus-B hepatitis agent, the closest relative of hepatitis C virus, in New World monkeys and chimpanzees. *Journal of Medical Virology* 65: 694–697.

Bukh, J. (2011). Hepatitis C homolog in dogs with respiratory illness. *PNAS* 108: 12563–12564.

Burbelo, P. D., Dubovi, E. J., Simmonds, P., Medina, J. L., Henriquez, J. A., Mishra, N., Wagner, J., Tokarz, R., Cullen, J. M., and other authors. (2012). Serology-Enabled Discovery of Genetically Diverse Hepaciviruses in a New Host. *Journal of Virology* 86: 6171–6178.

Candotti, D., Temple, J., Sarkodie, F. and Allain, J. P. (2003). Frequent recovery and broad genotype 2 diversity characterize hepatitis C virus infection in Ghana, West Africa. *Journal of Virology* 77: 7914–7923.

Cantaloube, J.-F., Gallian, P., Bokilo, A., Jordier, F., Biagini, P., Attoui, H., Chiaroni, J. and de Micco, P. (2010). Analysis of hepatitis C virus strains circulating in Republic of the Congo. *Journal of Medical Virology* 82: 562–567.

Chandriani, S., Skewes-Cox, P., Zhong, W., Ganem, D. E., Divers, T. J., Van Blaricum, A. J., Tennant, B. C. and Kistler, A. L. (2013). Identification of a previously undescribed divergent virus from the Flaviviridae family in an outbreak of equine serum hepatitis. *Proc Natl Acad Sci USA* 110: E1407–E1415.

Choo, Q. L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W. and Houghton, M. (1989). Isolation of a cDNA Clone Derived from a Blood-Borne Non-A, Non-B Viral Hepititis Genome. *Science* 244: 359–362.

Cocquerel, L., Voisset, C. and Dubuisson, J. (2006). Hepatitis C virus entry: potential receptors and their biological functions. *The Journal of General Virology* 87: 1075–1084.

Colina, R., Casane, D., Vasquez, S., García-Aguirre, L., Chunga, A., Romero, H., Khan, B. and Cristina, J. (2004). Evidence of intratypic recombination in natural populations of hepatitis C virus. *Journal of General Virology* 85: 31–37.

Combe, P., La Ruche, G., Bonard, D., Ouassa, T., Faye-Kett e, H., Sylla-Koko, F. and Dabis, F. (2001). Hepatitis B and C Infections, Human Immunodeficiency Virus and Other Sexually Transmitted Infections Among Women of Childbearing Age in Côte D'ivoire, West Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 95: 493–496.

Cristina, J. and Colina, R. (2006). Evidence of structural genomic region recombination in Hepatitis C virus. *Virology Journal* 3: 1–8.

Cunha, L., Plouzeau, C., Ingrand, P., Gudo, J. P. S., Ingrand, I., Mondlane, J., Beauchant, M. and Agius, G. (2007). Use of replacement blood donors to study the epidemiology of major blood-borne viruses in the general population of Maputo, Mozambique. *Journal of Medical Virology* 79: 1832-1840.

Darwish, M. A., Raouf, T. A., Rushdy, P., Constantine, N. T., Rao, M. R. and Edelman, R. (1993). Risk factors associated with a high seroprevalence of hepatitis C virus infection in Egyptian blood donors. *The American Journal of Tropical Medicine and Hygiene* 49: 440–447.

Deinhardt, F., Holmes, A. W., Capps, R. B. and Popper, H. (1967). Studies On The Transmission Of Human Viral Hepatitis to Marmoset Monkeys. *Journal of Experimental Medicine* 125: 673–688.

Delaporte, E., Thiers, V., Dazza, M. C., Romeo, R., Mlika-Cabanne, N., Aptel, I., Schrijvers, D., Bréchet, C. and Larouzé, B. (1993). High level of hepatitis C endemicity in Gabon, equatorial Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 87: 636–637.

Djoko, C. F., Rimoin, A. W., Vidal, N., Tamoufe, U., Wolfe, N. D., Butel, C., LeBreton, M., Tshala, F. M., Kayembe, P. K., Muyembe, J.-J., Edidi-Basepeo, S., Pike, B. L., Fair, J. N., Mbacham, W. F., Saylors, K. E., Mpoudi-Ngole, E., Delaporte, E., Grillo, M. and Peeters, M. (2011). High HIV Type 1 Group M polDiversity and Low Rate of Antiretroviral Resistance Mutations Among the Uniformed Services in Kinshasa, Democratic Republic of the Congo. *AIDS Research and Human Retroviruses* 27: 323–329.

Drexler, J. F., Corman, V. M., Muller, M. A., Lukashev, A. N., Gmyl, A., Coutard, B., Adam, A., Ritz, D., Leijten, L. M., van Riel, D., Kallies, R., Klose, S. M., Gloza-Rausch, F., Binger, T., Annan, A., Adu-Sarkodie, Y., Oppong, S., Bourgarel, M., Rupp, D., Hoffmann, B., Schlegel, M., Kümmerer, B. M., Krüger, D. H., Schmidt-Chanasit, J., Setién, A. A., Cottontail, V. M., Hemachudha, T., Wacharapluesadee, S., Osterrieder, K., Bartenschlager, R., Matthee, S., Beer, M., Kuiken, T., Reusken, C., Leroy, E. M., Ulrich, R. G., Drosten, C. (2013). Evidence for novel hepaciviruses in rodents. *PLoS Pathogens* 9: e1003438.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161: 1307-1320.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J. and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Bio* 4: e88.

Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7: 214.

Drummond, A. J., Suchard, M. A., Xie, D. and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969–1973.

Einav, S., Elazar, M., Danieli, T. and Glenn, J. S. (2004). A nucleotide binding motif in hepatitis C virus (HCV) NS4B mediates HCV RNA replication. *Journal of Virology* 78: 11288–11295.

Ellis, L. A., Brown, D., Conradie, J. D., Paterson, A., Sher, R., Millo, J., Theodossiadou, E. and Dusheiko, G. M. (1990). Prevalence of hepatitis C in South Africa: detection of anti-HCV in recent and stored serum. *Journal of Medical Virology* 32: 249–251.

Epstein, J. H., Quan, P.-L., Briese, T., Street, C., Jabado, O., Conlan, S., Ali Khan, S., Verdugo, D., Hossain, M. J., Hutchison, S. K., Egholm, M., Luby, S. P., Daszak, P. and Lipkin, W. I. (2010). Identification of GBV-D, a novel GB-like flavivirus from old world frugivorous bats (Pteropus giganteus) in Bangladesh. *PLoS Pathogens* 6: e1000972.

Evans, M. J., Hahn, von, T., Tscherne, D. M., Syder, A. J., Panis, M., Wölk, B., Hatziioannou, T., McKeating, J. A., Bieniasz, P. D. and Rice, C. M. (2007). Claudin-1 is a hepatitis C virus co-receptor required for a late step in entry. *Nature* 446: 801–805.

Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution* 35: 1229-1242.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164–166.

Féray, C., Caccamo, L., Alexander, G. J., Ducot, B., Gugenheim, J., Casanovas, T., Loinaz, C., Gigou, M., Burra, P., Barkholt, L., Esteban, R., Bizollon, T., Lerut, J., Minello-Franza, A., Bernard, P. H., Nachbaur, K., Botta-Fridlund, D., Bismuth, H., Schalm, S. W. and Samuel, D. (1999). European collaborative study on factors influencing outcome after liver transplantation for hepatitis C. European Concerted Action on Viral Hepatitis (EUROHEP) Group. *Gastroenterology* 117: 619–625.

Frank, C., Mohamed, M. K., Strickland, G. T., Lavanchy, D., Arthur, R. R., Magder, L. S., Khoby, El, T., Abdel-Wahab, Y., Aly Ohn, E. S., Anwar, W. and Sallam, I. (2000). The role of parenteral antischistosomal therapy in the spread of Hepatitis C Virus in Egypt. *Lancet* 355: 887–891.

Gallei, A., Pankraz, A., Thiel, H. J. and Becher, P. (2004). RNA recombination in vivo in the absence of viral replication. *Journal of Virology* 78: 6271–6281.

Gaudieri, S., Rauch, A., Pfafferott, K., Barnes, E., Cheng, W., McCaughan, G., Shackel, N., Jeffrey, G. P., Mollison, L., Baker, R., Furrer, H., Günthard, H. F., Freitas, E., Humphreys, I., Klenerman, P., Mallal, S., James, I., Roberts, S., Nolan, D. and Lucas, M. (2009). Hepatitis C virus drug resistance and immune-driven adaptations: relevance to new antiviral therapy. *Hepatology* 49: 1069-82.

Ghany, M. G., Strader, D. B., Thomas, D. L. and Seeff, L. B. (2009). Diagnosis, management, and treatment of hepatitis C: an update. *Hepatology* 49: 1335-1374

Gosert, R., Egger, D., Lohmann, V., Bartenschlager, R., Blum, H. E., Bienz, K. and Moradpour, D. (2003). Identification of the hepatitis C virus RNA replication complex in Huh-7 cells harboring subgenomic replicons. *Journal of Virology* 77: 5487–5492.

Gutierrez, R. A., Dawson, G. J., Knigge, M. F., Melvin, S. L., Heynen, C. A., Kyrk, C. R., Young, C. E., Carrick, R. J., Schlauder, G. G., Surowy, T. K., Dille, B. J., Coleman, P. F., Thiele, D. L., Lentino, J. R., Pachucki, C. and Mushahwar, I.K. (1997). Seroprevalence of GB virus C and persistence of RNA and antibody. *Journal of Medical Virology* 53: 167–173.

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22: 160-174.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97-109.

Hauri, A. M., Armstrong, G. L. and Hutin, Y. J. F. (2004). The global burden of disease attributable to contaminated injections given in health care settings. *International journal of STD and AIDS* 15: 7–16.

Heled, J. and Bouckaert, R. R. (2013). Looking for trees in the forest: summary tree from posterior samples. *BMC Evolutionary Biology* 13: 221.

Ilyas, J. A. and Vierling, J. M. (2011). An overview of emerging therapies for the treatment of chronic hepatitis C. *Clinics in liver disease* 15: 515–536.

Jeannel, D., Fretz, C., Traore, Y., Kohdjo, N., Bigot, A., Gamy, E. P., Jourdan, G., Kourouma, K., Maertens, G., Fumoux, F., Fournel, J. J. and Stuyver, L. (1998).

Evidence for high genetic diversity and long-term endemicity of hepatitis C virus genotypes 1 and 2 in West Africa. *Journal of Medical Virology* 55: 92–97.

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of Protein Molecules*. New York: Academic Press. pp 21-132.

Kageyama, S., Agdamag, D. M., Alesna, E. T., Leaño, P. S., Heredia, A. M. L., Abellanosa Tac An, I. P., Jereza, L. D., Tanimoto, T., Yamamura, J. and Ichimura, H. (2006). A natural inter-genotypic (2b/1b) recombinant of hepatitis C virus in the Philippines. *Journal of Medical Virology* 78: 1423–1428.

Kalinina, O., Norder, H., Mukomolov, S. and Magnius, L. O. (2002). A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg. *Journal of Virology* 76: 4034–4043.

Kane, A., Lloyd, J., Zaffran, M., Simonsen, L. and Kane, M. (1999). Transmission of hepatitis B, hepatitis C and human immunodeficiency viruses through unsafe injections in the developing world: model-based regional estimates. *Bulletin of the World Health Organization* 77: 801–807.

Kapoor, A., Simmonds, P., Gerold, G., Qaisar, N., Jain, K., Henriquez, J. A., Firth, C., Hirschberg, D. L., Rice, C. M., Shields, S. and Lipkin, W. I. (2011). Characterization of a canine homolog of hepatitis C virus. *PNAS* 108: 11608–11613.

Kapoor, A., Simmonds, P., Scheel, T. K., Hjelle, B., Cullen, J. M., Burbelo, P. D., Chauhan, L. V., Duraisamy, R., Sanchez Leon, M., Jain, K., Vandegrift, K. J., Calisher, C. H., Rice, C. M. and Lipkin, W. I. (2013a). Identification of rodent homologs of hepatitis C virus and pegiviruses. *mBio* 4: e00216-13.

Kapoor, A., Simmonds, P., Cullen, J. M., Scheel, T. K., Medina, J. L., Giannitti, F., Nishiuchi, E., Brock, K. V., Burbelo, P. D., Rice, C. M. and Lipkin, W. I. (2013). Identification of a pegivirus (GB virus-like virus) that infects horses. *Journal of Virology* 87: 7185–7190.

Keele, B. F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Mpoudi Ngole, E., Bienvenue, Y., Delaporte, E., Brookfield, J. F. Y., Sharp, P. M., Shaw, G. M., Peeters, M. and Hahn, B. H. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 313: 523-526.

Kim, J. P. and Fry, K. E. (1997). Molecular characterization of the hepatitis G virus. *Journal of Viral Hepatitis* 4: 77–79.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217: 624-626.

Kiser, J. J. and Flexner, C. (2013). Direct-acting antiviral agents for hepatitis C virus infection. *Annual Review of Pharmacology and Toxicology* 53: 427-449.

Korzaya, L. I., Lapin, B. A., Keburiya, V. V. and Chikobava, M. G. (2002). Spontaneous infection of lower primates with hepatitis C virus. *Bulletin of Experimental Biology and Medicine* 133: 178-181.

Kuiken, C., Yusim, K., Boykin, L. and Richardson, R. (2005). The Los Alamos hepatitis C sequence database. *Bioinformatics* 21: 379–384.

Kuo, G., Choo, Q. L., Alter, H. J., Gitnick, G. L., Redeker, A. G., Purcell, R. H., Miyamura, T., Dienstag, J. L., Alter, M. J. and Stevens, C. E. (1989). An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. *Science* 244: 362–364.

Lauck, M., Sibley, S. D., Lara, J., Purdy, M. A., Khudyakov, Y., Hyeroba, D., Tumukunde, A., Weny, G., Switzer, W. M., Chapman, C. A., Hughes, A. L., Friedrich, T. C., O'Connor, D. H. and Goldberg, T. L. (2013). A novel hepacivirus with an unusually long and intrinsically disordered NS5A protein in a wild Old World primate. *Journal of VIrology* 87: 8971–8981.

Lauer, G. M. and Walker, B. D. (2001). Hepatitis C virus infection. *The New England Journal of Medicine* 345: 41–52.

Laurent, C., Henzel, D., Mulanga-Kabeya, C., Maertens, G., Larouze, B. and Delaporte, E. (2001). Seroepidemiological survey of hepatitis C virus among commercial sex workers and pregnant women in Kinshasa, Democratic Republic of Congo. *International Journal of Epidemiology* 30: 872–877.

Leary, T. P., Muerhoff, A. S., Simons, J., Pilot-Matias, T., Erker, J., Chalmers, M., Schlauder, G., Dawson, G., Desai, S. and Mushahwar, I. K. (2005). Sequence and genomic organization of GBV-C: a novel member of the flaviviridae associated with human non-A-E hepatitis. *Journal of Medical Virology* 48: 60–67.

Lee, Y.-M., Lin, H.-J., Chen, Y.-J., Lee, C.-M., Wang, S.-F., Chang, K.-Y., Chen, T.-L., Liu, H.-F. and Chen, Y.-M. A. (2010). Molecular epidemiology of HCV genotypes among injection drug users in Taiwan: Full-length sequences of two new subtype 6w strains and a recombinant form 2b6w. *Journal of Medical Virology* 82: 57–68.

Legrand-Abravanel, F., Claudinon, J., Nicot, F., Dubois, M., Chapuy-Regaud, S., Sandres-Saune, K., Pasquier, C. and Izopet, J. (2007). New Natural Intergenotypic (2/5) Recombinant of Hepatitis C Virus. *Journal of Virology* 81: 4357.

Li, Y., Boehning, D. F., Qian, T., Popov, V. L. and Weinman, S. A. (2007). Hepatitis C virus core protein increases mitochondrial ROS production by stimulation of Ca2+ uniporter activity. *The FASEB Journal* 21: 2474–2485.

Lindenbach, B. D. and Rice, C. M. (2005). Unravelling hepatitis C virus replication from genome to function. *Nature* 436: 933–938.

Liu, H. M., Aizaki, H., Machida, K., Ou, J.-H. J. and Lai, M. M. C. (2012). Hepatitis C virus translation preferentially depends on active RNA replication. *PloS One* 7: e43600.

Liu, H.-F., Muyembe-Tamfum, J.-J., Dahan, K., Desmyter, J. and Goubau, P. (1999). High prevalence of GB virus C/hepatitis G virus in Kinshasa, Democratic Republic of Congo: A phylogenetic analysis. *Journal of Medical Virology* 60: 159–165.

Lok, A. S., Gardiner, D. F., Lawitz, E., Martorell, C., Everson, G. T., Ghalib, R., Reindollar, R., Rustgi, V., McPhee, F., Wind-Rotolo, M., Persson, A., Zhu, K., Dimitrova, D. I., Eley, T., Guo, T., Grasela, D. M. and Pasquinelli, C. (2012). Preliminary study of two antiviral agents for hepatitis C genotype 1. *New England Journal of Medicine* 366: 216-224.

Madhava, V., Burgess, C. and Drucker, E. (2002). Epidemiology of chronic hepatitis C virus infection in sub-Saharan Africa. *The Lancet Infectious Diseases* 2: 293–302.

Magiorkinis, G., Magiorkinis, E., Paraskevis, D., Ho, S. Y., Shapiro, B., Pybus, O. G., Allain, J. P., Hatzakis, A. (2009). The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. *PLoS Med* 6: e1000198.

Maheshwari, A., Ray, S. and Thuluvath, P. J. (2008). Acute hepatitis C. *Lancet* 372: 321–332.

Makuwa, M., Souquière, S., Telfer, P., Leroy, E., Bourry, O., Rouquet, P., Clifford, S., Wickings, E. J., Roques, P. and Simon, F. (2003). Occurerence of hepatitis viruses in wild-born non-human primates: a 3-year (1998-2001) epidemiological survey in Gabon. *Journal of Medical Primatology* 32: 307-314

Makuwa, M., Souquière, S., Telfer, P., Bourry, O., Rouquet, P., Kazanji, M., Roques, P. and Simon, F. (2006). Hepatitis viruses in non-human primates. *Journal of Medical Primatology* 35: 384–387.

Markov, P. V., Pépin, J., Frost, E., Deslandes, S., Labbé, A.-C. and Pybus, O. G. (2009). Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa. *The Journal of General Virology* 90, 2086–2096.

McCartney, E. M., Semendric, L., Helbig, K. J., Hinze, S., Jones, B., Weinman, S. A. and Beard, M. R. (2008). Alcohol Metabolism Increases the Replication of Hepatitis C Virus and Attenuates the Antiviral Action of Interferon. *The Journal of Infectious Diseases* 198: 1766–1775.

McHutchison, J. G., Gordon, S. C., Schiff, E. R., Shiffman, M. L., Lee, W. M., Rustgi, V. K., Goodman, Z. D., Ling, M. H., Cort, S. and Albrecht, J. K. (1998). Interferon alfa-2b alone or in combination with ribavirin as initial treatment for chronic hepatitis C. *The New England Journal of Medicine* 339: 1485–1492.

Mellor, J., Holmes, E. C., Jarvis, L. M., Yap, P. L. and Simmonds, P. (1995). Investigation of the pattern of hepatitis C virus sequence diversity in different geographical regions: implications for virus classification. *Journal of General Virology* 76: 2493–2507.

Messina, J. P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G. S., Pybus, O. G. and Barnes, E. (2014). Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology* doi: 10.1002/hep.27259

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. and Teller, A. H. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 21: 1087-1092.

Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics* 11: 31–46.

Mohd Hanafiah, K., Groeger, J., Flaxman, A. D. and Wiersma, S. T. (2013). Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. *Hepatology* 57: 1333-1342.

Moradpour, D., Penin, F. and Rice, C. M. (2007). Replication of hepatitis C virus. *Nature Reviews Microbiology* 5: 453–463.

Morel, V., Fournier, C., François, C., Brochot, E., Helle, F., Duverlie, G. and Castelain, S. (2011). Genetic recombination of the hepatitis C virus: clinical implications. *Journal of Viral Hepatitis* 18: 77–83.

Muerhoff, A. S., Smith, D. B., Leary, T. P., Erker, J. C., Desai, S. M. and Mushahwar, I. K. (1997). Identification of GB virus C variants by phylogenetic analysis of 5'-untranslated and coding region sequences. *Journal of Virology* 71: 6501–6508.

Muerhoff, A. S., Leary, T. P., Simons, J., Pilot-Matias, T., Dawson, G., Erker, J., Chalmers, M., Schlauder, G., Desai, S. and Mushahwar, I. K. (1995). Genomic organization of GB viruses A and B: two new members of the Flaviviridae associated with GB agent hepatitis. *Journal of Virology* 69: 5621–5630.

Murphy, D. G., Willems, B., Deschenes, M., Hilzenrat, N., Mousseau, R. and Sabbah, S. (2007a). Use of Sequence Analysis of the NS5B Region for Routine Genotyping of Hepatitis C Virus with Reference to C/E1 and 5' Untranslated Region Sequences. *Journal of Clinical Microbiology* 45: 1102–1112.

Murphy, D., Chamberland, J., Dandavino, R. and Sablon, E. (2007b). A New Genotype of Hepatitis C Virus Originating from Central Africa. *Hepatology* 46: 623A.

Nagalo, B. M., Bisseye, C., Sanou, M., Kienou, K., Nebié, Y. K., Kiba, A., Dahourou, H., Outtara, S., Nikiema, J. B., Moret, R., Zongo, J. D. and Simpore, J. (2012). Seroprevalence and incidence of transfusion-transmitted infectious diseases among blood donors from regional blood transfusion centres in Burkina Faso, West Africa. *Tropical Medicine and International Health* 17: 247-253.

Ndjomou, J. (2003). Phylogenetic analysis of hepatitis C virus isolates indicates a unique pattern of endemic infection in Cameroon. *The Journal of General Virology* 84: 2333–2341.

Ndong-Atome, G.-R., Makuwa, M., Njouom, R., Branger, M., Brun-Vézinet, F., Mahé, A., Rousset, D. and Kazanji, M. (2008). Hepatitis C virus prevalence and genetic diversity among pregnant women in Gabon, central Africa. *BMC infectious diseases* 8: 82.

Neal, K., Jones, D., Killey, D. and James, V. (1994). Risk factors for hepatitis C virus infection. A case-control study of blood donors in the Trent Region (UK). *Epidemiology and Infection* 112: 595–601.

Nerrienet, E., Pouillot, R., Lachenal, G., Njouom, R., Mfoupouendoun, J., Bilong, C., Mauclere, P., Pasquier, C. and Ayouba, A. (2005). Hepatitis C virus infection in Cameroon: A cohort-effect. *Journal of Medical Virology* 76: 208–214.

Neumann, A. U., Lam, N. P., Dahari, H., Gretch, D. R., Wiley, T. E., Layden, T. J. and Perelson, A. S. (1998). Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. *Science* 282: 103–107.

Njouom, R., Nerrienet, E., Dubois, M., Lachenal, G., Rousset, D., Vessière, A., Ayouba, A., Pasquier, C. and Pouillot, R. (2007). The hepatitis C virus epidemic in Cameroon: Genetic evidence for rapid transmission between 1920 and 1960. *Infection, Genetics and Evolution* 7: 361–367.

Njouom, R., Frost, E., Deslandes, S., Mamadou-Yaya, F., Labbé, A.-C., Pouillot, R., Mbélesso, P., Mbadingai, S., Rousset, D. and Pépin, J. (2009). Predominance of hepatitis C virus genotype 4 infection and rapid transmission between 1935 and 1965 in the Central African Republic. *Journal of General Virology* 90: 2452-2456.

Njouom, R., Caron, M., Besson, G., Ndong-Atome, G-R., Makuwa, M., Pouillot, R., Nkoghé, D., Leroy, E. and Kazanji, M. (2012). Phylogeography, risk factors and genetic history of hepatitis C virus in Gabon, central Africa. *PLoS One* 7: e42002.

Noppornpanth, S., Lien, T. X., Poovorawan, Y., Smits, S. L., Osterhaus, A. D. M. E. and Haagmans, B. L. (2006). Identification of a naturally occurring recombinant genotype 2/6 hepatitis C virus. *Journal of Virology* 80: 7569–7577.

Noubiap, J. J., Joko, W. Y., Nansseu, J. R., Teng, U. G. and Siaka, C. (2013). Seroepidemiology of human immunodeficiency virus, hepatitis B and C viruses, and syphilis infections among first-time blood donors in Edéa, Cameroon. *Int J Infect Dis* 17: e832-7.

Pavlović, D., Neville, D. C. A., Argaud, O., Blumberg, B., Dwek, R. A., Fischer, W. B. and Zitzmann, N. (2003). The hepatitis C virus p7 protein forms an ion channel that is inhibited by long-alkyl-chain iminosugar derivatives. *PNAS* 100: 6104.

Pépin, J. and Labbé, A.-C. (2008). Noble goals, unforeseen consequences: control of tropical diseases in colonial Central Africa and the iatrogenic transmission of blood-borne viruses. *Tropical Medicine and International Health* 13: 744–753.

Pépin, J., Labbé, A.-C., Mamadou-Yaya, F., Mbélesso, P., Mbadingai, S., Deslandes, S., Locas, M.-C. and Frost, E. (2010a). Iatrogenic Transmission of Human T Cell Lymphotropic Virus Type 1 and Hepatitis C Virus through Parenteral Treatment and Chemoprophylaxis of Sleeping Sickness in Colonial Equatorial Africa. *Clinical infectious Diseases* 51: 777–784.

Pépin, J., Lavoie, M., Pybus, O. G., Pouillot, R., Foupouapouognigni, Y., Rousset, D., Labbé, A.-C. and Njouom, R. (2010b). Risk Factors for Hepatitis C Virus Transmission in Colonial Cameroon. *Clinical infectious Diseases* 51: 768–776.

Pépin, J. (2011). *The Origins of AIDS*. Cambridge University Press.

Pilot-Matias, T. J., Carrick, R. J., Coleman, P. F., Leary, T. P., Surowy, T. K., Simons, J. N., Muerhoff, A. S., Buijk, S. L., Chalmers, M. L., Dawson, G. J., Desai, S. M. and Mushahwar, I. K. (1996). Expression of the GB virus C E2 glycoprotein using the Semliki Forest virus vector system and its utility as a serologic marker. *Virology* 225: 282–292.

Pondé, R. A. (2011). Hidden hazards of HCV transmission. *Medical Microbiology and Immunology* 200: 7–11.

Poynard, T., Bedossa, P. and Opolon, P. (1997). Natural history of liver fibrosis progression in patients with chronic hepatitis C. *Lancet* 349: 825–832.

Poynard, T., Marcellin, P., Lee, S. S., Niederau, C., Minuk, G. S., Ideo, G., Bain, V., Heathcote, J., Zeuzem, S., Trepo, C. and Albrecht, J. (1998). Randomised trial of

interferon alpha 2b plus ribavirin for 48 weeks or for 24 weeks versus interferon alpha 2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus. *Lancet* 352: 1426–1432.

Pybus, O. G., Rambaut, A. and Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155: 1429-1437.

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C. and Harvey, P. H. (2001). The epidemic behavior of the hepatitis C virus. *Science* 292: 2323–2325.

Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H. and Rambaut, A. (2003). The Epidemiology and Iatrogenic Transmission of Hepatitis C Virus in Egypt: A Bayesian Coalescent Approach. *Molecular Biology and Evolution* 20: 381–387.

Pybus, O. G., Markov, P. V., Wu, A. and Tatem, A. J. (2007). Investigating the endemic transmission of the hepatitis C virus. *International Journal for Parasitology* 37: 839–849.

Pybus, O. G., Barnes, E., Taggart, R., Lemey, P., Markov, P. V., Rasachak, B., Syhavong, B., Phetsouvanah, R., Sheridan, I., Humphreys, I. S., Lu, L., Newton, P. N. and Klenerman, P. (2008). Genetic History of Hepatitis C Virus in East Asia. *Journal of Virology* 83: 1071–1082.

Pybus, O. G. and Gray, R. R. (2013). The virus whose family expanded. *Nature* 498: 310-311.

Quan, P. L., Firth, C., Conte, J. M., Williams, S. H., Zambrana-Torrelio, C. M., Anthony, S. J., Ellison, J. A., Gilbert, A. T., Kuzmin, I. V., Niezgoda, M., Osinubi, M. O., Recuenco, S., Markotter, W., Breiman, R. F., Kalemba, L., Malekani, J.,

Lindblade, K. A., Rostal, M. K., Ojeda-Flores, R., Suzan, G., Davis, L. B., Blau, D. M., Ogunkoya, A. B., Alvarez Castillo, D. A., Moran, D., Ngam, S., Akaibe, D., Agwanda, B., Briese, T., Epstein, J. H., Daszak, P., Rupprecht, C. E., Holmes, E. C., Lipkin, W. I. (2013) Bats are a major natural reservoir for hepaciviruses and pegiviruses. *Proc Natl Acad Sci USA* 110: 8194–8199

Raghwani, J., Thomas, X. V., Koekkoek, S. M., Schinkel, J., Molenkamp, R., van de Laar, T. J., Takebe, Y., Tanaka, Y., Mizokami, M., Rambaut, A. and Pybus, O. G. (2012). Origin and Evolution of the Unique Hepatitis C Virus Circulating Recombinant Form 2k/1b. *Journal of Virology* 86: 2212–2220.

Ray, S. C., Arthur, R. R., Carella, A., Bukh, J. and Thomas, D. L. (2000). Genetic epidemiology of hepatitis C virus throughout Egypt. *The Journal of Infectious Diseases* 182: 698–707.

Salemi, M. and Vandamme A-M. (2002). Hepatitis C Virus Evolutionary Patterns Studied Through Analysis of Full-Genome Sequences. *Journal of Molecular Evolution* 54: 62-70.

Salminen, M. O., Carr, J. K., Burke, D. S. and McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Research and Human Retroviruses* 11: 1423–1425.

Serafino, A., Valli, M. B., Andreola, F., Crema, A., Ravagnan, G., Bertolini, L. and Carloni, G. (2003). Suggested role of the Golgi apparatus and endoplasmic reticulum for crucial sites of hepatitis C virus replication in human lymphoblastoid cells infected in vitro. *Journal of Medical Virology* 70: 31–41.

Shepard, C. W., Finelli, L. and Alter, M. J. (2005). Global epidemiology of hepatitis C virus infection. *The Lancet Infectious Diseases* 5: 558–567.

Simmonds, P., Holmes, E. C., Cha, T. A., Chan, S. W., McOmish, F., Irvine, B., Beall, E., Yap, P. L., Kolberg, J. and Urdea, M. S. (1993). Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. *The Journal of General Virology* 74: 2391–2399.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus--15 years on. *The Journal of General Virology* 85: 3173–3188.

Simmonds, P., Bukh, J., Combet, C., Deléage, G., Enomoto, N., Feinstone, S., Halfon, P., Inchauspé, G., Kuiken, C., Maertens, G., Mizokami, M., Murphy, D. G., Okamoto, H., Pawlotsky, J. M., Penin, F., Sablon, E., Shin-I, T., Stuyver, L. J., Thiel, H. J., Viazov, S., Weiner, A. J. and Widell, A. (2005). Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology* 42: 962–973.

Simmonds, P. (2013). The origin of the hepatitis C virus. *Curr Top Microbiol Immunol* 369: 1-15.

Simons, J. N., Pilot-Matias, T. J., Leary, T. P., Dawson, G. J., Desai, S. M., Schlauder, G. G., Muerhoff, A. S., Erker, J. C., Buijk, S. L. and Chalmers, M. L. (1995). Identification of two flavivirus-like genomes in the GB hepatitis agent. *PNAS* 92: 3401–3405.

Smith, D. B., Pathirana, S., Davidson, F., Lawlor, E., Power, J., Yap, P. L. and Simmonds, P. (1997). The origin of hepatitis C virus genotypes. *The Journal of General Virology* 78: 321–328.

Smith, B. D., Morgan, R. L., Beckett, G. A., Falck-Ytter, Y., Holtzman, D., Teo, C.-G., Jewett, A., Baack, B., Rein, D. B., Patel, N., Alter, M., Yartel, A. and Ward, J. W. (2012). Recommendations for the identification of chronic hepatitis C virus infection

among persons born during 1945-1965. *Morbidity and Mortality Weekly Report* 61 (RR04): 1-18.

Sousa, J. D. de, Müller, V., Lemey, P. and Vandamme, A.-M. (2010). High GUD Incidence in the Early 20th Century Created a Particularly Permissive Time Window for the Origin and Initial Spread of Epidemic HIV Strains. *PloS One* 5: e9936

Stapleton, J. T., Foung, S., Muerhoff, A. S., Bukh, J. and Simmonds, P. (2011). The GB viruses: a review and proposed classification of GBV-A, GBV-C (HGV), and GBV-D in genus Pegivirus within the family Flaviviridae. *The Journal of General Virology* 92: 233–246.

Strickland, G. T. (2010). An Epidemic of Hepatitis C Virus Infection While Treating Endemic Infectious Diseases in Equatorial Africa More than a Half Century Ago: Did It Also Jump-Start the AIDS Pandemic? *Clinical infectious Diseases* 51: 785–787.

Strimmer, K. and Pybus, O. G. (2001). Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol Biol Evol* 18: 2298-2305.

Stumpf, M. P. H. and Pybus, O. G. (2002). Genetic diversity and models of viral evolution for the hepatitis C virus. *FEMS Microbiology Letters* 214: 143-152.

Sun, C. A., Chen, H. C., Lu, C. F., You, S. L., Mau, Y. C., Ho, M. S., Lin, S. H. and Chen, C. J. (1999). Transmission of hepatitis C virus in Taiwan: prevalence and risk factors based on a nationwide survey. *Journal of Medical Virology* 59: 290–296.

Suzuki, T., Ishii, K., Aizaki, H. and Wakita, T. (2007). Hepatitis C viral life cycle. *Advanced Drug Delivery Reviews* 59: 1200–1212.

Szabo, G., Wands, J. R., Eken, A., Osna, N. A., Weinman, S. A., Machida, K. and Joe Wang, H. (2010). Alcohol and hepatitis C virus--interactions in immune dysfunctions and liver damage. *Alcohol Clin Exp Res* 34: 1675–1686.

Tacke, M., Schmolke, S., Schlueter, V., Sauleda, S., Esteban, J. I., Tanaka, E., Kiyosawa, K., Alter, H. J., Schmitt, U., Hess, G., Ofenloch-Haehnle, B. and Engel, A. M. (1997). Humoral immune response to the E2 protein of hepatitis G virus is associated with long-term recovery from infection and reveals a high frequency of hepatitis G virus exposure among healthy blood donors. *Hepatology* 26: 1626–1633.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* 17: 57-86.

Terrault, N. A. (2002). Sexual activity as a risk factor for hepatitis C. *Hepatology* 36: S99–S105.

Tess, B. H., Levin, A., Brubaker, G., Shao, J., Drummond, J. E., Alter, H. J. and O'Brien, T. R. (2000). Seroprevalence of hepatitis C virus in the general population of northwest Tanzania. *The American Journal of Tropical Medicine and Hygiene* 62: 138–141.

Theodore, D. and Lemon, S. M. (1998). GB Virus C, Hepatitis G Virus, or Human Orphan Flavivirus? *Hepatology* 25: 1285–1286.

Thomas, D. L., Villano, S. A., Riester, K. A., Hershow, R., Mofenson, L. M., Landesman, S. H., Hollinger, F. B., Davenny, K., Riley, L., Diaz, C., Tang, H. B. and Quinn, T. C. (1998). Perinatal transmission of hepatitis C virus from human immunodeficiency virus type 1- infected mothers. *The Journal of Infectious Diseases* 177: 1480–1488.

Tibbs, C. J., Palmer, S. J., Coker, R., Clark, S. K., Parsons, G. M., Hojvat, S., Peterson, D. and Banatvala, J. E. (1991). Prevalence of hepatitis C in tropical communities: The importance of confirmatory assays. *Journal of Medical Virology* 34: 143–147.

Tscherne, D. M., Evans, M. J., Hahn, von, T., Jones, C. T., Stamataki, Z., McKeating, J. A., Lindenbach, B. D. and Rice, C. M. (2007). Superinfection Exclusion in Cells Infected with Hepatitis C Virus. *Journal of Virology* 81: 3693–3703.

van Boheemen, S., de Graaf, M., Lauber, C., Bestebroer, T. M., Raj, V. S., Zaki, A. M., Osterhaus, A. D., Haagmans, B. L., Gorbalenya, A. E., Snijder, E. J. and Fouchier, R. A. (2012). Genomic characterisation of a newly discovered coronavirus assouciated with acute respiratory distress syndrome in humans. *mBio* 3: e00473-12

van der Poorten, D. and George, J. (2008). Disease-Specific Mechanisms of Fibrosis: Hepatitis C Virus and Nonalcoholic Steatohepatitis. *Clinics in liver disease* 12: 805–824.

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B. and Delaporte, E. (2000). Unprecedented Degree of Human Immunodeficiency Virus Type 1 (HIV-1) Group M Genetic Diversity in the Democratic Republic of Congo Suggests that the HIV-1 Pandemic Originated in Central Africa. *Journal of Virology* 74: 10498–10507.

Wakely, J. (2008). Coalescent Theory: An Introduction. Roberts and Company Publishers, Greenwood Village, Colorado.

Whelan, S., Liò, P. and Goldman, N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics* 17: 262-272.

Williams, C. F., Klinzman, D., Yamashita, T. E., Xiang, J., Polgreen, P. M., Rinaldo, C., Liu, C., Phair, J., Margolick, J. B., Zdunek, D., Hess, G. and Stapleton, J. T. (2004). Persistent GB virus C infection and survival in HIV-infected men. *The New England Journal of Medicine* 350: 981–990.

Woelk, C. H. and Holmes, E. C. (2002). Reduced Positive Selection in Vector-Borne RNA Viruses. *Molecular Biology and Evolution* 19: 2333–2336.

Xiang, J., Klinzman, D., McLinden, J., Schmidt, W. N., LaBrecque, D. R., Gish, R. and Stapleton, J. T. (1998). Characterization of hepatitis G virus (GB-C virus) particles: evidence for a nucleocapsid and expression of sequences upstream of the E1 protein. *Journal of Virology* 72: 2738–2744.

Zhang, J., Yamada, O., Ito, T., Akiyama, M., Hashimoto, Y., Yoshida, H., Makino, R., Masago, A., Uemura, H. and Araki, H. (1999). A single nucleotide insertion in the 5'-untranslated region of hepatitis C virus leads to enhanced cap-independent translation. *Virology* 261: 263–270.

Zwickl, D. J. (2006, April 27). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.

2 HEPATITIS C VIRUS INFECTIONS IN THE DEMOCRATIC REPUBLIC OF CONGO EXHIBIT A COHORT EFFECT

Published as:

Iles JC, Harrison GLA, Lyons S, Djoko CF, Tamoufe U, Lebreton M, Schneider BS, Fair JN, Tshala FM, Kayembe PK, Muyembe JJ, Edidi-Basepeo S, Wolfe ND, Klenerman P, Simmonds P, Pybus OG (2013) Hepatitis C virus infections in the Democratic Republic of Congo exhibit a cohort effect. Infection, Genetics and Evolution *19:386-94*

2.1 SUMMARY OF AUTHORSHIP

JI wrote this chapter, performed laboratory work testing and sequencing samples for HPgV, conducted phylogenetic analysis, analysed the demographic data of the samples and created all figures in this chapter. HCV serology, RT-PCR and sequencing was performed by GLAH. SL assisted with the HPgV testing and sequencing. PK, PS and OGP provided supervisory support. All others were involved in the collection and provision of the blood samples analysed.

2.2 ABSTRACT

The prevalence and genetic diversity of hepatitis C virus (HCV) and human pegivirus (HPgV) in many regions of sub-Saharan Africa is poorly characterized, including in the Democratic Republic of Congo – the largest country in the region and one of the most populous. To address this situation we conducted a molecular epidemiological survey of HCV and HPgV (previously named GB Virus C or hepatitis G virus) in samples collected in 2007 from 299 males from the DRC, whose ages ranged from 21 to 71 years old. Samples were tested for the presence of HCV antibodies by ELISA and reactive samples were subsequently tested for HCV RNA using RT-PCR in which both the HCV Core and NS5B genome regions were amplified. Remaining samples were tested for HPgV RNA and the HPgV NS3 genome region of positive samples was amplified. For HCV, 13.7% of the samples were seropositive (41/299) but only 3.7% contained detectable HCV RNA (11/299). HPgV RNA was found in 12.7% (33/259) of samples. HCV viremia was strongly associated with age; the percentage of samples that contained detectable HCV RNA was 0.5% in those younger than 50 and 13% in those older than 50. Our study represents the first systematic survey of HCV genetic diversity in the DRC. HCV sequences obtained belonged to diverse lineages of genotype 4, including subtypes 4c, 4k, 4l and 4r, plus one unclassified lineage that may constitute a new subtype. These data suggest that HCV in the DRC exhibits an age 'cohort effect', as has been recently reported in neighbouring countries, and are consistent with the hypothesis that HCV transmission rates were higher in the midtwentieth century, possibly as a result of parenteral, iatrogenic, or other unidentified factors. Different HCV subtypes were associated with individuals of different ages, implying that HCV infection in the DRC may have arisen through multiple separate HCV epidemics with different causes.

2.3 INTRODUCTION

Hepatitis C virus (HCV) is a globally-distributed human pathogen, present in approximately 130–170 million people worldwide, and an estimated 3–4 million people are thought to be infected with HCV each year (Shepard et al., 2005). The introduction of screening strategies for HCV following its discovery in 1989 (Choo et al., 1989) has greatly reduced the transmission of the virus through blood transfusion and blood products, and the main transmission route of HCV in developed countries is now the use of contaminated needles by injecting drug users (Pondé, 2011). In developing countries, non-sterile injections and other unsafe medical interventions are thought to contribute to continuing HCV incidence (Kane et al., 1999).

The HCV genome exhibits considerable sequence heterogeneity and is classified using phylogenetic methods into six confirmed genotypes, each of which is further subdivided into numerous subtypes (Simmonds, 2004). A seventh provisional genotype was isolated in Canada in 2007 from an individual originally from the Democratic Republic of Congo (DRC;Murphy et al., 2007a,b). Some regions, such as sub-Saharan Africa and South-East Asia, harbor unusually diverse HCV strains, likely reflecting the long-term endemic transmission of HCV in these locations (Simmonds, 2004; Pybus et al., 2007). For example, highly-diverse lineages of HCV genotypes 1 and 2 are present in West Africa (Jeannel et al., 1998; Candotti et al., 2003; Markov et al., 2009). Further east, HCV genotypes 1 and 4 are commonly found in central African countries such as Cameroon, Gabon and the Central African Republic (CAR), and strains of HCV genotype 4 are also found in Egypt and the Middle East (Ray et al., 2000; Pybus et al., 2003; Shepard et al., 2005; Njouom et al., 2009). Overall, considerable sequence diversity of HCV genotypes 1, 2 and 4 in sub-Saharan Africa has been observed, and it is likely these strains have been present there for at least

several centuries (Smith et al., 1997; Pybus et al., 2001; Pybus et al., 2007; Ndjomou et al., 2003).

Previous studies have reported high HCV seroprevalence in many central and West African countries and the WHO estimates that the region contains 18.8% of HCV infections worldwide (Kane et al., 1999; Berkes and Cotler, 2005). For example, 5.2% of blood donors in Ghana were seropositive (Candotti et al., 2003), as were 11.6% of blood donors in Cameroon (Nerrienet et al., 2005) and 6.5% of subjects in Gabon (Delaporte et al., 1993). In addition to high overall prevalence in the general population, several studies of central African populations have reported a significant increase in HCV prevalence with age, for example in Cameroon (Nerrienet et al., 2005; Pépin et al., 2010b), Gabon (Ndong-Atome et al., 2008), Equatorial Guinea (Basaras et al., 1999) and the Republic of the Congo (Cantaloube et al., 2010). The observation of high HCV prevalences in those aged 50 or older could be explained by a defined period of increased transmission in the past, possibly as a result of non-sterile medical interventions, as has been proposed for Cameroon between the 1920s and 1960s (Njouom et al., 2007). Support for this interpretation of the age distribution of HCV comes from Egypt, where variation in HCV prevalence among age groups and among locations closely matches the level of exposure of those groups to parenteral anti-schistosomiasis therapy, which was widely administered in the first half of the twentieth century (Frank et al., 2000).

Despite its large size and geographically central position, there have been few studies to date of HCV in the DRC. Tibbs et al. (1991) screened 173 samples from rural populations in local hospitals for anti-HCV antibodies and estimated HCV seroprevalence to be 6.4%. Laurent et al. (2001) surveyed pregnant women and commercial sex workers (CSW) in Kinshasa and reported seroprevalences of 6.6%

among CSW and 4.3% among pregnant women. Liu et al. (1999) also investigated pregnant women from Kinshasa (n= 97) and found that while 10.3% carried human pegivirus (HPgV) RNA, only 1% carried HCV RNA. Most recently, Batina Agasa et al. (2010) screened 140 patients with sickle-cell disease in Kisangani for anti-HCV antibodies and found 7.9% were seropositive, all of whom had received blood transfusions before the introduction of HCV screening in 2004. Most of these prevalence studies have been limited in size and study population.

In addition to our poor understanding of HCV epidemiology in the DRC, there is scant information about the genetic diversity of HCV in the country. No country-specific surveys have been published and the HCV sequence database (Kuiken et al., 2005) contains only 21 sub-genomic sequences from 9 isolates labeled as originating from the DRC. These sequences belong to various subtypes of genotype 4, although classification is uncertain, as many of these sequences are <500 nt in length. However, both patients in which the provisional genotype 7 was discovered were originally from the DRC (Murphy et al., 2007a, b) suggesting that the country may harbor further undetected diversity. The presence of high viral diversity in a region can aid the search for the geographic origin and source population of a virus, neither of which is known for HCV. The only other known viruses in the genus Hepacivirus are GB Virus B (GBV-B) and the Canine Hepacivirus/Non-Primate Hepacivirus (CHV/NPHV), both of which are highly divergent from HCV and not known to infect humans: GBV-B has not been found in a natural host, and to date CHV/NPHV has been found only in dogs and horses (Kapoor et al., 2011; Stapleton et al., 2011; Burbelo et al., 2012; Lyons et al., 2012).

In order to improve our knowledge of the molecular epidemiology and genetic diversity of HCV in the DRC 299 blood samples from the country were screened for

HCV using both serological assays and PCR. From those samples that contained HCV RNA we attempted to obtain viral sequences from the Core and NS5B genes.

In addition, the same samples were also screened for HPgV RNA using PCR and part of the NS3 gene was sequenced from HPgV positive samples. Note that HPgV was previously termed GB Virus C or hepatitis G virus and its new name has been provisionally approved by the ICTV Flavivirus Study Group. Using available demographic information the age distribution of HCV infections in our sample set were reconstructed and compared.

2.4 MATERIALS AND METHODS

2.4.1 Study population

EDTA blood samples were collected from informed consenting members of the uniformed services as part of a screening program for HIV and other infectious diseases. These samples have been studied previously for HIV-1 (Djoko et al., 2011, wherein full details of sample collection can be found) and human parvovirus 4 (Sharp et al., 2010). Collection took place in Kinshasa, capital of the DRC, between June and September 2007. The samples were anonymised although patient date of birth and date of collection were available for most. All samples were from male individuals. Mean age was 38.7 ± 1.28, range 21–71.

2.4.2 HCV serology, RT-PCR and sequencing

Serological tests for HCV were conducted using Ortho 3.0 Enhanced SAVe (Ortho Clinical Diagnostics) as per the manufacturer's instructions. All HCV samples that were reactive for anti-HCV antibodies were tested for HCV-RNA. Viral RNA was extracted from sera using the Qiagen miniprep kit (QIAGEN) as per the

manufacturer's instructions, with one modification: 500µl of sera was centrifuged at 6000xG for 1 hour, from which 360µl of sera was removed. RNA was subsequently extracted from the remaining 140µl. Three subgenomic regions were amplified (5' UTR, Core, and NS5B) using Superscript III (Invitrogen, Life Technologies) followed by nested PCR with the Fast Start High Fidelity PCR System (Roche Applied Science) using standard protocols. Controls were run in parallel at each step. Primers were designed from an alignment of representative genomes from each of the 7 HCV genotypes; primers are listed inTable 1. The internal primers were used for sequencing with BigDye Terminator v3.1 (Applied Biosystems). Traces were examined using Sequencher 5.0 (Gene Codes). The HCV sequences have accession numbers KC012607-KC12616 (Core sequences) and KC506766-KC506776 (NS5B sequences).

2.4.3 Phylogenetic analysis

Nucleotide sequences were aligned by hand using Se-Al v2.0 together with HCV and HPgV reference sequences, obtained from the HCV sequence database (Kuiken et al., 2005) and from GenBank. Phylogenies were estimated for each alignment using maximum likelihood (ML), as implemented in GARLI v0.951 (Zwickl, 2006) under a general time-reversible (GTR) nucleotide substitution model with gamma-distributed among-site rate variation. Statistical support for phylogenetic clustering was calculated using a ML bootstrap approach with 500 bootstrap replicates; bootstrap scores were summarized using the Consense package in PHYLIP (Felsenstein, 1989). Phylogenies were visualized and annotated using FigTree (http://tree.bio.ed.ac.uk/software/figtree).

2.4.4 HPgV RT-PCR and sequencing

All samples that were negative for HCV antibodies were subsequently screened for HPgV through the amplification and sequencing of a 268 nt amplicon located at positions 4126–4275 in the NS3 region relative to isolate H77. Primers are listed in

Table 2.1. RNA was extracted from 200µl plasma using QIAamp MinElute Virus Spin Kit (Qiagen) and converted to cDNA with random primers using Superscript III. PCR amplification was performed using Fast Start High Fidelity (Roche Applied Science) and the products were sequenced using Big Dye Terminator v3.1. Traces were examined using Sequencher 5.0. The HPgV sequences have accession numbers KC506736-KC506765.

Primer	*Primer sequences*	*Genotype*
5' UTREx400F	CCTTGTGGTACTGCCTGATAG	All
5' UTRIn405F	CTGATAGGGTGCTTGCGAGTG	All
CoreIn980R	AGTGCCARRAGGAAGATAGA	All
EIn1420R	CCAGTTCATCATCATGTCCCA	All
EEx1423R	GGRCTCCAGTTCATCATCATGTC	All
5' UTRExF1	CCCTGTGAGGAACTWCTGTCTTCACGC	All
5' UTRInF3	TCTAGCCATGGCGTTAGTRYGAG	All
CoreExR2	GGTGCACGGTCTACGAGACCT	All
CoreInR4	CACTCGCAAGCACCCTATCAGGCAGT	All
NS5BExF	TGGGGATCCCGTATGATACCCGCTGCTTTGA	1–5 and 7
NS5BExR	CGGAATTCCTGGTCATAGCCTCCGTGAA	1–5 and 7
NS5BExFG6	CCHATGGGGTTYTCCTAYGACAC	6
NS5BExRG6	GGNGCYGAGTAYCTGGTCATGGC	6
NS5BInF	GACACCCGCTGCTTTGACTC	All
NS5BInR	GAGTCTTCACGGAGGCTATGACNAGGTA	All
GBVNS3_4126a	GSGCNATGGGNCCNTAYATGGA	HPgV
GBVNS3_4735as	GTNACYTCVACNACCTCCTCYACCA	HPgV
GBVNS3_4275s	GTGGTNATHTGYGAYGAGTGYCA	HPgV
GBVNS3_4543as	TCRCACTCMRCCTTKGARTGRCARAA	HPgV

Table 2.1: *Primers used in the RT-PCR, PCR and sequencing reactions of the 5' UTR, Core and NS5B regions of HCV, and of the NS3 region of HPgV.*

2.5 RESULTS

2.5.1 HCV serology

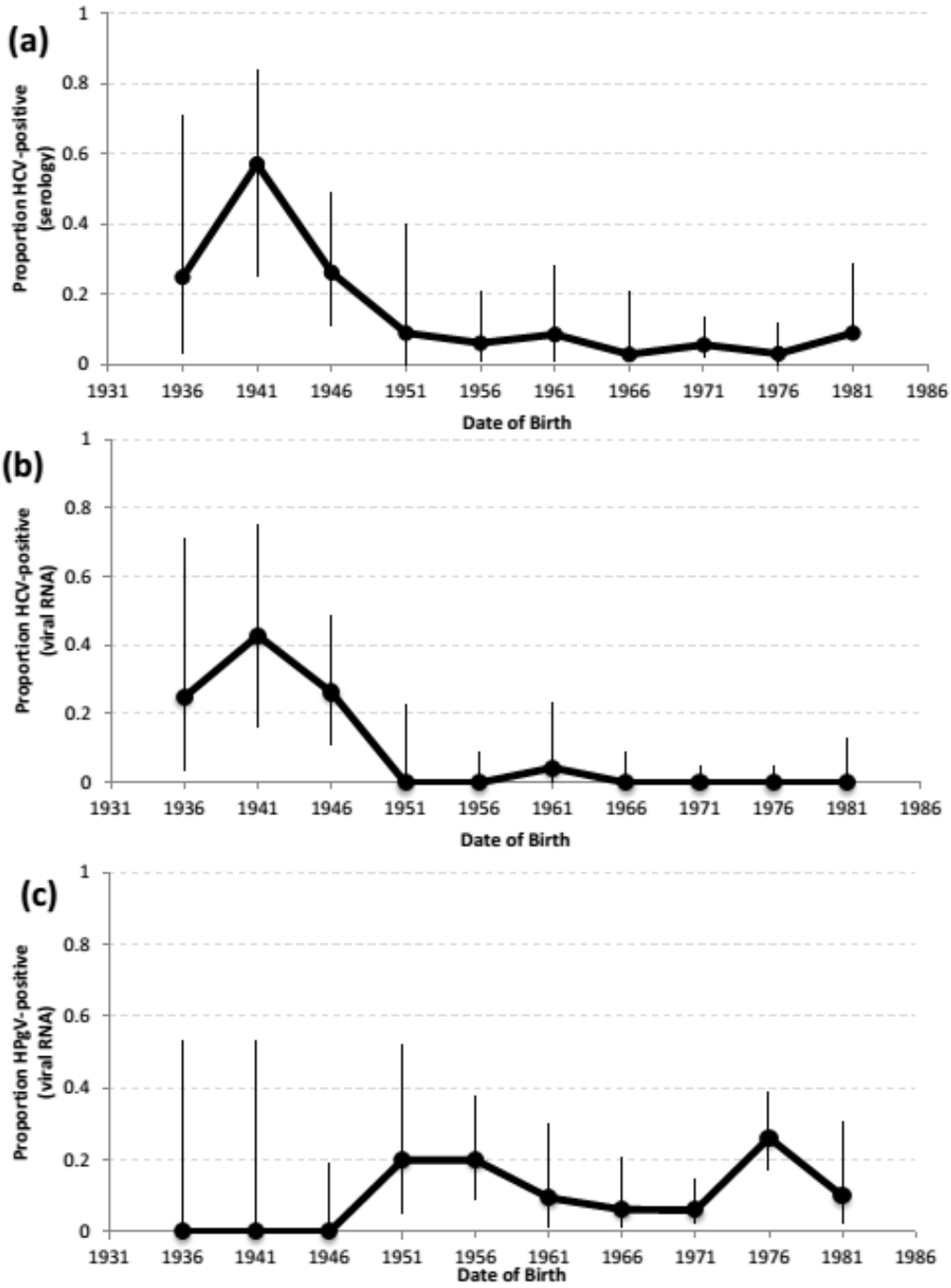
Of the samples screened for anti-HCV antibodies, 13.7% were reactive (41/299). Figure 2.1a shows the age distribution of the 41 HCV-seropositive samples. Samples from individuals born before 1950 were significantly more likely to be HCV

seropositive than those born after that date when tested against the assumption of linear increase in HCV prevalence with age ($p < 0.0001$; Chi Square goodness-of-fit test). The estimated seroprevalence in the three oldest age classes ranged between 25% and 57%, whereas in the remaining younger age classes the estimated seroprevalence was 6–19%. Of the 41 seropositive samples, only 24 had a signal to cut-off ratio >3.8 and might be considered to represent active infections (Alter et al., 2003).

2.5.2 HCV RT-PCR and sequencing

Only about one quarter of the HCV samples that were reactive for anti-HCV antibodies contained detectable HCV RNA: among all samples 3.7% (11 out of 299) were PCR-positive. Possible reasons for this substantial difference between seropositivity and RNA-positivity are explored in the discussion. There was, however, a relationship between the ELISA signal to cut-off ratios and RNA-positivity: no HCV RNA was recovered from samples with signal to cut-off ratios <3.6, whilst RNA was obtained from 42% (10 out of 24) samples with ratios >3.8. Four samples exhibited signal to cut-off ratios close to this threshold (between 3.0 and 3.8), of which one tested positive for HCV RNA.

Figure 2.1 *(next page)**. The age distribution among 299 blood samples from the DRC of (a) HCV seroprevalence (all reactive samples) and (b) HCV RNA prevalence (Core and/or NS5b sequence). (c) The age distribution of HPgV RNA prevalence among HCV seronegative samples. Samples were assigned to one of ten age categories by date of birth. 15 samples (all negative) did not have date of birth information and are thus not included in these numbers. The y-axis shows the proportion of samples in each age category that were positive. Error bars represent 95% confidence limits of this proportion, estimated using the Adjusted Wald method (Agresti and Coull 1998).*


(a)
Proportion HCV-positive (serology)
1
0.8
0.6
0.4
0.2
0
1931 1936 1941 1946 1951 1956 1961 1966 1971 1976 1981 1986
Date of Birth
(b)
Proportion HCV-positive (viral RNA)
1
0.8
0.6
0.4
0.2
0
1931 1936 1941 1946 1951 1956 1961 1966 1971 1976 1981 1986
Date of Birth
(c)
Proportion HPgV-positive (viral RNA)
1
0.8
0.6
0.4
0.2
0
1931 1936 1941 1946 1951 1956 1961 1966 1971 1976 1981 1986
Date of Birth

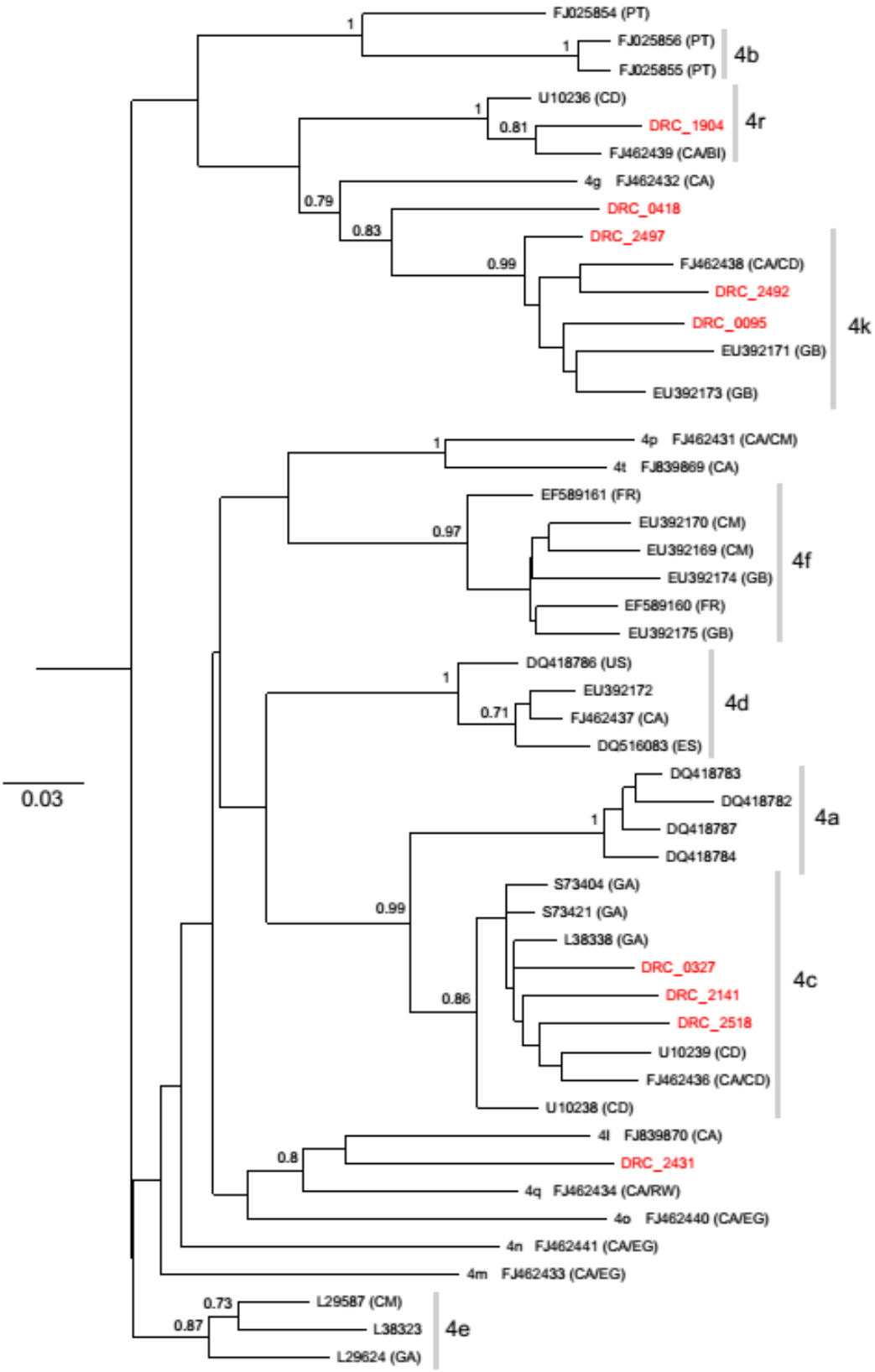

FJ025854 (PT)
FJ025856 (PT)
FJ025855 (PT)
4b
U10236 (CD)
DRC_1904
FJ462439 (CA/BI)
4r
4g FJ462432 (CA)
DRC_0418
DRC_2497
FJ462438 (CA/CD)
DRC_2492
DRC_0095
EU392171 (GB)
EU392173 (GB)
4k
4p FJ462431 (CA/CM)
4t FJ839869 (CA)
EF589161 (FR)
EU392170 (CM)
EU392169 (CM)
EU392174 (GB)
EF589160 (FR)
EU392175 (GB)
4f
DQ418786 (US)
EU392172
FJ462437 (CA)
DQ516083 (ES)
4d
0.03
DQ418783
DQ418782
DQ418787
DQ418784
4a
S73404 (GA)
S73421 (GA)
L38338 (GA)
DRC_0327
DRC_2141
DRC_2518
U10239 (CD)
FJ462436 (CA/CD)
U10238 (CD)
4c
4l FJ839870 (CA)
DRC_2431
4q FJ462434 (CA/RW)
4o FJ462440 (CA/EG)
4n FJ462441 (CA/EG)
4m FJ462433 (CA/EG)
L29587 (CM)
L38323
L29624 (GA)
4e

Figure 2.2 *(previous page). Maximum likelihood phylogeny of HCV genotype 4, estimated from the Core region sequences. New DRC isolates from this study are in red; other sequences are reference isolates from Genbank or the HCV Sequence Database. Numbers next to nodes represent maximum likelihood bootstrap support values; only values >0.7 are shown. Grey bars indicate the location of HCV subtypes. Reference isolates are labeled with accession number and, where available, the location of sampling (ISO 3166 two-letter country codes, in parenthesis). For some isolates two locations are noted: the first corresponds to the country of sampling and the second corresponds to the country of origin (information obtained from the primary literature). The phylogeny is midpoint rooted and the scale bar is in units of nucleotide substitutions per site.*

The discrepancy between seropositivity and RNA positivity is notable, and has been seen in other studies in Africa (Njouom *et al.,* 2012; Mullis *et al.,* 2013; King *et al.,* 2014). This may be indicative of the presence of another pathogen circulating in Africa with sufficient antigenic similarity to cause cross-reactivity in these serological assays.

For nine samples we obtained both a 1023 nt sequence from the 5' UTR-Core genome region and a 342 nt sequence from the NS5B genome region. For the remaining two samples (DRC2450 and DRC1424) we were only able to obtain a NS5B sequence. The age distribution of HCV seropositive samples (Figure 2.1a) matched that of the HCV RNA-positive samples (Figure 2.1b).

2.5.3 HCV phylogenetic analysis

The HCV sequences obtained were combined with reference sequences and subjected to phylogenetic analysis. The HCV RNA-positive samples from the DRC grouped with various subtypes and lineages of genotype 4. The phylogenies estimated from the Core

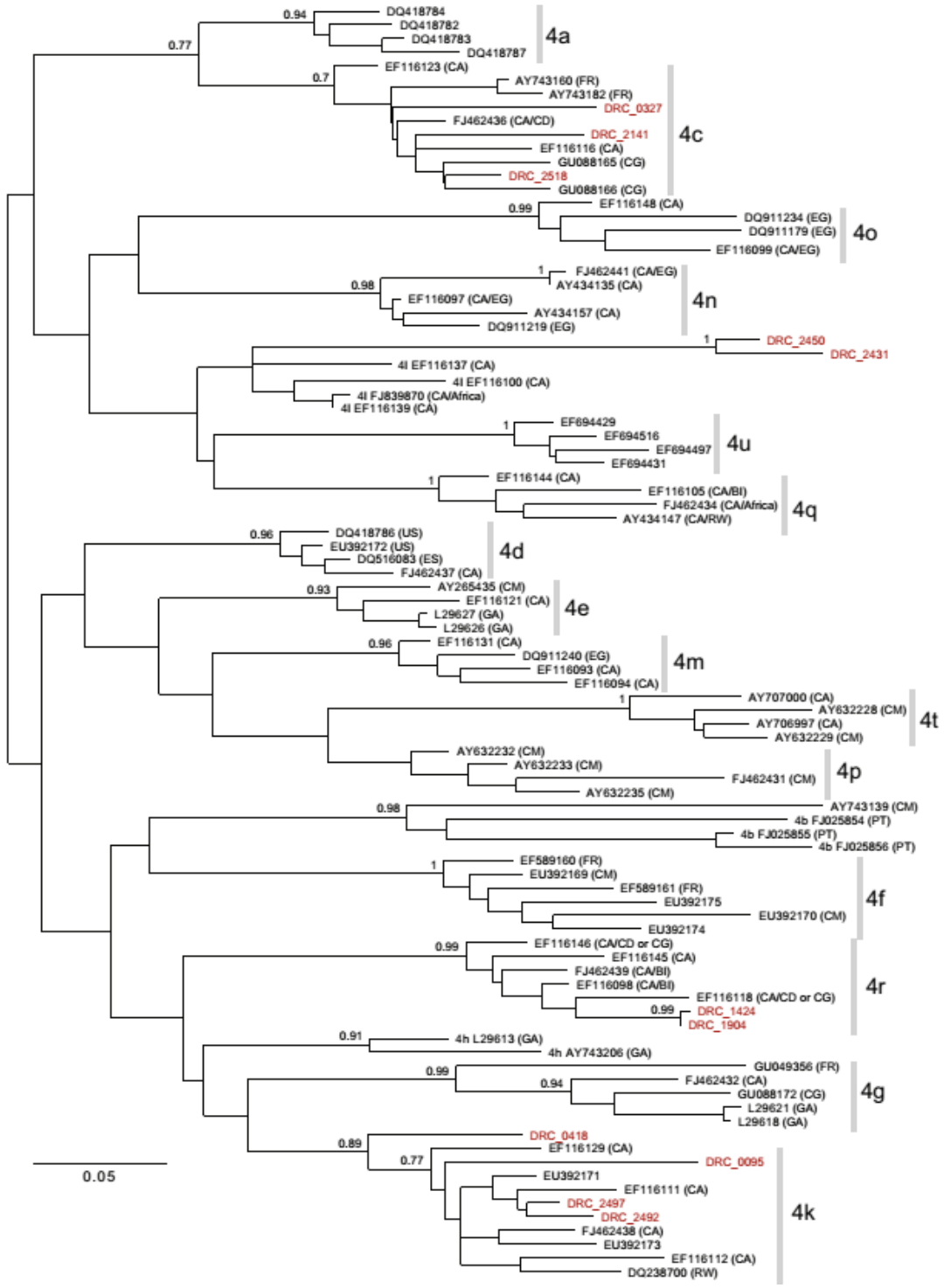


Figure 2.3. *Maximum-likelihood phylogeny of HCV genotype 4, estimated from NS5B region sequences. For further details see legend to Fig. 2.2.*

(Figure 2.2) and NS5B (Figure 2.3) sequences gave concordant results. Three samples grouped within subtype 4c (DRC0327, DRC2141, DRC2518), three grouped within

subtype 4k (DRC0095, DRC2492, DRC2497), and two grouped within subtype 4r (DRC1904, DRC1424). Isolate DRC0418 was placed immediately basal to subtype 4k in both trees (hereafter termed 4k-like). All of these clusters were supported with high maximum likelihood bootstrap scores (≥0.7). In contrast the two remaining isolates (DRC2431 and DRC2450) were closely related to each other (Fig. 3; bootstrap score = 1.0) but distinct from all currently-identified subtypes. This pair may therefore represent a new unclassified lineage, although further samples are required for confirmation.

There was a significant difference in the relative frequency of the different HCV subtypes according to the age of the infected individuals (Figure 2.4). Specifically, isolates classified as subtype 4k, 4l or 4c were only found in samples from individuals born before 1950, whereas those classified as subtype 4r were only found in samples from individuals born after 1956 (Fisher's exact test; p= 0.018).

2.5.4 HPgV RT-PCR, sequencing and analysis

The proportion of samples that were RNA-positive was higher for HPgV than for HCV, with 12.7% (33/259) of samples testing positive for HPgV RNA. For each of the positive samples we obtained a 268 nt sequence of the HPgV NS3 gene. The age distribution of HPgV RNA-positive samples (Figure 2.1c) was different to that observed for HCV (Figure 2.1b), as HPgV RNA was more likely to be found in samples from younger individuals than older ones. No HPgV RNA was found in individuals born before 1950 and the highest rate of HPgV positivity (>20%) was found in those born between 1976 and 1980. However, the HPgV prevalence results are difficult to interpret correctly, as only samples that were not seroreactive for HCV were available for analysis (see Discussion).

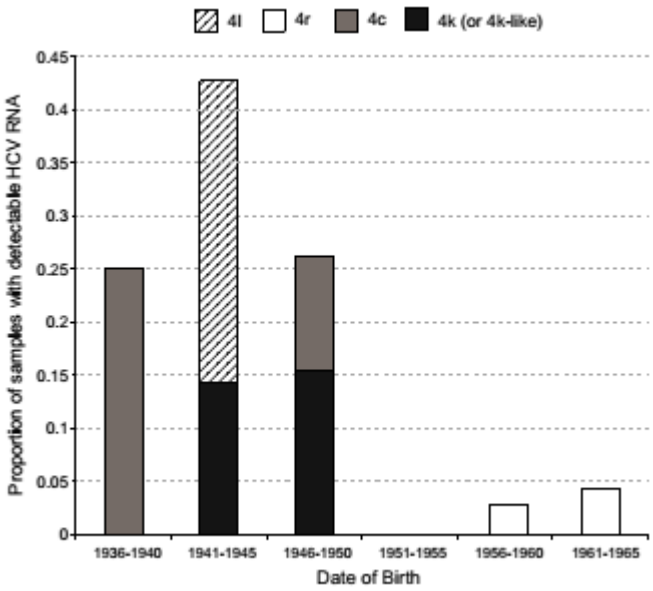


Figure 2.4. *The age distribution of HCV subtypes among samples with detectable HCV RNA in the Core and/or NS5B region. Samples from the oldest HCV-positive subjects belonged to subtype 4c. Subtypes 4k, 4l and 4r were most common in samples from subjects born between 1941 and 1950.*

Figure 2.5 shows the maximum likelihood phylogeny of the HPgV NS3 sequences obtained here. The bootstrap scores for most nodes in the tree were low and not all of the HPgV genotypes were resolved as reciprocally monophyletic groups, suggesting that the phylogenetic signal in the sequences is not strong. Most of the HPgV sequences from the DRC grouped together with genotype 1 strains. Four DRC samples were genetically distinct and clustered most closely with genotype 5 (DRC0101, DRC1220, DRC1520, DRC2908). However these groupings were not supported with high bootstrap support values.

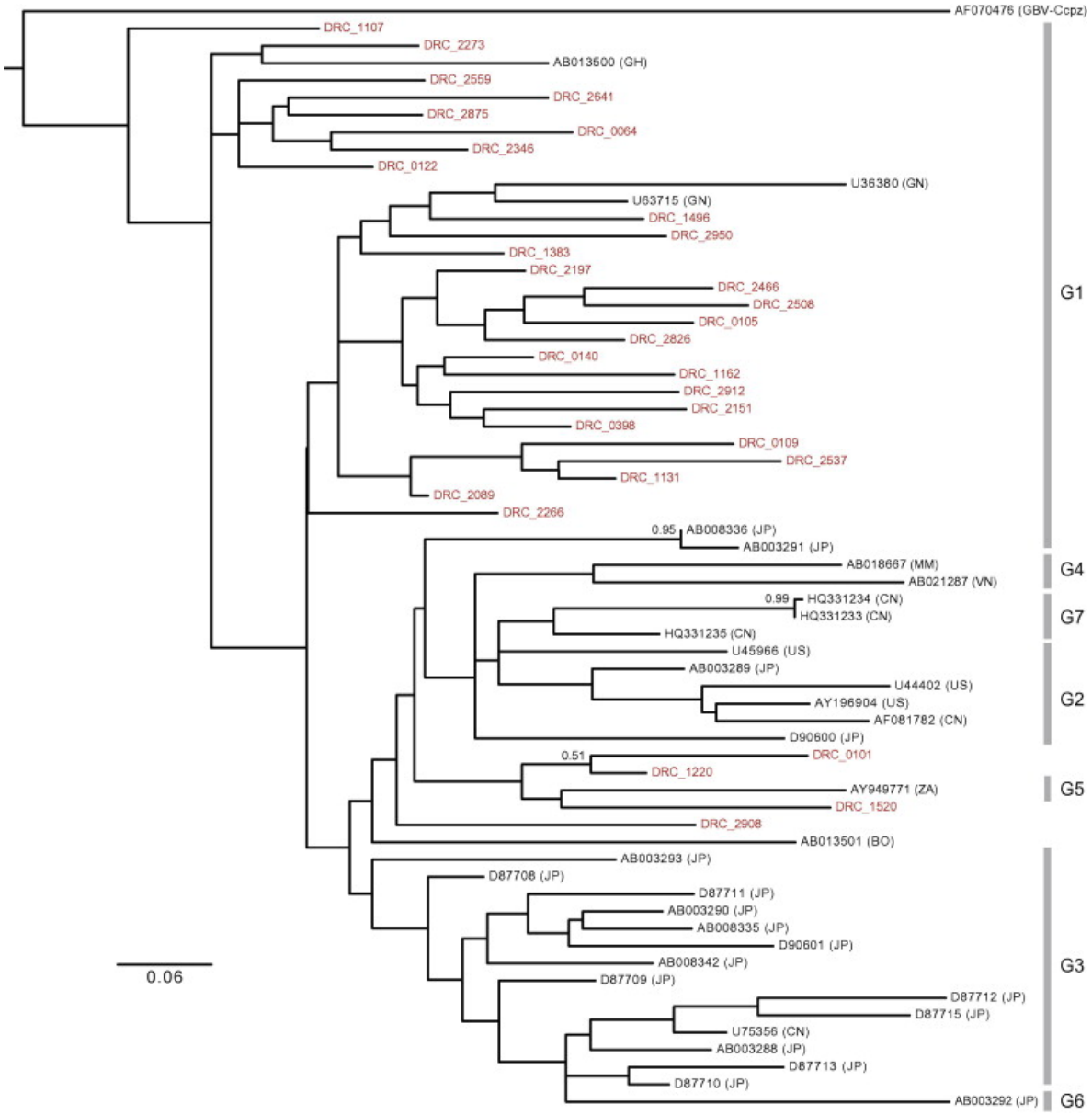


Figure 2.5. *Maximum-likelihood phylogeny of HPgV, estimated from partial NS3 sequences. New DRC isolates from this study are in red; other sequences are reference isolates from Genbank or the HCV Sequence Database. Numbers next to nodes represent maximum likelihood bootstrap support values; only values >0.5 are shown. Grey bars indicate HPgV genotypes 1–7; these labels are very approximate due to low phylogenetic resolution. Reference isolates are labeled with accession number and, where available, the location of sampling (ISO 3166 two-letter country codes, in parenthesis). The phylogeny is rooted using SPgVcpz (previously named GBV-Ccpz) and the scale bar is in units of nucleotide substitutions per site.*

2.6 DISCUSSION

In an attempt to better understand the molecular epidemiology and genetic diversity of HCV in the DRC we screened 299 blood samples from the country for HCV using serological assays and PCR. An important finding of this study is the raised HCV-seropositivity and HCV RNA-positivity among subjects aged 50 and over. Although overall HCV seropositivity (13.7%) was high compared to the global prevalence of HCV, previous studies have reported seropositivity of 6.4% in blood donors and 6.6% in commercial sex workers (CSWs) in Kinshasa, the capital of the DRC (Laurent et al., 2001; Tibbs et al., 1991). However, we observed a substantial difference between levels of HCV seropositivity (13.7%) and HCV RNA-positivity (3.7%), although the age profiles of these markers of infection were similar (Fig. 1a and b). A comparable difference was reported for HCV genotype 4 infections in Uganda (Biggar et al., 2006) and could result from non-specific ELISA reactivity, low detectable viraemia, or unusually high rates of clearance after infection. We did not undertake a confirmatory test (such as a RIBA immunoblot assay) to exclude non-specific reactivity against other pathogens (Callahan et al., 1993), hence we hereafter discuss only the age distribution of HCV seroprevalence, not its absolute value.

Laurent et al. (2001) noted that HCV prevalence in CSWs increased with age, from 2.8% in those aged <20–21.3% in those aged >40. Although our sample set is comparatively small in size, its wide age range enabled us to observe that HCV prevalence in men in the DRC rises most rapidly in those aged >50 (Fig. 1a and b). Similar age profiles have been identified in neighboring countries: in several locations in south and south-eastern Cameroon HCV seroprevalence rises rapidly with age, surpassing 50% in those aged 50 and older in a mixed population sample group (Nerrienet et al., 2005; Pépin et al., 2010a). HCV seropositivity also increases with age

among pregnant women in Gabon, although at much lower levels (<6%;Ndong-Atome et al., 2008), and among Bantu populations in the Republic of the Congo (Cantaloube et al., 2010). Since HCV infection is far more likely to reduce than to increase survival, these results suggest that rates of HCV transmission in such populations were higher in the past.

The strongest evidence that past HCV transmission can generate an age cohort-effect comes from Egypt, where parenteral antischistosomal therapy campaigns between 1930 and 1955 resulted in the repeated intravenous treatment of a large cross-section of the population with needles that were very likely incompletely sterilized (Strickland, 2010). Estimated levels of exposure to this treatment varied among locations and age groups, but closely matched HCV prevalence in each case (Frank et al., 2000). Today, approximately 10–20% of the Egyptian population is chronically infected by HCV and 90% of those infections are caused by subtype 4a (Arthur et al., 1997). The possible cause or causes of past iatrogenic HCV transmission in sub-Saharan Africa are more complex and less well understood (reviewed inPépin, 2011). In Ebolowa, Cameroon, Pépin et al. (2010b)found that HCV seropositivity was associated with past intravenous treatment for malaria, but not with antitreponemal treatment, as previously suspected (Pépin and Labbé, 2008). While HCV can be transmitted through sexual or intrafamilial routes, past iatrogenic transmission better explains why HCV seropositivity varies greatly among locations within a country (Frank et al., 2000; Nerrienet et al., 2005) as well as its elevation in those aged >50 in affected locations.

The HCV sequences obtained in this study were genetically diverse; among only eleven RNA-positive isolates we found three previously-recognised HCV subtypes (4c, 4r, 4k) and an unclassified group (comprising isolates DRC2431 and DRC2450) that may constitute a new subtype. The HCV subtypes detected in our survey matched

those previously observed for the DRC and surrounding countries. The three previously-reported Core gene sequences from the DRC (U10236, U10238 and U10239) belong to subtypes 4r and 4c (Fig. 2.2). Several other isolates classified as subtypes 4r, 4c and 4k were sampled in the Republic of Congo, Rwanda or Burundi, or were obtained from immigrants to Canada from those countries or from the DRC (Figs. 2.2 and 2.3). In contrast, HCV genotype 4 lineages described in the DRC are different from the genotype 4 lineages typically found in Egypt (e.g. subtypes 4a, 4m, 4n, 4o; Abdel Hamid et al., 2007). Overall, HCV genotype 4 genetic diversity is greater in Central Africa than in Egypt and North Africa, strengthening the hypothesis that the former represents the region of origin of genotype 4 (Ndjomou et al., 2003; Pybus et al., 2007).

Interestingly, the relatively frequency of the lineages found in the DRC varied with the age of the infected individual: subtypes 4c and 4k were found in those born before 1950, whereas subtype 4r only was recovered from those born after 1956. In combination with the age-distribution results, this suggests that different routes or events (iatrogenic or otherwise) may explain HCV transmission among infected individuals of different ages. Larger sample sizes and detailed geographic information will be needed to investigate this hypothesis and we hope to address these issues in future surveys. Past rates of HCV transmission in Africa have been estimated from contemporary sequence data using coalescent-based methods (Pybus et al., 2003; Njouom et al., 2007; 2009;Ndong-Atome et al., 2008; Pépin et al., 2010a,b;). However this technique requires minimal sample sizes of 15–20 isolates per subtype for reliable estimation and thus cannot be applied here. Historical research into specific parenteral treatment campaigns in the DRC will also be crucial in reconstructing the epidemic history of the HCV in the region (Pépin, 2011).

It is interesting to compare the HCV results with those obtained for human pegivirus (HPgV) from the same set of samples. Globally, 1–4% of healthy blood donors are viremic for HPgV and another 13% carry anti-HPgV antibodies (Blair et al., 1998; Gutierrez et al., 1997; Pilot-Matias et al., 1996; Tacke et al., 1997). Here, we found HPgV RNA in 12.7% of samples, consistent with a previous survey that reported viremia in 10.3% of pregnant women in Kinshasa (Liu et al., 1999). We could not assess the frequency of HCV/HPgV co-infection because due to limited sample volume only those samples that were HCV seronegative were available to be screened for HPgV. This also prevents us from drawing firm conclusions from the age-distribution of HPgV viremia in our samples (Fig. 2.1c): HCV and HPgV are both efficiently transmitted via infected blood (Bhattarai and Stapleton, 2012) and therefore the removal of HCV seropositive samples may have biased downwards our estimate of HPgV prevalence in those aged >50. Most of the HPgV viruses we obtained clustered with genotype 1, which is typical of African HPgV strains, whilst a minority was more closely related to genotype 5. These classifications are not conclusive, however, because our HPgV phylogeny was not supported with high bootstrap values.

Although the work here represents the largest study to date of HCV molecular epidemiology in the DRC – the largest country in sub-Saharan Africa – it is still limited by a comparatively small sample size and a restricted sample population. Through virus genome sequencing, in addition to serological testing, we were able to provide insights into the epidemiology of HCV. Despite the limitations of sample size, differences in HCV positivity among those born before and after 1950 were significant and we expect this cohort effect to be observed in more comprehensive surveys undertaken in future. The military population surveyed in this study limits the applicability of these results to the wider population; they are all male, may have

travelled much more within and outside the country than non-military personnel, and will have had higher exposure to risk factors such as blood transfusion and parenteral medical treatment. Therefore a study of a diverse civilian population within the DRC may be useful to confirm and expand on the patterns described in this chapter. Lastly, this study illustrates that the importance of the HCV epidemic in sub-Saharan Africa and the need to plan for the aging cohort. The prevalence and diversity of HCV and related viruses in some of the largest and most populous countries in Africa (such as the DRC, Nigeria, Sudan, Ethiopia, Kenya and Angola) remain poorly studied, highlighting the need for broader surveys in this important region.

2.7 ACKNOWLEDGEMENTS

We thank the Ministry of Public Health and the Ministry of National Defense of the Democratic Republic of Congo for permission to undertake this study and the U.S. Embassy of the DRC for their continued support. This project was financially supported by Global Viral. PK and GLAH were supported by The Wellcome Trust, Oxford Martin School, and NIHR Biomedical Research Centre, Oxford. CFD was supported in part by the NIH Fogarty International Center AIDS International Training and Research Program (2 D 43 TW000010-16/17). NDW is supported by the NIH Director's Pioneer Award (DP1-OD000370). Metabiota and Global Viral are graciously supported by the U.S. Department of Defense Armed Forces Health Surveillance Center, Division of Global Emerging Infections, Surveillance Operations (AFHSC GEIS), The Defense Threat Reduction Agency Cooperative Biological Engagement Program (DTRA-CBEP), Google.org, The Skoll Foundation, DoD HIV/ AIDS Prevention Program (DHAPP), the Henry M. Jackson Foundation for the Advancement of Military Medicine, and the U.S. Agency for International

Development (USAID) Emerging Pandemic Threats Program PREDICT project, under the terms of Cooperative Agreement Number GHN-A-OO-09-00010-00.

2.8 REFERENCES

Abdel-Hamid, M., El-Daly, M., Molnegren, V., El-Kafrawy, S., Abdel-Latif, S., Esmat, G., Strickland, G.T., Loffredo, C., Albert, J. and Widell, A. (2007). Genetic diversity in hepatitis C virus in Egypt and possible association with hepatocellular carcinoma. *Journal of General Virology* 88: 1526–1531.

Alter, M.J., Kuhnert, W.L. and Finelli, L. (2003). Guidelines for laboratory testing and result reporting of antibody to hepatitis C virus. *Centers for Disease Control and Prevention – Morbidity and Mortality Weekly Report Recommendations and Reports* 52: 1–13.

Arthur, R.R., Hassan, N.F., Abdallah, M.Y., El-Sharkawy, M.S., Saad, M.D., Hackbart, B.G. and Imam, I.Z. (1997). Hepatitis C antibody prevalence in blood donors in different governorates in Egypt. *Transactions of the Royal Society of Tropical Medicine and* Hygiene 91: 271–274.

Basaras, M., Santamaría, A., Sarsa, M., Gutiérrez, E., de Olano, Y. and Cisterna, R. (1999). Seroprevalence of hepatitis B and C, and human immunodeficiency type 1 viruses in a rural population from the republic of equatorial Guinea. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 93: 250–252.

Batina Agasa, S., Dupont, E., Kayembe, T., Molima, P., Malengela, R., Kabemba, S., Andrien, M., Lambermont, M., Cotton, F., Vertongen, F. and Gulbis, B. (2010). Multiple transfusions for sickle cell disease in the Democratic Republic of Congo: the importance of the hepatitis C virus. *Transfusion Clinique et Biologique* 17: 254–259.

Berkes, J. and Cotler, S.J. (2005). Global epidemiology of HCV infection. *Current Hepatitis Reports* 4: 125–130.

Bhattarai, N. and Stapleton, J.T. (2012). GB virus C: the good boy virus? *Trends in Microbiology* 20: 124–130.

Biggar, R.J., Ortiz-Conde, B.A., Bagni, R.K., Bakaki, P.M., Wang, C.D., Engels, E.A., Mbulaiteye, S.M. and Ndugwa, C.M. (2006). Hepatitis C virus genotype 4 in Ugandan children and their mothers. *Emerging Infectious Diseases* 12: 1440–1443.

Blair, C., Davidson, F., Lycett, C., McDonald, D., Haydon, G., Yap, P.L., Hayes, P., Simmonds, P. and Gillon, J. (1998). Prevalence, incidence, and clinical characteristics of hepatitis G virus/GB virus C infection in Scottish blood donors. Journal of Infectious Diseases 178: 1779–1782.

Burbelo, P.D., Dubovi, E.J., Simmonds, P., Medina, J.L., Henriquez, J.A., Mishra, N., Wagner, J., Tokarz, R., Cullen, J.M., Iadarola, M. J., Rice, C. M., Lipkin, W. I. and Kapoor, A. (2012). Serology-enabled discovery of genetically diverse hepaciviruses in a new host. *Journal of Virology* 86: 6171–6178.

Callahan, J.D., Constantine, N.T., Kataaha, P., Zhang, X., Hyams, K.C. and Bansal, J. (1993). Second generation hepatitis C virus assays: performance when testing African sera. *Journal of Medical Virology* 41: 35–38.

Candotti, D., Temple, J., Sarkodie, F. and Allain, J.P. (2003). Frequent recovery and broad genotype 2 diversity characterize hepatitis C virus infection in Ghana, West Africa. *Journal of Virology* 77: 7914–7923.

Cantaloube, J.-F., Gallian, P., Bokilo, A., Jordier, F., Biagini, P., Attoui, H., Chiaroni, J. and de Micco, P. (2010). Analysis of hepatitis C virus strains circulating in Republic of the Congo. *Journal of Medical Virology* 82: 562–567.

Choo, Q.L., Kuo, G., Weiner, A.J., Overby, L.R., Bradley, D.W. and Houghton, M. (1989). Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science* 244: 359–362.

Delaporte, E., Thiers, V., Dazza, M.C., Romeo, R., Mlika-Cabanne, N., Aptel, I., Schrijvers, D., Bréchet, C. and Larouzé, B. (1993). High level of hepatitis C endemicity in Gabon, equatorial Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 87: 636–637.

Djoko, C.F., Rimoin, A.W., Vidal, N., Tamoufe, U., Wolfe, N.D., Butel, C., LeBreton, M., Tshala, F.M., Kayembe, P.K., Muyembe, J. J., Edidi-Basepeo, S., Pike, B. L., Fair, J. N., Mbacham, W. F., Saylors, K. E., Mpoudi-Ngole, E., Delaporte, E., Grillo, M. and Peeters, M. (2011). High HIV type 1 group M pol diversity and low rate of antiretroviral resistance mutations among the uniformed services in Kinshasa, Democratic Republic of the Congo. *AIDS Research and Human Retroviruses* 27: 323–329.

Felsenstein, J. (1989). PHYLIP – phylogeny inference package (version 3.2). *Cladistics* 5: 164–166.

Frank, C., Mohamed, M.K., Strickland, G.T., Lavanchy, D., Arthur, R.R., Magder, L.S., El Khoby, T., Abdel-Wahab, Y., Aly Ohn, E.S., Anwar, W. and Sallam, I. (2000). The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. *Lancet* 355: 887–891.

Gutierrez, R.A., Dawson, G.J., Knigge, M.F., Melvin, S.L., Heynen, C.A., Kyrk, C.R., Young, C.E., Carrick, R.J., Schlauder, G.G., Surowy, T. K., Dille, B. J., Coleman, P. F., Thiele, D. L., Lentino, J. R., Pachucki, C. and Mushahwar, I. K. (1997).

Seroprevalence of GB virus C and persistence of RNA and antibody. *Journal of Medical Virology* 53: 167–173.

Jeannel, D., Fretz, C., Traore, Y., Kohdjo, N., Bigot, A., Pê Gamy, E., Jourdan, G., Kourouma, K., Maertens, G., Fumoux, F., Fournel, J. J. and Stuyver, L. (1998). Evidence for high genetic diversity and long-term endemicity of hepatitis C virus genotypes 1 and 2 in West Africa. *Journal of Medical Virology* 55: 92–97.

Kapoor, A., Simmonds, P., Gerold, G., Qaisar, N., Jain, K., Henriquez, J., Firth, C., Hirschberg, D., Rice, C., Shields, S. and Lipkin, W. (2011). Characterization of a canine homolog of hepatitis C virus. *Proceedings of the National Academy of Sciences of the United States of America* 108: 11608–11613.

Kane, A., Lloyd, J., Zaffran, M., Simonsen, L. and Kane, M. (1999). Transmission of hepatitis B, hepatitis C and human immunodeficiency viruses through unsafe injections in the developing world: model-based regional estimates. *Bulletin of the World Health Organization* 77: 801–807.

King, S., Adjei-Asante, K., Appiah, L., Adinku, D., Beloukas, A., Atkins, M., Sarfo, S. F., Chadwick, D., Phillips, R. O. and Geretti, A. M. (2014). Antibody screening tests variably overestimate the prevalence of hepatitis C virus infection among HIV-infected adults in Ghana. *Journal of Viral Hepatitis* ePub doi: 10.1111/jvh.12354

Kuiken, C., Yusim, K., Boykin, L. and Richardson, R. (2005). The Los Alamos hepatitis C sequence database. *Bioinformatics* 21: 379–384.

Laurent, C., Henzel, D., Mulanga-Kabeya, C., Maertens, G., Larouze, B. and Delaporte, E. (2001). Seroepidemiological survey of hepatitis C virus among commercial sex workers and pregnant women in Kinshasa, Democratic Republic of Congo. *International Journal of Epidemiology* 30: 872–877.

Liu, H.-F., Muyembe-Tamfum, J.-J., Dahan, K., Desmyter, J. and Goubau, P. (1999). High prevalence of GB virus C/hepatitis G virus in Kinshasa, Democratic Republic of Congo: a phylogenetic analysis. *Journal of Medical Virology* 60: 159–165.

Lyons, S., Kapoor, A., Sharp, C., Schneider, B.S., Wolfe, N.D., Culshaw, G., Corcoran, B., McGorum, B.C. and Simmonds, P. 2012. Nonprimate Hepaciviruses in Domestic Horses, United Kingdom. *Emerging Infectious Diseases* 18: 1976–1982.

Markov, P.V., Pépin, J., Frost, E., Deslandes, S., Labbe, A.-C. and Pybus, O.G. (2009). Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa. *Journal of General Virology* 90: 2086–2096.

Mullis, C. E., Laeyendecker, O., Reynolds, S. J., Ocama, P. Quinn, J., Boaz, I., Gray, R. H., Kirk, G. D., Thomas, D. L., Quinn, T. C. and Stabinski, L. (2013). High frequency of false-positive hepatitis C virus enzyme-linked immunosorbent assay in Rakai, Uganda. *Clinical Infectious Diseases* 57: 1747-50.

Murphy, D.G., Willems, B., Deschenes, M., Hilzenrat, N., Mousseau, R. and Sabbah, S. (2007a). Use of sequence analysis of the NS5b region for routine genotyping of hepatitis C virus with reference to C/E1 and 5' untranslated region sequences. *Journal of Clinical Microbiology* 45: 1102–1112.

Murphy, D., Chamberland, J., Dandavino, R. and Sablon, E. (2007b). A new genotype of hepatitis C virus originating from Central Africa. *Hepatology* 46: 623A.

Ndjomou, J., Pybus, O.G. and Matz, B. (2003). Phylogenetic analysis of hepatitis C virus isolates indicates a unique pattern of endemic infection in Cameroon. *Journal of General Virology* 84: 2333–2341.

Ndong-Atome, G.-R., Makuwa, M., Njouom, R., Branger, M., Brun-Vézinet, F., Mahé, A., Rousset, D. and Kazanji, M. (2008). Hepatitis C virus prevalence and genetic diversity among pregnant women in Gabon, Central Africa. *BMC Infectious Diseases* 8: 82.

Nerrienet, E., Pouillot, R., Lachenal, G., Njouom, R., Mfoupouendoun, J., Bilong, C., Mauclere, P., Pasquier, C. and Ayouba, A. (2005). Hepatitis C virus infection in Cameroon: a cohort-effect. *Journal of Medical Virology* 76: 208–214.

Njouom, R., Frost, E., Deslandes, S., Mamadou-Yaya, F., Labbé, A.-C., Pouillot, R., Mbélesso, P., Mbadingai, S., Rousset, D. and Pépin, J. (2009). Predominance of hepatitis C virus genotype 4 infection and rapid transmission between 1935 and 1965 in the Central African Republic. *Journal of General Virology* 90: 2452–2456.

Njouom, R., Nerrienet, E., Dubois, M., Lachenal, G., Rousset, D., Vessière, A., Ayouba, A., Pasquier, C. and Pouillot, R. (2007). The hepatitis C virus epidemic in Cameroon: genetic evidence for rapid transmission between 1920 and 1960. *Infection, Genetics and Evolution* 7: 361–367.

Njouom, R., Caron, M., Besson, G., Ndong-Atome, G-R., Makuwa, M., Pouillot, R., Nkoghé, D., Leroy, E. and Kazanji, M. (2012). Phylogeography, risk factors and genetic history of hepatitis C virus in Gabon, central Africa. *PLoS One* 7: e42002.

Pépin, J. (2011). The Origins of AIDS. Cambridge University Press, UK.

Pépin, J. and Labbé, A.-C. (2008). Noble goals, unforeseen consequences: control of tropical diseases in colonial Central Africa and the iatrogenic transmission of blood-borne viruses. *Tropical Medicine and International Health* 13: 744–753.

Pépin, J., Labbé, A.-C., Mamadou-Yaya, F., Mbélesso, P., Mbadingai, S., Deslandes, S., Locas, M.-C. and Frost, E. (2010a). Iatrogenic transmission of human T cell lymphotropic virus type 1 and hepatitis C virus through parenteral treatment and chemoprophylaxis of sleeping sickness in colonial equatorial Africa. *Clinical Infectious Diseases* 51: 777–784.

Pépin, J., Lavoie, M., Pybus, O.G., Pouillot, R., Foupouapouognigni, Y., Rousset, D., Labbé, A.-C. and Njouom, R. (2010b). Risk factors for hepatitis C virus transmission in colonial Cameroon. *Clinical Infectious Diseases* 51: 768–776.

Pilot-Matias, T.J., Carrick, R.J., Coleman, P.F., Leary, T.P., Surowy, T.K., Simons, J.N., Muerhoff, A.S., Buijk, S.L., Chalmers, M.L., Dawson, G. J., Desai, S. M. and Mushahwar, I. K. (1996). Expression of the GB virus C E2 glycoprotein using the semliki forest virus vector system and its utility as a serologic marker. *Virology* 225: 282–292.

de Pondé, R.A. (2011). Hidden hazards of HCV transmission. *Medical Microbiology and Immunology* 200: 7–11.

Pybus, O.G., Drummond, A.J., Nakano, T., Robertson, B.H. and Rambaut, A. (2003). The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Molecular Biology and Evolution* 20: 381–387.

Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C. and Harvey, P.H. (2001). The epidemic behavior of the hepatitis C virus. *Science* 292: 2323–2325.

Pybus, O.G., Markov, P.V., Wu, A. and Tatem, A.J. (2007). Investigating the endemic transmission of the hepatitis C virus. *International Journal for Parasitology* 37: 839–849.

Ray, S.C., Arthur, R.R., Carella, A., Bukh, J. and Thomas, D.L. (2000). Genetic epidemiology of hepatitis C virus throughout Egypt. *Journal of Infectious Diseases* 182: 698–707.

Sharp, C.P., Vermeulen, M., Nébié, Y., Djoko, C.F., LeBreton, M., Tamoufe, U., Rimoin, A.W., Kayembe, P.K., Carr, J.K., Servant-Delmas, A., Laperche, S., Harrison, G. L., Pybus, O. G., Delwart, E., Wolfe, N. D., Saville, A., Lefrère, J. J. and Simmonds, P. (2010). Changing epidemiology of human parvovirus 4 infection in sub-Saharan Africa. *Emerging Infectious Diseases* 16: 1605–1607.

Shepard, C.W., Finelli, L. and Alter, M.J., (2005). Global epidemiology of hepatitis C virus infection. *Lancet Infectious Diseases* 5: 558–567.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus - 15 years on. *Journal of General Virology* 85: 3173–3188.

Smith, D.B., Pathirana, S., Davidson, F., Lawlor, E., Power, J., Yap, P.L. and Simmonds, P. (1997). The origin of hepatitis C virus genotypes. *Journal of General Virology* 78: 321–328.

Stapleton, J.T., Foung, S., Muerhoff, A.S., Bukh, J. and Simmonds, P. (2011). The GB viruses: a review and proposed classification of GBV-A, GBV-C (HGV), and GBV-D in genus Pegivirus within the family Flaviviridae. *Journal of General Virology* 92: 233–246.

Strickland, G.T. (2010). An epidemic of hepatitis C virus infection while treating endemic infectious diseases in equatorial Africa more than a half century ago: did it also jump-start the AIDS pandemic? *Clinical infectious Diseases* 51: 785–787.

Tacke, M., Schmolke, S., Schlueter, V., Sauleda, S., Esteban, J.I., Tanaka, E., Kiyosawa, K., Alter, H.J., Schmitt, U., Hess, G., Ofenloch-Haehnle, B. and Engel, A. M. (1997). Humoral immune response to the E2 protein of hepatitis G virus is associated with long-term recovery from infection and reveals a high frequency of hepatitis G virus exposure among healthy blood donors. *Hepatology* 26: 1626–1633.

Tibbs, C.J., Palmer, S.J., Coker, R., Clark, S.K., Parsons, G.M., Hojvat, S., Peterson, D. and Banatvala, J.E. (1991). Prevalence of hepatitis C in tropical communities: the importance of confirmatory assays. *Journal of Medical Virology* 34: 143–147.

Zwickl, D.J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.

3 PHYLOGEOGRAPHY AND EPIDEMIC HISTORY OF HEPATITIS C VIRUS GENOTYPE 4 IN AFRICA

Published as:

Iles JC, Raghwani J, Harrison GLA, Pepin J, Djoko CF, Tamoufe U, Lebreton M, Schneider BS, Fair JN, Tshala FM, Kayembe PK, Muyembe JJ, Edidi-Basepeo S, Wolfe ND, Simmonds P, Klenerman P, Pybus OG (2014) Phylogeography and molecular epidemiology of hepatitis C virus genotype 4 in Africa. Virology *464-465: 233-243*

3.1 SUMMARY OF AUTHORSHIP

JI performed all laboratory analysis and computational analysis, created all figures and wrote this chapter. JR provided assistance and advice with BEAST coalescent analysis. GLAH provided assistance and advice with HCV plasma extraction, RT-PCR and sequencing. PK and OGP provided supervisory support. All others were involved in the collection and provision of the blood samples analysed.

3.2 ABSTRACT

HCV genotype 4 is prevalent in many African countries, yet little is known about the genotype's epidemic history on the continent. We present a comprehensive study of the molecular epidemiology of genotype 4. To address the deficit of data from the Democratic Republic of the Congo (DRC) we PCR amplified 60 new HCV isolates from the DRC, resulting in 33 core- and 48 NS5B-region sequences. Our data, together with genotype 4 database sequences, were analysed using Bayesian phylogenetic approaches. We find three well-supported intra-genotypic lineages and estimate that the genotype 4 common ancestor existed around 1733 (1650-1805). We show that genotype 4 originated in central Africa and that multiple lineages have been exported to north Africa since ~1850, including subtype 4a which dominates the epidemic in Egypt. We speculate on the causes of the historical intra-continental spread of genotype 4, including population movements during World War 2.

3.3 INTRODUCTION

Hepatitis C virus (HCV) is a major human pathogen that causes substantial morbidity and mortality worldwide. It is estimated that more than 185 million people are seropositive for HCV and that there are 3-4 million new infections each year (Mohd Hanafiah *et al.*, 2013). Infection with the virus is typically asymptomatic or unspecific in the initial stages, but once it progresses to long-term chronic infection it can lead to liver cirrhosis, fibrosis, and sometimes hepatocellular carcinoma (Lauer and Walker, 2001).

HCV is a genetically diverse virus that is classified into seven genotypes (1-7) with an average of 35% nucleotide divergence between strains belonging to different genotypes. All genotypes except 5 and 7 are subdivided into numerous subtypes (1a, 1b, 1c, 2a, 2b etc.) and the average nucleotide divergence between subtypes of the same genotype is around 25% (Murphy *et al.*, 2007a; Simmonds *et al.*, 1993; Smith *et al.*, 2014).

There are two major categories to HCV strains: first, the 'global epidemic' subtypes (e.g. 1a, 1b, 2a and 3a) that cause the majority of HCV infections worldwise and spread rapidly during the twentieth century (Magiorkinis *et al.*, 2009; Pybus *et al.*, 2005; Simmonds, 2004), and secondly highly divergent 'endemic' strains that are typically found in a restricted geographic area, indicating the presence of their genotype in that location for hundreds or thousands of years (Simmonds, 2004). For example, HCV genotype 2 has many endemic strains in West Africa, genotypes 1 and 4 in Central Africa and the Middle East and genotype 6 in East Asia (Candotti *et al.*, 2003; Jeannel *et al.*, 1998; Mellor *et al.*, 1995; Ndjomou, 2003; Pybus *et al.*, 2009).

The epidemiology of HCV prior to the discovery of the virus is poorly understood. Documentary evidence of past HCV transmission is difficult to establish as symptoms during acute infection are unspecific and HCV incidence before the widespread use of injections was likely too low to create notable outbreaks of disease. Further, samples available for retrospective screening that were archived before the 1970s are exceptionally rare (Gray et al. 2013). As a consequence, evolutionary analyses of contemporary HCV gene sequences using phylogenetic and coalescent-based methods have been utilised to estimate dates of viral divergence and to estimate the product of effective number of infections and generation time through time. Additionally, previous studies of genotype 2 in Africa (Markov *et al.*, 2009) and of genotype 6 in Asia (Pybus *et al.*, 2009) employed phylogeographic and molecular clock methods and provided insights into the historical geographic spread of HCV, the age of HCV genotypes and subtypes, and their recent transmission history.

To date there has been no systematic phylogeographic or evolutionary study of HCV genotype 4 as a whole. This genotype is common throughout most of Central Africa and parts of the Middle East. Recent estimates indicate that there are ~8 million people infected with HCV in Central and Eastern Sub-Saharan African, and >15 million people infected across North Africa and the Middle East (Mohd Hanafiah *et al.*, 2013). Genotype 4 (and subtype 4a in particular) dominates the HCV epidemic in Egypt, where 15% of adults are antibody-positive for HCV, with a much higher prevalence seen in older cohorts (El-Zanaty *et al.,* 2008). HCV in Egypt has been described as a 'local epidemic', whereby the transmission of one or a few subtypes rises rapidly within a region, but without the international dissemination observed for the 'global epidemic' subtypes such as 1a and 1b. High HCV seroprevalences and local epidemics associated with other subtypes of genotype 4 and have also been reported in many sub-

Saharan African countries. For example, 11.2% of people screened in rural Gabon were seropositive for HCV (Njouom *et al.*, 2012) of whom 92% were infected with genotype 4. In that study major risk factors for HCV infection were past injections, hospital admissions, and age greater than 55. In Cameroon, HCV seroprevalence was 11% in a group of high-HIV risk individuals and 16% of the HCV infections were classified as genotype 4 (Ndjomou, 2003). In a separate Cameroonian cohort comprising individuals aged over 60, HCV seroprevalence was 56% and 54% of infections were genotype 4 (Pépin *et al*, 2010a). In each of these studies HCV seroprevalence was strongly associated with age and subtypes 4a and 4r were observed.

The evolution and genetic history of genotype 4 is worthy of investigation for several reasons. Firstly, together with genotype 1, genotype 4 responds less well to interferon-based anti-HCV drug treatment than genotypes 2 and 3, especially in patients of African descent (Chen *et al.*, 2012; Rose *et al.*, 2013) and it has been hypothesised that this phenotype is a consequence of the long-term presence of genotypes 1 and 4 in Central African populations (Rose *et al.*, 2013). Secondly, there are a number of unanswered questions concerning the origin and spread of HCV genotype 4 within Africa. For example, the current distribution and past spread of genotype 4 strains among countries is unclear and the geographic source of the HCV lineages present in Egypt and the Middle East are currently unknown. Additionally, in recent years there has been rapid growth in the prevalence of HCV subtypes 4a and 4d in Europe, particularly among injecting drug users (IDUs), hence a comprehensive overview of genotype 4 diversity may prove useful for public health assessments outside Africa (Ciccozzi *et al.*, 2012; de Bruijne *et al.*, 2009; van Asten *et al.*, 2004).

The investigation of HCV evolution in Central Africa is hampered by a lack of information about its epidemiology and genetic diversity in the region. Mohd Hanafiah *et al*. (2013) define the evidentiary support for HCV prevalence in the region as 'very limited'. Although recent surveillance studies have explored the genetic diversity of HCV in Cameroon, Gabon and the Republic of Congo (Cantaloube *et al.*, 2010; Ndong-Atome *et al.*, 2008; Nerrienet *et al.*, 2005; Njouom *et al.*, 2007; 2012) there is little information about the diversity of the virus in the Democratic Republic of the Congo (DRC). This country is the second largest in Africa and has 67 million residents, making it the third most populous in the continent. However, the ongoing conflict there since 1996 has made disease surveillance difficult. The large size and central position of the DRC within Central Africa mean that phylogeographic studies of HCV in the region will be incomplete without a comprehensive survey of viral diversity in the country. Further, it is possible that the DRC harbors previously-undetected variants of the virus: the only published isolate of HCV genotype 7 was isolated from a Canadian resident who had emigrated from the DRC (Murphy *et al.*, 2007b).

At present there is little information about the genetic diversity of HCV infections in the DRC. In a previous small-scale survey of blood samples from the country (see Chapter 2) we detected HCV RNA in eleven individuals. Phylogenetic analysis of HCV core and NS5B region sequences from these samples indicated that they belonged to several classified and unclassified subtypes of genotype 4. In this study we address the deficit of HCV genetic information from the DRC with the screening and sequencing of 1999 blood samples from the country. We combine these new data with genotype 4 sequences gathered from online databases and originating from countries across Africa and worldwide. This enables us to analyse the DRC samples in context

with the larger diversity of HCV genotype 4 viruses and to investigate the long-term evolutionary history of the virus within the African continent.

3.4 METHODS

3.4.1 Study Population

A total of 1999 EDTA blood samples were collected from informed consenting members of the uniformed services as part of a screening program for infectious diseases. This collection has been studied previously for HIV-1 (Djoko *et al.*, 2011), human parvovirus 4 (Sharp *et al.*, 2010) and human pegiviruses (Chapter 2). A preliminary small scale survey (*n*=299) of this collection for HCV discovered HCV RNA in ~4% of samples (Chapter 2). Collection took place during 2007 in Kinshasa, capital of the DRC. The samples were anonymised although patient year of birth were available for most. All samples were from male individuals, whose mean age was 44 (range 22–77 years).

3.4.2 Screening, RT-PCR and sequencing

All samples were tested for HCV RNA. Viral RNA was extracted from sera using the Nucleospin 96 RNA kit (Macherey-Nagel) as per the manufacturer's instructions. The reaction product was screened for HCV RNA with a one-step RT-PCR amplification of the 5'UTR region using Superscript III with Platinum *taq* (Invitrogen, Life Sciences). Samples positive for HCV RNA were subsequently amplified and sequenced in the core and NS5B regions using the same enzymes as used for the 5'UTR, noted above. Controls were run in parallel at each step. Primers were obtained from previous studies or were designed using a large alignment of whole HCV genome sequences that included subtypes belonging to all 7 genotypes (Table 3.1). The internal primers were used for sequencing with BigDye Terminator (Applied Biosystems). Traces were

examined using 4Peaks (Nucleobytes). A total of 34 core sequences (accession numbers KF813071-KF813095, KJ408429-KJ408436 and KJ416140) and 48 NS5B sequences (KF826150-KF826197) were obtained in this study.

Primer Name	Source	Sequence (5'-3')	Position
Murphy 5'UTR F	(Murphy *et al.*, 2007b)	GAAAGCGTCTAGCCATGGCGTTAGT	71-95
Murphy 5'UTR R	(Murphy *et al.*, 2007b)	CTCGCAAGCACCCTATCAGG	311-292
5' UTR Ex 400F	This study	CCTTGTGGTACTGCCTGATAG	279-299
CHV core 980 Rex	This study	AGTGCCARRAGGAACATAGA	883-864
5'UTR In 405 F	This study	CTGATAGGGTGCTTGCGAGTG	293-313
CHV core 973 Rin	This study	AGTGCCARRAGGAAGATAGARAA	883-861
NS5B Ex 8274 F	This study	TGGGGATCCCGTATGATACCCGCTGCTTTGA	8245-8275
NS5B Ex 8616 R	This study	CGGAATTCCTGGTCATAGCCTCCGTGAA	8643-8616
NS5B In 8378 for	This study	GACACCCGCTGCTTTGACTC	8259-8278
NS5B In 8611 rev	This study	GAGTCTTCACGGAGGCTATGACNAGGTA	8638-8611

Table 3.1: *Details of primers used in this study. Position is numbered relative to isolate H77 (Genbank accession number AF009606).*

3.4.3 Sequence Collation

All available genotype 4 sequences were downloaded from the Los Alamos HCV sequence database (Kuiken *et al.*, 2005) and from GenBank. Sequences were retained if they spanned either of the two subgenomic regions sequenced in this study: core (positions 342-1265 relative to H77) or NS5B (positions 8265-8624). Only one sample per region from each infected individual was retained and sequences from non-human subjects were excluded, as were sequence fragments shorter than 200 nucleotides. We noted a disproportionate number of sequences from subtype 4a, largely resulting from the high number of published studies concerning the Egyptian HCV epidemic. To

bring the number of 4a sequences approximately in line with those of other subtypes we randomly removed sequences separately from each of the three main sub-genotypic lineages of the 4a phylogeny, thereby maintaining the full genetic diversity of the subtype in our data set. We also reduced in a similar manner the disproportionate number of subtype 4d sequences sampled in France.

The reference database sequences were collated with the new sequences obtained in this study, resulting in a total of 806 isolates across all genome regions. All sequences were aligned by hand using Se-Al v2.0 (available from http://tree.bio.ed.ac.uk), resulting in a 'core alignment' containing 177 core sequences and an 'NS5B alignment' containing 765 NS5B sequences. Subsequently, a 'concatenated alignment' was created by combining and concatenating core and NS5B sequences if they were sampled from the same individual and covered both genome regions. The resulting joint alignment contained 136 taxa. In order to best estimate the branching order among HCV subtypes within genotype 4, we also compiled a 'whole genome' alignment that contained all genotype 4 reference genome sequences described in Smith *et al.* (2014).

For each isolate we surveyed online databases and the primary literature for two pieces of information: year of sampling and country of origin. Most isolates (83%) were sampled from African or Middle Eastern countries. A search of the primary literature revealed that some HCV strains sampled in Europe or North America represent infections from recent immigrants from Africa or the Middle East. In these instances, the 'country of origin' of the infection is defined as the country from which the individual emigrated. A summary of the geographic distribution of the sequences used in this study is provided in Figure 3.1.

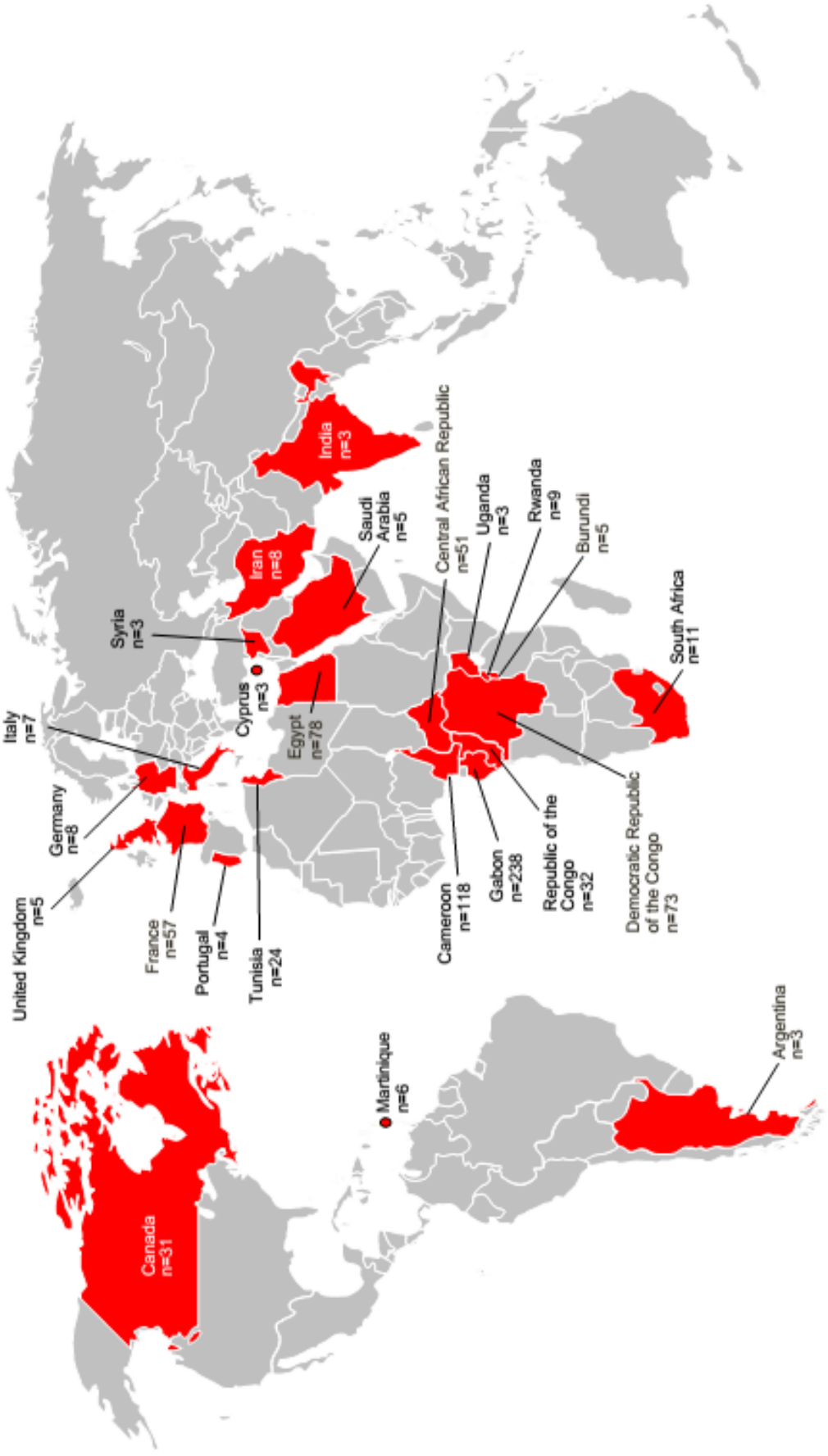

India n=3
Saudi Arabia n=5
Iran n=8
Syria n=3
Cyprus n=3
Egypt n=78
Italy n=7
Germany n=8
United Kingdom n=5
France n=57
Portugal n=4
Tunisia n=24
Central African Republic n=51
Uganda n=3
Rwanda n=9
Burundi n=5
South Africa n=11
Cameroon n=118
Gabon n=238
Republic of the Congo n=32
Democratic Republic of the Congo n=73
Martinique n=6
Canada n=31
Argentina n=3

Figure 3.1 (previous page). *Map of the world showing the country of origin and sample size of core and NS5B samples used in this study. Where a single isolate yielded multiple sequences, they were only counted as one sample on this figure. Countries contributing less than three samples are not shown.*

3.4.4 Phylogenetic Analysis

Phylogenies were estimated for the core, NS5B and whole genome alignments using maximum likelihood (ML) as implemented in GARLI v0.951 (Zwickl, 2006). The analysis used a General Time-Reversible (GTR) nucleotide substitution model, estimated base frequencies, and a gamma distribution model of among-site rate variation. Statistical support for phylogenetic clustering was calculated using an ML bootstrap approach with 500 bootstrap replicates; bootstrap scores were summarized using TreeAnnotator (http://beast.bio.ed.ac.uk/TreeAnnotator). Phylogenies were visualized and annotated using FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree). Newly-generated sequences were classified by computing p-distances to the HCV subtype reference sequences provided in Smith *et al.*, (2014). A p-distance threshold of <0.1 was used to assign subtypes. P-distances between sequences were calculated using DNAdist in the Phylip package (Felsenstein, 1989).

3.4.5 Calibration of the molecular clock

Molecular clock models can be used to reconstruct the evolutionary history of HCV genotype 4 on a timescale of years. As in previous studies, we cannot directly estimate reliable HCV evolutionary rates from the alignments in hand as the range of sampling times is too narrow (Pybus *et al.*, 2009; Salemi and Vandamme, 2002). Therefore we estimated evolutionary rates for the core and NS5B regions using independent sets of HCV sequences that have been shown to contain good temporal information. These

estimates were then used as informative prior distributions for evolutionary rate parameters in all subsequent Bayesian evolutionary analyses (see next section).

Gray *et al.* (2011) undertook a comprehensive analysis of HCV evolutionary rates and we use their alignments to estimate rates that are specific to the core and NS5B genome regions sequenced here (positions 342-945 and 8265-8624, respectively). We analysed two alignments, comprising 65 subtype 1a sequences and 54 subtype 1b sequences, respectively (Gray et al. 2011). These rates are likely to be accurate for our genotype 4 study because genotypes 1 and 4 are more closely related than other subtypes (Salemi and Vandamme, 2002) and because HCV evolutionary rates vary considerably more between genome regions than they do between genotypes and subtypes (Gray *et al.*, 2011). Evolutionary rates were estimated using the Bayesian Markov Chain Monte Carlo (MCMC) inference method implemented in BEAST v1.7.5 (Drummond *et al.*, 2012). These analyses employed a SDR06 nucleotide substitution model (two independent HKY+Γ substitution models – one for the first and second codon positions, and one for the third), an uncorrelated lognormal relaxed molecular clock model, and a Bayesian skyline plot coalescent model (Shapiro *et al.*, 2006). Nucleotide frequencies were estimated from the data.

Rates of molecular evolution were estimated for three different partitions of the whole genome alignment: (i) a core partition (sites 342-945), (ii) a NS5B partition (sites 8265-8624), and (iii) a concatenated core+NS5B partition (sites 342-945 plus 8265-8624). The rate parameters estimated for these three partitions are shown in Table 3.2. Each MCMC analysis was run for at least 100,000,000 states. Table 3.2 also includes evolutionary rate estimates for whole genome sequences, which were taken directly from Gray et al. (2011).

Genome region	Genome positions	Estimated nucleotide substitution rate (subs/site/year)	95% credible region
(i) Core	342-944	5.39×10^{-4}	$3.41 – 7.46 \times 10^{-4}$
(ii) NS5B	8274-8612	9.87×10^{-4}	$6.74 – 14.4 \times 10^{-4}$
(iii) Concatenated (Core + NS5B)	342-944 and 8274-8612	7.43×10^{-4}	$4.91 – 10.4 \times 10^{-4}$
(iv) Complete genome	342-9374	13.5×10^{-4}	$9.97 – 17.0 \times 10^{-4}$

Table 3.2: *Estimated evolutionary rate parameters. Genome positions are numbered relative to isolate H77 (Genbank accession number AF009606).*

3.4.6 Bayesian Evolutionary Analysis

To estimate the epidemic and movement history of HCV genotype 4 we analysed the 'whole genome' and 'concatenated' alignments using the Bayesian Markov Chain Monte Carlo (MCMC) inference method implemented in BEAST v. 1.7.5 (Drummond *et al.*, 2012).

As with the evolutionary rate estimation analyses described above, we used the SDR06 substitution model, an uncorrelated lognormal relaxed molecular clock, and a Bayesian Skyline coalescent model with 10 groups. For both the 'whole genome' and 'concatenated' data sets, Bayesian model selection tests showed that the SDR06 substitution model substantially outperformed the GTR+Γ model (Bayes Factor >100; calculated using Tracer v1.5). For both data sets, a normal prior distribution was placed on the mean evolutionary rate parameter, such that the mean and variance of the prior distribution matched the 'concatenated' and 'whole genome' rate estimates shown in Table 2. Each MCMC run contained 200 million states, sampled once every 5000

states; trees were sampled every 50000 states. Multiple MCMC runs were calculated to ensure convergence and were combined to increase the accuracy of parameter estimates. MCMC convergence and effective sample sizes were monitored using Tracer v. 1.5. Maximum clade credibility trees were calculated and annotated using TreeAnnotator 1.7.5 (Drummond *et al.*, 2012). FigTree v1.3.1 was used to colour lineages according to their sampling location using the parsimony criterion. We deliberately chose not to apply more sophisticated Bayesian discrete state phylogeographic models (*e.g.* Lemey *et al.* 2009) to our data. Such models are highly parametric and are unlikely to be informative when applied to phylogenies with comparatively few location state changes and no sampled sequences close to the phylogeny root (such as the tree presented in Figure 3.7).

3.5 RESULTS

3.5.1 Age distribution

Of the 1999 samples tested, 3% (*n*=60) were positive for HCV 5'UTR RNA. Of these 60 samples, 33 produced core sequence and 48 produced NS5B sequence. These results are broadly consistent with our pilot study which detected HCV RNA in 3.7% (*n*=11) of samples and yielded 9 core sequences and 11 NS5B sequences (Chapter 2).

Figure 3.2 shows the age distribution of HCV RNA positivity in our population, (including data from the pilot study). Samples were scored as HCV RNA positive if sequence was obtained from at least one of the 5'UTR, core or NS5B regions. When grouped according to date of birth, samples from older individuals are more likely to contain HCV RNA than younger ones; the highest prevalence (9.2%) was observed in the 1930-1945 cohort. Samples from patients born in or before 1945 were significantly more likely to contain HCV RNA than those born after (p<0.001 using Fisher's exact

test), and in addition significantly more likely than would be expected with linear prevalence growth proportional to age ($p<0.0001$ using chi-square goodness-of-fit test).

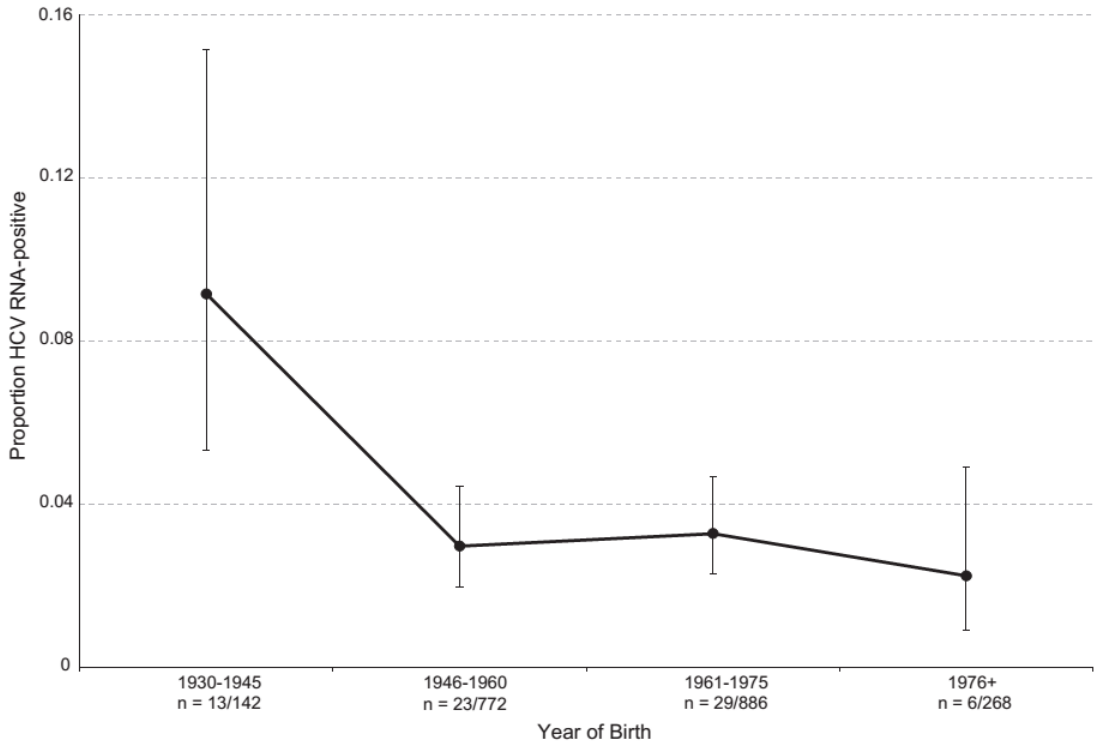


Figure 3.2. *The age distribution of HCV RNA positivity among 2298 blood samples from the DRC. Samples were assigned to one of four age categories by year of birth. Numbers below each category indicate the number of positive samples/total number of samples. Fifteen samples (all negative) did not have date of birth information and are not included. The y-axis shows the proportion of samples in each age category that were HCV RNA-positive. The errors bars represent 95% confidence limits of this proportion, estimated using the adjusted Wald method (Agresti and Coull 1998).*

Subtype	Samples (Core)	Samples (NS5B)	Most common sampling location of Core and NS5B sequences
4a	11	46	Egypt
4b	1	10	Democratic Republic of the Congo
4c	35	80	Gabon
4d	17	45	France
4e	29	142	Gabon
4f	16	106	Cameroon
4g	4	7	Gabon
4h	1	15	Cameroon
4k	23	77	Democratic Republic of the Congo
4l	1	10	Egypt
4m	2	14	Egypt
4n	1	7	Egypt
4o	3	24	Egypt
4p	1	13	Cameroon
4q	2	6	Rwanda
4r	12	40	Democratic Republic of the Congo
4t	9	31	Cameroon
4u	1	14	Egypt
4v	2	7	Rwanda
4w	5	3	Portugal
4car	0	26	Central African Republic
4drc	0	3	Democratic Republic of the Congo
Unclassified	0	45	

***Table 3.3**. Summary of subtypes seen in samples gathered in this study.*

3.5.2 Phylogenetic Analysis

Maximum-likelihood phylogenies estimated from the core and NS5B alignments are too large to display here and are thus provided in Supplementary Figures 7.3.1 and 7.3.2. Sequences from all formally defined subtypes of genotype 4 (Smith *et al.* 2014) were present in the alignments, as well as sequences from subtypes that are not formally defined due to a lack of whole-genome reference sequences (e.g. 4e, 4h, 4u). The numbers of sequences assigned to each subtype and the most common locations of sampling of each subtype are shown in Table 3.3.

A total of 74 samples were genetically too divergent to be assigned to a known subtype, 26 of which appear to belong to a provisional subtype circulating in the Central African Republic (referred to here as 4car). Sequences from five isolates

(including one obtained in this study) were discordant, i.e. they grouped into different subtypes in the core and NS5B alignments. The HCV-positive samples sequenced in this study were genetically diverse and were classified as belonging to subtypes 4c (n=17), 4h (n=2), 4k (n=18), and 4r (n=8). One sample (DRC0387) could not be classified into any known subtype but grouped with two unclassified isolates (DRC2431 and DRC2450) from our pilot study (Chapter 2). This cluster of three strains is denoted 4drc here and represents a potentially new subtype (see Table 3.3). One sample (isolate DRC1427) generated a core sequence that was classified as 4q and a NS5B sequence classified as 4c. Few nodes in the core and NS5B maximum likelihood trees had high bootstrap support (see Figures 7.3.1 and 7.3.2), but that is not unexpected for phylogenies estimated from these short subgenomic regions (as previously noted in Pybus *et al.* 2009 and elsewhere).

We discerned four clusters (denoted C1, C2, C3 and C4) that contained multiple sequences from our study population (see Figures 7.3.1 and 7.3.2). Cluster C1 was present in subtype 4c in both the core and NS5B trees and contained ten new DRC isolates (plus one from the pilot study). Cluster C2 was also found in subtype 4c and comprised six isolates from this study in the NS5B tree (only four of which are present in the core tree). Cluster C3 contained 12 subtype 4k samples from this study (plus two from the pilot study). C3 also included 15 samples from Tunisia in the NS5B tree, whereas in the core tree it included seven sequences from Gabon. Finally, cluster C4 was present in subtype 4r. In the NS5B tree C4 contained eight sequences from this study (plus two from the pilot study).

Samples from our study population also appeared outside of these four clusters. Specifically, two further isolates were placed inside subtype 4c, nine within 4k, two within 4h, one within 4q and three grouped in the unclassified lineage 4drc (see

above). The pair of subtype 4h sequences grouped significantly, together with a strain from Brazzaville (Republic of the Congo; GU088141).

3.5.3 Whole Genome Phylogenies

Figure 3.3 shows a maximum clade credibility (MCC) phylogeny of HCV genotype 4, obtained from the Bayesian molecular clock analysis of the whole genome sequences. An equivalent maximum likelihood phylogeny is provided in Figure 3.4. Up to three genomes per subtype were included and no complete genome sequences were available for some subtypes (subtypes 4e, 4h, 4s and 4u). As Figure 3.3 indicates, there is a great deal of phylogenetic structure above the subtype level. This structure is supported by high posterior probability values in the Bayesian phylogeny and by high bootstrap values in the ML phylogeny (Figure 3.3 and 3.4).

Three intra-genotypic lineages can be discerned and are denoted here L1, L2, and L3. These lineages correspond to the three clades closest to the root of genotype 4 with strong statistical support. Lineage L1 contains subtypes 4a, 4c, 4d, 4l, 4m, 4n, 4o, 4q and 4v, L2 contains subtypes 4b and 4w and L3 contains subtypes 4g, 4k, and 4r. In the ML phylogeny, L2 is present as an outgroup (Figure 3.4) whereas in the Bayesian tree it is placed as a sister group to L3, albeit with a comparatively low posterior probability of 0.78 (Figure 3.3).

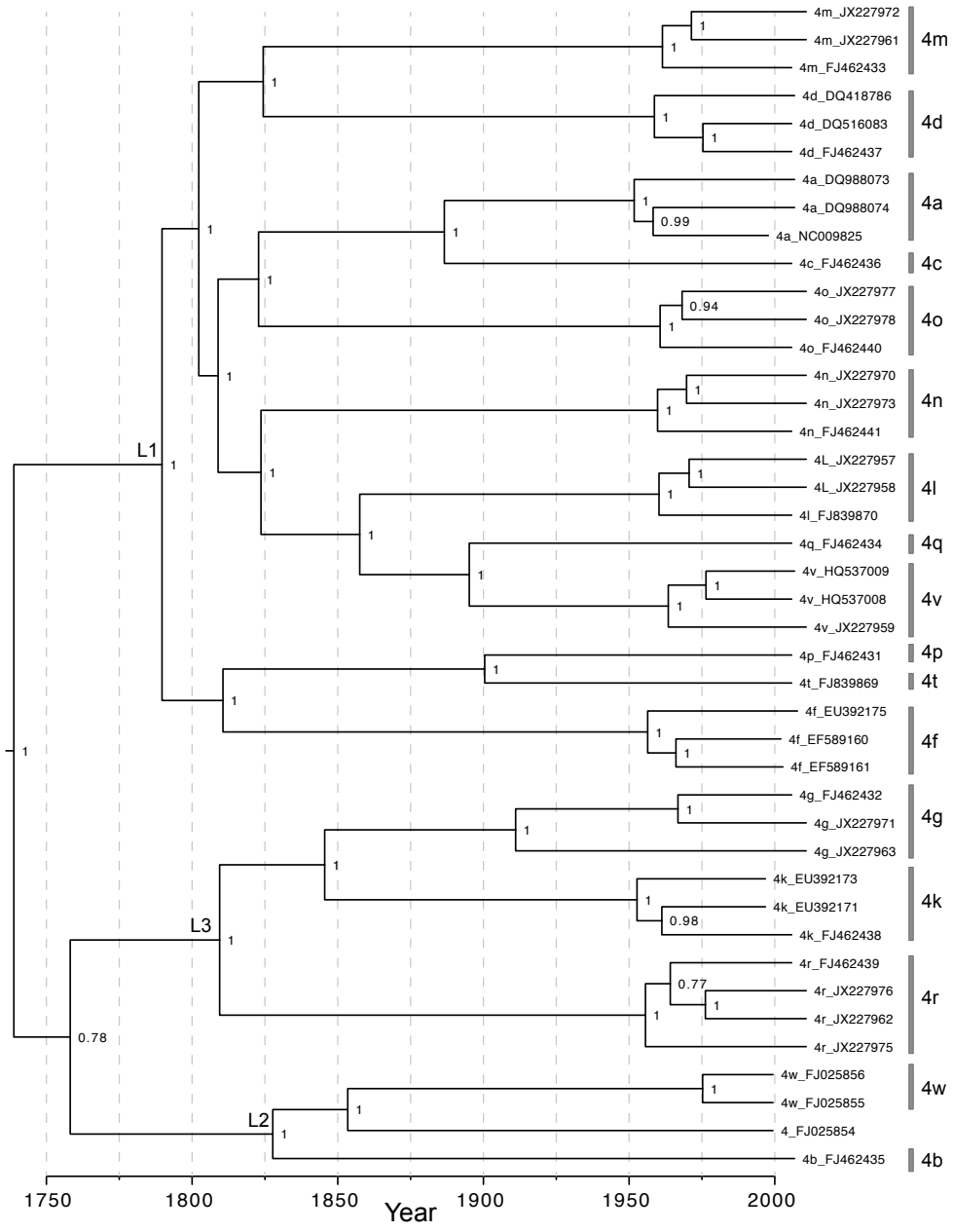


Figure 3.3. *Maximum clade credibility molecular clock phylogeny, estimated from the whole genome alignment. Branch lengths represent time (see scale bar at the bottom of the figure). Posterior clade probabilities are shown next to each node. Sequences are labelled with their subtype and accession number. Subtypes are indicated on the right side of the diagram. The three intra-genotypic lineages discussed in the main text are labelled L1, L2, and L3.*

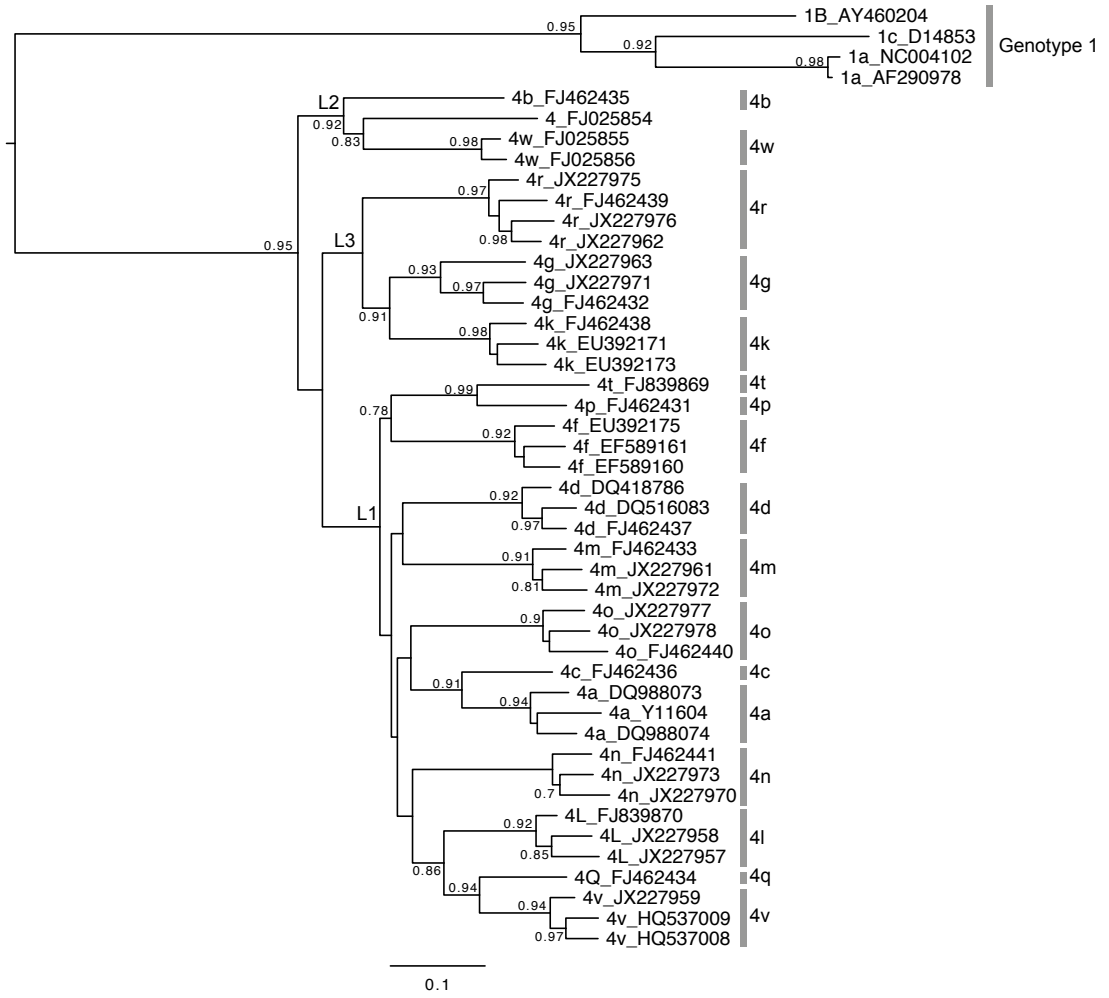


***Figure 3.4**. Estimated maximum likelihood phylogeny for the whole genome alignment. Bootstrap scores >70% are shown next to each node and the phylogeny is rooted using a genotype 1 outgroup. Branch lengths are in units of expected substitutions per site (see scale bar at bottom of figure). Sequences are labelled with their subtype and accession number. Subtypes are indicated on the right side of the diagram. The three intra-genotypic lineages discussed in the main text are labelled L1, L2, and L3.*

The Bayesian molecular clock analysis provided an estimate of the date of the most recent common ancestor (MRCA) of genotype 4, which was 1733 (95% HPD credible region = 1650-1805). This date is more recent than some previous estimates for the origin of genotype 4. Pybus *et al*. (2001) estimated HCV evolutionary rates from a

small data set of dated sequences and used these to infer that the MRCA of genotype 4 existed about 350 years ago. Njouom *et al.* (2007) used the same rate estimates during a more comprehensive Bayesian phylogenetic analysis and dated the MRCA of genotype 4 to 1541 (95% CIs: 1343-1698). The more recent date estimated here is likely to be more accurate because (i) it is based on whole genome sequences rather than small subgenomic fragments and (ii) it employs new HCV evolutionary rates that were estimated using larger data sets of dated sequences and more powerful methods of analysis (Gray et al. 2011). We note that the time to the MRCA provided here could be underestimated as a result of strong purifying selection (Wertheim and Kosakovsky Pond 2011) or overestimated due to strong among-branch rate variation (Wertheim et al. 2012). However, Markov et al. (2012) showed that molecular clock estimates of HCV lineage movement between Africa and the Americas matched the known timeframe of the trans-Atlantic slave trade. This suggests that among-subtype HCV divergence dates within a genotype can be estimated with reasonable accuracy.

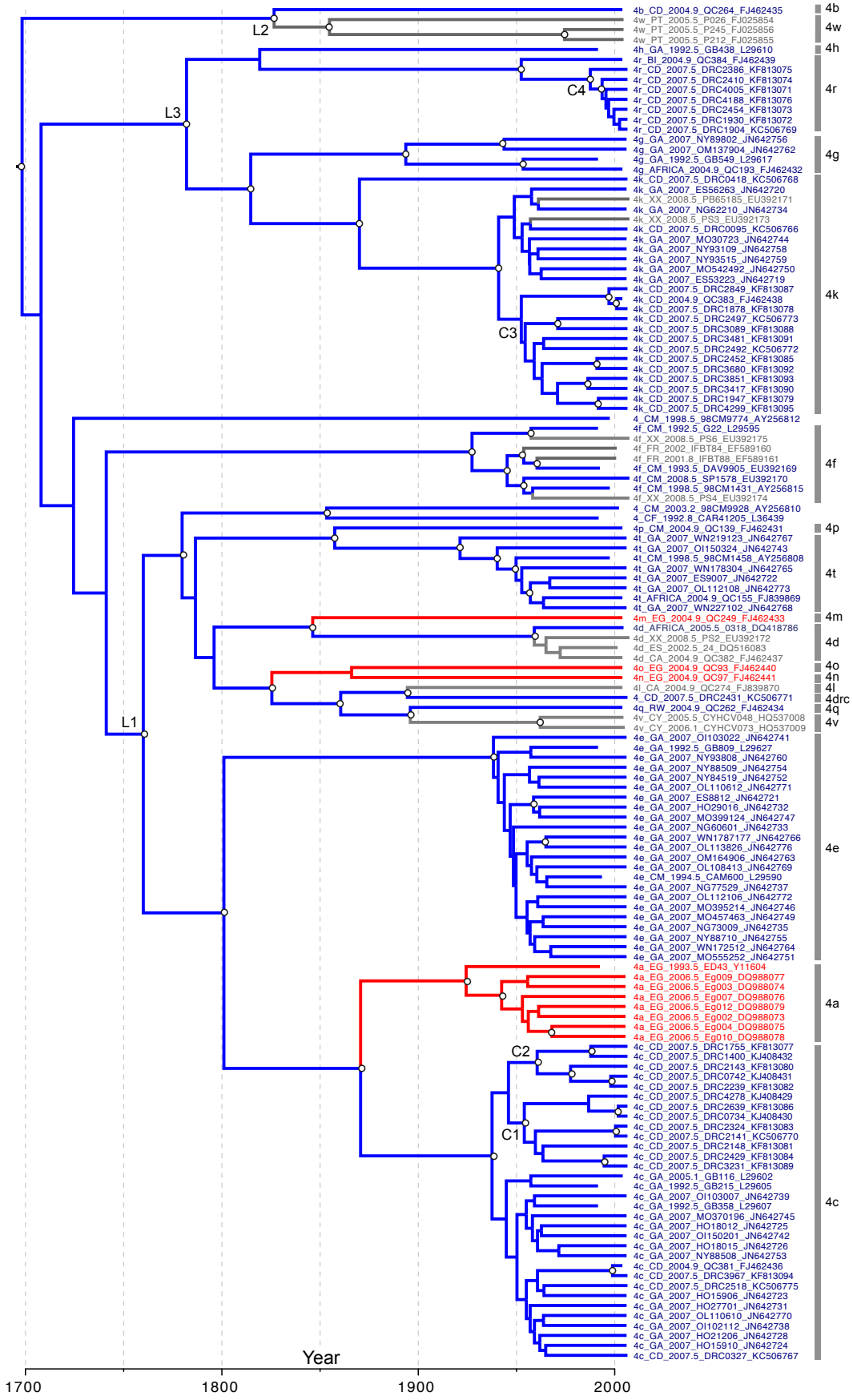
3.5.4 Concatenated alignment Phylogenies

Figure 3.5 shows the MCC phylogeny obtained from the Bayesian molecular clock analysis of the concatenated sequences (core plus NS5B). As in the whole genome phylogenies, all subtypes were monophyletic, and three well-supported intra-genotypic lineages were again observed. The greater number of subtypes present in concatenated analysis indicates that L1 contains subtypes 4a, 4c, 4d, 4e, 4l, 4m, 4n, 4o, 4p, 4q, 4t and 4v, L2 contains subtypes 4b and 4w, and lineage L3 contains subtypes 4g, 4h, 4k and 4r. L2 is again placed in a basal position (as in the full-genome ML phylogeny; Figure 3.4). However the weak posterior probability for this placement (0.21) suggests that L2 may in reality be a sister group to L3 (as in the full genome Bayesian phylogeny; Figure 3.3). Subtype 4f plus one divergent isolate from Cameroon

(98CM9774) did not group into any of these three well supported intra-genotypic lineages.

The concatenated alignment yielded an estimate of the date of the genotype 4 common ancestor of 1687 (95% HPD credible region: 1538, 1811). This estimate overlaps considerably with the estimate from the whole genome alignment and the upper 95% HPD limits of both are almost identical. The credible region for the concatenated alignment is wider than that for the whole genome alignment, likely reflecting the reduced phylogenetic information in the former.

Figure 3.5 (next page). *Maximum clade credibility molecular clock phylogeny, estimated from the concatenated alignment. Taxa are coloured according to location of sequence origin (blue = sub-Saharan Africa; red = Middle East and North Africa; grey = rest of the world). The locations of internal branches were inferred using parsimony and are coloured similarly. Branch lengths represent time (see scale bar at the bottom of the figure). Nodes with a posterior probability >0.9 are labelled with a white circle. Sequences are labelled as follows: subtype, sampling location using two-letter country codes (ISO 3166; see Table 3.4), sampling date, isolate name, accession number. XX represents an unknown location. Subtypes are indicated on the right side of the diagram. The three intra-genotypic lineages discussed in the main text are labelled L1, L2, and L3. The four clusters of samples obtained in this study discussed in the main text are labelled C1, C2, C3, and C4.*


4b_CD_2004.9_QC264_FJ462435
4w_PT_2005.5_P026_FJ025854
4w_PT_2005.5_P245_FJ025856
4w_PT_2005.5_P212_FJ025855
4h_GA_1992.5_GB438_L29610
4r_BI_2004.9_QC384_FJ462439
4r_CD_2007.5_DRC2386_KF813075
4r_CD_2007.5_DRC2410_KF813074
4r_CD_2007.5_DRC4005_KF813071
4r_CD_2007.5_DRC4188_KF813076
4r_CD_2007.5_DRC2454_KF813073
4r_CD_2007.5_DRC1930_KF813072
4r_CD_2007.5_DRC1904_KC506769
4g_GA_2007_NY89802_JN642756
4g_GA_2007_OM137904_JN642762
4g_GA_1992.5_GB549_L29617
4g_AFRICA_2004.9_QC193_FJ462432
4k_CD_2007.5_DRC0418_KC506768
4k_GA_2007_ES56263_JN642720
4k_XX_2008.5_PS65185_EU392171
4k_GA_2007_NG62210_JN642734
4k_XX_2008.5_PS3_EU392173
4k_CD_2007.5_DRC0095_KC506766
4k_GA_2007_MO30723_JN642744
4k_GA_2007_NY93109_JN642758
4k_GA_2007_NY93515_JN642759
4k_GA_2007_MO542492_JN642750
4k_GA_2007_ES53223_JN642719
4k_CD_2007.5_DRC2849_KF813087
4k_CD_2004.9_QC383_FJ462438
4k_CD_2007.5_DRC1878_KF813078
4k_CD_2007.5_DRC2497_KC506773
4k_CD_2007.5_DRC3089_KF813088
4k_CD_2007.5_DRC3481_KF813091
4k_CD_2007.5_DRC2492_KC506772
4k_CD_2007.5_DRC2452_KF813085
4k_CD_2007.5_DRC3680_KF813092
4k_CD_2007.5_DRC3851_KF813093
4k_CD_2007.5_DRC3417_KF813090
4k_CD_2007.5_DRC1947_KF813079
4k_CD_2007.5_DRC4299_KF813095
4_CM_1998.5_98CM9774_AY256812
4f_CM_1992.5_G22_L29595
4f_XX_2008.5_PS6_EU392175
4f_FR_2002_IFBT84_EF589160
4f_FR_2001.8_IFBT88_EF589161
4f_CM_1993.5_DAV9905_EU392169
4f_CM_2008.5_SP1578_EU392170
4f_CM_1998.5_98CM1431_AY256815
4f_XX_2008.5_PS4_EU392174
4_CM_2003.2_98CM9928_AY256810
4_CF_1992.8_CAR41205_L36439
4p_CM_2004.9_QC139_FJ462431
4t_GA_2007_WN219123_JN642767
4t_GA_2007_OI150324_JN642743
4t_CM_1998.5_98CM1458_AY256808
4t_GA_2007_WN178304_JN642765
4t_GA_2007_ES9007_JN642722
4t_GA_2007_OL112108_JN642773
4t_AFRICA_2004.9_QC155_FJ839869
4t_GA_2007_WN227102_JN642768
4m_EG_2004.9_QC249_FJ462433
4d_AFRICA_2005.5_0318_DQ418786
4d_XX_2008.5_PS2_EU392172
4d_ES_2002.5_24_DQ516083
4d_CA_2004.9_QC382_FJ462437
4o_EG_2004.9_QC93_FJ462440
4n_EG_2004.9_QC97_FJ462441
4l_CA_2004.9_QC274_FJ839870
4_CD_2007.5_DRC2431_KC506771
4q_RW_2004.9_QC262_FJ462434
4v_CY_2005.5_CYHCV048_HQ537008
4v_CY_2006.1_CYHCV073_HQ537009
4e_GA_2007_OI103022_JN642741
4e_GA_1992.5_GB809_L29627
4e_GA_2007_NY93808_JN642760
4e_GA_2007_NY88509_JN642754
4e_GA_2007_NY84519_JN642752
4e_GA_2007_OL110612_JN642771
4e_GA_2007_ES8812_JN642721
4e_GA_2007_HO29016_JN642732
4e_GA_2007_MO399124_JN642747
4e_GA_2007_NG60601_JN642733
4e_GA_2007_WN1787177_JN642766
4e_GA_2007_OL113826_JN642776
4e_GA_2007_OM164906_JN642763
4e_GA_2007_OL108413_JN642769
4e_CM_1994.5_CAM600_L29590
4e_GA_2007_NG77529_JN642737
4e_GA_2007_OL112106_JN642772
4e_GA_2007_MO395214_JN642746
4e_GA_2007_MO457463_JN642749
4e_GA_2007_NG73009_JN642735
4e_GA_2007_NY88710_JN642755
4e_GA_2007_WN172512_JN642764
4e_GA_2007_MO555252_JN642751
4a_EG_1993.5_ED43_Y11604
4a_EG_2006.5_Eg009_DQ988077
4a_EG_2006.5_Eg003_DQ988074
4a_EG_2006.5_Eg007_DQ988076
4a_EG_2006.5_Eg012_DQ988079
4a_EG_2006.5_Eg002_DQ988073
4a_EG_2006.5_Eg004_DQ988075
4a_EG_2006.5_Eg010_DQ988078
4c_CD_2007.5_DRC1755_KF813077
4c_CD_2007.5_DRC1400_KJ408432
4c_CD_2007.5_DRC2143_KF813080
4c_CD_2007.5_DRC0742_KJ408431
4c_CD_2007.5_DRC2239_KF813082
4c_CD_2007.5_DRC4278_KJ408429
4c_CD_2007.5_DRC2639_KF813086
4c_CD_2007.5_DRC0734_KJ408430
4c_CD_2007.5_DRC2324_KF813083
4c_CD_2007.5_DRC2141_KC506770
4c_CD_2007.5_DRC2148_KF813081
4c_CD_2007.5_DRC2429_KF813084
4c_CD_2007.5_DRC3231_KF813089
4c_GA_2005.1_GB116_L29602
4c_GA_1992.5_GB215_L29605
4c_GA_2007_OI103007_JN642739
4c_GA_1992.5_GB358_L29607
4c_GA_2007_MO370196_JN642745
4c_GA_2007_HO18012_JN642725
4c_GA_2007_OI150201_JN642742
4c_GA_2007_HO18015_JN642726
4c_GA_2007_NY88508_JN642753
4c_CD_2004.9_QC381_FJ462436
4c_CD_2007.5_DRC3967_KF813094
4c_CD_2007.5_DRC2518_KC506775
4c_GA_2007_HO15906_JN642723
4c_GA_2007_HO27701_JN642731
4c_GA_2007_OL110610_JN642770
4c_GA_2007_OI102112_JN642738
4c_GA_2007_HO21206_JN642728
4c_GA_2007_HO15910_JN642724
4c_CD_2007.5_DRC0327_KC506767
4b
4w
4h
4r
4g
4k
4f
4p
4t
4m
4d
4o
4n
4l
4drc
4q
4v
4e
4a
4c
L1
L2
L3
C1
C2
C3
C4
Year
1700
1800
1900
2000

The molecular clock analysis was also used to estimate the date of origin of the four sequence clusters (C1-4) that contained isolates from our study population (see above). The estimated date of the MRCA of cluster C1 was 1953 (95% HPD: 1938, 1982). For cluster C2 the MRCA was dated to 1960 (95% HPD: 1930, 1976) and for cluster C3 it was 1952 (95% HPD: 1928, 1975). The estimated MRCA for cluster C4 was somewhat more recent and was dated to 1987 (1976, 2000). While these clusters had low bootstrap scores in the ML trees (Supp. Figs 7.3.1 and 7.3.2), in this analysis C1, C2 and C4 are all supported with posterior probabilities >0.9, while C3 has a posterior probability of 0.79. This increased statistical support is likely due to the combined phylogenetic signal in the concatenated analysis as compared to that available in the ML reconstruction of individual genome regions.

The taxa and branches of Figure 3.5 have been coloured according to the known or inferred country of origin for each isolate. The majority of strains are from sub-Saharan Africa (blue). Sequences from North Africa/Middle East (red) and from the rest of the world (grey) cluster together *within* the greater diversity of genotype 4 from sub-Saharan Africa. Thus it is clear that genotype 4 originated in central Africa before disseminating elsewhere. Further, some isolates without location information or which were sampled outside Africa (e.g. subtype 4f and 4k strains closely related to those from Cameroon and Gabon) may also represent infections that were acquired in Central Africa. Almost all isolates sampled *outside* central Africa belong to the intra-genotypic lineage L1 (Figure 3.5). Sequences from Egypt are commonly found in subtypes 4a, 4l, 4m, 4n and 4u (Supp. Figs 7.3.1 and 7.3.2; Table 3.3). Subtype 4d is currently found in many countries, especially in Western Europe, where it is prevalent among some injecting drug user populations (de Bruijne 2009).

Figure 3.5 provides some clues as to the origin of the most common subtypes of genotype 4. For example, subtype 4a is found worldwide but is highly prevalent in Egypt, where it was likely spread by widespread injection treatment campaigns during the mid-twentieth century (e.g. Strickland 2006; Pybus et al. 2003). Using our results we can explore the question of from where the Egyptian HCV epidemic originated. Our analysis indicates that subtypes 4a and 4c diverged from each other around 1870 and that subtype 4c mainly comprises sequences from Gabon and the DRC. In the core and NS5B trees (Figures 3.3 and 3.4) we can observe strains sampled from Cameroon and the Central African Republic that are phylogenetically immediately basal to subtypes 4a, 4n and 4o. Figure 3.5 indicates that these subtypes (and subtype 4m) arrived in Egypt from Central Africa no earlier than 1825.

Fifteen subtype 4k isolates from Tunisia cluster together within the much greater diversity of subtype 4k sampled from central Africa. The Tunisian strains are closely related to DRC isolates from this study belonging to cluster C3. Although these Tunisian samples were not included in the concatenated analysis (because only NS5B sequences were available) we can combine Figure 7.3.2 and Figure 3.5 to conclude that the Tunisian subtype 4k infections originated from central Africa (possibly from the DRC) at around the time of the MRCA of cluster C3, which was 1952 (1928, 1975). A second cluster of six unclassified Tunisian isolates can be observed in the NS5B phylogeny and form a sister lineage to subtype 4o (Figure 7.3.2).

3.5.5 Bayesian Skyline Analysis

The epidemic history of genotype 4 was investigated by estimating a Bayesian skyline plot from the concatenated alignment (Figure 3.6). This plot represents the effective number of HCV infections through time, back to the estimated TMRCA of the genotype. Figure 3.6 shows that prior to the twentieth century there was an extended

period of low-level transmission of genotype 4. There is a sharp transition from this minimal transmission to rapid exponential growth starting around the 1950s. The epidemic history of genotype 4 after 1975 is much harder to discern due to very large confidence limits: either scenario after 1975 of continued growth or a stabilisation of prevalence are statistically compatible with the data. Previous studies of genotype 4 in several African countries (Central African Republic, Republic of Congo, Gabon, Cameroon) have noted similar, but earlier, transitions to rapid expansion, between 1930 and 1960 (Cantaloube *et al.*, 2010; Njouom *et al.*, 2007; 2009; 2012).

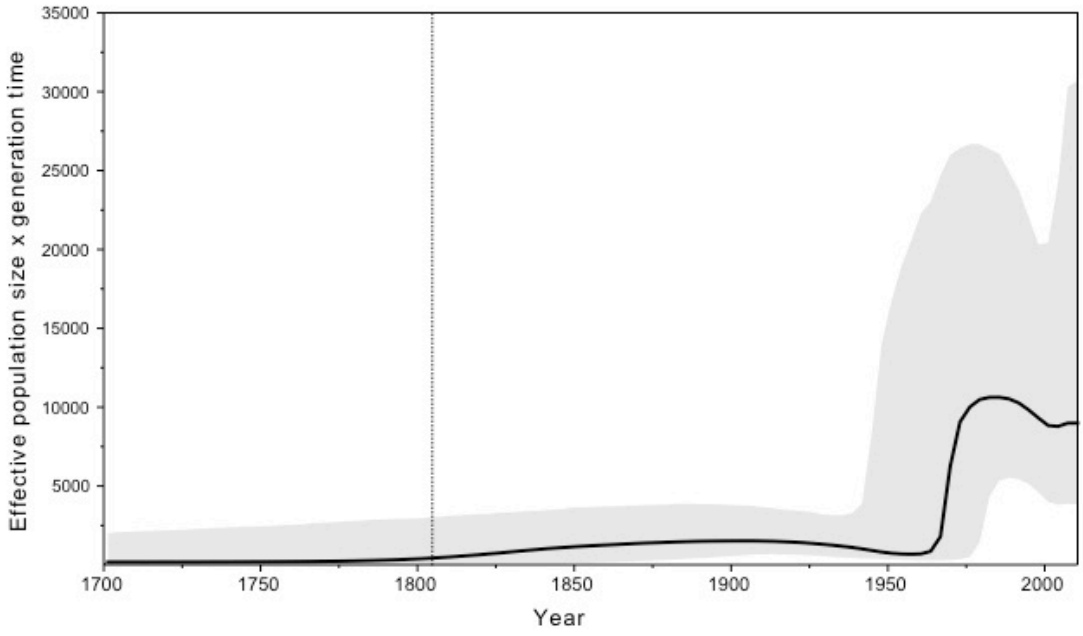


Figure 3.6. *Bayesian skyline plot, showing the epidemic history of genotype 4 estimated from the concatenated alignment. The black line represents the estimated effective number of infections through time. The blue lines represent the 95% highest posterior density confidence intervals of this estimate. The earliest date in the plot is the median estimate of the TMRCA of genotype 4, while the dotted line shows the upper 95% highest posterior density confidence interval of this date.*

The slightly later date of the epidemic transition reported here is likely explained by the new evolutionary rates used in our analysis, which are faster than those employed previously (see above for discussion). The overall shape of the epidemic curve presented in Figure 3.6 closely resembles that previously estimated for genotype 6 in Asia (Pybus et al. 2009). However, as discussed in that paper, skyline plots that span entire HCV genotypes should be interpreted cautiously, as they exhibit geographic structure and combine subtypes that have experienced different rates of growth during the twentieth century (see Pybus et al. 2009 for further details). Thus only the broad qualitative trend of a twentieth-century transition to epidemic growth is likely to be robust.

3.6 DISCUSSION

In this study we report the first large-scale survey of HCV genetic diversity in the DRC, one of the geographically largest and most populous African countries. The previously unsurveyed state of the DRC represented a significant gap in our understanding of HCV in Africa and our results complement recent surveys of HCV in neighbouring countries, including Gabon, Cameroon, Republic of Congo and Central African Republic (e.g. Cantaloube *et al.*, 2010, Njouom *et al.*, 2007; 2009; 2012, Pépin *et al.,* 2008, 2010a, 2010b). We observed a high level of HCV genotype 4 diversity in our study population; isolates were classified as belonging to subtypes 4c, 4k, 4h and 4r, as well as to a potential new subtype circulating in the DRC (denoted 4drc). One isolate (DRC1427) was classified discordantly (4c/4q) in two sub-genomic regions sequenced and may represent a previously undetected recombinant strain. This diversity, together with the absence of globally-prevalent subtypes (e.g. 1a, 1b, 3a), suggests that genotype 4 has been present in the DRC for a long period of time.

However, the genetic diversity of HCV in the DRC is not as high as that in other central African countries where various subtypes of multiple HCV genotypes are sometimes observed (e.g. genotypes 1, 2 and 4 in Cameroon; Ndjomou et al. 2003, Pepin 2010a). Further, we failed to detect any genotype 7 isolates despite using the same primers as those used to initially discover the strain (Murphy *et al.,* 2007b), indicating that the prevalence of genotype 7 within the DRC is low. We conclude that the overall pattern of genetic diversity of HCV in the DRC is fundamentally different to that of HIV-1: the diversity of HIV-1 in the DRC is as large as that observed worldwide, implying that the virus originated in central Africa and its diversity elsewhere is reduced due to founder effects (Vidal *et al.* 2000, Rambaut *et al*. 2001), whereas we observed only one HCV genotype in the DRC.

Approximately 3% of our samples screened contained detectable HCV RNA, slightly less than the 5.2% and 6.4% found in HCV screening surveys in Gabon and Central African Republic respectively (Njouom *et al.,* 2009; 2012). Past studies of HCV in the DRC have only assayed seroprevalence, and have found anti-HCV in 6.4% of blood donors (Tibbs *et al,* 1991), 6.6% of sex workers and 4.3% of pregnant women (Laurent *et al.*, 2001). These seroprevalence estimates are again somewhat less than those found in the Central African Republic (10.5%) and Gabon (11.2%) (Njouom *et al.,* 2009; 2012) Despite the large size of our survey, sampling was limited to males and therefore our prevalence estimates may not equal those of the general population. Cantaloube *et al.,* (2010) found that men were 43% more likely to be seropositive than women in the Republic of the Congo, and it is possible that a similar pattern is true in the DRC. We confirmed that HCV positivity is significantly associated with older age in our study population (Figure 3.2; Chapter 2). Although the long-term persistence of most HCV infections means that older subjects are in general more likely to be infected, the

notable increase in HCV prevalence in those born before 1945 suggests a distinct historical event. This 'cohort effect' has also been reported for HCV genotype 4 in Gabon, Cameroon and Egypt and is thought to be due to past iatrogenic transmission resulting from public health campaigns during the twentieth century that involved injections (Frank *et al.,* 2000, Nerrienet et al. 2005; Njouom *et al.*, 2007; 2012). In the absence of any epidemiological data we cannot speculate further on the possible causes in our study population of the age effect. However, our reconstruction of the epidemic history of genotype 4 (Figure 3.8) clearly shows a transition from low to high prevalence, likely representing the combined effects of iatrogenic transmission events in many different African countries during the middle of the twentieth century (Pepin *et al.,* 2008).

Figure 3.2 shows that those born before 1945 had the highest level of HCV RNA positivity, while those born between 1946 and 1960 have the same HCV prevalence as the later age categories. Figure 3.6 on the other hand, shows that the majority of transmissions happened between 1960 and 1975. This implies that the route by which the transmission occurred did not spread the virus to those aged 15 and below, although the lack of subjects born before 1930 means I cannot say whether transmission was contained within the 1930-1945 age group (aged 15-30 at the time) or instead affected the entire adult population.

The molecular clock and phylogeographic results reported here clearly indicate that genotype 4 epidemics in North Africa, the Middle East and Europe all originated from central Africa (Figure 3.5). Figure 3.5 can be also used to estimate the time of exportation of lineages from central Africa to elsewhere; however, the ages of these movement events will often be overestimated due to the exclusion from the concatenated alignment of many central African strains that were sequenced in only

the core or NS5B regions (Figures 3.3 and 3.4). Genotype 4 lineages present in Egypt are of particular interest, as the country has the highest prevalence of HCV worldwide (Frank *et al.,* 2000, El-Zanaty *et al*., 2008). We estimate that subtype 4a, which accounts for the majority of the Egyptian epidemic, arrived there from central Africa sometime between 1860 and 1925 (the estimated date of the MRCA of subtype 4a). This time frame precedes World War II, during which an 8000-man contingent of the Force Publique (an army composed of soldiers from the Congo serving in the Free Belgian Forces) was stationed in Egypt between 1943 and 1944 (Nigel, 1991). However, these dates can be reconciled if two or more genetically-distinct subtype 4a lineages arrived in Egypt during World War II. It is clear that the Egyptian HCV epidemic originated from more than one Central Africa strain, as subtypes 4a, 4m, 4n and 4o each independently moved to Egypt. Stronger evidence for the role of historical events in the trans-African movement of genotype 4 comes from the presence of Tunisian isolates closely related to strains from the DRC, within subtype 4k. Our molecular clock results date the MRCA of the Tunisian 4k strains to between 1928 and 1975. Tunisian troops were deployed in the DRC as part of the UN peacekeeping forces during the Congo crisis of 1960-65 (Mays, 2010) and may therefore have transported the virus to Tunisia upon their return. The movements of large numbers of troops among populations, combined with a high likelihood of blood transfusions and/or parenteral medical treatments, provided an ideal scenario for the spread of HCV out of central Africa.

Our NS5B phylogeny contains a very high diversity of genotype 4 sequences sampled from France (Figure 3.4). This can be explained by the historical presence of the French colonial empire in central Africa, followed by more recent immigration of individuals to France from former colonies (see also Nicot et al. 2004). We also noted

that subtype 4w has, to date, been found only in Portugal. Portugal was one of the first nations to have a foothold in the Congo region (Appiah and Gates, 1999) and was present in both mainland Angola and in the Cabinda exclave that lies between the DRC and the Republic of the Congo. Our phylogenetic results suggest that subtype 4w is a sub-lineage of 4b, the latter being a highly diverse subtype found in several countries, including countries that neighbour Cabinda (Figures 3.4, 3.5).

Other previously unrecognised geographic trends can be discerned from our analysis. For example, the majority of isolates belonging to subtypes 4q and 4v (9 out of 14) were sampled from Rwanda and Burundi, and samples from these two countries are rarely found in other subtypes. This implies that there has been little further viral dissemination from these countries following the introduction of subtypes 4q and 4v there. The reverse pattern is observed in several adjacent countries of central Africa, specifically the DRC, Republic of the Congo, Cameroon and Gabon. Multiple subtypes of genotype 4 are present in each of these countries, and sequences from them appear in all three of the intra-genotypic lineages L1, L2 and L3. This indicates that historical viral movement among these locations was comparatively common. This might be explained by historical demographic movements and trading links during the African colonial era. However, genotype 4 does not appear to have spread to the Americas via the slave trade, unlike genotype 2 which is present across west Africa (Markov *et al.*, 2009). This may reflect the various ways that European colonial powers exploited different African regions; the western Atlantic coast, including the Gold Coast, was the main hub for the transatlantic slave trade, whereas Central and Eastern Africa were more commonly used to provide natural resources and labour (Suret-Canale, 1971).

Approximately 9% (74 out of 806) of all isolates examined in our study are not classified into a currently defined subtype, and all unclassified isolates were from

central African countries. Detection of divergent yet uncommon strains is to be expected if HCV genotype 4 originated and has circulated in central Africa for several centuries. Under this hypothesis the more prevalent subtypes of genotype 4 likely represent those few lineages that by chance were amplified by changes in human behavior during the previous century (as previously suggested for subtypes 1a, 1b, 3a globally, and for genotype 6 in Asia; e.g. Simmonds, 2004; Magiorkinis *et al.*, 2009; Pybus et al. 2009). We noted two lineages (4car and 4drc) that represent candidates for possible new subtypes. Full genome sequencing of isolates from 4car and 4drc would confirm their subtype status. Lineages that are currently rare and geographically-restricted could potentially spread epidemically and internationally if introduced into high risk groups, as presumably occurred to subtype 4d (de Bruijne, 2009, Ciccozzi, 2012). This risk is of particular relevance to genotype 4 which, like genotype 1, is more refractory to standard anti-viral drug treatment than other genotypes (e.g. Chen *et al.*, 2012). There are no large scale surveys of HCV diversity in several highly populous African countries – Ethiopia, Tanzania, Angola and Chad among them – and some other countries are represented only by surveys with small sample sizes. Further screening of HCV genetic diversity in Africa is required to help plan effective treatment strategies in the region and inform future vaccine design.

3.7 ACKNOWLEDGEMENTS

We thank the Ministry of Public Health and the Ministry of National Defense of the Democratic Republic of Congo for permission to undertake this study and the U.S. Embassy of the DRC for their continued support. This project was financially supported by Global Viral and Metabiota via funding from the U.S. Department of Defense Armed Forces Health Surveillance Center, Division of Global Emerging Infections, Surveillance Operations (AFHSC GEIS), the Defense Threat Reduction Agency Cooperative Biological Engagement Program (DTRA-CBEP), Google.org, the Skoll Foundation, the DoD HIV/AIDS Prevention Program (DHAPP), and was made possible by the generous support of the American people through the United States Agency for International Development (USAID) Emerging Pandemic Threats PREDICT. PK is an NIHR Senior Investigator and he and GLAH were supported by The Wellcome Trust (WT 091663MA), Oxford Martin School, and NIHR Biomedical Research Centre, Oxford. CFD was supported in part by the NIH Fogarty International Center AIDS International Training and Research Program (2 D 43 TW000010-16/17). NDW is supported by the NIH Directors Pioneer Award (DP1-OD000370). The contents are the responsibility of the authors and do not necessarily reflect the views of the DoD, USAID, or the United States Government.

3.8 REFERENCES

Agresti A. and Coull B. A. (1998) Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician* 52: 19-126.

Arthur, R. R., Hassan, N. F., Abdallah, M. Y., El-Sharkawy, M. S., Saad, M. D., Hackbart, B. G. and Imam, I. Z. (1997). Hepatitis C antibody prevalence in blood donors in different governorates in Egypt. *Transactions of the Royal Society of Tropical Medicine,* 91: 271–274.

Appiah, K.W. and Gates, H.L. (1999). *Africana: The Encyclopedia of the African and African American Experience*. p. 1105. Philadelphia, PA: Running Press.

Candotti, D., Temple, J., Sarkodie, F. and Allain, J. P. (2003). Frequent recovery and broad genotype 2 diversity characterize hepatitis C virus infection in Ghana, West Africa. *Journal of Virology,* 77: 7914–7923.

Cantaloube, J.-F., Gallian, P., Bokilo, A., Jordier, F., Biagini, P., Attoui, H., Chiaroni, J. and de Micco, P. (2010). Analysis of hepatitis C virus strains circulating in Republic of the Congo. *Journal of Medical Virology,* 82: 562–567.

Chen, Y., Xu, H. X., Wang, L. J., Liu, X. X., Mahato, R. I. and Zhao, Y. R. (2012). Meta-analysis: IL28B polymorphisms predict sustained viral response in HCV patients treated with pegylated interferon-a and ribavirin. *Alimentary Pharmacology and Therapeutics,* 36: 91–103.

Ciccozzi, M., Equestre, M., Costantino, A., Marascio, N., Quirino, A., Presti, A. L., Cella, E., Bruni, R., Liberto, M. C., Foca A., Pisani, G., Zehended, G. and Ciccaglione, A.R. (2012). Hepatitis C virus genotype 4d in Southern Italy: Reconstruction of its

origin and spread by a phylodynamic analysis. *Journal of Medical Virology,* 84: 1613–1619.

De Bruijne, J., Schinkel, J., Prins, M., Koekkoek, S. M., Aronson, S. J., van Ballegooijen, M. W., Reesink, H. W., Molenkamp, R. and van de Laar, T. J. W. (2009). Emergence of Hepatitis C Virus Genotype 4: Phylogenetic Analysis Reveals Three Distinct Epidemiological Profiles. *Journal of Clinical Microbiology,* 27: 3832-3838

Djoko, C. F., Rimoin, A. W., Vidal, N., Tamoufe, U., Wolfe, N. D., Butel, C., LeBreton, M., Tshala, F. M., Kayembe, P. K., Muyembe, J. J., Edidi-Basepeo, S., Pike, B. L., Fair, J. N., Mbacham, W. F., Saylors, K. E., Mpoudi-Ngole, E., Delaporte, E., Grillo, M. and Peeters, M. (2011). High HIV type 1 group M pol diversity and low rate of antiretroviral resistance mutations among the uniformed services in Kinshasa, Democratic Republic of the Congo. *AIDS Research and Human Retroviruses* 27: 323–329.

Drummond, A. J., Suchard, M. A., Xie, D. and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969–1973.

El-Zanaty, F. and Way, A. (2009) Egypt Demographic and Health Survey 2008 – Final Report. Rockville, MD: The DHS Program.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5, 164–166.

Frank, C., Mohamed, M. K., Strickland, G. T., Lavanchy, D., Arthur, R. R., Magder, L. S., El Khoby, T., Abdel-Wahab, Y., Aly Ohn, E. S., Anwar, W. and Sallam, I. (2000).

The role of parenteral antischistosomal therapy in the spread of Hepatitis C Virus in Egypt. *Lancet* 355, 887–891.

Gray, R. R., Parker, J., Lemey, P., Salemi, M., Katzourakis, A. and Pybus, O. G. (2011). The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evolutionary Biology* 11: 131.

Gray, R. R., Tanaka, Y., Takebe, Y., Magiorkinis, G., Buskell, Z., Seeff, L., Alter, H. and Pybus, O. G. (2013) Evolutionary analysis of hepatitis C virus gene sequences from 1953. *Philosophical Transactions of the Royal Society Series B* 368:20130168

Iles, J. C., Abby Harrison, G. L., Lyons, S., Djoko, C. F., Tamoufe, U., LeBreton, M., Schneider, B. S., Fair, J. N., Tshala, F. M., Kayembe, P. K., Muyembe, J. J., Edidi-Basepeo, S., Wolfe, N. D., Klenerman, P., Simmonds, P. and Pybus, O. G. (2013). Hepatitis C virus infections in the Democratic Republic of Congo exhibit a cohort effect. *Infection, Genetics and Evolution* 19: 386–394.

Jeannel, D., Fretz, C., Traore, Y., Kohdjo, N., Bigot, A., Gamy, E. P., Jourdan, G., Kourouma, K., Maertens, G., Fumoux, F., Fournel, J. J. and Stuyver, L. (1998). Evidence for high genetic diversity and long-term endemicity of hepatitis C virus genotypes 1 and 2 in West Africa. *Journal of Medical Virology* 55: 92–97.

Kuiken, C., Yusim, K., Boykin, L. and Richardson, R. (2005). The Los Alamos hepatitis C sequence database. *Bioinformatics* 21: 379–384.

Lauer, G. M. and Walker, B. D. (2001). Hepatitis C virus infection. *New England Journal of Medicine* 345: 41–52.

Laurent, C., Henzel, D., Mulanga-Kabeya, C., Maertens, G., Larouze, B. and Delaporte, E. (2001). Seroepidemiological survey of hepatitis C virus among

commercial sex workers and pregnant women in Kinshasa, Democratic Republic of Congo. *International Journal of Epidemiology* 30:872-7.

Lemey, P., Rambaut, A., Drummond, A.J. and Suchard, M.A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology* 5:e1000520

Magiorkinis, G., Magiorkinis, E., Paraskevis, D., Ho, S. Y. W., Shapiro, B., Pybus, O. G., Allain, J-P. and Hatzakis, A. (2009). The Global Spread of Hepatitis C Virus 1a and 1b: A Phylodynamic and Phylogeographic Analysis. *PLoS Medicine* 6: e1000198

Markov, P. V., Pépin, J., Frost, E., Deslandes, S., Labbé, A. C. and Pybus, O. G. (2009). Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa. *Journal of General Virology* 90: 2086–2096.

Markov, P. V., van de Laar, T. J., Thomas, X. V., Aronson, S. J., Weegink, C. J., van den Berk, G. E., Prins, M., Pybus, O. G. and Schinkel, J. (2012) Colonial history and contemporary transmission shape the genetic diversity of hepatitis C genotype 2 in Amsterdam. *Journal of Virology* 86:7677-87

Mays, T. M. (2010). *Historical Dictionary of Multinational Peacekeeping*. p. 321. Washington DC: Scarecrow Press.

Mellor, J., Holmes, E. C., Mellor, J., Jarvis, L. M., Yap, P. L., Holmes, E. C., Simmonds, P., Jarvis, L. M., Yap, P. L. and Simmonds, P. (1995). Investigation of the pattern of hepatitis C virus sequence diversity in different geographical regions: implications for virus classification. The International HCV Collaborative Study Group. *Journal of General Virology* 76: 2493–2507.

Mohd Hanafiah, K., Groeger, J., Flaxman, A. D. and Wiersma, S. T. (2013). Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. *Hepatology* 57: 1333–1342.

Murphy, D. G., Chamberland, J., Dandavino, R. and Sablon, E. (2007a). A New Genotype of Hepatitis C Virus Originating from Central Africa. *Hepatology* 46: 623A.

Murphy, D. G., Willems, B., Deschenes, M., Hilzenrat, N., Mousseau, R. and Sabbah, S. (2007b). Use of Sequence Analysis of the NS5B Region for Routine Genotyping of Hepatitis C Virus with Reference to C/E1 and 5' Untranslated Region Sequences. *Journal of Clinical Microbiology* 45: 1102–1112.

Ndjomou, J., Pybus, O. G. and Matz, B. (2003). Phylogenetic analysis of hepatitis C virus isolates indicates a unique pattern of endemic infection in Cameroon. *Journal of General Virology* 84: 2333–2341.

Ndong-Atome, G., Makuwa, M., Njouom, R., Branger, M., Brun-Vézinet, F., Mahé, A., Rousset, D. and Kazanji, M. (2008). Hepatitis C virus prevalence and genetic diversity among pregnant women in Gabon, central Africa. *BMC infectious diseases* 8: 82.

Nerrienet, E., Pouillot, R., Lachenal, G., Njouom, R., Mfoupouendoun, J., Bilong, C., Mauclere, P., Pasquier, C. and Ayouba, A. (2005). Hepatitis C virus infection in Cameroon: A cohort-effect. *Journal of Medical Virology* 76: 208–214.

Nicot, F., Legrand-Abravanel, F., Sandres-Saune, K., Boulestin, A., Dubois, M., Alric, L., Vinel, J. P., Pasquier, C. and Izopet, J. (2005) Heterogeneity of hepatitis C virus genotype 4 strains circulating in south-western France. *Journal of General Virology.* 86: 107-14.

Nigel, T. (1991). *Foreign Volunteers of the Allied Forces, 1939–45.* p.17. London: Osprey.

Njouom, R., Nerrienet, E., Dubois, M., Lachenal, G., Rousset, D., Vessière, A., Ayouba, A., Pasquier, C. and Pouillot, R. (2007). The hepatitis C virus epidemic in Cameroon: genetic evidence for rapid transmission between 1920 and 1960. *Infection, Genetics and Evolution* 7: 361–367.

Njouom, R., Frost, E., Deslandes, S., Mamadou-Yaya, F., Labbé, A.-C., Pouillot, R., Mbélesso, P., Mbadingai, S., Rousset, D. and Pépin, J. (2009). Predominance of hepatitis C virus genotype 4 infection and rapid transmission between 1935 and 1965 in the Central African Republic. *The Journal of General Virology* 90: 2452–2456.

Njouom, R., Caron, M., Besson, G., Ndong-Atome, G.-R., Makuwa, M., Pouillot, R., Nkoghé, D., Leroy, E. and Kazanji, M. (2012). Phylogeography, risk factors and genetic history of hepatitis C virus in Gabon, central Africa. *PloS One* 7: e42002

Pépin, J. and Labbé, A-C. (2008). Noble goals, unforeseen consequences: control of tropical diseases in colonial Central Africa and the iatrogenic transmission of blood-borne viruses. *Tropical Medicine and International Health* 13:744-753

Pépin, J., Lavoie, M., Pybus, O. G., Pouillot, R., Foupouapouognigni, Y., Rousset, D., Labbé, A-C. and Njouom, R. (2010a). Risk factors for Hepatitis C virus transmission in colonial Cameroon. *Clinical Infectious Diseases* 51: 768-776.

Pépin, J., Labbé, A. C., Mamadou-Yaya, F., Mbélesso, P., Mbadingai, S., Deslandes, S., Locas, M. C. and Frost, E. (2010b). Iatrogenic transmission of human T cell lymphotropic virus type 1 and hepatitis C virus through parenteral treatment and chemoprophylaxis of sleeping sickness in colonial Equatorial Africa. *Clinical Infectious Diseases* 51: 777-784.

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C. and Harvey, P.H. (2001). The Epidemic Behavior of the Hepatitis C Virus. *Science* 292: 2323-2325.

Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H. and Rambaut, A. (2003). The Epidemiology and Iatrogenic Transmission of Hepatitis C Virus in Egypt: A Bayesian Coalescent Approach. *Molecular Biology and Evolution* 20: 381–387.

Pybus, O. G., Cochrane, A., Holmes, E. C. and Simmonds, P. (2005). The hepatitis C virus epidemic among injecting drug users. *Infection, Genetics and Evolution* 5: 131–139.

Pybus, O. G., Barnes, E., Taggart, R., Lemey, P., Markov, P. V., Rasachak, B., Syhavong, B., Phetsouvanah, R., Sheridan, I., Humphreys, I. S., Lu, L., Newton, P. N. and Klenerman, P. (2009). Genetic History of Hepatitis C Virus in East Asia. *Journal of Virology* 83: 1071–1082.

Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M. and Holmes, E. C. (2001). Human Immunodeficiency virus. Phylogeny and the origin of HIV-1. *Nature* 410: 1047-8

Rose, R., Markov, P. V., Lam, T. T. and Pybus, O. G. (2013). Viral evolution explains the associations among hepatitis C virus genotype, clinical outcomes, and human genetic variation. *Infection, Genetics and Evolution* 20: 418–421.

Salemi, M. and Vandamme, A.-M. (2002). Hepatitis C Virus Evolutionary Patterns Studied Through Analysis of Full-Genome Sequences. *Journal of Molecular Evolution* 54: 62–70.

Shapiro, B., Rambaut, A. and Drummond, A.J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution* 23:7-9

Sharp, C. P., Vermeulen, M., Nébié, Y., Djoko, C. F., LeBreton, M., Tamoufe, U., Rimoin, A. W., Kayembe, P. K., Carr, J. K., Servant-Delmas, A., Laperche, S., Harrison, G. L. A., Pybus, O. G., Delwart, E., Wolfe, N. D., Saville, A., Lefrère, J-J. and Simmonds, P. (2010). Epidemiology of Human Parvovirus 4 Infection in Sub-Saharan Africa. *Emerging Infectious Diseases* 16: 1605–1607.

Simmonds, P., Holmes, E. C., Cha, T. A., Chan, S. W., McOmish, F., Irvine, B., Beall, E., Yap, P. L., Kolberg, J. and Urdea, M. S. (1993). Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. *Journal of General Virology* 74: 2391–2399.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus - 15 years on. *Journal of General Virology* 85: 3173–3188.

Smith, D. B., Bukh, J., Kuiken, C., Muerhoff, A. S., Rice, C. M., Stapleton, J. T. and Simmonds, P. (2014). Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: Updated criteria and genotype assignment web resource. *Hepatology* 59: 318–327.

Strickland, G.T. (2006) Liver disease in Egypt: hepatitis C superseded schistosomiasis as a result of iatrogenic and biological factors. *Hepatology* 43: 915-22

Suret-Canale, J. (1971) *French Colonialism in tropical Africa 1900-1945*. pp 462-484. London: C. Hurst and Company.

Tibbs, C. J., Palmer, S. J., Coker, R., Clark, S. K., Parsons, G. M., Hojvat, S., Peterson, D. and Banatvala, J. E. (1991). Prevalence of hepatitis C in tropical communities: the importance of confirmatory assays. *Journal of Medical Virology* 34:143-7

van Asten, L., Verhaest, I., Lamzira, S., Hernandez-Aguado, I., Zangerle, R., Boufassa, F., Rezza, G., Broers, B., Robertson, J. R., Brettle, R. P., McMenamin, J., Prins, M., Cochrane, A., Simmonds, P. and Coutinho, R.A. (2004). Spread of Hepatitis C Virus among European Injection Drug Users Infected with HIV: A Phylogenetic Analysis. *Journal of Infectious Diseases* 189: 292–302.

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B. and Delaporte, E. (2000). Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *Journal of Virology* 74: 10498-507.

Wertheim, J.O., Fourment, M. and Kosakovsky Pond, S.L. (2012). Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. *Molecular Biology and Evolution* 29:451-6.

Wertheim, J.O. and Kosakovsky Pond, S.L. (2011) Purifying selection can obscure the ancient age of viral lineages. *Molecular Biology and Evolution* 28:3355-65.

Zwickl, D. J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.

4 COALESCENT RECONSTRUCTION OF THE TRANSMISSION HISTORY OF THE HEPATITIS C VIRUS IN THE DRC

Contribution by collaborators: Dr. Jacques Pépin and Dr. Eric Frost performed the collection, screening and sequencing of the Kinshasa samples.

4.1 INTRODUCTION

Though the Hepatitis C virus (HCV) infects many millions worldwide, its prevalence is not equally distributed geographically. Certain subtypes, for example 1a, 1b and 3a, are considered to be epidemic worldwide, whist other subtypes, even those closely related to epidemic strains, are only found in limited geographic regions and often at much lower prevalences. Understanding the historical spread of HCV thus requires the investigation of factors that might have caused some lineages to have expanded rapidly and spread globally, while others remained at their historically-low endemic levels.

HCV lineages fall into three broad categories; endemic, local epidemic, and global epidemic (Stumpf and Pybus, 2002). A small group of the most well-known HCV subtypes (1a, 1b, 2a, 2b and 3a) are globally epidemic; they are prevalent in many countries around the world and responsible for the majority of HCV infections seen in clinics in both developed and developing countries (Simmonds, 2004). In contrast to the global epidemic subtypes, the majority of HCV lineages are comparatively uncommon and may exhibit high genetic diversity due to a long history of endemic transmission in a particular region, although the size of the region in which an endemic lineage is found can vary greatly (Pybus *et al.,* 2007). The routes of transmission of endemic HCV are not well understood, and have been hypothesized to include tattooing, scarification, mother-child transmission, or mechanical transmission on the mouthparts of biting insects (Pybus *et al.,* 2007). A third category of HCV lineage arises when the virus is amplified to a high prevalence within a certain population, but without significant spread to other locations. The risk factors that give rise to these 'local epidemic' lineages may include needle-sharing injecting drug use (IDU) or past parenteral treatment campaigns undertaken in specific locations whose sterilization procedures were insufficient.

To understand what factors led to the genesis of local epidemic lineages it is necessary to understand the transmission of HCV in the past. Several sources of information about past transmission can be envisaged. First, historical medical records and the accounts of healthcare practitioners might indicate the existence and prevalence of a disease by reporting recognisable symptoms. Second, by applying epidemiological models to surveys of people currently infected with HCV it is possible to statistically infer risk factors for infection and past incidence. Third, biological samples (sera or tissue) from infected individuals that were collected in the past and subsequently stored can provide evidence that the virus was present at a specific place and time, and genetic analysis of viral gene sequences from archived samples can calibrate estimates of viral evolutionary rates and divergence times. Fourth, statistical methods based on coalescent theory (Kingman, 1982) can be used to infer historical transmission rates from virus gene sequences collected in the modern day (Pybus and Rambaut, 2009).

Under the right circumstances each of these four source of information can be very powerful, but for HCV neither historical medical records nor archived samples are likely to be helpful. In most subjects (~80%) initial HCV infection is asymptomatic and even acute infection shows non-specific symptoms such as nausea, fever and fatigue; meanwhile, chronic infection takes decades to manifest symptoms such as liver cirrhosis (WHO 2014). In addition, the virus was only discovered in 1989 (Choo *et al.,* 1989). This means that medical records more than 25 years old will not contain attempts to diagnose HCV (except as 'non-A non-B hepatitis') and will not report symptoms that can indicate HCV with certainty. Serum samples were also less likely to be deliberately archived before the virus was discovered and consequently very few HCV gene sequences from before 1980s have been reported. The oldest HCV sequences reported to date were obtained from samples collected between 1948 and

1954 as part of routine blood testing of the US military (Gray *et al.,* 2013, Seeff *et al.,* 2000). The level of laboratory infrastructure required for the continual preservation of HCV from the 1950s during the HCV epidemic is probably rare, and especially so in developing regions such as sub-Saharan Africa.

Modern diagnostic methods such as serological assays and RT-PCR are able to accurately detect current or past HCV infection in individual patients (Daniels *et al.,* 2009, Araujo, 2012) and active monitoring of HCV infection is now undertaken by public health and blood transfusion agencies worldwide (Selvarajah and Busch, 2012). These incidence data can be used to infer past HCV transmission history and can be combined with information obtained from patient questionnaires or medical records to investigate potential historical risk factors for HCV infection. This approach was used to great effect in Frank *et al.,* 2000, in order to understand the HCV epidemic in Egypt. Frank *et al.* (2000) compared contemporary HCV prevalences in each Egyptian age cohort in different locations, with the exposure of each cohort to parenteral antischistosomal treatment (PAT), which was estimated from Egyptian government records. They demonstrated that prevalence and PAT exposure were closely correlated, suggesting that the Egyptian HCV epidemic was spread iatrogenically by the PAT campaigns. Other studies, such as Deuffic *et al.,* (1999) and Sypsa *et al.,* (2004), have used mathematical methods and estimates of epidemiological parameters to infer the past population dynamics of HCV infection from contemporary patterns of prevalence.

Sampled HCV gene sequences provide an alternative source of information about past transmission history. HCV epidemic history can been inferred using coalescent-based approaches (Pybus *et al.,* 2001). The coalescent is a stochastic process that models how the timings of coalescence events between lineages in a phylogeny are determined by the population dynamic history of the sampled population (Kingman, 1982).

Coalescent-based approaches can estimate the demographic history of a population (i.e. its change in effective population size through time) and other evolutionary parameters from sets of sampled virus sequences (Fu and Li, 1999). Several different coalescent-based methods have been applied to HCV (e.g. Tanaka *et al.*, 2002; Pybus *et al.,* 2003, 2009; Mizokami and Tanaka, 2004; Nakano *et al.,* 2004; Dearlove and Wilson, 2013). Modern studies use Bayesian Markov Chain Monte Carlo (MCMC) inference, as implemented in the BEAST 1.8 software package (Drummond *et al.,* 2012), which combines molecular clock and coalescent models whilst also incorporating statistical uncertainty arising from phylogenetic error. Bayesian MCMC methods simultaneously generate posterior probabilities for evolutionary and demographic parameters and the sample phylogeny.

Coalescent methods that have been applied to viral epidemic history can be classified into two general categories (see Ho and Shapiro, 2011 for a review): parametric and non-parametric. Parametric coalescent methods assume particular functions for the history of population size, such as constant size, exponential growth, or logistic growth (Weiss and von Haeseler, 1998; Pybus et al. 2001). Each function is defined by a small number of demographic parameters that can be estimated from sampled viral sequences. Parametric models can represent complex population dynamic histories by combining multiple growth phases. For example, Pybus *et al.,* 2003 developed a model that contained three phases: (i) constant population size in the past, (ii) a period of exponential growth and (iii) a period of constant size near the present. The five parameters of this model include the exponential growth rate and the transition times between growth phases. Pybus *et al*., 2003 applied this model to the Egyptian HCV epidemic investigated by Frank *et al.,* 2000 and were able to reconstruct an estimated growth curve for the epidemic from modern sequence data. The exponential growth

phase estimated by this analysis took place during the parenteral antischistosomal therapy (PAT) campaign suggested as a cause for the epidemic by Frank *et al.,* 2000, showing that this method can provide independent verification of risk factors suggested by epidemiological investigations.

The second category of coalescent methods comprises so-called 'non-parametric' approaches, such as the Bayesian skyline plot (Strimmer *et al.,* 2001; Drummond *et al.,* 2005) and the Bayesian skyride/skygrid approach (Minin, Bloomquist and Suchard, 2008; Gill *et al.,* 2013). These methods define demographic history using very general, flexible functions, such as piecewise constant functions with many changepoints or steps (see Figure 4.1 for an example of the Bayesian skyline plot). The key benefit of nonparametric models is that they make minimal assumptions about a population's dynamic history. They can be used to suggest which parametric models might be suitable for a given data set and can provide independent verification of epidemic histories generated from modern seroprevalence data, as described above.

One disadvantage of non-parametric models as compared to parametric ones is that they do not provide direct estimates of specific parameters such as exponential growth rates, and so a parametric analysis may be most powerful once a non-parametric analysis has determined a suitable population model to use (Pybus *et al.,* 2000).

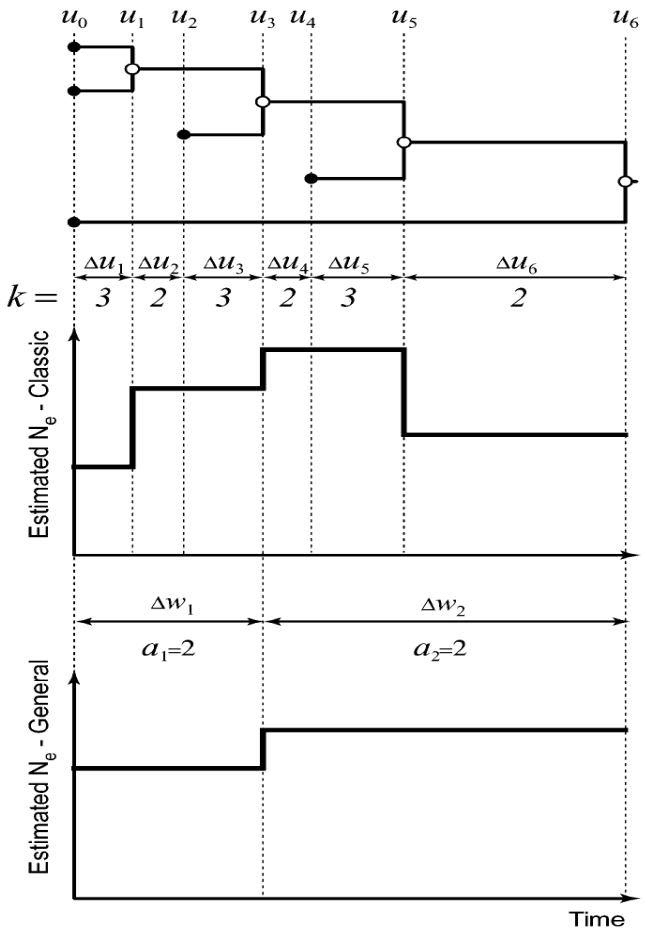


Figure 4.1. *An illustration of the skyline plot approach, reproduced from Drummond et al., 2005. Top graph: a genealogy of five sequences (n=5) sampled at three different time points. Middle graph: the classic skyline plot for the genealogy, which shows the estimated effective population size (Ne) calculated as a function of the number of extant lineages (k) multiplied by the window duration (Δu). Ne changes at each coalescent event, resulting in a stepwise function with n-2 change points and n-1 Ne values. Bottom graph: the generalised skyline plot for the genealogy. In the generalised skyline plot, adjacent coalescent windows are gathered into a smaller number of groups (2 in this case). Each group has duration Δw and the Ne for each group is calculated. In the Bayesian skyline analysis, different genealogies are sampled to represent phylogenetic uncertainty, and their resultant generalised skyline plots are combined into a single plot (e.g. as in Figures 4.6 and 4.7).*

The skyline and skyride/skygrid differ in how the number of steps in the demographic history is defined, and how they are combined and smoothed. The skyline model uses a multiple change-points system to estimate the effective population size dynamics in a Bayesian framework; the number of change points in population size is fixed *a priori* and changepoints must coincide with lineage coalescence times (Drummond *et al.*, 2005). The Skygrid model, by contrast, places a Gaussian Markov random field prior distribution on the piecewise-constant demographic model, smoothing the population trajectory. In addition, the user specifies fixed points in real time where the population size trajectory can change, allowing the Gaussian Markov random field prior to be independent of the genealogy (Gill *et al.*, 2013).

Two sources of information about past HCV transmission – epidemiological incidence data and coalescent inference from viral gene sequences – were combined by Pepin *et al.,* 2010 to analyse the history of HCV spread in Cameroon. Following prior studies that demonstrated high HCV prevalence among those born before 1945 in Cameroon (Nerrienet *et al.,* 2005), Pepin *et al.* undertook a cross-sectional study of inhabitants of Ebolowa aged ≥60 years, issued subjects with an epidemiological questionnaire pertaining to potential risk factors for HCV infection, and tested them for anti-HCV antibodies. Viral gene sequences were obtained using RT-PCR from 71% of the 252 HCV-positive sera and Bayesian skyline plots were estimated from the resulting sequences. The analysis exposed intravenous antimalarial treatment, blood transfusion and age as risk factors for HCV seropositivity, while the skyline plot analysis showed a period of epidemic growth in genotypes 1, 2 and 4 between 1900 and 1960. Together the results suggest a role for intravenous antimalarial treatment in the past transmission of HCV in Ebolowa, as blood transfusion was not available during the earlier stages of

the epidemic and age being a risk factor can be explained from a cohort effect whose only survivors are elderly.

In Chapter 1 I showed that there was an age structure to HCV infection in the DRC, such that individuals born before 1950 were significantly more likely to be infected with HCV, similar to the pattern observed in several neighboring countries such as Cameroon, the Republic of the Congo and Gabon (Nerrienet *et al.,* 2005; Cantaloube *et al.,* 2010; Njouom *et al.,* 2012). Further, in Chapter 2 I found that ~53% of HCV genotype 4 sequences obtained from a large cohort of individuals from the DRC military grouped into DRC-specific clusters within genotypes 4c, 4k and 4r, and those clusters had most recent common ancestors (MRCAs) that existed around 50 years ago. Taken together, these results suggest that the Democratic Republic of the Congo (DRC) may contain local epidemic lineages similar to those previously observed in Cameroon and Gabon (Njouom *et al.,* 2007, 2012; Pépin *et al.,* 2010) which may be associated with increased rates of transmission during the 20th century. However, Chapters 1 and 2 reported an insufficient number of DRC sequences from each subtype to perform a reliable coalescent analysis, which might corroborate the hypothesis of rapid transmission in the past.

In this chapter I aim to directly test the hypothesis that HCV underwent rapid transmission in the DRC, particularly Kinshasa, halfway through the 20th century. A cohort of elderly individuals in Kinshasa were recruited and screened for HCV seropositivity; those samples that were seropositive HCV RNA were amplified and sequenced (this work was undertaken by collaborators). I combined these new sequences with DRC HCV sequences from Genbank and from Chapters 1 and 2. The resulting alignments for subtypes 4k and 4r were large enough for coalescent-based inference, and I analysed them using a variety of coalescent analyses in order to

estimate the epidemic history of HCV in the DRC over the twentieth century. As I used a large number of different coalescent methods to analyse the data, this chapter additionally provides an opportunity to compare the performance of different coalescent methods when applied to HCV subtype alignments. The time periods for the HCV epidemic indicated in Chapters 1 and 2 overlap with an estimated time of high HIV-1 transmission in Kinshasa (Worobey *et al*., 2008), and so the generated estimates for HCV transmission are examined in light of their implications for HIV-1 transmission based on shared risk factors for HCV and HIV-1.

4.2 METHODS

4.2.1 Study Population

A sample of participants were recruited who were aged 70 or more and had lived in Kinshasa for at least 30 years. Participants with dementia, aphasia or an inability to understand Lingala, the local *lingua franca*, were excluded. Participants provided written consent, filled out a questionnaire in Lingala and provided dry blood spots. In total 839 samples were collected. The samples were tested for HCV using a two-step procedure; first, each sample was tested with Monolisa anti-HCV Plus V2 (Bio-Rad). Those samples with an optical density/cutoff ratio of lower than 0.9 were dismissed as negative, while those samples with a ratio between 0.9 and 5.0 went through a second round of testing with the Innolia immunoblot technique (Innogenetics). In samples with a ratio higher than 5.0, PCR amplification was attempted in the 5'UTR region using primers KY78 and KY80 (see Table 4.1 for primer details and Figure 4.2 for a diagram of this procedure).

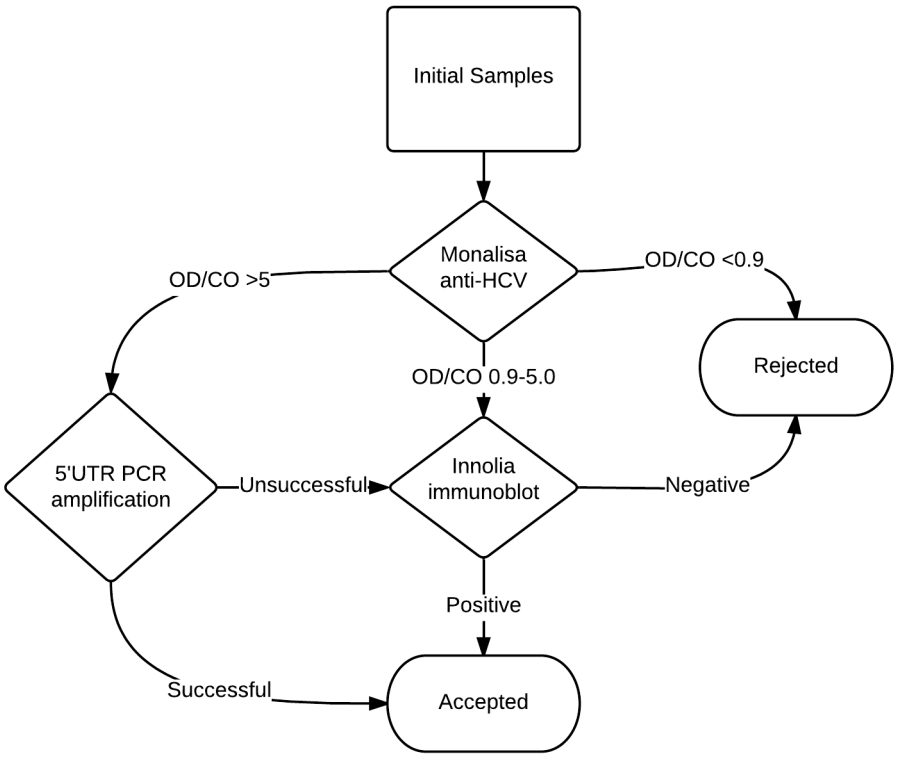


Figure 4.2. Schematic of the procedure for determining which samples were HCV positive or negative. OD/CO = optical density/cutoff ratio.

If the amplification was successful, the sample was treated as positive, and otherwise Innolia immunoblot was used to confirm if the sample was seropositive. HCV RNA-positive samples were subjected to two further PCR amplifications: a 1020-nt fragment of the core and E1 genes (positions 293-1313 relative to isolate H77) and a 382-nt fragment of the NS5B gene (positions 8256-8638); primers used are given in Table 4.1. 217 participants (26.1%) were seropositive for HCV, and this study was able to extract HCV sequence from 118 samples with 56 sequenced in the core region and 105 sequenced in the NS5B region. The fieldwork and laboratory work described in this section was undertaken by my collaborators, Dr. Jacques Pepin and Dr. Eric Frost.

Primer Name	Sequence	Start*	End*
KY80	GCAGAAAGCGTCTAGCCATGGCGT	56	79
KY78	CTCGCAAGCACCCTATCAGGCAGT	299	276
405F	CTGATAGGGTGCTTGCGAGTG	293	313
HCVE1Ri	CATCATCATGTCCCANGCCAT	1313	1293
HCVNS5F2o	TATGATACCCGCTGCTTTGACTCNAC	8256	8281
HCVNS5Rnb	TACCTGGTCATAGCCTCCGTGAAGGCTC	8638	8611

***Table 4.1**. Primers used in this study. Start and end positions relative to isolate H77.*

4.2.2 Sequence Collation

All HCV genotype 4 gene sequences gathered from the DRC or from individuals known to have emigrated from DRC (where stated as such in the literature) were downloaded from the Los Alamos HCV sequence database (Kuiken *et al.*, 2005) and from GenBank. Sequences were retained if they spanned either of the two subgenomic regions sequenced in this study and analysed in Chapters 1 and 2: core (positions 345-1287 relative to H77) and NS5B (positions 8322-8624). Only one sample for each sub-genomic region was retained from each infected individual. These database sequences were combined with the new sequences obtained in this study (see Section 4.2.1) and aligned by hand using Se-Al v2.0 (available from http://tree.bio.ed.ac.uk), resulting in a core alignment of 105 sequences, an NS5B alignment of 173 sequences and a concatenated alignment containing the 82 samples that had sequences in both regions.

4.2.3 Phylogenetic Analysis

Phylogenies were estimated for the core and NS5B alignments using maximum likelihood (ML) as implemented in GARLI v0.951 (Zwickl, 2006). The analysis used a General Time-Reversible (GTR) nucleotide substitution model, estimated base

frequencies, and a gamma distribution model of among-site rate variation. Statistical support for phylogenetic clustering was calculated using an ML bootstrap approach with 500 bootstrap replicates; bootstrap scores were summarized using TreeAnnotator (http://beast.bio.ed.ac.uk/TreeAnnotator). Phylogenies were visualized and annotated using FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree). Newly-generated sequences were classified by adding the HCV subtype reference sequences provided in Smith *et al.*, (2014) to the alignment and computing p-distances to them, resulting in a core alignment containing 91 samples and an NS5B alignment containing 140 samples. A p-distance threshold of <0.15 was used to assign a newly generated sequence to a previously defined subtype. P-distances between sequences were calculated using DNAdist in the Phylip package (Felsenstein, 1989).

Alignments with few sequences may not contain sufficient genetic information to reliably estimate a viral epidemic history using coalescent theory, although a precise minimum sample size is difficult to determine (see Dearlove and Wilson, 2013). A rule of thumb has emerged, advising that the number of groups used in skyline analyses should = (number of sequences)/4 rounded up to the nearest integer. Although this rule does not appear in the literature it does reflect the observation that too many groups can increase estimation error (Heled and Drummond, 2008). In this analysis, the population size trajectory of HCV in the DRC is hypothesized to have three stages – low endemic prevalence, rapid exponential growth, and high current prevalence. Since this trajectory is unlikely to be detectable if less than four population size values are used, only those alignments that contain 16 or more sequences were specified for further study using coalescent methods. For each of these subtypes 6 alignments were created; two each in the core, NS5B and concatenated regions. For each of these subgenomic regions, one alignment contained only sequences gathered in this study

whilst the other alignment contained those, plus all other DRC sequences from that subtype gathered from online databases (See Table 4.2).

Subtype	Core	NS5B	Concatenated
4k (new samples from this study only)	22	44	18
4k (new samples plus all database sequences)	38	68	31
4r (new samples from this study only)	19	35	14*
4r (new samples plus all database sequences)	28	51	21

*Table 4.2. Number of DRC HCV sequences in the alignments used for coalescent analyses. *Not used for coalescent analysis due to sample size <16.*

4.2.4 Molecular Clock Calibration

The sequences gathered in this study do not contain enough temporal information to directly estimate reliable HCV evolutionary rates from my alignments (Pybus *et al.,* 2009; Salemi and Vandamme, 2002). Therefore, following the approach used in Chapter 2, I estimated evolutionary rates for the subgenomic regions used in this study using independent alignments of HCV sequences that contain sufficient temporal information (Gray *et al.* 2011), and used these rate estimates as prior distributions during the Bayesian coalescent analyses.

As in Chapter 2, I used two alignments of heterochronous HCV sequences from Gray *et al.* (2011) that contain 65 subtype 1a sequences and 54 subtype 1b sequences, respectively. Evolutionary rates were estimated using the Bayesian Markov Chain Monte Carlo (MCMC) inference method implemented in BEAST v1.8 (Drummond *et al.*, 2012). These analyses employed a HKY nucleotide substitution model, an uncorrelated lognormal relaxed molecular clock model, a gamma site heterogeneity model with four categories, and a Bayesian skyline plot coalescent model. Base

frequencies were estimated from the data. Four partitions of the whole genome alignment were defined, each of which had independently specified substitution and molecular clock models: (1) a core partition (sites 345-1287), (2) a NS5B partition (sites 8322-8624), (3) a concatenated partition (sites 345-1287 plus 8322-8624), and (4) the remainder of the genome excluding UTRs (sites 1288-8321 plus sites 8625-9374). The rate parameters estimated for partitions 1-3 are shown in Table 4.3. Each MCMC analysis was run for at least 100,000,000 states. This is the same method as that used in Chapter 2, except that the positions of the subgenomic regions analysed are slightly different, so as to include the new HCV sequences generated here.

Genome Region	Mean evolutionary rate (95% HPD)
Core (sites 345-1287)	7.21×10^{-4} (5.01×10^{-4}, 9.69×10^{-4})
NS5B (sites 8322-8624)	9.91×10^{-4} (7×10^{-4}, 1.34×10^{-3})
Concatenated (sites 345-1287, 8322-8624)	7.97×10^{-4} (5.38×10^{-4}, 1.05×10^{-3})

Table 4.3. *Evolutionary rate parameters.*

4.2.5 Bayesian Evolutionary Analysis

Using the Bayesian Markov Chain Monte Carlo (MCMC) inference method implemented in BEAST v1.8 (Drummond *et al.*, 2012) I estimated the epidemic history of HCV in the DRC under various coalescent models. Table 4.4 details the 12 alignments that were subjected to coalescent analysis. To ensure that coalescent inferences were reliable, I only considered alignments that contained more than 15 sequences. Consequently no analysis was performed on the concatenated 4r alignment that contained only sequences generated in this study (see Table 4.2).

In my analysis, as in previous studies (e.g. Pybus *et al.*, 2009), I used an SDR06 nucleotide substitution model (two independent HKY+Γ substitution models – one for

the first and second codon positions, and one for the third), a gamma model of among-site rate heterogeneity with 4 categories, and an uncorrelated lognormal relaxed molecular clock. Some analyses were repeated using a strict molecular clock in order to test which clock model provided the better fit, as assessed using Akaike's information criterion through MCMC (AICM) (Baele *et al.,* 2012).

I used three different coalescent methods for estimating epidemic history in my analyses; (i) the Bayesian Skyline model (Drummond *et al.,* 2015), (ii) the Bayesian Skygrid model (Gill *et al.,* 2013), and a parametric demographic model called Con-Exp-Con (CEC) which implements a three phase demographic history (Pybus *et al.,* 2003). The Skyline and Skygrid models are both nonparametric models that use a piecewise-constant model of population size to estimate the trajectory of effective population size through time. The CEC model is a parametric model that assumes the analysed population starts at a certain population size in the present (N_1), continues at this population size up to a certain time (t_1), then starts growing or shrinking exponentially at a set rate (r) for a certain length of time (dt), up to time t_0, after which the population is again constant with size N_0; see Figure 4.3 for a graphical definition of the model. The CEC model was first used in Pybus *et al.*, 2003, although this study is the first time this model has been implemented in BEAST. The BEAST implementation was carried out by Prof. Oliver Pybus and Prof. Andrew Rambaut.

The number of groups in the Bayesian skyline analyses was calculated as n/4 rounded to the nearest integer where n is the number of sequences in the alignment, with a maximum of 10 groups. The effective population size for each group was given a vague uniform prior distribution, bounded at 0 and 10^{100}. Bayesian skygrid analyses were performed using n-1 grid points, as suggested in Gill *et al.*, 2012. For the CEC model, prior distributions for the four model parameters (N_0, t_1, r, dt) were based on

the skyline plot results. N_0 was given a lognormal prior distribution (Log(mean)= 9.5 and Log(standard deviation)=1.25), while r was given a Laplace distribution (mean=0 and scale=1). t_1 was given a normal prior distribution (mean=40 and standard deviation=15), while dt was given a normal prior distribution (mean=20 and standard deviation=10).

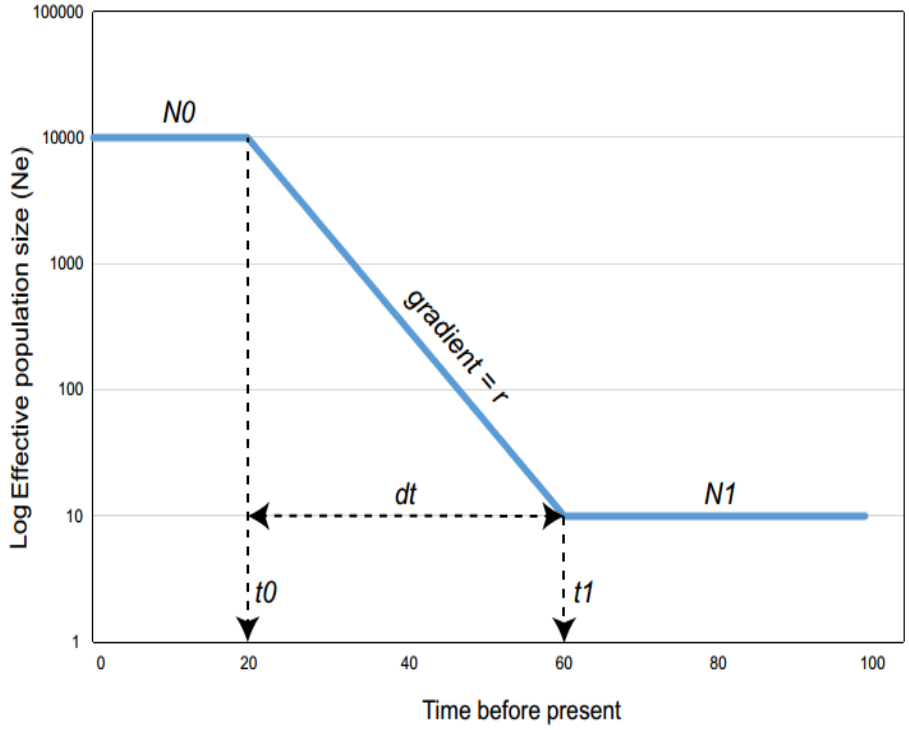


Figure 4.3. Illustration of the Constant-Exponential-Constant (CEC) demographic model. The model has four parameters: effective population size at present (N0), the amount of time before present the exponential growth ends (t0), the duration of the exponential period (dt) and the exponential growth rate (r). Also labeled are the initial population size (N1) and the time when exponential growth starts (t1), although these parameters are not used in the CEC model. The y axis is NeT, the effective number of infections multiplied by the generation time, and is on a log scale.

In all analyses, a normal prior distribution was placed upon the mean evolutionary rate parameter according to the genome region in question, as detailed in Table 4.3. Each MCMC run contained at least 100 million states, with parameters and trees sampled once every 10000 states. MCMC convergence and effective sample sizes were monitored using Tracer v. 1.6.1, and the same program was used to generate plots of effective population size against time from each analysis. Maximum clade credibility trees were calculated and annotated using TreeAnnotator 1.7.5 (Drummond *et al.*, 2012). To compare the output of different coalescent models, I estimated AICM for each model using the method-of-moments estimator (Baele *et al.,* 2012), as implemented in Tracer v1.6.1. This was used instead of Bayes Factors estimated with the smoothed harmonic mean estimator, because the latter method provides less reliable results for model testing (Baele *et al.,* 2012).

4.3 RESULTS

4.3.1 Genotyping of newly sequenced samples

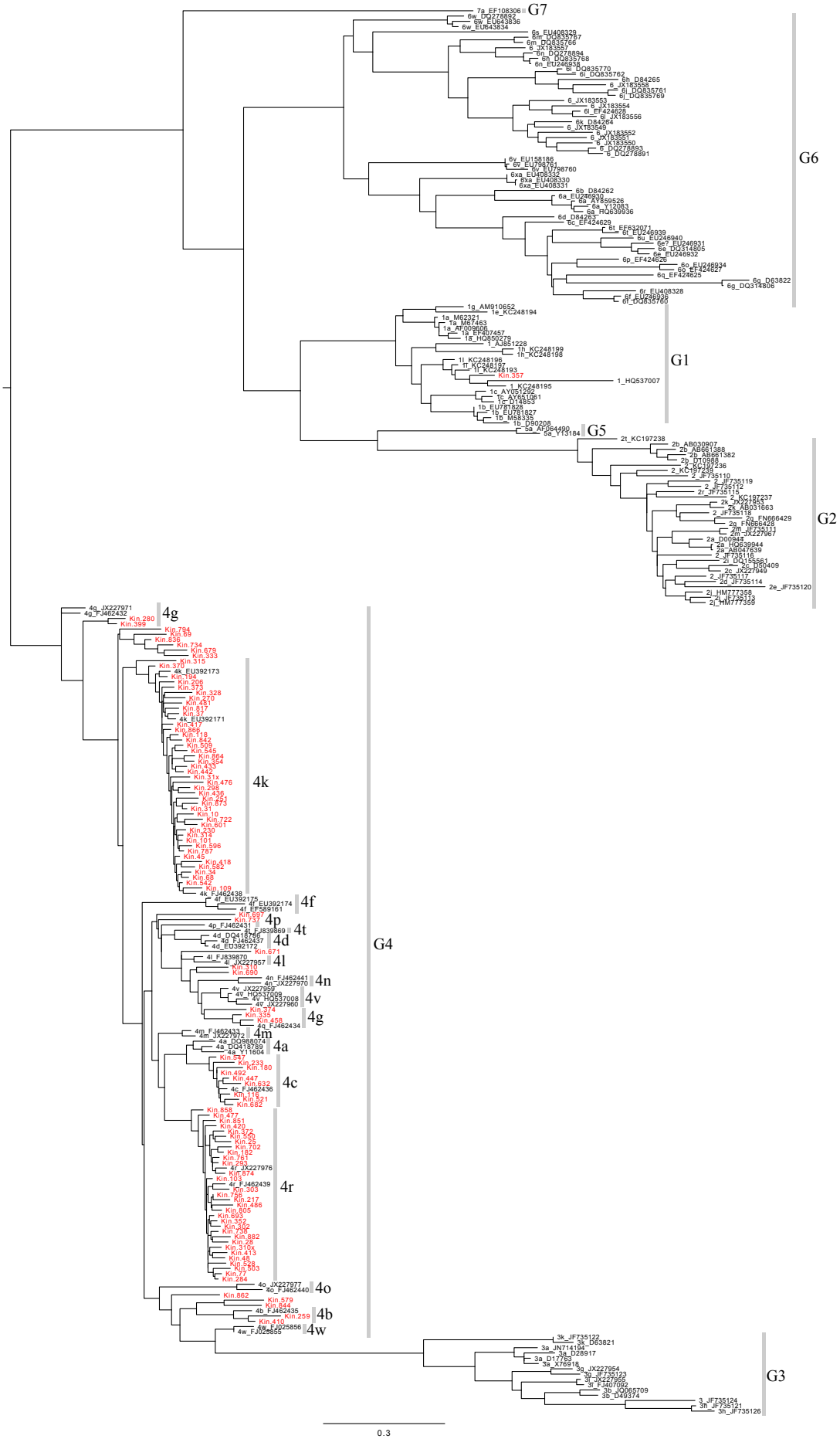
Figures 4.4 and 4.5 show maximum likelihood trees estimated from core and NS5B sequences, respectively. Samples newly sequenced in this study are shown in red, whilst all other sequences represent genotype/subtype reference genomes. This analysis classified one sample (Kin.357) as genotype 1, one (Kin.619) as genotype 7 and 116 as genotype 4. Within genotype 4, the new samples were classified as belonging to subtypes 4a, 4b, 4c, 4g, 4k, 4l, 4n, 4q, 4r, 4v and 4w. Three samples were too divergent to be classified within any currently-defined subtype (Kin.697, Kin.737, Kin.844) and one sample (Kin.476) was phylogenetically discordant, belonging to 4r in the core region and 4k in the NS5B region. For a full list of subtype classification results, see Table 4.4. The majority of new isolates belonged to subtypes 4k (n=47;

40% of generated sequences) and 4r (n=39; 33%). The genetic diversity of these isolates (one subtype of genotype 1, eleven subtypes of genotype 4 and one genotype 7 strain) is larger than that seen in several previous studies of HCV in the DRC and neighbouring countries; for example, Cantaloube *et al.* (2010) surveyed the Republic of Congo and found five genotype 4 subtypes and one genotype 2 subtype, while Njouom *et al.* (2007) compiled four surveys of HCV in Cameroon and found five genotype 1 subtypes, one genotype 2 subtype and three genotype 4 subtypes. In Chapters 1 and 2 I reported subtypes 4c, 4h, 4k, 4q and 4r among a total of 2298 samples from the DRC; the higher diversity seen in this study among only 850 samples may be due to this study collecting samples from a different and possibly more diverse elderly population, whereas the samples analysed in chapters 1 and 2 were only from male military recruits.

This study includes only the second isolate of HCV genotype 7 sample yet discovered. The first genotype 7 virus (isolate QC69) was isolated in Canada from a DRC migrant (Murphy *et al.*, 2007), and the finding of a second isolate in a DRC resident supports the hypothesis that the DRC is the site of origin of genotype 7.

Figure 4.4 and 4.5 (next pages). Estimated maximum-likelihood phylogeny of all HCV sequences from the DRC, together with reference sequences. Figure 4.3 is estimated from the core region alignment, while Figure 4.4 is estimated from the NS5B alignment. Nodes are highlighted with a red circle if they have bootstrap support >70%, and the phylogeny is midpoint rooted. Taxa names of new samples obtained in this study are highlighted in red. Reference sequences are labelled with their subtype and accession number, separated by an underscore.


G7
G2
G1
G3
G6
G5
4m
4o
4d
4l
4q
4v
4n
4a
4c
4t
4p
4f
4w
4b
G4
4r
4g
4k
0.2


G7
G6
G1
G5
G2
4g
4k
4f
4p
4t
4d
4l
4n
4v
4g
4m
4a
4c
4r
4o
4b
4w
G4
G3
0.3

Genotype/ subtype	Identifiers of samples obtained in this study
7	619
1	357
4?	697, 737, 844
4a	333
4b	259, 410
4c	116, 180, 233, 271, 447, 492, 521, 547, 632, 682, 876
4g	280, 399
4k	10, 31, 34, 37, 45, 68, 101, 109, 118, 148, 194, 206, 230, 251, 270, 285, 297, 298, 314, 315, 328, 354, 370, 373, 401, 417, 418, 433, 436, 442, 476*, 481, 509, 542, 545, 582, 596, 601, 722, 787, 794, 817, 836, 842, 859, 864, 866, 873
4l	310, 690
4n	671
4q	335, 458
4r	25, 28, 48, 69, 77, 103, 146, 182, 217, 284, 293, 301, 302, 303, 310B, 352, 372, 413, 420, 476*, 477, 486, 503, 528, 550, 571, 679, 693, 698, 702, 734, 738, 756, 761, 802, 805, 851, 858, 874, 882
4v	374
4w	579, 862

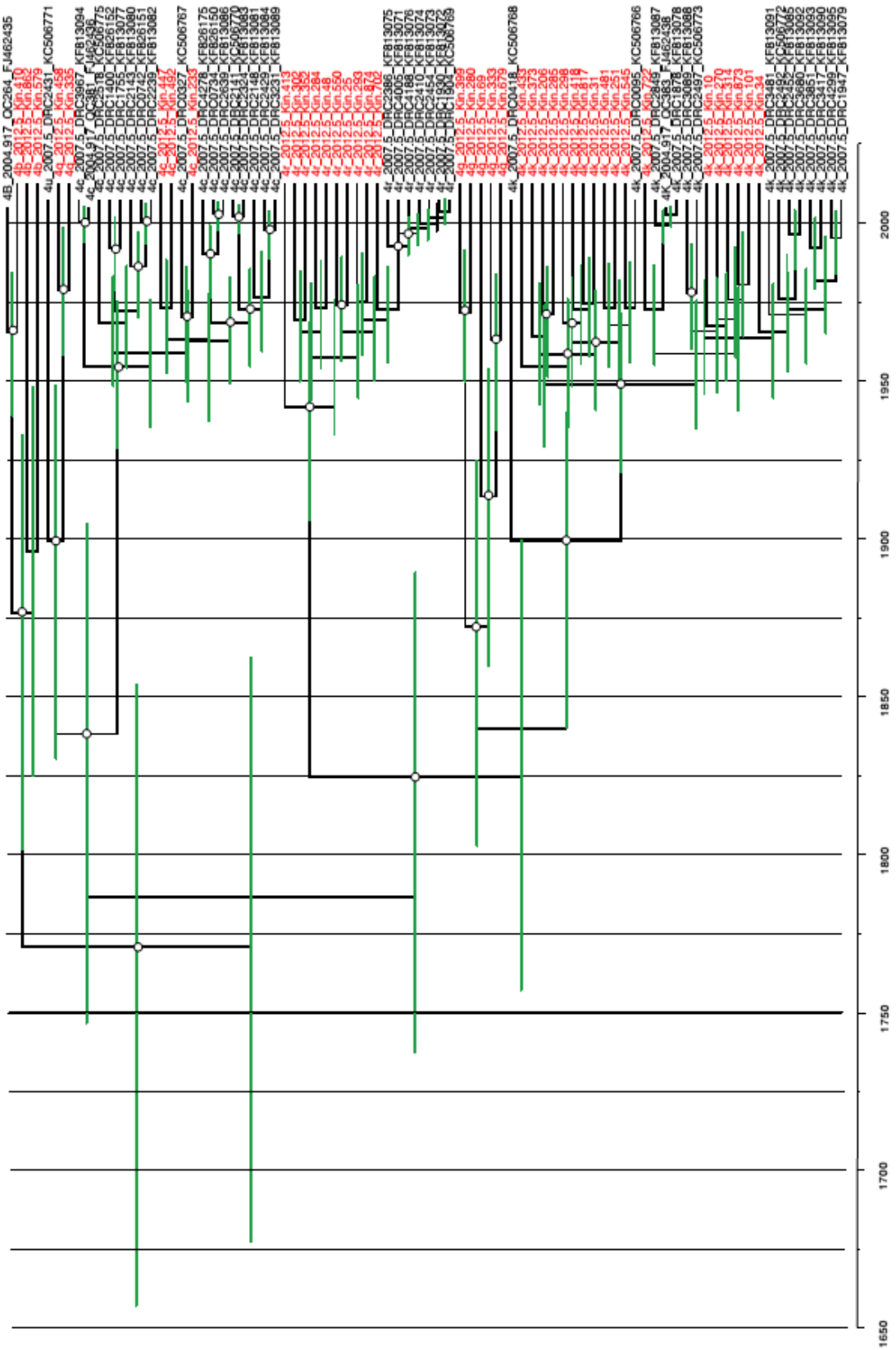
Table 4.4: *subtype classification of samples obtained in this study. Samples designated 4? were classified as genotype 4 but do not belong to a known subtype. Sample 476 (marked with a *) contained 4r sequence in the core region and 4k sequence in the NS5B region.*

The p-distance between the previously discovered isolate QC69 and the genotype 7 virus discovered here (Kin.619) is 0.19, indicating that the two isolates are likely to be different subtypes of genotype 7 and suggesting that there is likely to be a great deal more diversity of HCV genotype 7 as yet undiscovered in the DRC. Further, one sample that was not classified into any subtype (Kin.690) was very similar to the 3 samples that constitute the provisional subtype named 4drc that I discovered in chapter 2, giving further support to the classification of this lineage as a new subtype circulating in the DRC.

4.3.2 Coalescent Analysis of DRC samples

Figure 4.6 shows a molecular clock phylogeny estimated from the concatenated alignment of all genotype 4 isolates sampled in the DRC or from known DRC migrants. This analysis estimates that the most recent common ancestor (MRCA) of genotype 4 in the DRC existed around 1771 (95% HPD: 1652, 1849), very similar to the date estimated in chapter 2 for the MRCA of genotype 4 strains from all countries, which was 1733 (95% HPD: 1650, 1805). This supports the hypothesis that genotype 4 originated in or around the DRC, as there is no evidence of any founder effects reducing the MRCA in the DRC.

Figure 4.6 (next page). Maximum clade credibility molecular clock phylogeny, estimated from the concatenated alignment of all HCV genotype 4 sequences from the DRC. Branch lengths represent time (see scale bar at the bottom of the figure). Nodes with a posterior probability >0.9 are labelled with a white circle, and each node has a bar showing the highest posterior density 95% credible regions for its age. Samples newly generated in this study are marked in red. Sequences are labelled with their subtype, date of collection and accession number.


2000
1950
1900
1850
1800
1750
1700
1650

Compared to other analyses of DRC isolates (and Chapter 2 in particular), the sequences gathered in this study were widely dispersed within each subtype, indicating that HCV in the sample population is highly diverse and may be more representative of overall HCV diversity in the DRC than previous studies. HCV strains isolated from subjects who had migrated from the DRC to Canada are very genetically similar to DRC samples, suggesting that those individuals were infected in the DRC. Strains QC383 and QC381, sampled in Canada, diverged from their closest African relatives after 2000. QC264 diverged from its closest relative in 1965 (95% HPD: 1939, 1984), although the high diversity and low prevalence of subtype 4b may have led to under-sampling of that subtype's genetic diversity.

4.3.3 Coalescent analyses at the genotype level

Figure 4.7 shows coalescent-based reconstructions of the demographic history of HCV genotype 4 in the DRC using the Bayesian skyline, skygrid and CEC coalescent models. The graphs plot effective population size multiplied by the generation time (*NeT*) through time. The alignment comprised concatenated Core+NS5B sequences of all DRC genotype 4 isolates that have been sequenced in both these regions, henceforth called the comprehensive concatenated alignment (n=82).

The CEC model (green) shows the simplest pattern; the virus' ancestral *NeT* starts at a low level of roughly 300 with a comparatively small 95% credible region. Effective population size stays constant until roughly 1960, at which time there is rapid exponential growth until around 1975 that increases *NeT* to around 2000 and this value is maintained until the present. The Bayesian skyline analysis shows a similar pattern, with an ancestral *NeT* of ~200 to ~400. Fluctuations in *NeT* are not significant (given the credible regions) until around 1960, after which the population undergoes exponential growth until around the year 2000. Unlike the CEC analysis, the Bayesian

skyline plot shows a second phase of population growth around the year 2000, during which it approximately doubles in size. This second period of growth may represent a real demographic trend undetected by the CEC model (which, by definition, can only exhibit one phase of exponential growth).

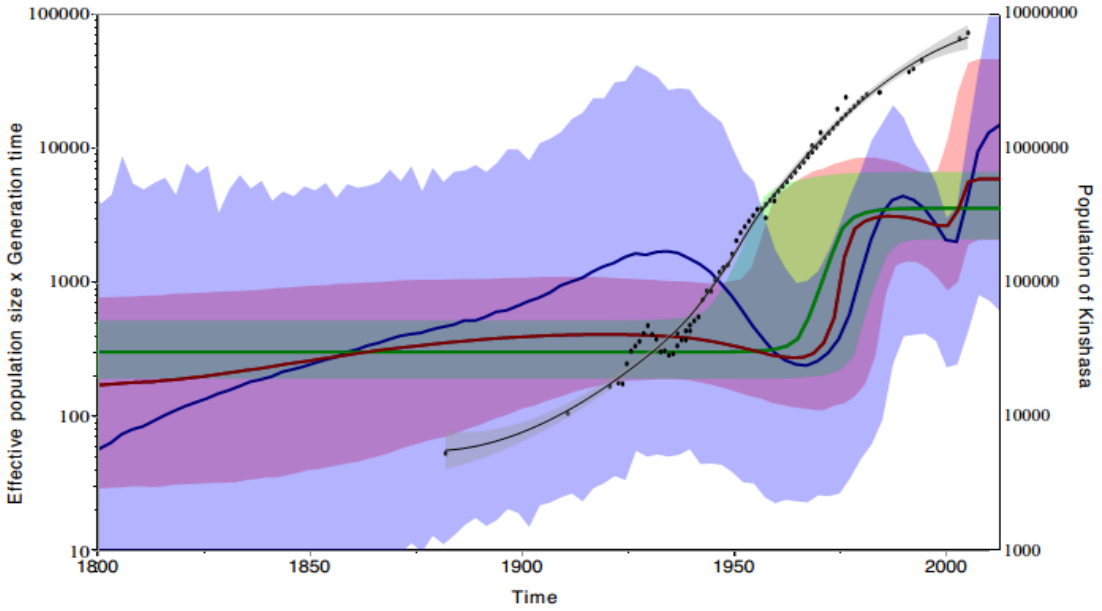


*Figure 4.7. Reconstructed population growth curves of HCV genotype 4 through time, estimated from the concatenated comprehensive alignment using different coalecent models. Each graph shows the median (thick line) and 95% highest posterior density intervals (shaded area) of NeT (defined as effective population size*generation time). NeT is shown on a log scale. This plot combines the results from three different models, colour coded as follows: skyline model = red, CEC model = green, skygrid model = blue. Black dots show the population of Kinshasa, also on a log scale, and the black line is a Lowess curve fitted to the population with shaded areas showing 95% confidence intervals (population data from Faria et al., 2014).*

Alternatively, the second growth phase may be an analysis artifact, arising from the inclusion of (i) multiple viral lineages with different epidemic histories into a single analysis, superimposing multiple curves upon each other and giving the appearance of multiple peaks, and/or (ii) a subset of samples that are epidemiologically linked from a recent epidemic and which therefore cannot be said to be 'randomly sampled' from the study population and thus falsely exaggerate recent growth in *NeT* (Holmes *et al.,* 1999; Pybus *et al.,* 2009). The skyline plot has much larger confidence limits than the CEC model in the pre-1950 and post-2000 periods, but for the first growth period the two models show similar levels of uncertainty. Comparison of the CEC and Bayesian skyline results illustrates their respective advantages and disadvantages. The former has smaller credible regions but may fail to fully reconstruct more complex population size trajectories.

The skygrid plot (blue) exhibits far larger 95% credible regions than either of the other two models. The trajectory of median *NeT* through time follows approximately the same pattern exhibited by the CEC and Bayesian skyline methods, but with the addition of substantial oscillations; every large change in median Ne is followed by a change in the opposite direction. This is particularly pronounced during 1850-1950 and exceeds the 95% credible region of the skyline plot estimate during this time. These fluctuations are probably a result of the way in which the skyride/skygrid models "smooth" changes in Ne through time, such that they are less suitable to populations that have undergone very rapid changes in Ne. Simulated data provided by Ho and Shapiro, 2011 suggest that this smoothing effect may mask very rapid changes in *NeT*. Figure 4.8 shows the data presented by Ho and Shapiro, 2011, which suggests that the GRMF smoothing used in the skyride and skygrid methods underestimates the rate of population growth when *NeT* is subjected to an instantaneous 'step' increase.

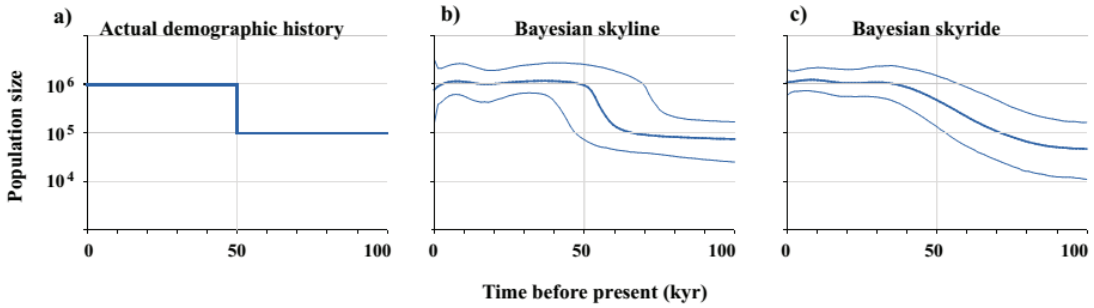


Figure 4.8. Estimated population growth curves from different skyline plot methods on a data set simulated under a model of step-wise population growth. Figure is reproduced from Ho and Shapiro, 2011 (see therein for simulation and analyses details). Panel (a) shows the actual demographic history of the simulated population, while the other plots show the population growth curves reconstructed by the (b) Bayesian skyline and (c) skyride models. In (b) and (c) the middle line represents the median estimate of Ne and the upper and lower lines represent the 95% highest posterior density intervals of that estimate. Ne is shown on a log scale.

Although the skygrid is thought, in general, to show less uncertainty than the skyline or skyride models (Gill *et al.,* 2013), on our data its credible regions were considerably larger. Thus the skygrid model is perhaps not optimal when applied to single-locus data sets of comparatively small size such as that analysed here; when applied to multiple independent loci the skygrid model appears to show more accuracy than other models (Gill *et al.,* 2013).

Finally, the black dots show the estimated population of Kinshasa, and the black line shows a Lowess curve fitted to the population with the shaded area representing 95% confidence intervals. This data, reproduced from Faria *et al.,* 2014, clearly demonstrates that while the population of Kinshasa went through a great increase halfway through the past century its period of growth does not match with that of

HCV, and so it is unlikely that the exponential growth in HCV was simply caused by the growing population of the city.

4.3.4 Coalescent analyses at the subtype level

As discussed in section 4.3.3, coalescent inferences obtained by combining different HCV subtypes (i.e. Fig 4.7) are difficult to interpret, as some lineages may undergo exponential growth at different times and at different rates, while others may never leave the endemic state. It is therefore more usual for coalescent methods to be applied separately to individual HCV subtypes. Thus to learn more about the HCV epidemic in Kinshasa, I performed separate coalescent analyses on subtype 4k and 4r alignments. Only these two subtypes equaled or exceeded the minimum of 16 samples needed for reliable analysis (see section 4.2.3).

For each subtype I analysed six alignments – two each for the core, NS5B and concatenated core+NS5B regions, containing the samples generated in this study and all sequences gathered from the DRC respectively. Additionally, I used multiple analyses to explore the robustness of the results to model specification. Each alignment was analysed using four models, representing all combinations of Bayesian skyline vs CEC demographic models, and strict vs relaxed molecular clock models. Following the results in section 4.3.3, the Bayesian skygrid model was determined to be unsuitable for these data and was not used.

Figure 4.9 shows the results of these coalescent analyses. Each plot of epidemic history is scaled to fit on the same axes and organized by coalescent method and alignment type. Although there are some differences, detailed below, it is notable that all subtypes, alignment types and analysis methods give largely congruent epidemic growth patterns. Prior to 1900, HCV subtypes 4k and 4r were circulating at an *NeT* of

~100. Between ~1925 and ~1965 *NeT* values start to increasing exponentially, reaching a plateau of ~1000 to ~10,000 sometime between ~1960 and ~1980. As in the previous section, the CEC model generally exhibits smaller credible regions than the Bayesian skyline model, which is to be expected due to its low number of parameters.

Both strict and relaxed clock skyline models show similar levels of uncertainty, with the exception of the concatenated alignments, for which the strict clock model gives larger 95% credible regions. It is not obvious what might cause this increased uncertainty; the sample sizes of the concatenated alignments are generally smaller. For the CEC coalescent model, the relaxed clock analyses had credible regions that were slightly but consistently larger than those produced by the strict clock model.

When the same coalescent method is applied to subtypes 4k and 4r, there is remarkably little difference seen between the estimated population trajectories for the two subtypes. The initial size of the HCV population is similar for each, and the period of exponential growth seems to start at the same time. The second brief period of exponential growth around 2000, noted in section 4.3.3, is present here (particularly in the skyline analyses of NS5B alignments that contain sequences from the online databases) but is less pronounced and not significant given the width of the credible regions. It is possible that some NS5B sequences collected from online databases may be epidemiologically linked and responsible for the appearance of a second upturn in the plot near the present.

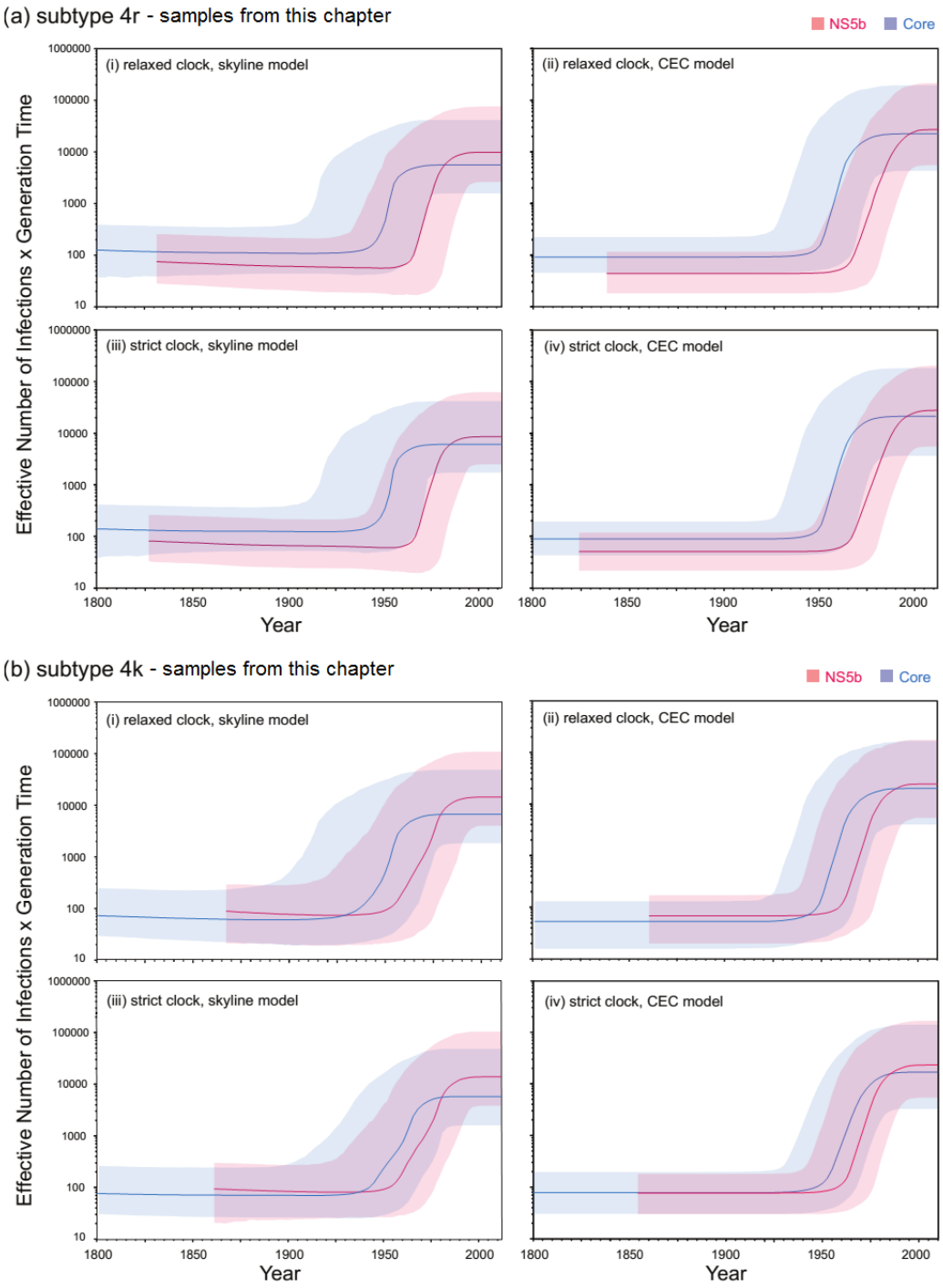
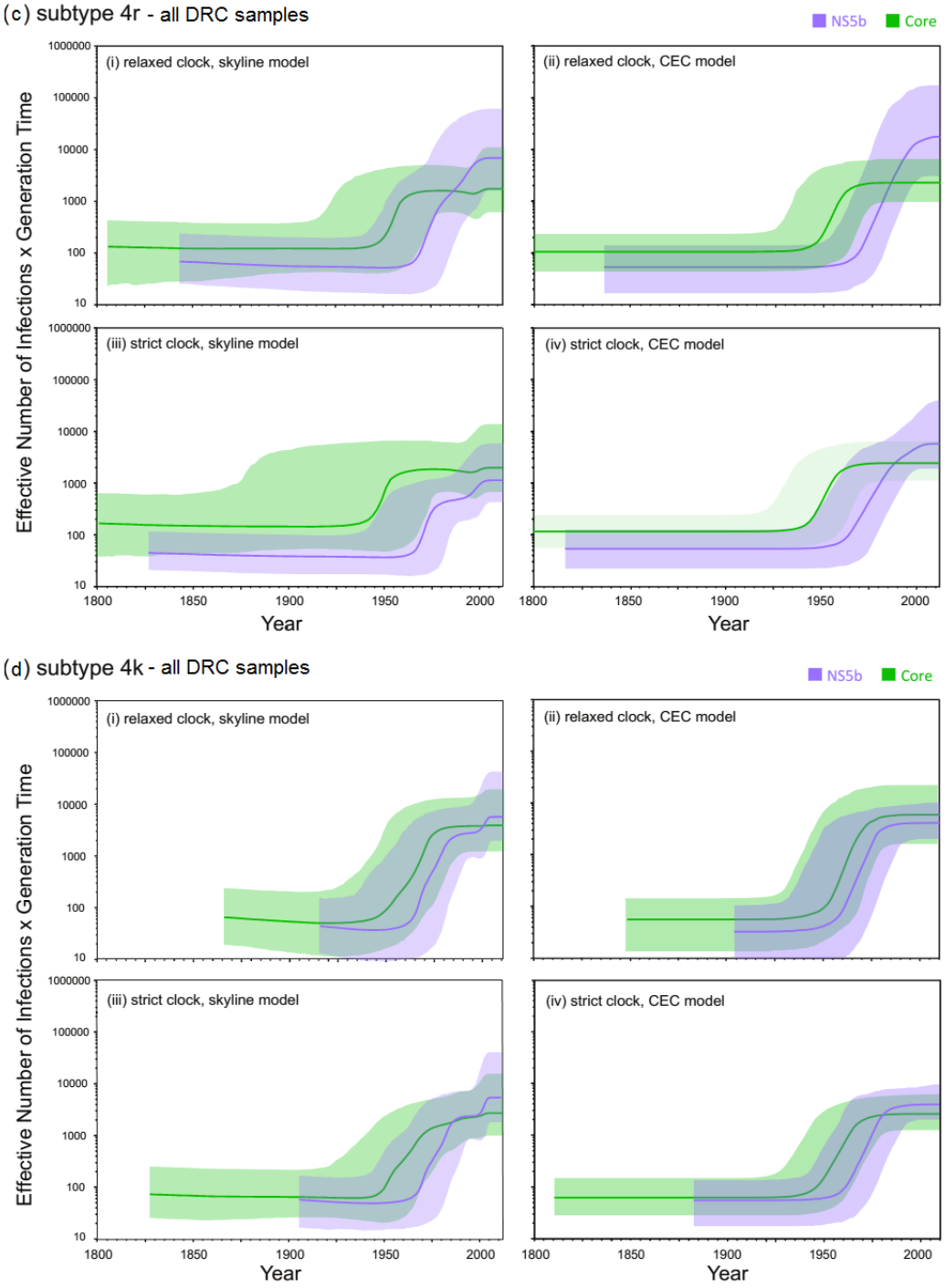


Figure 4.9. Reconstructed population growth curves over time of HCV subtypes 4r and 4k. Each graph plots the median (black line) and 95% highest posterior density intervals (shaded area) of Ne. Ne is shown on a log scale, and each plot is scaled to the same axes to aid comparison.

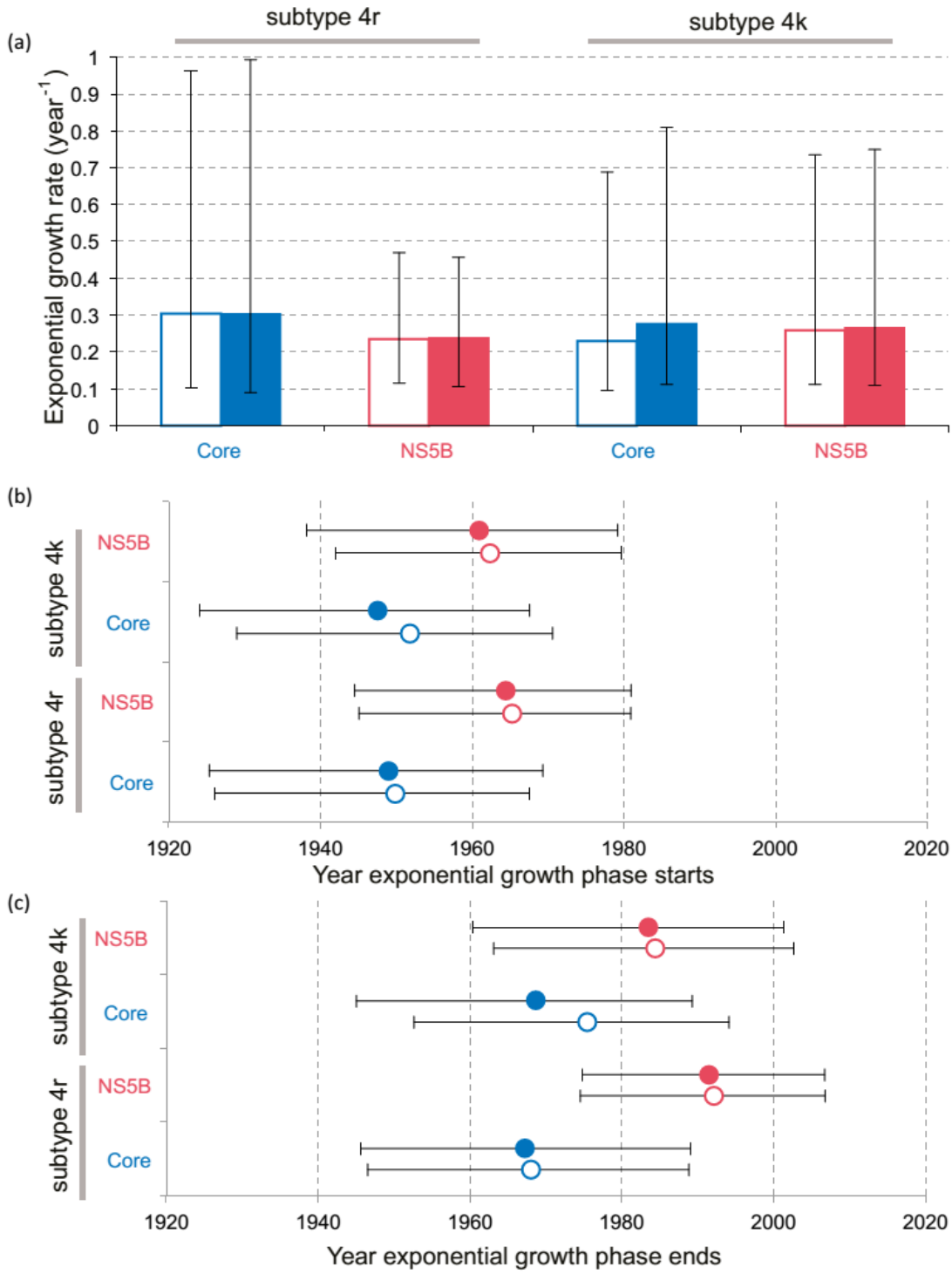


Each column displays the results from a particular demographic model, while each row represents a particular clock model. Each plot combines analyses from multiple genome regions (corresponding to the columns in Table 4.2), as indicated on the colour key for each section.

4.3.5 Estimated Epidemic Parameters

To quantify the three phase epidemic history revealed in Figure 4.9, I investigated the estimated parameter values of the CEC coalescent model. Figure 4.10a and d show the median estimates and 95% credible regions for *r* (the exponential growth rate) obtained from different data sets and models. Figures 4.10b, c, e and f show the corresponding values for the start and end times of the exponential growth phase, respectively.

The estimated median *r* values for subtype 4r are between 0.201 and 0.311, and for subtype 4k are between 0.198 and 0.266. These *r* values are faster than those previously estimated for epidemic subtypes; Pybus *et al.,* 2001 and 2005 gave *r* estimates of between *r*=0.079 and *r*=0.104 for 1a, 1b and 3a, while Nakano *et al.,* 2004 estimated an *r* for 1a of between 0.078 and 0.185 and for 1b of between 0.04 and 0.119, varying based on country of sampling. In fact, the values generated in this sample are most similar to the *r* values estimated for the Egyptian HCV epidemic (between r=0.237 and r=0.264; Pybus *et al*., 2003). This is consistent with the idea that a parenteral HCV transmission event similar to that experienced in Egypt may have occurred in the DRC; it is not obvious what other factors could explain this high rate of growth. The analyses place large credible regions on *r* and in many cases much higher growth rates cannot be rejected, but even at their lowest values these rates represent rapid growth. Additionally, there is a difference between estimates of *r* obtained from strict and relaxed clock analyses, with the strict clock estimates consistently higher than those from the relaxed clock analyses.


(a)
subtype 4r
subtype 4k
Exponential growth rate (year^{-1})
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Core
NS5B
Core
NS5B
(b)
subtype 4k
NS5B
Core
subtype 4r
NS5B
Core
1920
1940
1960
1980
2000
2020
Year exponential growth phase starts
(c)
subtype 4k
NS5B
Core
subtype 4r
NS5B
Core
1920
1940
1960
1980
2000
2020
Year exponential growth phase ends

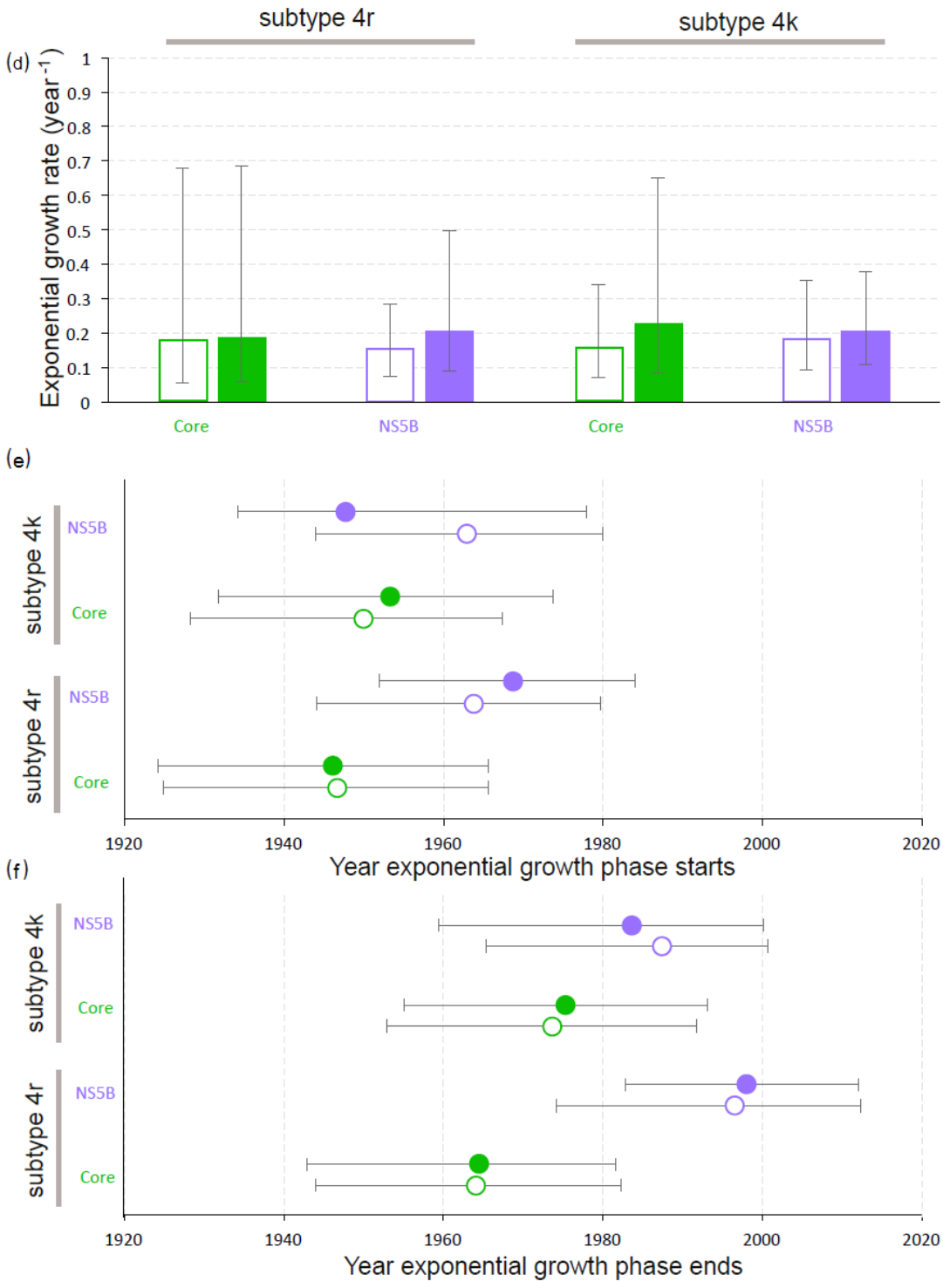

subtype 4r
subtype 4k
(d)
Exponential growth rate (year-1)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Core
NS5B
Core
NS5B
(e)
subtype 4k
NS5B
Core
subtype 4r
NS5B
Core
1920
1940
1960
1980
2000
2020
Year exponential growth phase starts
(f)
subtype 4k
NS5B
Core
subtype 4r
NS5B
Core
1920
1940
1960
1980
2000
2020
Year exponential growth phase ends

Figure 4.10 (previous page). Estimated parameters of the CEC demographic model, for the subtype 4r and 4k data sets. Boxes and points represent the median estimate for each parameter, while error bars show the 95% HPD confidence intervals. Open boxes and points represent values estimated using a strict molectular clock model; solid boxes and points represent those estimates using a relaxed molecular clock model. The graphs are dived according to the sample group used: graphs (a)-(c) display estimates obtained from alignments containing only the isolates obtained in this study, while graphs (d)-(f) display estimates obtained from alignments containing the isolates from this study, plus all other HCV 4r and 4k sequences from the DRC available in GenBank. (a) and (d) show the estimates for r, the exponential growth rate; (b) and (e) show estimates for the start time of the exponential phase; (c) and (d) show estimates for end of the exponential phase.

Figures 4.10b, c, e and f show the estimates for the start and end times of the exponential growth phase. Averaging across analyses, the results indicate that the exponential phase began in 1955 (with estimates ranging between 1946 and 1969) and ended in 1973 (with estimates ranging between 1964 and 1998). The estimated epidemic start date is similar to that seen for HIV-1 in the DRC, which appears to have moved from a constant population size to exponential growth in the 1950s (Worobey *et al*., 2008), and whose expansion has been hypothesized to be caused by urban population growth in the DRC (Worobey et al. 2008) and by intravenous syphilis treatment, performed on 7500 people between 1949 and 1954 (Pépin, 2012). This treatment may have transmitted a hepatitis-causing agent (most likely hepatitis B virus) that caused the 'inoculation hepatitis' outbreak seen in 1951-1952 in Kinshasa (then Léopoldville) and reported in Beheyt, 1953, and so it is possible that HCV was transmitted in this manner too.

4.3.6 Model Comparison

Figure 4.11 shows the AICM values for each of the models used, on seven of the DRC alignments. Lower values represent a better fit to the data. For the comprehensive concatenated alignment of all samples from the DRC, the relaxed clock skyline model is the best fit. The skyline model might be expected to provide a better fit for this data set, as its estimated population trajectory was more complicated than the CEC model allows for (Fig 4.7), and the skyline model is thought to outperform the skygrid model when only one locus is used and the population trajectory is complex (Gill *et al.,* 2013).

For the subtype 4k and 4r alignments, the model selection results seem to vary according to the number of taxa in the alignment. In the alignments with most taxa (NS5B) the strict clock is the clear best fit, with the CEC model slightly preferred for subtype 4k and the skyline model preferred for 4r. With the smaller alignments (concatenated and core), the preferred clock model seems to vary by subtype; for subtype 4k, the relaxed clock is a better fit than the strict clock in all analyses, while for 4r the strict clock is universally preferred for both CEC and skyline models in all alignments. The choice of skyline and CEC models has less impact on the model's AICM than the choice of clock model, implying that the two coalescent models work similarly well on single-lineage data sets.

The differing model selection scores for the relaxed and strict clock models may be due to greater rate heterogeneity in the subtype 4k core and concatenated alignments than in the 4r alignments; the relaxed clock coefficient of variation (CoV) estimated from the 4k concatenated alignment was 0.387 (95% CI: 0.214, 0.591), compared to 0.108 (95% CI: ~0, 0.253) for the concatenated 4r skyline analysis.

Figure 4.11. Summary of the Akaike's information criterion through MCMC (AICM) values estimated for each alignment. Lower AICM values indicate a better model fit, and the models named in bold are those with the lowest AICM value for each alignment. Bars represent the estimated value, while error bars represent the estimate's standard error generated through 200 bootstrap replicates.

The same pattern exists in the core relaxed clock skyline analyses, with CoV for subtypes 4k and 4r being 0.494 (95% CI: 0.319, 0.682) and 0.159 (95% CI: 0.0009, 0.33) respectively. In every subtype 4k relaxed clock analysis the CoV estimated excluded zero, whereas relaxed clock analyses of subtype 4r were much less likely to reject a strict clock (i.e. CoV~0). This suggest that when the data contains significant rate heterogeneity among branches, the increased fit of a relaxed clock model to the data more than compensates for increased model complexity.

4.4 DISCUSSION

The first goal of this study was to phylogenetically analyse new HCV sequence data from the DRC to generate a more comprehensive assessment of HCV diversity in the country, and was successful in this aim. I found HCV isolates in the DRC from three different genotypes, including only the second genotype 7 strain to have been reported. Only one sample belonged to genotype 1, but I found eleven different subtypes from genotype 4, including one potential recombinant isolate and three samples too divergent to be classified into any subtype. This high level of diversity provides evidence that HCV is long-established in this region. Of the three samples too divergent to fall into any established subtype, one associated closely with the three samples termed 4drc in Chapter 2, but as all four of these samples were only sequenced in the NS5B region there is not yet enough evidence to claim they form a new subtype. As more sequences are gathered from the DRC and are sequenced in other genome regions, this group may develop into clade that meet all the criteria for defining a new subtype.

The new sequences generated in this study make possible the use of coalescent methods that could not be previously applied due to low sample numbers. The

coalescent analyses reported here provide the first estimates of the epidemic history of HCV lineages in the DRC. In Chapter 2 I reported a 'cohort effect', whereby older individuals from the DRC military were more likely to be infected with or have antibodies to HCV, particularly those born before 1950. The plots in Figures 4.6 and 4.8, provide a historical explanation for this age distribution; rapid growth in the effective population size of HCV subtypes 4k and 4r in the DRC place between ~1950 and ~1973, and individuals too young to be exposed to HCV or not yet born would not have been infected during this growth period.

Although it is not possible to determine the cause of the epidemic from sequence data alone, it is notable that subtypes 4k and 4r show near-identical changes in *NeT*, strongly suggesting the influence of a population-level factor that increased the transmissability of all HCV strains during the mid to late twentieth century, rather than a viral genetic factor specific to one HCV subtype. Additionally, I know of no evidence of a cultural or behavioural shift in injecting drug use in the DRC; any causative factor must be large enough to explain the epidemic yet sufficiently short-lived that those born later would have lower HCV prevalence. Medical parenteral interventions such as the intravenous syphilis treatment performed on a significant fraction of the Léopoldville population between 1949 and 1954 (Pépin, 2012) could have spread HCV in a manner consistent with my results. While that particular public health campaign finished too early to coincide with the entire exponential growth phase of subtypes 4k and 4r, it may have amplified HCV to a large enough prevalence that other transmission routes (iatrogenic and non-iatrogenic) may have been responsible for the rest of the growth. My coalescent results indicate that HCV subtypes 4k and 4r were rapidly growing during the Congo Crisis of 1960-65, when large numbers of foreign troops were deployed in the DRC. Blood transfusions and/or

parenteral medical treatments may have been more frequent as a result of the crisis, and this could have extended the exponential growth phase beyond 1954. The epidemic histories estimated for the Kinshasa-only and "all DRC" alignments show very similar growth, implying that the HCV epidemic in Kinshasa encompassed the majority of HCV growth in the DRC in this time period or that the dynamics of HCV transmission outside and inside of Kinshasa were very similar. Information on where in the DRC patients were infected or samples were collected was missing for the great majority of sequences in this analysis, however, so it may simply be that this analysis includes few sequences originating from outside Kinshasa.

This study illustrates the use of HCV prevalence and diversity as a marker for the past transmission of blood-borne viruses, and it is interesting to consider the implications for the Human Immunodeficiency Virus (HIV) of the results reported here, especially as the majority of HCV sequences analysed originated from Kinshasa. My estimates of HCV's epidemic growth phase cover the period during which HIV-1 is predicted to have arrived in Kinshasa, undergone rapid expansion, and spread to the rest of the world (Pépin, 2013). It is difficult to use epidemiological survey data to study this period of the HIV epidemic, as those infected with HIV-1 in the 1950s are unlikely to have survived to today, and so HCV provides our best chance to evaluate the risk of parenteral HIV-1 transmission in the past. HCV evolves very rapidly (albeit slightly slower than HIV), and HCV recombination is rare; consequently, HCV samples collected in the present can provide information about viral transmission dynamics 60 years ago or more, as demonstrated by studies of HCV in Egypt (Frank *et al.,* 2000; Pybus *et al.,* 2003). Further, different lineages of HCV (i.e. subtypes 4k and 4r) can be used as independent observations of past epidemic dynamics in the region, and while other subtypes collected in this study such as 4c had sample sizes too small for full

analysis they showed a similar pattern. The CEC model enable hypotheses concerning epidemic history to be tested directly by estimating the parameters of the exponential phase directly from modern sequence data, with appropriate confidence limits. The estimated dates when growth began and ceased are invaluable when attempting to interpret viral gene sequences in the context of other data, for example risk factor questionnaires and historical research that generate information on which risk factors might be responsible. This approach could be extended to infer the historical transmission of a wide range of other blood-borne viruses, such as the hepatitis B virus (HBV) and the human T-lymphotrophic virus type 1 (HTLV-1).

This study also provides an opportunity to contrast the output of different coalescent methods and models on alignments of limited size. Analyses using small numbers of sequences can struggle to generate results with narrow confidence limits but in many situations (e.g. samples gathered during an ongoing epidemic, or population surveys of viruses that are at low prevalence) larger alignments may not be available. In such cases choosing the most appropriate coalescent model may improve the reliability of the results obtained. This issue is particularly significant for HCV because recombination is rare, hence analyses cannot use multiple independent loci to boost the amount of information contained in a set of sequences about population dynamic history (Heled and Drummond, 2008), as is possible for influenza A (Rambaut *et al.,* 2008).

In general when analyzing small data sets, the most precise results may be gained by using a parametric model that matches the demographic history of the sampled population if the history is known. If it is unknown then the skyline plot model (rather than the skyride/skygrid model) is likely the best choice to determine the trajectory of *NeT*, after which an appropriate parametric model may be used. Even when model

selection results indicate that a parametric model is a worse fit than the skyline model, it may still be helpful in providing estimates of population dynamics parameters such as exponential growth rates. The model test results also suggest that both relaxed and strict clocks may be applicable, as which was preferred seemed to vary among the alignments analysed. HCV can show a large degree of rate heterogeneity between branches (Salemi and Vandamme, 2002), and so it may be best to use a relaxed clock first and only use a strict clock if the coefficient of variation in the relaxed clock analysis fails to exclude zero, meaning that there is no evidence for rate variation among lineages. Lastly, it is important to note that even methods that give weak statistical support can still be useful in a Bayesian framework as they can demonstrate that different prior distributions and model specifications yield similar results. Here, almost every alignment and model analysed provided similar estimates of the epidemic history of HCV in Kinshasa.

4.5 REFERENCES

Araujo, A. C. (2012). Antibody- and genome-based identification of recent HCV infection. *Antiviral Therapy* 17: 1459-1464.

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A. and Alekseyenko, A. V. (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* 29:2157-67.

Beheyt, P. (1953). Contribution à l'étude des hépatites en Afrique L'hépatite épidémique et l'hépatite par inoculation. *Ann Soc Belge Med Trop,* 33: 297-340.

Cantaloube, J.-F., Gallian, P., Bokilo, A., Jordier, F., Biagini, P., Attoui, H., Chiaroni, J. and de Micco, P. (2010). Analysis of hepatitis C virus strains circulating in Republic of the Congo. *Journal of Medical Virology,* 82: 562–567.

Choo, Q. L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W. and Houghton, M. (1989). Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science* 244, 359–362.

Daniels, D., Grytdal, S. and Wasley, A. (2009). Centers for Disease Control and Prevention. Surveillance for acute viral hepatitis – United States, 2007. *MMWR Surveill Summ* 58:1-27

Dearlove, B. and Wilson, D. J. (2013). Coalescent inference for infectious disease: meta-analysis of hepatitis C. *Phil Trans R Soc B* 368: 20120314.

Deuffic, S., Buffat, L., Poynard, T. and Valleron, A-J. (1999). Modeling the hepatitis C virus epidemic in France. *Hepatology* 29: 1596-601.

Drummond, A. J., Rambaut, A., Shapiro, B. and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* 22:1185-1192.

Drummond, A. J., Suchard, M. A., Xie, D. and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969–1973.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedfore, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pepin, J., Posada, D., Peeters, M., Pybus, O. G. and Lemey, P. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science* In Press.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164–166.

Frank, C., Mohamed, M. K., Strickland, G. T., Lavanchy, D., Arthur, R. R., Magder, L. S., Khoby, El, T., Abdel-Wahab, Y., Aly Ohn, E. S., Anwar, W. and Sallam, I. (2000). The role of parenteral antischistosomal therapy in the spread of Hepatitis C Virus in Egypt. *Lancet* 355:887–891.

Fu, Y-X. and Li, W-H. (1999). Coalescing into the 21st Century: an overview and prospects of coalescent theory. *Theoretical Population Biology* 56: 1-10.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B. and Suchard, M. A. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution* 30: 713-724.

Gray, R. R., Parker, J., Lemey, P., Salemi, M., Katzourakis, A. and Pybus, O. G. (2011). The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evolutionary Biology* 11: 131.

Gray, R. R., Tanaka, Y., Takebe, Y., Magiorkinis, G., Buskell, Z., Seeff, L., Alter, H. J. and Pybus, O. G. (2013). Evolutionary analysis of hepatitis C virus gene sequences from 1953. *Philosophical Transactions of the Royal Society* 368: 20130168.

Heled, J and Drummond, A. J. (2008). Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology* 8: 289.

Ho, S. Y. W. and Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol Ecol Resour* 11: 423-34.

Holmes, E. C., Pybus, O. G. and Harvey, P. H. (1999). The molecular population dynamics of HIV-1. In The Evolution of HIV (KA Crandall, editor) pp. 177-207. John Hopkins University Press

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability* 19: 27-43.

Kuiken, C., Yusim, K., Boykin, L. and Richardson, R. (2005). The Los Alamos hepatitis C sequence database. *Bioinformatics* 21: 379–384.

Minin, V. N., Bloomquist, E. W. and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution* 25: 1459-1471.

Mizokami, M. and Tanaka, Y. (2004). Molecular evolutionary analysis predicts the incidence of hepatocellular carcinoma in the United States and Japan. *Cancer Chemother Pharmacol* 54: S83-S86.

Murphy, D. G., Willems, B., Deschenes, M., Hilzenrat, N., Mousseau, R. and Sabbah, S. (2007). Use of Sequence Analysis of the NS5B Region for Routine Genotyping of Hepatitis C Virus with Reference to C/E1 and 5' Untranslated Region Sequences. *Journal of Clinical Microbiology* 45: 1102–1112.

Nakano, T., Lu, L., Liu, P. and Pybus, O. G. (2004). Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. *The Journal of Infectious Diseases* 190: 1098-1108.

Nerrienet, E., Pouillot, R., Lachenal, G., Njouom, R., Mfoupouendoun, J., Bilong, C., Mauclere, P., Pasquier, C. and Ayouba, A. (2005). Hepatitis C virus infection in Cameroon: a cohort-effect. *Journal of Medical Virology* 76: 208-214.

Njouom, R., Nerrienet, E., Dubois, M., Lachenal, G., Rousset, D., Vessière, A., Ayouba, A., Pasquier, C. and Pouillot, R. (2007). The hepatitis C virus epidemic in Cameroon: genetic evidence for rapid transmission between 1920 and 1960. *Infection, Genetics and Evolution* 7: 361–367.

Njouom, R., Caron, M., Besson, G., Ndong-Atome, G. R., Makuwa, M., Pouillot, R., Nkoghé, D., Leroy, E. and Kazanji, M. (2012). Phylogeography, risk factors and genetic history of hepatitis C virus in Gabon, central Africa. *PLoS One* 7: e42002.

Pépin, J., Lavoie, M., Pybus, O. G., Pouillot, R., Foupouapouognigni, Y., Rousset, D., Labbé, A. C. and Njouom, R. (2010). Risk factors for hepatitis C virus transmission in colonial Cameroon. *Clin Infect Dis* 51: 768-76.

Pépin, J. (2012). The expansion of HIV-1 in colonial Léopoldville, 1950s: driven by STDs or STD control? *Sex Transm Infect* 88:307-312.

Pepin, J. (2013). The origins of AIDS: from patient zero to ground zero. *J Epidemiol Community Health* 67: 473-5.

Pybus, O. G., Rambaut, A. and Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155:1429-37

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C. and Harvey, P.H. (2001). The Epidemic Behavior of the Hepatitis C Virus. *Science* 292: 2323-2325.

Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H. and Rambaut, A. (2003). The epidemiology and iatrogenic transmission of Hepatitis C Virus in Egypt: a Bayesian coalescent approach. *Mol Biol Evol* 20: 381-387.

Pybus, O. G., Cochrane, A., Holmes, E. C. and Simmonds, P. (2005). The hepatitis C virus epidemic among injecting drug users. *Infection, Genetics and Evolution* 5: 131–139.

Pybus, O. G., Markov, P., Wu, A. and Tatem, A. J. (2007) Investigating the endemic transmission of the hepatitis C Virus. *International Journal of Parasitology* 37:839-49

Pybus, O. G., Barnes, E., Taggart, R., Lemey, P., Markov, P. V., Rasachak, B., Syhavong, B., Phetsouvanah, R., Sheridan, I., Humphreys, I. S., Lu, L., Newton, P. N. and Klenerman, P. (2009). Genetic History of Hepatitis C Virus in East Asia. *Journal of Virology* 83: 1071–1082.

Pybus, O. G. and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics* 10: 540-50.

Rambaut, A., Pybus, O., Nelson, M., Viboud, C., Taubenberger, J. and Holmes, E. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature* 453: 615-619

Salemi, M. and Vandamme, A.-M. (2002). Hepatitis C Virus Evolutionary Patterns Studied Through Analysis of Full-Genome Sequences. *Journal of Molecular Evolution* 54: 62–70.

Seeff, L. B., Miller, R. N., Rabkin, C. S., Buskell-Bales, Z., Straley-Eason, K. D., Smoak, B. L., Johnson, L. D., Lee, S. R. and Kaplan, E. L. (2000). 45-year follow-up of hepatitis C virus infection in healthy young adults. *Annals of Internal Medicine* 132: 105-11.

Selvarajah, S. and Busch, M. P. (2012). Transfusion transmission of HCV, a long but successful road map to safety. *Antiviral Therapy* 17: 1423-9.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus - 15 years on. *Journal of General Virology* 85: 3173–3188.

Smith, D. B., Bukh, J., Kuiken, C., Muerhoff, A. S., Rice, C. M., Stapleton, J. T. and Simmonds, P. (2014). Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: Updated criteria and genotype assignment web resource. *Hepatology* 59: 318–327.

Strimmer, K. and Pybus, O. G. (2001). Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution* 18: 2298-305.

Stumpf, M. P. H. and Pybus, O. G. (2002). Genetic diversity and models of viral evolution for the hepatitis C virus. *FEMS Microbiology Letters* 214: 143-152.

Sypsa, V., Touloumi, G., Tassopoulos, N. C., Ketikoglou, I., Vafiadis, I., Hatzis, G., Tsantoulas, D., Akriviadis, E., Delladetsima, J., Demonakou, M. and Hatzakis, A. (2004). Reconstructing and predicting the hepatitis C virus epidemic in Greece: increasing trends of cirrhosis and hepatocellular carcinoma despite the decline in incidence of HCV infection. *Journal of Viral Hepatitis* 11: 366-374.

Tanaka, Y., Hanada K., Mizokami, M., Yeo, A. E., Shih, J. W., Gojobori, T. and Alter, H. J. (2002). A comparison of the molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence in the United States will increase over the next two decades. *Proc Natl Acad Sci USA* 99: 15584-9.

Weiss, G. and von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics* 149: 1539-1546.

World Health Organization. Hepatitis C. Fact sheet N°164. Revised April 2014. http://www.who.int/mediacentre/factsheets/fs164/en/. Accessed July 29th, 2014.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J-J., Kabongo, J-M. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T. P. and Wolinsky, S. M. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455: 661-664.

Zwickl, D. J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.

5 DISCOVERY AND CHARACTERISATION OF A HCV RECOMBINANT CIRCULATING IN CAMEROON

Contributions by collaborators: Richard Njouom provided samples and performed initial genotyping (section 5.2.2). Nick Wong performed sample preparation for the Illumina RNAseq sequencing run. Tommy Lam wrote a computer script used to identify HCV-positive reads from the Illumina NGS data.

5.1 INTRODUCTION

The genetic diversity of HCV is exceptionally high. The virus is classified into seven genotypes (1-7) with an average of 35% nucleotide divergence between strains belonging to different genotypes (Smith *et al.,* 2013). All genotypes except 5 and 7 are subdivided into numerous subtypes (1a, 1b, 1c, 2a, 2b etc.), with about 25% divergence within a genotype and 15% within a subtype. Despite this large variation in nucleotide sequence, there has been little evidence that recombination plays more than a minor role in HCV evolution, particularly in comparison to HIV-1 (Galli and Bukh, 2014). It is only recently that natural HCV recombinants have been shown to be circulating in human populations; the first to be discovered was a subtype 2k/1b recombinant found in a number of injecting drug users in St. Petersburg, Russia, in 2002 (Kalinina et al., 2002; Morel et al., 2010; Raghwani et al., 2012). Since then more naturally-occurring recombinants have been found and in 16 cases the recombinant's full genome has been sequenced. Details of the HCV recombinants reported to date are shown in Table 5.1.

It is important to note the difference between intra-genotypic recombinants (recombinants between different subtypes within the same genotype) and inter-genotypic recombinant (those whose parental lineages belong to two different genotypes). While break points for intra-genotypic recombinants are seen across the genome, inter-genotypic recombinant break points are only seen in the NS2 gene or near the NS2/NS3 boundary (Table 5.1). Additionally, all inter-genotypic recombinants discovered so far are genotype 2-derived in the 5' end of the genome, although the genotype of the 3' end varies.

Subtypes	Estimated Recombination Break Point(s)		Source
	Gene regions	Nucleotide position	
1a/1c	E1, E2	1407, 2050	Cristina and Colina, 2006
1a/1c	Core, E1, E2, NS2, NS3	801, 1261, 2181, 3041, 3781	Ross *et al.,* 2008
1b/1a	NS5b	8320	Colina *et al.*, 2004
1b/1a	Core	387	Moreno *et al.*, 2009
2/5*	NS2/NS3	3420-3440	Legrand-Abravanel et al., 2007
2b/1a*	NS2/NS3	3405-3416	Bhattacharya *et al.,* 2011.
2b/1b*	NS2/NS3	3443	Yokoyama *et al.,* 2011.
2b/1b*	NS2/NS3	3399	Kageyama *et al.*, 2006
2b/1b*	NS2	3298-3305	Hoshino *et al.*, 2012
2b/6w*	NS2/NS3	3429	Lee *et al.,* 2010
2i/6p*	NS2/NS3	3405-3464	Noppornpanth *et al.*, 2006
2k/1b*	NS2/NS3	3175	Kalinina *et al.*, 2002
4d/4a	Between E2 and NS5A	Unknown	Calado *et al.,* 2011
6a/6o	NS5b	8345	Shi *et al.,* 2012
6e/6h	NS5b	8356	Shi *et al.,* 2012
6e/6o	NS5b	8358	Shi *et al.,* 2012
6n/6o	NS5b	8372	Shi *et al.,* 2012

Table 5.1. *Published HCV recombinants. Inter-genotypic recombinants are denoted with an asterisk. Nucleotide positions are relative to isolate H77.*

HCV recombinants are also classified according to their potential epidemiological significance. Unique recombinant forms (URFs) are those that have been found only in one patient (or in closely linked patients), while circulating recombinant forms (CRFs) are recombinant strains that have been found in multiple patients. Only the 2k/1b strain found by Kalinina *et al* (2002) and further investigated by Morel *et al., (*2010) and Raghwani *et al* (2012) has been confirmed as a HCV CRF. Although the 2b/1b recombinant reported by Hoshino *et al.*, (2012) was detected in two different patients, they both attended the same clinic and so it cannot be taken as definite

confirmation of a CRF. This implies that the epidemiological significance of recombination in HCV is low, alternatively, the prevalence of HCV may be underestimated due to the use of genotyping methods that are unlikely to detect recombination, such as serotyping and single-locus sequencing (Morel *et al.,* 2010).

With the exception of the recombinants discovered by Cristina and Colina, 2006 and Ross *et al.,* 2008, all HCV recombinants discovered have only one break point. This is a clear difference to recombination in other viruses, e.g. HIV, where CRFs can have multiple break points and recombination between CRFs can lead to increasingly complex patterns of genome inheritance (Lau and Wong, 2013).

Determining the mechanics of recombination in HCV and providing better tools to detect recombinant strain would be useful: different genotypes and subtypes can have different clinical profiles and respond differently to treatment (Chen *et al.* 2012). Additionally, recombination may accelerate the evolution of anti-drug resistance by bringing together multiple independent mutations and boosting the rate at which viruses adapt to combination antiviral therapy (Lai, 1992). The recent development of direct-acting anti-HCV drugs (Conteduca *et al.,* 2013) increases the importance of a potential role for HCV recombination in drug resistance evolution. There is already some evidence that the 2k/1b strain is less responsive to antiviral therapy than pure genotype 2 or 3 strains (Morel *et al.,* 2010). Determining the factors that underlie the frequency of recombination in HCV is therefore important to the on-going development of antiviral therapies, as will the development of protocols to correctly identify recombinant strains.

There has been much speculation about how HCV recombination might occur *in vivo*, especially as superinfection exclusion via CD-81 down-regulation is thought to make

infection of a single cell by multiple HCV strains unlikely (Tscherne *et al.*, 2007). Combined with the difficulty of studying HCV *in vivo*, this means that we have little information about the mechanisms that lead to recombination, but *in vitro* methods and studies of other positive-sense RNA viruses can provide clues.

A common model of recombination in RNA viruses is the copy choice mechanism, whereby the viral RNA-dependent RNA polymerase (RdRp) pauses during replication and switches from a donor template RNA to an acceptor template RNA without releasing the nascent strand being produced (Worobey and Holmes, 1999; Morel *et al.*, 2011). This mechanism may depend on specific and conserved sequences and RNA secondary structures in both strands, which could explain why there are trends in the location of viral recombination break points. For HCV, it could be that there is greater sequence similarity between different HCV genotypes in the NS2 region than elsewhere in the genome, limiting successful among-genotype recombination to that region (Table 5.1). Sequences belonging to the same genotype have more sequence similarity across the whole genome and therefore intra-genotype recombination may be permitted at a wider number of genomic locations.

A second model of RNA virus recombination posits recombination via RdRp-independent breakage and re-joining, whereby two strands of different strains broken via restriction enzymes or mechanical stress are fused together by self-ligation or cellular ligases (Galli and Bukh, 2014). This mechanism has been shown to create homologous and nonhomologous recombinants in Bovine viral diarrhea virus, a pestivirus in the same viral family as HCV, but the presence of this mechanism in HCV is yet to be shown (Gallei *et al.*, 2004).

The simplest method for detecting recombination in HCV is to amplify and sequence the sample in two separate genomic locations and construct phylogenetic trees of the sample and reference genomes for each location. If recombination has occurred then the sample will associate with different reference genomes at each location and statistical support for this association can be quantified with phylogenetic bootstrap scores (Felsenstein, 1985). While this is the simplest method, it is also the most vulnerable to false positive results; if two different HCV strains are present in the same sample then polymerase chain reaction (PCR) conditions and primers might favour amplifying one strain in one region and the other in a different genomic location (Walsh, Erlich and Higuchi, 1992), giving phylogenetically discordant results for the two genomic regions. Additionally, when two different strains are present in a sample the PCR enzymes might switch from one template to the other mid-amplification, especially if there is significant sequence similarity (Pääbo, Irwin and Wilson, 1990). A potential solution to these hurdles is to isolate individual sequences with bacterial plasmid cloning, and thus ensure that only one sequence is present in each reaction. If multiple clones have been sequenced and independently show support for recombination, you can be much more certain that recombination has occurred. Alternatively, primers can be designed such that they can amplify only one subtype of the virus (Worobey and Holmes, 1999).

Once a full genome of the recombinant has been obtained then the recombinant structure of the genome can be characterising using the phylogenetic bootscanning approach (Salminen *et al.*, 1995). This method divides the viral genome into small sliding windows (200-500nt) then estimates the genetic distances between sequences in each window using a maximum-likelihood criterion, repeated for 100-1000 bootstrap replicates. The replicates are then assessed to find the sequence the study genome is

most often closest to in each window, creating a plot of the relative relatedness of the target genome to a set of reference genomes over the length of the genome. If recombination has occurred then a dramatic switch between reference sequences is seen in the plot at the position of the recombination breakpoint. This method improves on two-locus phylogenetic analysis by indicating precisely where in the genome recombination has occurred, although it typically requires a whole genome sequence to be obtained.

Next-generation sequencing methods represent a promising alternative to traditional sequencing approaches. With these methods (e.g. RNAseq by Illumina) all RNA present in a sample is amplified non-preferentially after being digested into short strands, resulting in tens of millions of 100nt 'reads' (Metzker, 2010). This technique has the benefit of having no steps that may preferentially amplify one strain of the virus over another, as even cDNA synthesis takes place using random primers, and the comprehensive sequencing allows the generation of not only a majority consensus sequence but also the sequence of all minor variants and quasispecies (Mancuso *et al.,* 2011). As RNA sequencing is highly non-specific, many reads correspond to host and bacterial and therefore this method requires extensive computational analysis to extract, assemble and interpret the viral-specific reads.

In this chapter, I have applied all of these methods to two putative recombinants strains that were isolated in Cameroon in the pursuit of two objectives. First, I aim to confirm that these two samples are indeed recombinants and characterise their mosaic genome structure. Second, I aim to compare the different methods available to detect recombination and develop new tools and techniques for detecting viral recombination from short-read next-generation sequence data.

5.2 METHODS

5.2.1 Study Population

The putative recombinant strains were obtained during a survey of HCV in Ebolowa, an urban centre of southern Cameroon (Pépin *et al.,* 2010). Briefly, individuals were included in the survey if they were aged 60 years or more and willing to consent, and excluded in the case of dementia or inability to speak a language known to the interviewers. A venous blood sample was obtained from each individual, and identified only by a study number. See Pépin *et al.,* (2010) for further details.

5.2.2 Initial sample preparation, RT-PCR and sequencing (all work in this section was performed by collaborators in Cameroon)

Samples were first serologically assessed for the presence of anti-HCV antibodies using Monolisa anti-HCV plus version 2 (Biorad). Samples with an optical density/cutoff value ratio of ≥ 6 were then retested with AxSYM version 3 (Abbott Laboratories), with a ratio of ≥20 treated as positive. Viral RNA was extracted from positive samples using the Qiamp Viral RNA kit (QIAGEN). HCV genotyping was performed on a 382-nt fragment of the NS5B gene amplified in a one-step RT-PCR with Superscript III (Life Technologies) with primers Pr3 and Pr5, followed by additional amplification with primers Pr4 and Pr5 and Platinum *Taq* (Life Technologies) - see Table 5.2 for primer details. This result was confirmed by converting RNA to cDNA with AMV-RT (Promega) and random hexamer primers, and then amplifying a 360-nt fragment of the core gene with primers CoreOS and CoreIS and then CoreIS and CoreIAS (Njouom *et al.* 2012). Successful amplicons were sequenced using BigDye Terminator Cycle sequencing (Applied Biosystems) and primers Pr3 and Pr5 for the NS5B region and CoreIS and CoreIAS for the core region.

Two samples (EBW 034 and EBW 436) generated core and NS5B sequences that appeared to be most closely related to different genotypes (specifically, subtypes 4f, 1e and 1l). These sequences, plus the original sample material, were then sent to Oxford for analysis.

5.2.3 Initial phylogenetic analysis

To perform a phylogenetic analysis of the Core and NS5B sequences of samples EBW 034 and EBW 436 I created separate core and NS5B region alignments. These alignments contained the query sequences plus representative sequences of all known subtypes of genotype 1 and genotype 4. From these alignments I estimated Maximum-Likelihood (ML) trees using the method implemented in GARLI v0.951 (Zwickl, 2006). The analysis used a General Time-Reversible (GTR) nucleotide substitution model, estimated base frequencies, and a gamma distribution model of among-site rate variation. Statistical support for phylogenetic clustering was calculated using an ML bootstrap approach with 500 bootstrap replicates; bootstrap scores were summarized using TreeAnnotator (http://beast.bio.ed.ac.uk). Phylogenies were visualized and annotated using FigTree v1.4 (http://tree.bio.ed.ac.uk).

5.2.4 Subtype-Specific Amplification

The initial sequence data generated by collaborators in Cameroon (section 5.2.2) was not sufficient to prove that samples EBW 034 and EBW 436 were recombinant, since they could also represent co-infection with two genotypes followed by preferential amplification of one genotype during PCR. To combat this problem I developed genotype-specific primers that were designed to only amplify one of the genotypes and could be confirmed with genotype 1 and genotype 4 controls. Thus if a sample was amplified in a particular region by the genotype 1 primer but not by the genotype 4 primer then the sample could be classed as genotype 1 in that region. This method

would confirm or rule out co-infection: if either region shows sequence from multiple genotypes then the sample is dually-infected, while if both regions only showed a single sequence then I would have support for recombination. Samples previously genotyped as genotype 1a and genotype 4f in other investigations were used as the genotype 1 and genotype 4 controls respectively.

To design subtype 4f-specific primers, I created an alignment of all published subtype 4f genomes. However this approach could not be used to design genotype 1 primers for subtypes 1e and 1l because there were no full genomes and very few sequences published at the time for subtypes 1e and 1l. Therefore I constructed an alignment containing all genotype 1 genomes, with no more than four genomes per subtype, which was subsequently used to design primers for regions that were conserved within genotype 1 but different in subtype 4f. The designed primers are shown in Table 5.2. In the core region, samples EBW 034 and EBW 436 were first amplified with the generic primers 5'UTR-Ex-400F and Gg-767-Rex-Core and then amplified with 5'UTR-In-405F and either G1-667-Rin-Core or G4-668-Rin-Core.

Due to the very high genetic diversity of the HCV NS3/NS4A region, it was difficult to find sites shared among subtypes within a genotype, but not between genotypes. Primers designed to target these few sites were too unstable to amplify the samples. This problem was exacerbated by the degree of genetic similarity between genotype 1 and genotype 4, which are more closely related than any other pair of HCV genotypes (Salemi and Vandamme, 2002). As the genotype-specific approach failed to work in the 3' end of the genome, I instead used generic primers in the NS5b region to confirm the results obtained in section 5.3.2. Primers used were NS5B-Ex-Fwd and NS5B-Ex-Rev, followed by NS5B-In-Fwd and NS5B-In-Rev.

The laboratory work in Oxford began with extraction from the plasma samples using QIAamp MinElute Virus Spin kits (QIAGEN). The viral RNA was then converted to cDNA using random primers and Superscript III (Invitrogen). The sequences were amplified using Faststart High Fidelity (Roche) and the primers described above. Positive samples were sequenced using BigDye Terminator (Applied Biosystems), and resulting sequences traces were assessed with 4Peaks (Nucleobytes) and aligned with Se-Al 2.0 (available from http://tree.bio.ed.ac.uk).

5.2.5 Cloning

To isolate individual HCV strands and remove the potential for PCR template jumping or preferential amplification, I attempted to clonally isolate strands using One Shot TOP10 chemically competent *E. coli* cells (Invitrogen). Although the procedure was repeated multiple times, in each case the resultant bacterial plasmids did not contain the HCV sequence of interest. While this issue may have been resolved with further attempts I decided to conserve the remaining material and minimize further sample freezing and thawing, and instead concentrated on attempts to obtain whole genome sequences (see below).

5.2.6 Whole-genome Illumina sequencing

All RNA present in the EBW 034 and EBW 436 samples was sequenced using Illumina RNAseq next-generation sequencing (Wang, Gerstein and Snyder, 2009). This method was chosen for two reasons: (i) the remaining sample volume was too limited to use a method such as primer walking, and (ii) unlike other methods, it avoids using primers that might bias which sequences are amplified, thereby giving a more accurate picture of all HCV sequences contained in the sample.

Sample processing was performed by Nick Wong as described in Batty *et al.,* (2013). Briefly, sequencing libraries were constructed from 100 ng of total RNA using the NEBNext mRNA Sample Prep Kit 1 (New England Biolabs) following the manufacturer's guidelines with minor modifications. This procedure breaks the mRNA into fragments, generates cDNA from the mRNA with random primers, repairs the ends of the cDNA library and appends a dA-tail, ligates adaptors to the ends of the cDNA fragments and finally uses PCR to enrich the adaptor-ligated cDNA library, resulting in inserts with a median length of 200 nt. Remaining adapter dimers and mRNA were removed using Agencourt Ampure RNAClean XP beads (Beckman Coulter). Amplicons were quantified and quality assessed using Quant-IT Qubit dsDNA High Sensitivity Assay (Invitrogen) and 1% E-gel (Invitrogen) respectively, and then sequenced on an Illumina HiSeq 2000 creating 100 nt paired end reads following standard Illumina protocols.

5.2.7 Genome assembly and analysis

The sequence reads from the Illumina sequencing run were extracted using Bam2Fastq (github.com/jts/bam2fastq). The resulting reads were then screened for HCV-related sequences using the BlastN algorithm (Altschul *et al.,* 1990). Specifically, I compared each pair of reads to an alignment of 90 HCV genomes that included all genotypes and retained all pairs where at least one sequence matched one or more of the HCV reference sequences with an e-value equal to or less than 0.001. The retained pairs were then assembled into contigs using Velvet (Zerbino and Birney, 2008) with a kmer size of k=65. The contigs were assembled into a draft genome using Se-Al 2.0 (tree.bio.ed.ac.uk/software/seal). Lastly, a final genome sequence was created by mapping all the HCV-related reads back onto the draft genome using Stampy (Lunter and Goodson, 2011).

Primer Name	Sequence	Genotype	Position*
5'UTR-Ex-400F	CCT TGT GGT ACT GCC TGA TAG	Generic	282-299
5'UTR-In-405F	CCT GAT AGG GTG CTT GCG AG	Generic	295-311
Gg-767-Rex-Core	CAY GTR AGG GTA TCG ATG AC	Generic	721-705
G1-667-Rin-Core	GTC ABT GGG GCC CCA ACT AG	G1	671-655
G4-668-Rin-Core	ATC ATT TGG RCC CCA AGA C	G4	671-656
NS5b-Ex-Fwd	TGG GGT TCT CRT ATG AYA CCC GCT GYT TTG	Generic	8248-8274
NS5b-Ex-Rev	AAT ACC TVG TCA TAG CCT CCG TGA	Generic	8637-8617
NS5b-In-Fwd	GAY ACC CGC TGY TTT GAC TC	Generic	8262-8278
NS5b-In-Rev	TAC CTN GTC ATA GCC TCC GTG AAG ACT C	Generic	8635-8611
CoreOS	ACT GCC TGA TAG GGT GCT TGC GAG	Generic	291-311
CoreOAS	ATG TAC CCC ATG AGG TCG GC	Generic	748-732
CoreIS	AGG TCT CGT AGA CCG TGC ATC ATG	Generic	324-344
CoreIAS	CAY GTR AGG GTA TCG ATG AC	Generic	721-705
Pr3	TAT GAY ACC CGC TGY TTT GCT C	Generic	8259-8278
Pr4	GCN GAR TAY CTV GTC ATA GCC TC	Generic	8641-8622
Pr5	GCT AGT CAT AGC CTC CGT	Generic	8633-8619

Table 5.2. Details of primers used in this study. Location numbering is relative to isolate H77 (Genbank accession number AF009606).

The resulting BAM file of aligned reads was analysed with Tablet (Milne *et al.*, 2013) to confirm the quality of the assembly and to add polymorphism information to the draft genome; where a non-consensus nucleotide was present in four or more reads at a site, the consensus nucleotide was replace with an ambiguity code as detailed in Table 5.3. Further, if the consensus nucleotide was present in less than four reads, or if 3 or 4 different nucleotides were present at a site, then the site was denoted with the ambiguity code N.

The completed genome was then subtyped using the Oxford HCV Automated Subtyping Tool (de Oliveira *et al.,* 2005). This tool (i) aligns a submitted sequence against genomes from all HCV genotypes using clustalw (ii) creates a Neighbour-Joining tree with the HKY distance method, and (iii) performs a bootscanning analysis using a sliding window of 400 nts moving in steps of 50 nts scoring each window according to the proportion of replicates that fall within each reference genome (de Oliveira *et al.,* 2005).

Bases at site	Ambiguity Code	Bases at site	Ambiguity Code
A and G	R	G and T	K
C and T	Y	A, C and G	V
A and C	M	A, G and T	D
G and C	S	A, C and T	H
A and T	W	C, G and T	B

Table 5.3. Ambiguity bases used in generating a consensus genome.

5.2.8 Deep Simplot

While the Oxford HCV Automated Subtyping Tool provides an assessment of whether the generated consensus genome is recombinant, it is still possible that genotype 1 and genotype 4 reads are present across the whole genome (as a result of co-infection) but are only present in very low numbers in certain regions. This would give an appearance of recombination in the consensus genome and hide the true state of the sample. To counter this, I devised a new computational method, which I term the Deep Simplot approach, to analyse and genotype each HCV-related read.

To carry out this technique, an alignment of the consensus genome and two reference genomes (one for each putative ancestor of the recombinant) is first constructed. Since

the genome assembly procedure (Section 5.3.6) assigns to each read a position in relation to the consensus genome, this information can be used to determine the region of the *reference* genomes to which each read corresponds. The number of nucleotide differences between the read and each of the two reference genomes is calculated. These values are then converted to a proportion to provide a standardized measure genetic distance to the two reference strains (i.e. the number of differences between the read and reference X is divided by the total number of differences between the read and references X and Y). This computation is repeated for every read in a sample and the resulting proportions are plotted against the genomic position of each read, thereby generating a graph of the relative similarity of reads to each reference that can distinguish between dual infection and recombination.

5.3 RESULTS

5.3.1 Initial subtyping

Figures 5.1 and 5.2 show the maximum-likelihood trees estimated from the core and NS5B alignments (containing reference genomes plus the sequences generated in Cameroon for the two putative recombinants - EBW 034 and EBW 436). In the core phylogeny (Fig 5.1), EBW 034 and EBW 436 both fall within subtype 4f, while in the NS5b phylogeny (Fig 5.2), EBW 034 falls within subtype 1l and EBW 436 is closest to subtype 1e. Although the subtype assignment for the two samples in the core region has weak bootstrap support due to the limited variation in that region, there is strong support for their placement within genotype 4. This contrasts with the strong support for placement within genotype 1 in the NS5B alignment. These results are consistent with, but do not confirm, the hypothesis that the samples are recombinant, and so further testing was performed.

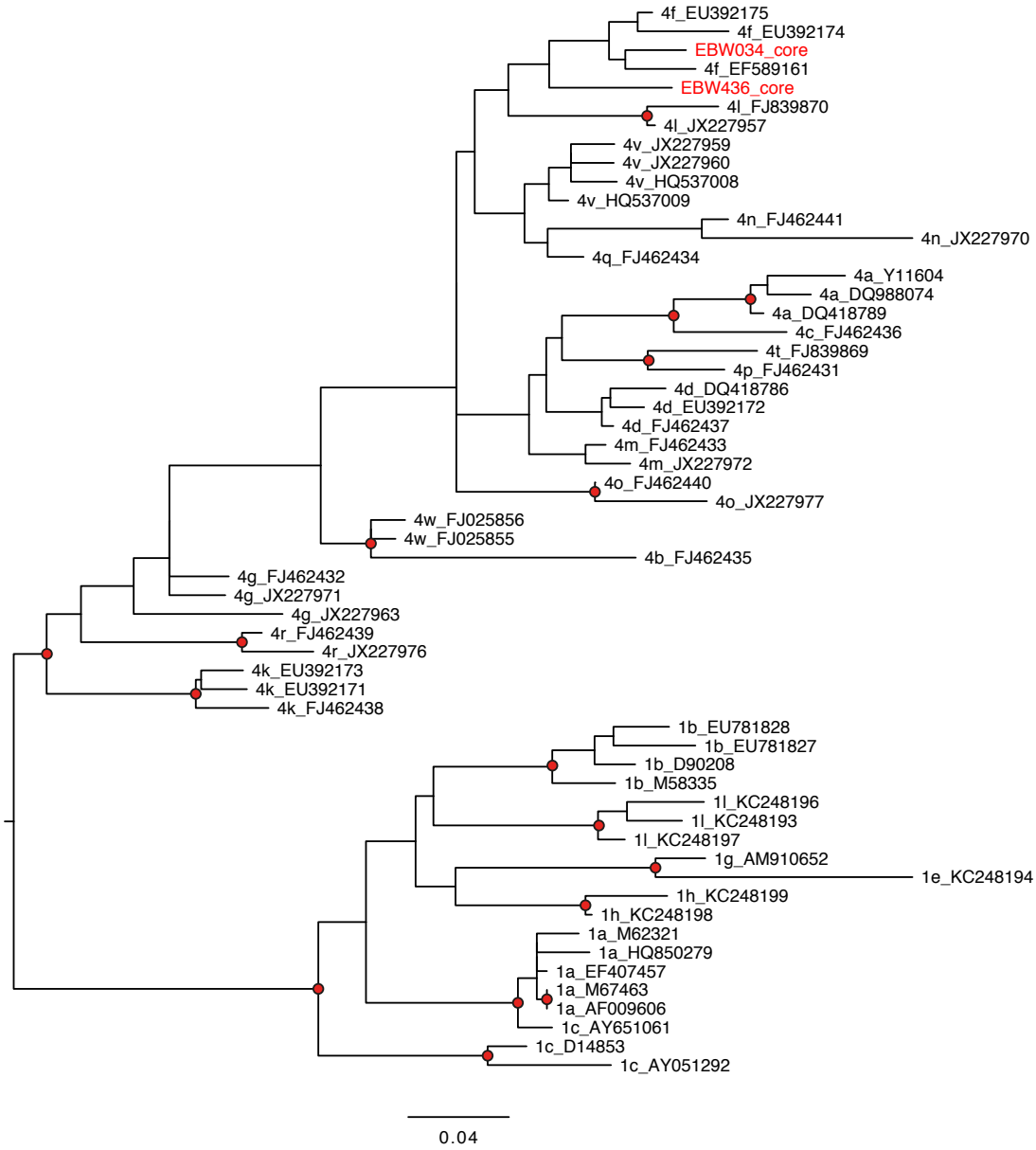


Figure 5.1. *Estimated maximum-likelihood midpoint-rooted phylogeny for the Core alignment. Nodes with a bootstrap support >70% are labeled with a red diamond. The study sequences generated in Cameroon are in red. Branch lengths are in units of expected substitutions per site (see scale bar at bottom of figure). Reference sequences are labelled with subtype and accession number. Subtypes are labeled on the right side of the figure.*

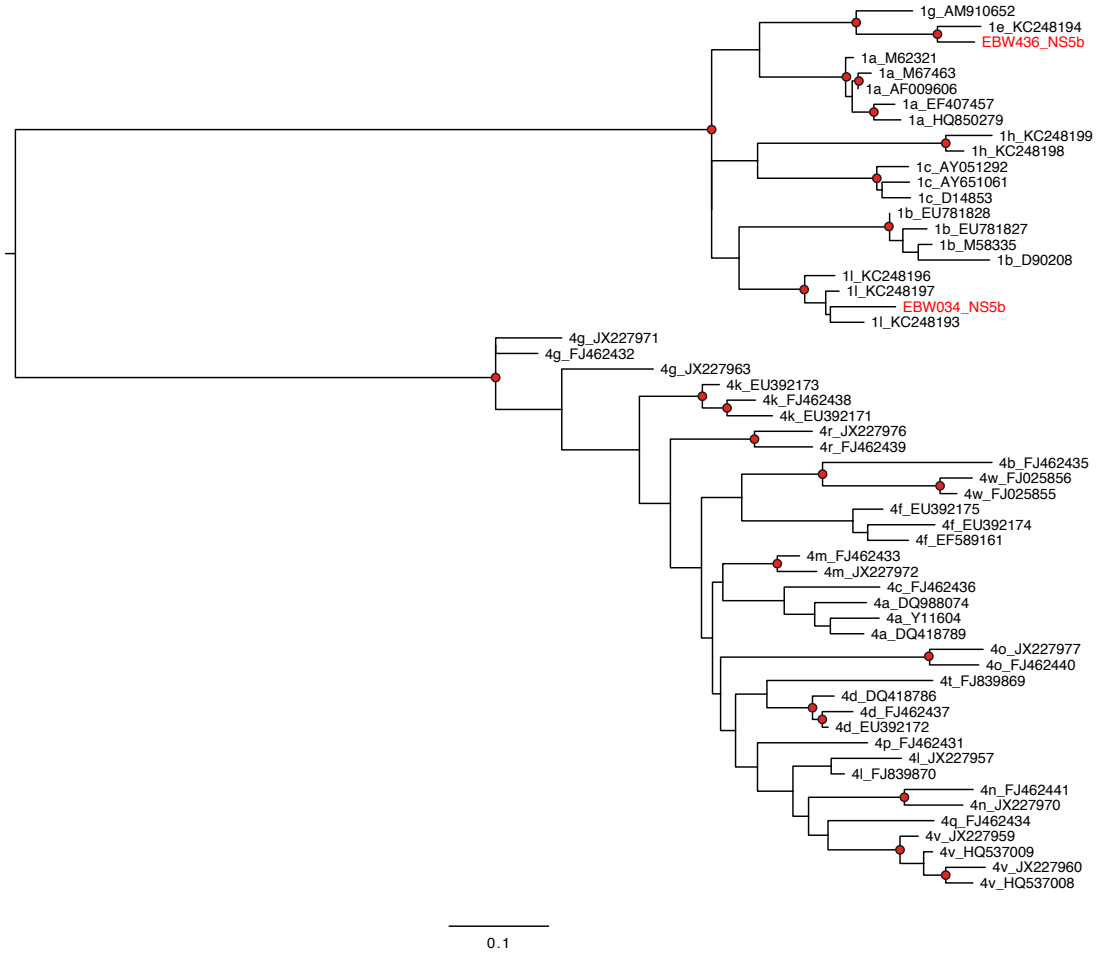


***Figure 5.2**. Estimated maximum-likelihood midpoint-rooted phylogeny for the NS5b alignment. See the legend of Figure 5.1 for further details.*

5.3.2 Genotype-specific amplification

Figure 5.3 shows a photograph of gel electrophoresis of the reaction product from the two core-region PCR amplification reactions with genotype-specific primers. The gel shows that the genotype 1 primers amplify only the genotype 1 control, whereas the genotype 4 primers amplify only the genotype 4 control and the sample EBW 436.

While I was unable to repeat this procedure in the NS5B region, I was able to amplify the NS5B region of the EBW 436 sample with generic primers. This generated a sequence that confirmed the results of the initial subtyping analysis (Figure 5.4). High levels of variation within NS5B may be responsible for the difficulty in creating genotype-specific primers, i.e. primers that amplify all samples within a certain genotype but do not amplify samples within another.

The genotype-specific PCR amplification created much less product for the EBW 034 sample, possibly due to the lower viral load seen in this sample (EBW 034 had a viral load of 1.47×10^5 IU/ml, while EBW 436 had a load of 4.4×10^6 IU/ml). Even though no product from the amplification of sample EBW 034 could be seen in Figure 5.3 a core sequence could still be obtained from this sample, implying issues in the loading of the gel. Figure 5.4 shows a core region phylogeny containing both the sequences obtained during initial subtyping (section 5.3.1) and those produced by genotype-specific amplification. For both samples the two sequences are almost identical and fall within genotype 4.

To summarise, the genotype-specific amplifications showed that in the core region both samples EBW 034 and EBW 436 produced genotype 4 sequence and no genotype 1. Amplification of the NS5B region of the EBW 436 sample produced a genotype 1

sequence. These results, although incomplete, are again consistent with the hypothesis of recombination.

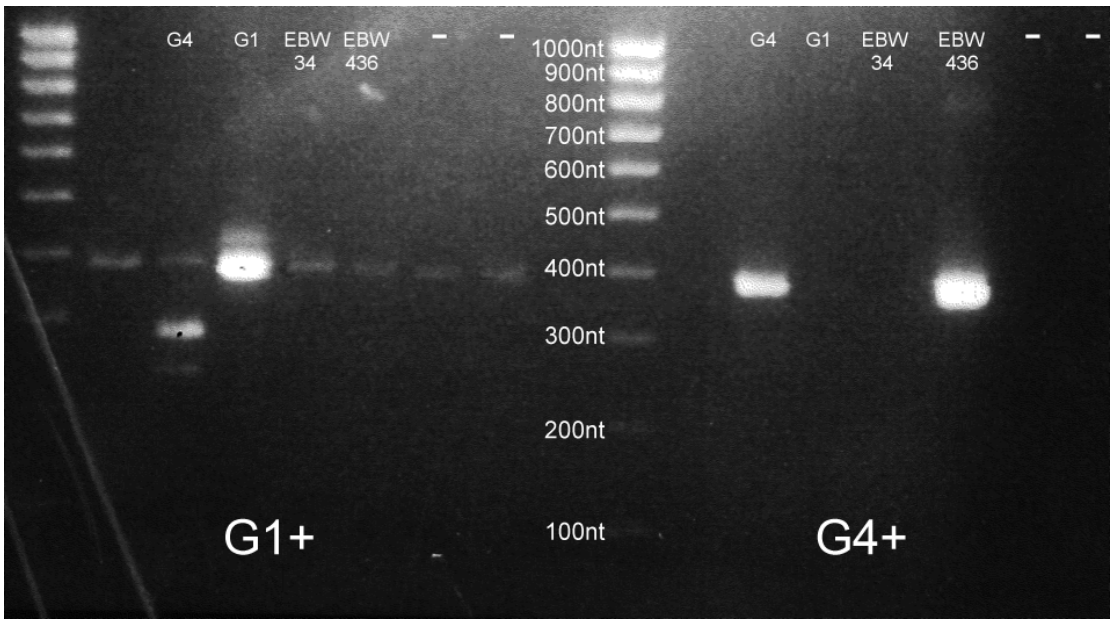


Figure 5.3. *Results of gel electrophoresis of the products of the genotype-specific amplification PCR reaction for the core region. Fragment size measuring is provided by Hyperladder IV; these bands have been labeled with the corresponding fragment size. The left side of the gel shows the result of the Genotype 1-specific primers, while the right side of the gel shows the result of the Genotype 4-specific primers. The sample in each column in as follows: column G1 contains product from a sample previously identified as 1a, column G4 contains product from a sample identified as 4f, columns EBW 34 and EBW 436 contains product from the respective samples, and blank columns (designated with -) contain the product from reactions using water instead of genetic material as a negative control.*

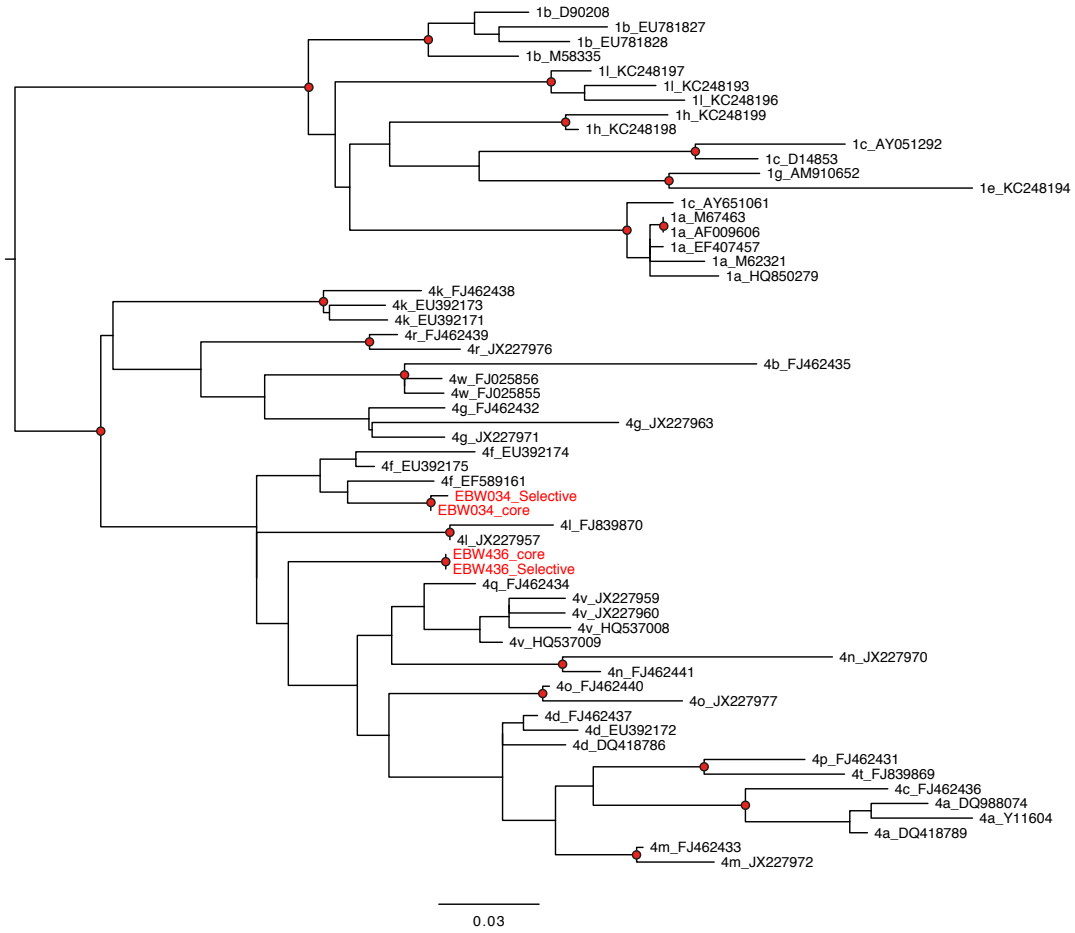


Figure 5.4. *Estimated maximum-likelihood midpoint-rooted phylogeny for the core region. EBW034_core and EBW436_core are sequences generated in Cameroon (section 5.3.1), while EBW034_selective and EBW436_selection are sequences generated using genotype-specific PCR (this section). See the legend of Figure 5.1 for further details.*

5.3.3 Whole-genome sequencing

The Illumina sequencing run for sample EBW 034 generated 5,415,510 reads, each 100 nucleotides long. Only 14 (0.00000258%) of these had significant similarity to our HCV database. An assembly of these reads using Velvet was not possible due to limited or no read overlap. Even so, these reads covered 13% of the EBW 034 genome (1400 nt), so this method still produced more genetic information than I was able to obtain from traditional Sanger sequencing. The generated reads covered sections of the NS2, NS3, NS4B, NS5A and NS5B genes. I used BlastN to determine the HCV genotype to which the reads were most closely related, which in each case was subtype 1l, confirming the results in Figure 5.2. The Illumina sequencing run for sample EBW 436 generated 9,861,538 reads, again each 100 nucleotides long. Of these, 3084 (0.000312%) were HCV-related. Using Velvet I assembled these reads into 12 contigs that provided complete coverage of the EBW 436 genome and ranged in length from 66 to 2958 nucleotides long. Coverage ranged from 1 to 62 reads per locus, with an average coverage of 28.1 reads per locus. When all EBW 436 reads were mapped back to the draft genome created from the contigs, I detected 85 single nucleotide polymorphisms (SNPs), and no read diverged from the consensus sequence at more than one locus.

Figure 5.5 shows the result of the HCV Automated Subtyping Tool on the EBW 436 consensus genome. This shows clear support for a mosaic genome structure - from the 5' UTR to the start of P7 the genome is closely related to genotype 4, and then quickly switches to being closely related to genotype 1 until the 3' end of the genome. Figure 5.5 suggests the break point in the EBW 436 genome is between positions 2450 and 2750 nt (relative to H77) near the 5' end of P7.

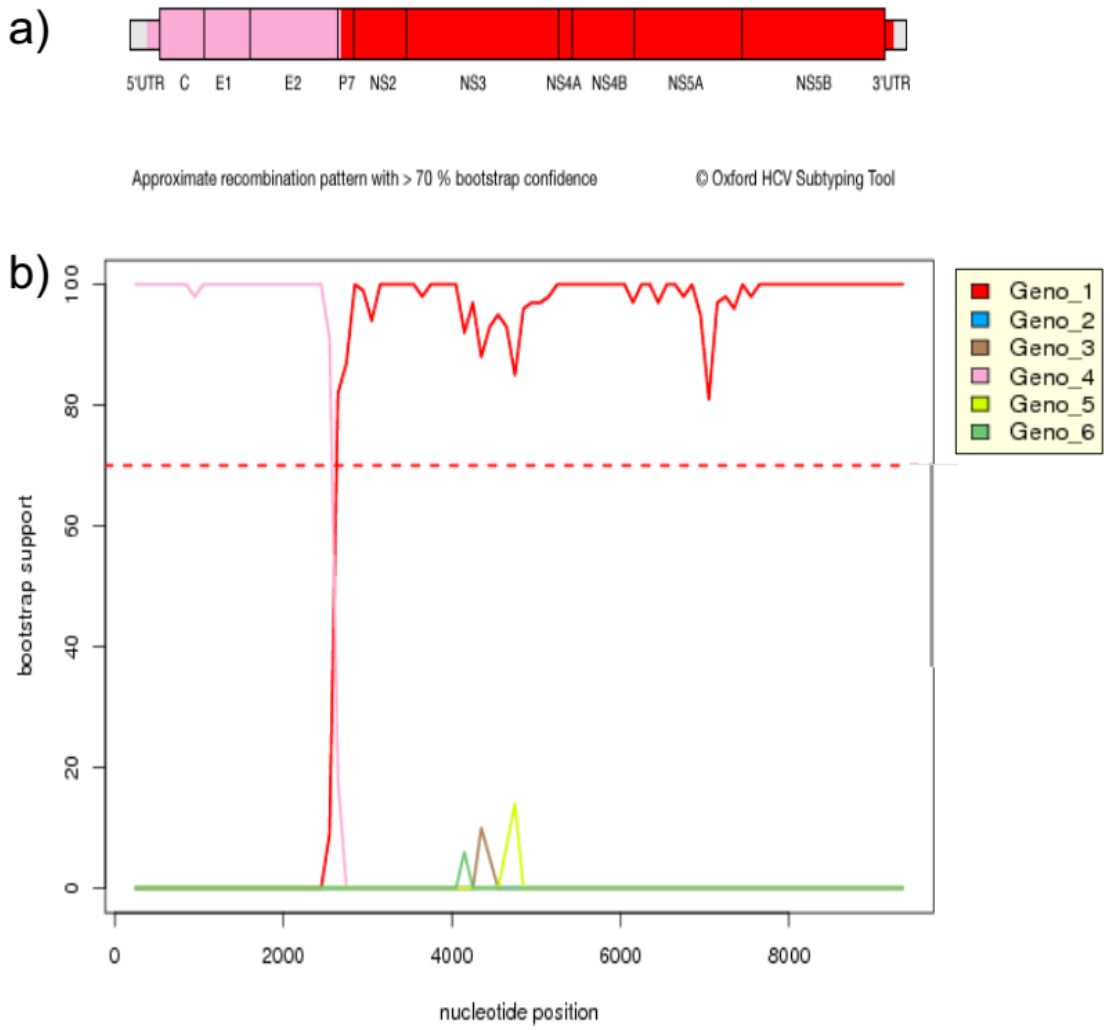


Figure 5.5. *Bootscan plot of the EBW 436 consensus genome generated by the Oxford HCV Automated Subtyping Tool. (a) Schematic of the HCV genome with gene regions labelled. Genome regions with an association to a particular genotype with stronger than 70% bootstrap support are coloured according to the genotype in question, see the key on the right. (b) Bootscan plot, showing bootstrap support for clustering of EBW 436 with the best supported reference genome in each 400nt sliding window. Genome similarity is coloured again according to the key on the right.*

5.3.4 Deep Simplot Analysis

Although highly unlikely, even the chimeric genome shown in Figure 5.5 may have occurred in the absence of recombination, if reads from one genotype were predominant in one region of the genome and in the minority in another region. To rule this out, I developed the Deep Simplot approach (see section 5.2.8) and applied this to the 3084 HCV reads obtained from sample EBW 436. The Deep Simplot analysis was initially attempted with subtypes 4f and 1e as genome reference sequences, but difficulties with the core region were experienced, where the reads showed low similarity to both reference genomes. Therefore more closely related reference strains for the core region were sought. This was done by estimating a maximum-likelihood tree (with 400 bootstrap replicates) from an alignment containing representative genomes of every subtype of genotype 4, as well as the 2400-nt long section of the EBW 436 consensus genome before the break point identified in section 5.3.3. The ML tree (not shown) did not provide strong support for sample EBW 436's placement within any defined subtype, although 4q was the closest subtype. Following this result, the Deep Simplot technique was carried out twice using reference genomes from subtypes 4f and 1e, and 4q and 1e.

Figure 5.6a shows the raw proportion of nucleotide sites that differ between each read and the two reference genomes, whilst Figure 5.6b shows the same values presented as a proportional of the total divergence to both reference genomes. These graphs provide strong support for recombination and clearly show the recombination break point at genome position 2530 nt (position 2645 in H77). Only 68 out of 633 reads (10.7%) between the start of the genome and position 2530 are closer to subtype 1e than subtype 4q, while in the 4f/1e analysis 83 reads (13%) are closer to 1e than 4f. Further, no reads after the breakpoint in any plot are closer to genotype 4 than genotype 1. This

difference between the regions is likely due to the high degree of sequence conservation in the core region, such that a very small number of mutations can make a G4 read appear more like G1. This analysis, unlike all previous analyses, definitively rules out the hypothesis of dual infection. If the sample was dually-infected then no breakpoint would be seen and every region of the genome would contain at least some reads that are 1e-like and some that were 4o-like.

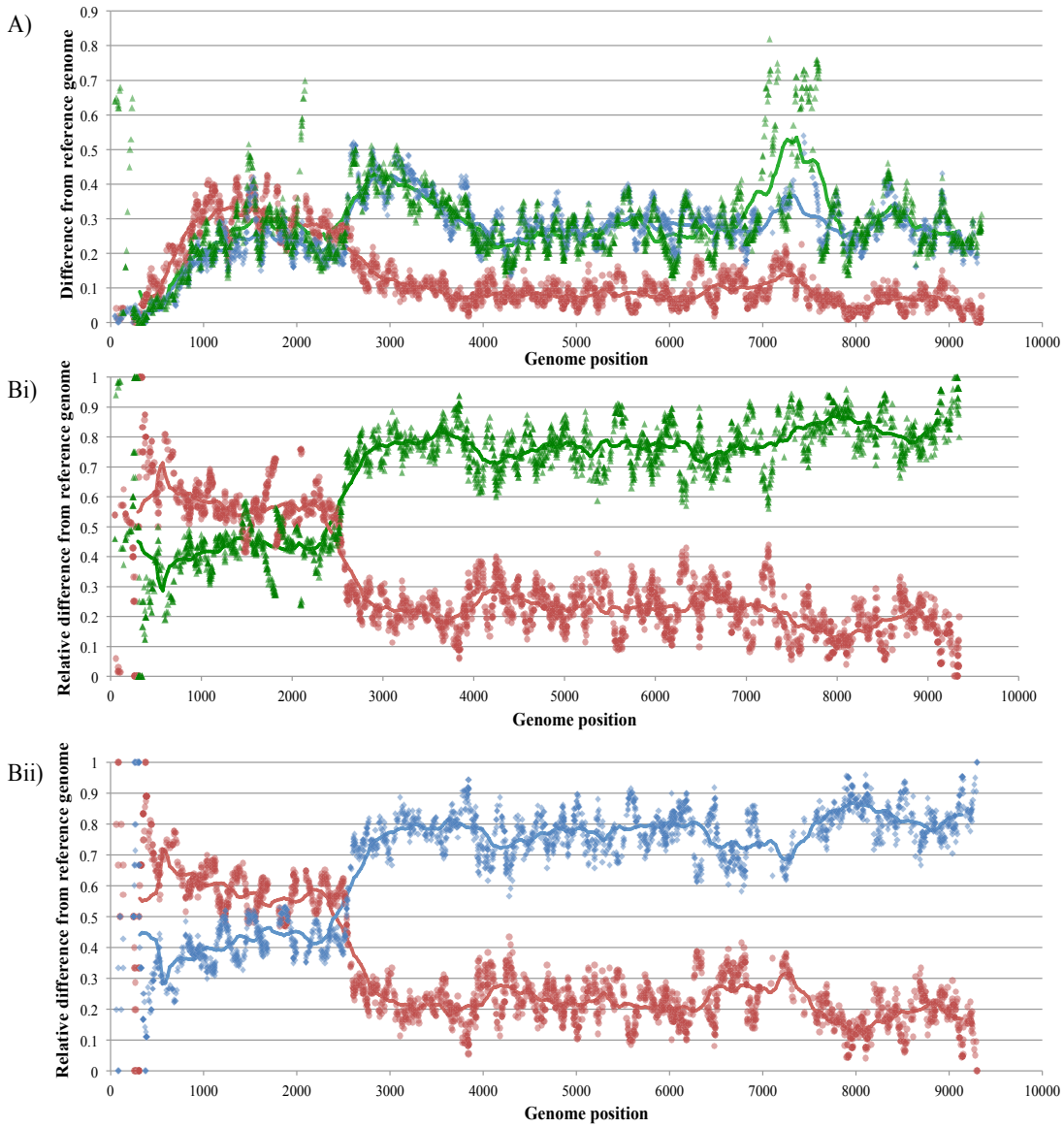


Figure 5.6. *Deep Simplot of the EBW 436 mapped reads compared to subtype 1e (red circle), 4f (green triangle) and 4q (blue diamond) reference genomes. In all diagrams, lower values represent greater sequence similarity to the reference genome. (a) Raw no. of differences from the reference genome as a proportion of the read. Lines show the average result for each 500nt window of the genome. (bi) Relative difference to one reference genome over another (calculated as no. differences to reference genome/(no. differences to 1e reference + no. differences to 4f reference)). (bii) As bi, but using 1e and 4o as reference genomes.*

5.4 DISCUSSION

The first goal of this study was to confirm recombination in the two samples EBW 436 and EBW 036. No analysis in this chapter was inconsistent with the hypothesis of recombination for either sample. The mosaic genome structure of EBW 436 was demonstrated unambiguously by the Illumina RNA seq results. Even though the RNAseq results for sample EBW 036 were incomplete, the results from initial subtyping and genotype-specific amplification of EBW 036, as well as the small amount of HCV sequence obtained from the RNAseq run, were consistent with it being a mosaic strain. If further material from the patient infected with EBW 034 became available then further attempts to sequence EBW 034 using RNAseq would be useful to provide more information the breakpoint location and the precise nature of the ancestral lineages.

The parent lineages are also worth investigation; 1e is endemic in Cameroon, and has only been isolated in two other countries - Vietnam and Canada (Murphy *et al.,* 2007). The 5' end of the recombinant, on the other hand, seems to originate from an unclassified subtype of genotype 4; while subtypes 4f and 4o are the closest, even that association does not have strong bootstrap support. BLAST searches found that the closest match to this section of the recombinant is isolate QC215 (accession number JF735132), an unsubtyped genotype 4 sample collected in Quebec (Murphy *et al.,* 2007). This implies that there is transfer of low-prevalence endemic HCV subtypes between Cameroon and Quebec, and the fact these HCV variants were found in a recombinant may give some clues as to how this transfer is occurring, as the co-infection required for recombination to take place is much more likely if the patient in question is an IDU (Schröter *et al.,* 2003). The presence of subtype 1e and this unsubtyped genotype 4 variant in a recombinant imply that they may be circulating in

Cameroon among cohorts of IDUs as is the case for subtype 4d (van Asten *et al.,* 2004), although further screening will be needed to state this for certain.

Another goal of this chapter was to investigate the efficacy of different molecular methods of detecting recombination versus the null hypothesis of dual infection. In this study four different approaches were employed: (i) Sanger sequencing of two genome regions with standard primers (ii) Sanger sequencing of the core region with subtype-specific primers (iii) genome assembly of RNAseq-generated reads and (iv) Deep Simplot mapping of the RNAseq-generated reads. Of these methods, the genome assembly of RNAseq reads provided the most information about the samples' genomes and of the recombination breakpoint in the genome of sample EBW 436. The Deep Simplot method provides the most definite confirmation of recombination as it is capable of excluding the possibility of dual infection. Based on these results, I would recommend that RNAseq or a similar next-generation sequencing method be used to confirm HCV recombination in the future, especially as the time required and cost of this sequencing technology decreases.

Many of the tests performed in this analysis go beyond that required in the literature to confirm recombination. For example, in Calado *et al.,* 2011, the putative recombinant was only Sanger sequenced in the E1l and NS5B regions. It is debatable whether this constitutes sufficient evidence to conclude that a sample contains a recombinant virus. In many applications, amplification and sequencing of the core and NS5B regions, followed by confirmatory sequencing of clonally-isolated strands should be enough to support recombination. The changing role of HCV genotypes in determining the course of treatment is worth discussing; while genotypes 1 and 4 are more resistant to the old interferon-alpha and ribavirin treatment (Chen *et al.,* 2012), no similar pattern has been seen with the addition of direct-acting antiviral drugs to the treatment regimen (Koff,

2014). Instead, the movement towards drugs that specifically target HCV proteins (e.g. boceprevir and telaprevir) may provide a greater advantage to strains permissive to recombination, as they will have a greater ability to accumulate resistance mutations. If recombinants with greater resistance to drug treatment become widespread, accurate detection of recombination without risk of false positives will be essential. If HCV recombinant strains become more epidemiologically relevant in the future, I would argue that clinics that subtype HCV by serology, or by the sequencing of only one genome region (as discussed in Morel *et al.,* 2011), move to multiple-region or whole genome next-generation sequencing instead. On the other hand, the decreasing cost and increasing read length of next-generation sequencing platforms may lead to next-generation sequencing being used routinely in clinics for genotype identification without any need for an incentive from clinically-significant recombinants.

This results in this chapter confirm the pattern in breakpoint location shown by previous recombinant: for inter-genotypic recombinants the recombination break point tends to occur at the genomic position that separates the HCV structural and nonstructural genes. The EBW 436 breakpoint is earlier in the genome than most, occurring in P7, and this may be due to the fact that all other inter-genotypic recombinants were classified as genotype 2 at the 5' end (see Table 5.1). It is possible that viable breakpoints for inter-genotypic recombination vary depending on the genotypes involved; HCV has extensive RNA secondary structure that is sensitive to mutation and can have a high impact on virus viability (Simmonds, Tuplin and Evans, 2004). Further analysis of HCV recombination breakpoint positions may help understand what factors determine the placement of the break point, as has been investigated in HIV by Ramirez *et al.,* 2008. As in all other HCV recombinants so far detected, the EBW436 recombinant is homologous, rather than non-homologous,

meaning that the genome structure of the virus is preserved. The lack of non-homologous recombinants may be due to the recombination mechanism only (or mainly) working on regions with high sequence identity or due to the low fitness of non-homologous recombinants such that they are not detected in clinical screening. Since non-homologous recombinants can be created *in vitro* via an RdRp-independent breakage and rejoining mechanism (Scheel *et al.,* 2013), the latter explanations seems more likely, although replicative recombination via the copy-choice mechanism may involve different factors.

The Deep Simplot graphs in Figure 5.6 highlight the genetic diversity of HCV present in a single clinical sample. At any given position in the HCV genome overlapping reads might differ in diversity to the reference genome by more than 15%, a similar level of diversity to that seen between isolates obtained from different individuals infected with the same subtype. The short length of reads generated by the RNAseq method makes it difficult to detect genetic linkage between variants and generate a complete picture of the viral quasispecies present in the host, but further work may be able to take advantage of the fact that paired-end reads can have a gap of up to 400 nt between them, linking more distant regions of the genome. Additionally, RNA sequencing technology continuing to improve and read length will likely increase in future; RNAseq platforms are already available that generate 250 nt paired-end reads (Jünemann *et al.,* 2013), greatly increasing our ability to infer linkage between mutations.

I would recommend that further analysis be performed on the Deep Simplot technique to fully understand its requirements, power, and limitation. Important future questions are (i) how should the best reference genomes be selected for the analysis and (ii) whether the analysis can be performed with more than two reference genomes.

Additionally the tool should be tested against artificial dual-infections (generated by deliberately mixing two samples in different ratios) and against simulated data in order to evaluate its performance.

One of the difficulties in characterising highly variable RNA virus sequences is the possibility that screening techniques, particularly PCR, are biased towards some virus variants over others. This problem was mitigated here through the use of Illumina RNAseq sequencing. No RNA is removed or filtered out during the sample preparation procedure, and the majority of sample EBW 436 was sequenced dozens of times over. Indeed, the sequencing depth created a new problem, in that the amount of sequence generated was so great that it was computationally time-consuming to identify the small fraction of reads belonging to the virus of interest. The BLAST-based filtering algorithm used to identify HCV reads from the sequencing output required a long processing time (130 hours to process 9.8 million reads for EBW 436), likely because the algorithm compared each pair of reads individually against a small alignment of reference genomes, while the BLAST algorithm works better comparing a small number of sequences to a much larger database. Rewriting this pipeline would therefore be very beneficial for future analysis. Even though the filtering of non-HCV reads was very lenient, only 14 of the 5.4 million reads generated from sample EBW 034 were categorised as HCV, emphasizing that even next-generation sequencing methods can extract only a limited amount of sequence data from samples with a low viral load or which have been repeatedly frozen and thawed. Next-generation sequencing is becoming more cost-efficient per kilobase of sequence than traditional methods of genotyping (Metzker, 2010), and as computational analysis methods improve it will increasingly become the best option for genotyping novel HCV variants or even for routine clinical genotyping. With its lack of sequence specificity and

capability to amplify all RNA present in a sample I believe that next generation sequencing will greatly increase our knowledge of HCV diversity, and will provide a much clearer picture of the rate of HCV recombination and the role it plays in the virus' evolution.

5.5 REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215: 403-10.

Batty, E. M., Wong, T. H. N., Trebes, A., Argoud, K., Attar, M., Buck, D., Ip, C. L. C., Golubchik, T., Cule, M., Bowden, R., Manganis, C., Klenerman, P., Barnes, E., Walker, A. S., Wyllie, D. H., Wilson, D. J., Dingle, K. E., Peto, T. E. A., Crook, D. W. and Piazza, P. (2013). A modified RNA-Seq approach for whole genome sequencing of RNA viruses from faecal and blood samples. *PLoS One,* 8:e66129.

Bhattacharya, D., Accola, M. A., Ansari, I. H., Striker, R. and Rehrauer, W. M. (2011). Naturally occurring genotype 2b/1a hepatitis C virus in the United States. *Virology Journal* 8: 458.

Calado, R. A., Rocha, M. R., Parreira, R., Piedade, J., Venenno, T. and Esteves, A. (2011). Hepatitis C virus subtypes circulating among intravenous drug users in Lisbon, Portugal. *Journal of Medical Virology* 83: 608-615.

Chen, Y., Xu, H. X., Wang, L. J., Liu, X. X., Mahato, R. I. and Zhao, Y. R. (2012). Meta-analysis: IL28B polymorphisms predict sustained viral response in HCV patients treated with pegylated interferon-a and ribavirin. *Alimentary Pharmacology and Therapeutics,* 36: 91–103.

Colina, R., Casane, D., Vasquez, S., García-Aguirre, L., Chunga, A., Romero, H., Khan, B. and Cristina, J. (2004). Evidence of intratypic recombination in natural populations of hepatitis C virus. *Journal of General Virology* 85: 31–37.

Conteduca, V., Sansonno, D., Russi, S., Pavone, F. and Dammacco, F. (2013). Therapy of chronic hepatitis C virus infection in the era of direct-acting and host-targeting antiviral agents. *Journal of Infection* 68: 1-20.

Cristina, J. and Colina, R. (2006). Evidence of structural genomic region recombination in Hepatitis C virus. *Virology Journal* 3: 1–8.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783-791.

Gallei, A., Pankraz, A., Thiel, H. J. and Becher, P. (2004). RNA recombination in vivo in the absence of viral replication. *Journal of Virology* 78, 6271–6281.

Galli, A. and Bukh, J. (2014). Comparative analysis of the molecular mechanisms of recombination in hepatitis C virus. *Trends in Microbiology* 22: 354-364.

Hoshino, H., Hino, K. and Miyakawa, H. (2012) Inter-geneotypic recombinant hepatitis C virus strains in Japan noted by discrepancies between immunoassay and sequencing. *Journal of Medical Virology* 84: 1018-1024.

Jünemann, S., Sedlazeck, F. J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., Mellmann, A., Goesmann, A., von Haeseler, A., Stoye, J. and Harmsen, D. (2013). Updating benchtop sequencing performance comparison. *Nature Biotechnology* 31: 294-296.

Kageyama, S., Agdamag, D. M., Alesna, E. T., Leaño, P. S., Heredia, A. M. L., Abellanosa Tac An, I. P., Jereza, L. D., Tanimoto, T., Yamamura, J. and Ichimura, H. (2006). A natural inter-genotypic (2b/1b) recombinant of hepatitis C virus in the Philippines. *Journal of Medical Virology* 78: 1423–1428.

Kalinina, O., Norder, H., Mukomolov, S. and Magnius, L. O. (2002). A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg. *Journal of Virology* 76, 4034–4043.

Koff, R. S. (2014). The efficacy and safety of sofosbuvir, a novel, oral nucleotide NS5B polymerase inhibitor, in the treatment of chronic hepatitis C virus infection. *Alimentary Pharmacology and Therapeutics* 39: 478-487.

Lai, M. M. (1992). RNA recombination in animal and plant viruses. *Microbiological Reviews* 56: 61-79.

Lau, K. A. and Wong, J. J. (2013). Current trends of HIV recombination worldwide. *Infectious Disease Reports* 5: e1.

Lauer, G. M. and Walker, B. D. (2001). Hepatitis C virus infection. *New England Journal of Medicine* 345: 41–52.

Lee, Y.-M., Lin, H.-J., Chen, Y.-J., Lee, C.-M., Wang, S.-F., Chang, K.-Y., Chen, T.-L., Liu, H.-F. and Chen, Y.-M. A. (2010). Molecular epidemiology of HCV genotypes among injection drug users in Taiwan: Full-length sequences of two new subtype 6w strains and a recombinant form 2b6w. *Journal of Medical Virology* 82: 57–68.

Legrand-Abravanel, F., Claudinon, J., Nicot, F., Dubois, M., Chapuy-Regaud, S., Sandres-Saune, K., Pasquier, C. and Izopet, J. (2007). New Natural Intergenotypic (2/5) Recombinant of Hepatitis C Virus. *Journal of Virology* 81: 4357.

Lunter, G and Goodson, M. (2010). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequencing reads. *Genome Research* 21: 936-39.

Mancuso, N., Tork, B., Skums, P., Ganova-Raeva, L., Mandoiu, I and Zelikovsky, A. (2011). Reconstructing viral quasispecies from NGS amplicon reads. *In Silico Biology* 11: 237-249.

Metzker, M. L. (2010). Sequencing technologies – the next generation. *Nature Reviews Genetics* 11: 31-46.

Mohd Hanafiah, K., Groeger, J., Flaxman, A. D. and Wiersma, S. T. (2013). Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. *Hepatology* 57: 1333–1342.

Morel, V., Descamps, V., François, C., Fournier, C., Brochot, E., Capron, D., Duverlie, G. and Castelain, S. (2010). Emergence of a genomic variant of the recombinant 2k/1b strain during a mixed Hepatitis C infection: a case report. *Journal of Clinical Virology* 47: 382-386.

Morel, V., Fournier, C., François, C., Brochot, E., Helle, F., Duverlie, G. and Castelain, S. (2011). Genetic recombination of the hepatitis C virus: clinical implications.

Moreno P, Alvarez M, Lopez L et al. (2009). Evidence of recombination in Hepatitis C Virus populations infecting a hemophiliac patient. *Virology Journal* 6: 203.

Murphy, D. G., Willems, B., Deschênes, M., Hilzenrat, N., Mousseau, R. and Sabbah, S. (2007). Use of sequence analysis of the NS5B region for routine genotyping of hepatitis C virus with reference to C/E1 and 5' untranslated region sequences. *Journal of Clinical Microbiology* 45:1102-1112.

Njouom, R., Pasquier, C., Ayouba, A., Gessain, A., Froment, A., Mfoupouendoun, J., Pouillot, R., Dubois, M., Sandres-Saune, Thonnon, J., Izopet, J. and Nerrienet, E.

(2003). High rate of hepatitis C virus infection and predominance of genotype 4 among elderly inhabitants of a remote village of the rain forest of South Cameroon. *Journal of Medical Virology* 71: 219-225.

Njouom, R., Caron, M., Besson, G., Ndong-Atome, G.-R., Makuwa, M., Pouillot, R., Nkoghé, D., Leroy, E. and Kazanji, M. (2012). Phylogeography, risk factors and genetic history of hepatitis C virus in Gabon, central Africa. *PloS One* 7: e42002

Noppornpanth, S., Lien, T. X., Poovorawan, Y., Smits, S. L., Osterhaus, A. D. M. E. and Haagmans, B. L. (2006). Identification of a naturally occurring recombinant genotype 2/6 hepatitis C virus. *Journal of Virology* 80: 7569–7577.

Pääbo, S., Irwin, D. M. and Wilson, A. C. (1990). DNA damage promotes jumping between templates during enzymatic amplification. *The Journal of Biological Chemistry* 265: 4718-21.

Pépin, J., Lavoie, M., Pybus, O. G., Pouillot, R., Foupouapouognigni, Y., Rousset, D., Labbé, A.-C. and Njouom, R. (2010). Risk Factors for Hepatitis C Virus Transmission in Colonial Cameroon. *Clinical infectious Diseases* 51, 768–776.

Raghwani, J., Thomas, X. V., Koekkoek, S. M., Schinkel, J., Molenkamp, R., van de Laar, T. J., Takebe, Y., Tanaka, Y., Mizokami, M., Rambaut, A. and Pybus, O. G. (2012). Origin and Evolution of the Unique Hepatitis C Virus Circulating Recombinant Form 2k/1b. *Journal of Virology* 86: 2212–2220.

Ramirez, B. C., Simon-Loriere, E., Galetto, R. and Negroni, M. (2008). *Virus Research* 134: 64-73.

Ross R, Verbeeck J, Viazov S, Lemey P, Van Ranst M, Roggendorf M. (2008). Evidence for a complex mosaic genome pattern in a full-length hepatitis C virus sequence. *Evol Bioinform Online* 4: 249–254.

Salemi, M. and Vandamme A-M. (2002). Hepatitis C Virus Evolutionary Patterns Studied Through Analysis of Full-Genome Sequences. *Journal of Molecular Evolution* 54: 62-70.

Salminen, M. O., Carr, J. K., Burke, D. S. and McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Research and Human Retroviruses* 11: 1423–1425.

Scheel, T. K. H., Galli, A., Li, Y-P., Mikkelsen, L. S., Gottwein, J. M. and Bukh, J. (2013). Productive homologous and non-homologous recombination of hepatitis C virus in cell culture. *PLoS Pathogens* 9: e1003228

Schröter, M., Feucht, H-H., Zöllner, B., Schäfer, P., Laufs, R. (2003). Multiple infections with different HCV genotypes: prevalence and clinical impact. *Journal of Clinical Virology* 27: 200-204.

Simmonds, P. (2004). Genetic diversity and evolution of hepatitis C virus -15 years on. *The Journal of General Virology* 85: 3173–3188.

Simmonds, P., Tuplin, A. and Evans, D. J. (2004). Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: implications for virus evolution and host persistence. *RNA* 10: 1337-1351.

Van Asten, L., Verhaest, I., Lamzira, S., Hernandez-Aguado, I., Zangerle, R., Boufassa, F., Rezza, G., Broers, B., Robertson, J. R., Brettle, R. P., McMenamin, J., Prins, M., Cochrane, A., Simmonds, P. and Coutinho, R. A. (2004). Spread of

Hepatitis C Virus among European injection drug users infected with HIV: a phylogenetic analysis. *The Journal of Infectious Diseases* 189: 292-302.

Walsh, P. S., Erlich, H. A. and Higuchi, R. (1992). Preferential PCR amplification of alleles: mechanisms and solutions. *Genome Research* 1: 241-250.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57-63.

Worobey, M. and Holmes, E. C. (1999). Evolutionary aspects of recombination in RNA viruses. *Journal of General Virology* 80: 2535-2543.

Yokoyama, K., Takahashi, M., Nishizawa, T., Nagashima, S., Jirintai, S., Yotsumoto, S., Okamoto, H. and Momoi, M. Y. (2011). Identification and characterization of a natural inter-genotypic (2b/1b) recombinant hepatitis C virus in Japan. *Archives of Virology* 156: 1591-1601.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18: 821-29.

6 CONCLUSIONS

6.1 SIGNIFICANCE OF THE DRC

As the hepatitis C virus (HCV) surged to global prominence over the past century, a small number of strains became responsible for the majority of HCV infections worldwide. This means that while the past 25 years have seen a great deal of research performed on HCV, much of the overall diversity of the virus is yet to be understood. The work performed in this thesis aims to redress this by gathering HCV gene sequences from a previously poorly surveyed region of the world (Mohd Hanafiah *et al.,* 2013) and using phylogenetic and phylodynamic techniques, to estimate the history of HCV in this area and across Africa.

This thesis has shown the importance of the Democratic Republic of the Congo (DRC) to the study of HCV genotype 4, on three separate levels. First, the present state of HCV in the DRC: the screening programs described in chapters 2, 3 and 4 show that HCV is prevalent in the DRC. The male study population screened in chapters 2 and 3 exhibited an overall seroprevalence of 13.7% and 3.1% of samples contained detectable HCV RNA. Further, I found evidence of a cohort effect: those born after 1945 showed a seroprevalence of 6.9% and an RNA prevalence of 2.9%, while those born before then had a seroprevalence of 45.5% and an RNA prevalence of 11.1%. These findings are complemented by those in chapter 4, which investigated an elderly civilian cohort in Kinshasa, all born before 1938, which exhibited a seroprevalence of 26.2% and RNA prevalence of 14.1%. These values are comparable those previously reported for Cameroon; samples taken in the Centre Pasteur du Cameroun in the nation's capital showed an overall seroprevalence of 15.7%, with a seroprevalence of 11.6% in those born after 1953 and 32.8% in those born before then (Nerrienet *et al.,* 2005). Furthermore, in chapter 4 I report a high diversity of HCV in Kinshasa, with samples from genotypes 1, 4 and 7 present. Eleven different subtypes of genotype 4

seen, and across chapters 3 and 4 I provide evidence for a new subtype circulating in the DRC (denoted subtype 4drc in this thesis).

Chapter 4's investigation of the epidemic history of HCV in Kinshasa shows that HCV's recent transmission history is another level on which the DRC is important. The clear epidemiological transition shown in the skyline plot analyses, from slow growth to exponential growth, is consistent with a man-made cause for HCV prevalence there, and fits into a pattern recently observed in many countries across Africa – Egypt, the Central African Republic, Cameroon and Gabon all have HCV epidemics that have been hypothesised to have been caused in whole or in part by mass treatment campaigns (Pybus *et al.,* 2003; Njouom *et al.,* 2009; Pépin *et al.,* 2010; Njouom *et al.,* 2012). The amount of sequence information gathered from Kinshasa in chapters 2, 3 and 4 allowed for a more detailed analysis than was performed in many of those papers, enabling me to independently estimate the parameters of epidemic growth for independent subtypes. Parallel changes in past growth rates seen two or more separate HCV lineages in one location are strong evidence for a significant past change in the virus' epidemiological circumstances. Possible iatrogenic causes of past HCV transmission in the DRC are consistent with the results of Faria *et al.,* 2014, who investigated HIV-1 sequence data from the DRC and Republic of Congo and found that HIV-1 group M underwent a period of increased epidemic growth in the middle of the 20th century, one possible cause of which was injectable public health campaigns undertaken during the Belgian colonial era. Further epidemiological work needs be undertaken to investigate whether past injections are associated with the past transmission of blood borne viruses in the DRC. My collaborators are collecting demographic data and past medical history from the population studied in chapter 4 with the intention of calculating odds ratios to indicate factors that are particular risks

for HCV seropositivity. Another avenue for future research would be to revisit the Egyptian epidemic; far more sequences have been generated in Egypt since the 2003 Pybus *et al.* study, and new and better estimates of the rate of evolution of HCV are now available.

HCV's endemic transmission during the three centuries before 1900 is the third level on which the DRC is significant. The continental-scale analysis performed in chapter 3 indicates a central African origin for HCV genotype 4, resolving the uncertainty arising from high levels of genetic diversity in both central Africa and the Middle East. Phylogenetic and molecular clock analyses make it clear that the genotype originated in central Africa and was later introduced into the Middle East on at least three occasions (see figure 3.5). In addition, in chapter 4 I found that HCV in the DRC was cosmopolitan – every subtype seen within the DRC was also reported in other countries, and sequences from the DRC included members of each of three major sub-genomic lineages. The major genotype 4 lineages diverged more than 200 years ago, meaning that HCV has been present in the DRC for a long time.

In summary, this thesis provides new significant results concerning the evolutionary and epidemic history of HCV in Africa, which also illustrate how the shift from endemic to epidemic transmission may have occurred, thereby explaining current patterns of HCV prevalence and diversity.

6.2 FACTORS SURROUNDING EMERGENCE

While HCV genotype 4 is now found across the world and is most prevalence in Africa (Messina *et al.,* 2014), chapters 3 and 4 indicate that only a hundred years ago it was circulating at much lower frequencies. Over the past hundred years the effective population size of HCV has grown by multiple orders of magnitude, but as figure 4.6

demonstrates, this cannot be solely explained by growth of the African population. Instead, the growth in HCV seems to be the result of a change in how HCV transmission occurred, as other routes supplanted the still poorly characterized endemic routes of transmission.

HCV can be used to illustrate factors in viral emergence. Firstly, a disease can be present in human populations for a long time and remain comparatively uncommon so long as the transmission routes available to it are inefficient; as shown in figures 3.5 and 3.6, HCV genotype 4 appears to have been present in human populations for centuries while remaining at low prevalence. Pybus *et al.,* 2001, used coalescent methods to estimate that the basic reproductive number (R_0) of endemic genotype 4 was 1.68. Secondly, even if R_0 is low external factors can amplify an endemic virus to high prevalences. This can be seen by the increased viral growth rate during the exponential growth phase between 1950 and 1980 in figures 4.6 and 4.8. Following Pybus et al. (2003), we can use the general relationship R_0 = 1 + (exponential growth rate)*(mean duration of infectiousness), to estimate R_0 during the growth phase. Using the results in chapter 4, this gives estimates of R_0 in the DRC during the mid-twentieth century of 4.96 to 7.22, comparable to the R_0 values of 3-7 previously estimated for the Egyptian HCV epidemic using genetic data (Pybus *et al.,* 2003). HCV growth rates in the DRC subsequently fell, likely because sterile techniques improved and the number of injections given declined. However, because of the long duration of HCV chronic infection, HCV prevalence remains higher than before the 1950s. It remains to be seen how long HCV prevalences will remain at this elevated level – the more closely studied epidemic in Egypt may provide the best indication of the future in sub-Saharan Africa.

Lastly, as discussed in Pépin and Labbé, 2008, ignorance can lead to unforeseen consequences from even the noblest goals; the 50-year lag between the development of antiparisitic drugs and the understanding of the transmission of blood-borne viruses resulted in the high prevalence of HCV in central Africa and potentially the establishement of the group M HIV-1 pandemic (Pépin, 2011; Faria *et al.,* 2014). When planning a public health intervention, then, it is important to consider the social and ecological dynamics of the population, but unintended consequences are always possible. The most effective way of preventing emergence may instead be to monitor the population afterwards for perturbations indicative of new diseases.

6.3 UNCOVERING THE DIVERSITY OF HCV

One of the goals of this thesis was to characterise the diversity of HCV in central Africa. Here I have reported HCV sequences from three different genotypes and demonstrate the presence of uncharacterized subtypes circulating in the Central African Republic and the DRC. Significantly, I also note the second sample ever found from HCV genotype 7 (chapter 4). Sequencing a complete genome from this sample will be an important next step in the taxonomic confirmation of the genotype. As the two genotype 7 sequences so far discovered were found in a native of the DRC and a DRC emigrant living in Canada (Murphy *et al.*, 2007) it is likely that further genotype 7 diversity will be found in the DRC, not to mention currently unknown genotypes of HCV.

A significant minority of sequences gathered from across Africa in chapter 4 did not belong to any subtype, illustrating the large number of HCV lineages that are too rare or too divergent to be assigned a subtype. The ability of a HCV lineage to spread through human populations seems to be determined by the behaviour of the

populations in which it is present it infects, rather than any genetic adaptations the virus may have (Pybus *et al.,* 2001). However HCV genotypes can vary widely in their responsiveness to antiviral treatments (Chen *et al.*, 2012; Rose *et al.*, 2013), hence the monitoring of viral diversity may be important effective future treatment of HCV in Africa.

The proportion of discordant samples (belonging to different subtypes or lineages in the core and NS5B regions) was unexpectedly high. In total 2.8 % (6/218) of the samples gathered in chapters 2, 3 and 4 and sequenced in both regions showed signs of recombination, although in each case both sequences were classified as belonging to genotype 4. It is possible that some of these results are in fact due to dual infection or highly divergent sequences being placed with poor bootstrap support into different clades. Sentandreu *et al.,* 2008 estimated a recombination rate of 10.7 % in patients undergoing HCV treatment, but as their study group included a large proportion of IDUs their estimates may not be applicable to a wider population.

As discussed in chapter 5, techniques required to reliably detect and confirm recombination are not routinely applied, and so the rate of recombination in HCV may be underestimated. Alternatively, discordant subtyping due to dual infection could be mischaracterised as recombinants, overestimating the rate of recombination. Next-generation sequencing technology is likely to be an important tool in the resolution of both of these problems as it can sequence the complete population of variants within a sample quickly, cost-effectively and without bias (Metzker, 2010). As the use of this technology increases, much more information about HCV diversity will likely become available, but great changes will need to be made in how sequence data is stored, catalogued and analysed in order to make the most of the data generated. The Deep Simplot technique developed in chapter 5 may be an important tool in the future

analysis of next-generation sequence data; while its main purpose is to distinguish between dual infection and recombination, and estimate the position of recombination breakpoints, it could also be used to assess the level of variation within a sample, or to identify the portions of a genome that have diverged the most from a reference. Further work is needed to refine the tool; its ability to distinguish between recombination and dual infection could be tested by mixing control positive samples before sequencing in various ratios and contrasting the results with that from a known recombinant, while the selection of reference genomes is a process that currently requires prior knowledge of the variation within the sample, and should be investigated further.

The history of HCV is in many ways the history of our ability to detect and characterise viruses. In the 1940s and 1950s, our ignorance of its existence likely led to its mass amplification in Africa via the use of inadequate sterile techniques. In the 1960s and 1970s, serological tests were developed to discriminate between Hepatitis A and Hepatitis B, and led to the realisation that a third non-A non-B pathogen was causing a significant proportion of hepatitis. In the 1989, HCV was discovered through the creation of a random-primed complementary DNA library, and in the decade following its discovery newly-developed serological tests effectively eliminated HCV from the blood banks of developed countries. In the past decade phylogenetic and screening techniques have greatly improved, leading to the discovery of a new genotype, circulating inter-genotypic recombinants, and closely related viruses in dogs, horses and rodents, as well as the reconstruction of HCV's transmission dynamics hundreds of years into the past. With the use of exponentially more sophisticated sequencing methods, far greater connectivity of clinics across the world, and ever-improving phylodynamic techniques such as those explored within this thesis, the

future promises to immeasurably expand our understanding of HCV and other pathogens worldwide.

6.4 REFERENCES

Chen, Y., Xu, H. X., Wang, L. J., Liu, X. X., Mahato, R. I. and Zhao, Y. R. (2012). Meta-analysis: IL28B polymorphisms predict sustained viral response in HCV patients treated with pegylated interferon-a and ribavirin. *Alimentary Pharmacology and Therapeutics,* 36: 91–103.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peters, M., Pybus, O. G. and Lemey, P. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346: 56-61.

Metzker, M. L. (2010). Sequencing technologies – the next generation. *Nat Rev Genet* 11: 31-46.

Mohd Hanafiah, K., Groeger, J., Flaxman, A. D. and Wiersma, S. T. (2013). Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. *Hepatology* 57: 1333-42.

Murphy, D. G., Willems, B., Deschenes, M., Hilzenrat, N., Mousseau, R. and Sabbah, S. (2007). Use of Sequence Analysis of the NS5B Region for Routine Genotyping of Hepatitis C Virus with Reference to C/E1 and 5' Untranslated Region Sequences. *Journal of Clinical Microbiology* 45: 1102–1112.

Nerrienet, E., Pouillot, R., Lachenal, G., Njouom, R., Mfoupouendoun, J., Bilong, C., Mauclere, P., Pasquier, C. and Ayouba, A. (2005). Hepatitis C virus infection Cameroon: a cohort-effect. *Journal of Medical Virology* 76: 208-214.

Njouom, R., Frost, E., Deslandes, S., Mamadou-Yaya, F., Labbé, A.-C., Pouillot, R., Mbélesso, P., Mbadingai, S., Rousset, D. and Pépin, J. (2009). Predominance of

hepatitis C virus genotype 4 infection and rapid transmission between 1935 and 1965 in the Central African Republic. *Journal of General Virology* 90: 2452-2456.

Njouom, R., Caron, M., Besson, G., Ndong-Atome, G-R., Makuwa, M., Pouillot, R., Nkoghé, D., Leroy, E. and Kazanji, M. (2012). Phylogeography, risk factors and genetic history of hepatitis C virus in Gabon, central Africa. *PLoS One* 7: e42002.

Pépin, J. and Labbé, A. C. (2008). Noble goals, unforeseen consequences: control of tropical diseases in colonial Central Africa and the iatrogenic transmission of blood-borne viruses. *Trop Med Int Health* 13: 744-53.

Pépin, J., Lavoie, M., Pybus, O. G., Pouillot, R., Foupouapouognigni, Y., Rousset, D., Labbé, A.-C. and Njouom, R. (2010). Risk Factors for Hepatitis C Virus Transmission in Colonial Cameroon. *Clinical infectious Diseases* 51: 768–776.

Pépin, J. (2011). *The Origins of AIDS*. Cambridge University Press.

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C. and Harvey, P. H. (2001). The epidemic behaviour of the hepatitis C virus. *Science* 292: 2323-25.

Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H. and Rambaut, A. (2003). The Epidemiology and Iatrogenic Transmission of Hepatitis C Virus in Egypt: A Bayesian Coalescent Approach. *Molecular Biology and Evolution* 20: 381–387.

Rose, R., Markov, P. V., Lam, T. T. and Pybus, O. G. (2013). Viral evolution explains the associations among hepatitis C virus genotype, clinical outcomes, and human genetic variation. *Infection, Genetics and Evolution* 20: 418–421.

Sentandreu, V., Jiménez-Hernández, N., Torres-Puente, M., Bracho, M. A., Valero, A., Gosalbes, M. J., Ortega, E., Moya, A. and González-Candelas, F. (2008). Evidence of recombination in intrapatient populations hepatitis C virus. *PLoS One* 18: e3239.

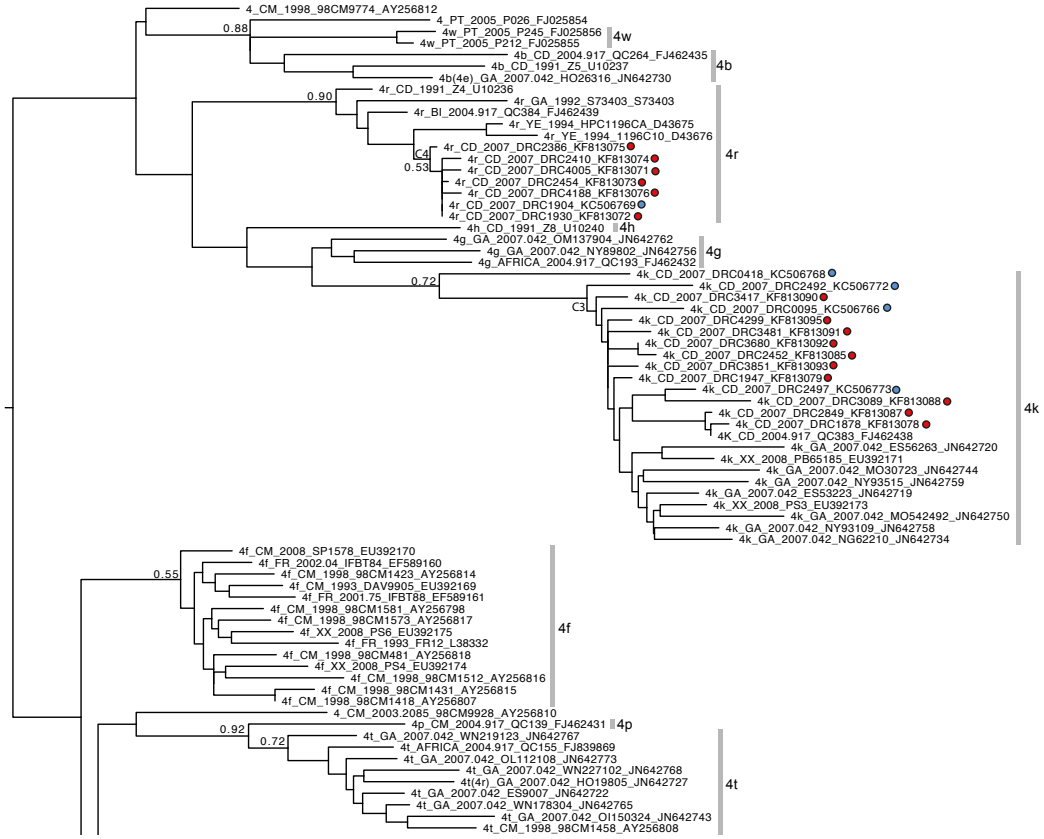
Williams, B. G., Gouws, E. (2014). R0 and the elimination of HIV in Africa: will 90-90-90 be sufficient? *arXiv*:1304.3720.

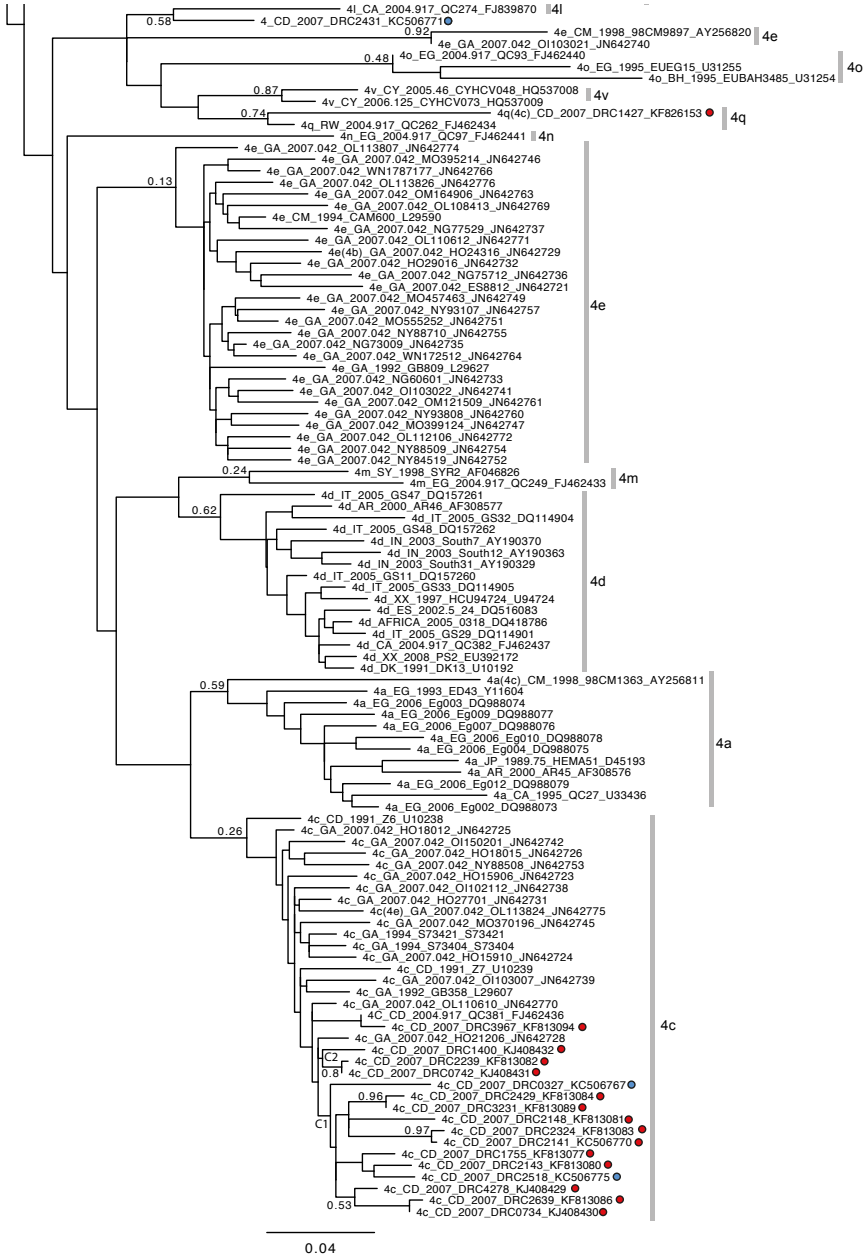
7 APPENDIX: SUPPLEMENTARY INFORMATION

7.3 CHAPTER 3

7.3.1 Core ML Phylogeny

Estimated maximum likelihood phylogeny for the core gene alignment. Bootstrap scores are shown for each subtype cluster and for clades containing samples obtained in this study, and the phylogeny is midpoint rooted. Branch lengths are in units of expected substitutions per site (see scale bar at bottom of figure). Sequences obtained in this study are marked with a red circle; those obtained from the same population during a pilot study (see Chapter 2) are marked with a blue circle. Sequences are labelled as follows: subtype, sampling location using two-letter country codes (ISO 3166), sampling date, isolate name, accession number. XX represents an unknown location. Subtypes are indicated with grey bars on the right side of the diagram. Subtypes in parentheses denote isolates whose core and NS5B subtypes are discordant. The subtype in parenthesis represents the subtype of the corresponding NS5B sequence. The four clusters of samples obtained in this study discussed in the main text are labelled C1, C2, C3, and C4.


4_CM_1998_98CM9774_AY256812
4_PT_2005_P026_FJ025854
0.88
4w_PT_2005_P245_FJ025856
4w_PT_2005_P212_FJ025855
4w
4b_CD_2004.917_QC264_FJ462435
4b_CD_1991_Z5_U10237
4b
4b(4e)_GA_2007.042_HO26316_JN642730
4r_CD_1991_Z4_U10236
0.90
4r_GA_1992_S73403_S73403
4r_BI_2004.917_QC384_FJ462439
4r_YE_1994_HPC1196CA_D43675
4r_YE_1994_1196C10_D43676
C4
4r_CD_2007_DRC2386_KF813075
0.53
4r_CD_2007_DRC2410_KF813074
4r_CD_2007_DRC4005_KF813071
4r_CD_2007_DRC2454_KF813073
4r_CD_2007_DRC4188_KF813076
4r_CD_2007_DRC1904_KC506769
4r_CD_2007_DRC1930_KF813072
4r
4h_CD_1991_Z8_U10240
4h
4g_GA_2007.042_OM137904_JN642762
4g_GA_2007.042_NY89802_JN642756
4g_AFRICA_2004.917_QC193_FJ462432
4g
4k_CD_2007_DRC0418_KC506768
0.72
4k_CD_2007_DRC2492_KC506772
4k_CD_2007_DRC3417_KF813090
C3
4k_CD_2007_DRC0095_KC506766
4k_CD_2007_DRC4299_KF813095
4k_CD_2007_DRC3481_KF813091
4k_CD_2007_DRC3680_KF813092
4k_CD_2007_DRC2452_KF813085
4k_CD_2007_DRC3851_KF813093
4k_CD_2007_DRC1947_KF813079
4k_CD_2007_DRC2497_KC506773
4k_CD_2007_DRC3089_KF813088
4k_CD_2007_DRC2849_KF813087
4k_CD_2007_DRC1878_KF813078
4K_CD_2004.917_QC383_FJ462438
4k_GA_2007.042_ES56263_JN642720
4k_XX_2008_PB65185_EU392171
4k_GA_2007.042_MO30723_JN642744
4k_GA_2007.042_NY93515_JN642759
4k_GA_2007.042_ES53223_JN642719
4k_XX_2008_PS3_EU392173
4k_GA_2007.042_MO542492_JN642750
4k_GA_2007.042_NY93109_JN642758
4k_GA_2007.042_NG62210_JN642734
4k
4f_CM_2008_SP1578_EU392170
4f_FR_2002.04_IFBT84_EF589160
0.55
4f_CM_1998_98CM1423_AY256814
4f_CM_1993_DAV9905_EU392169
4f_FR_2001.75_IFBT88_EF589161
4f_CM_1998_98CM1581_AY256798
4f_CM_1998_98CM1573_AY256817
4f_XX_2008_PS6_EU392175
4f_FR_1993_FR12_L38332
4f_CM_1998_98CM481_AY256818
4f_XX_2008_PS4_EU392174
4f_CM_1998_98CM1512_AY256816
4f_CM_1998_98CM1431_AY256815
4f_CM_1998_98CM1418_AY256807
4f
4_CM_2003.2085_98CM9928_AY256810
4p_CM_2004.917_QC139_FJ462431
4p
0.92
4t_GA_2007.042_WN219123_JN642767
0.72
4t_AFRICA_2004.917_QC155_FJ839869
4t_GA_2007.042_OL112108_JN642773
4t_GA_2007.042_WN227102_JN642768
4t(4r)_GA_2007.042_HO19805_JN642727
4t_GA_2007.042_ES9007_JN642722
4t_GA_2007.042_WN178304_JN642765
4t_GA_2007.042_OI150324_JN642743
4t_CM_1998_98CM1458_AY256808
4t


4l_CA_2004.917_QC274_FJ839870
4l
0.58
4_CD_2007_DRC2431_KC506771
0.92
4e_CM_1998_98CM9897_AY256820
4e
4e_GA_2007.042_OI103021_JN642740
0.48
4o_EG_2004.917_QC93_FJ462440
4o_EG_1995_EUEG15_U31255
4o
4o_BH_1995_EUBAH3485_U31254
0.87
4v_CY_2005.46_CYHCV048_HQ537008
4v
4v_CY_2006.125_CYHCV073_HQ537009
0.74
4q(4c)_CD_2007_DRC1427_KF826153
4q
4q_RW_2004.917_QC262_FJ462434
4n_EG_2004.917_QC97_FJ462441
4n
4e_GA_2007.042_OL113807_JN642774
4e_GA_2007.042_MO395214_JN642746
0.13
4e_GA_2007.042_WN1787177_JN642766
4e_GA_2007.042_OL113826_JN642776
4e_GA_2007.042_OM164906_JN642763
4e_GA_2007.042_OL108413_JN642769
4e_CM_1994_CAM600_L29590
4e_GA_2007.042_NG77529_JN642737
4e_GA_2007.042_OL110612_JN642771
4e(4b)_GA_2007.042_HO24316_JN642729
4e_GA_2007.042_HO29016_JN642732
4e_GA_2007.042_NG75712_JN642736
4e_GA_2007.042_ES8812_JN642721
4e_GA_2007.042_MO457463_JN642749
4e_GA_2007.042_NY93107_JN642757
4e_GA_2007.042_MO555252_JN642751
4e_GA_2007.042_NY88710_JN642755
4e_GA_2007.042_NG73009_JN642735
4e_GA_2007.042_WN172512_JN642764
4e_GA_1992_GB809_L29627
4e_GA_2007.042_NG60601_JN642733
4e_GA_2007.042_OI103022_JN642741
4e_GA_2007.042_OM121509_JN642761
4e_GA_2007.042_NY93808_JN642760
4e_GA_2007.042_MO399124_JN642747
4e_GA_2007.042_OL112106_JN642772
4e_GA_2007.042_NY88509_JN642754
4e_GA_2007.042_NY84519_JN642752
4e
0.24
4m_SY_1998_SYR2_AF046826
4m_EG_2004.917_QC249_FJ462433
4m
4d_IT_2005_GS47_DQ157261
4d_AR_2000_AR46_AF308577
0.62
4d_IT_2005_GS32_DQ114904
4d_IT_2005_GS48_DQ157262
4d_IN_2003_South7_AY190370
4d_IN_2003_South12_AY190363
4d_IN_2003_South31_AY190329
4d_IT_2005_GS11_DQ157260
4d_IT_2005_GS33_DQ114905
4d_XX_1997_HCU94724_U94724
4d_ES_2002.5_24_DQ516083
4d_AFRICA_2005_0318_DQ418786
4d_IT_2005_GS29_DQ114901
4d_CA_2004.917_QC382_FJ462437
4d_XX_2008_PS2_EU392172
4d_DK_1991_DK13_U10192
4d
0.59
4a(4c)_CM_1998_98CM1363_AY256811
4a_EG_1993_ED43_Y11604
4a_EG_2006_Eg003_DQ988074
4a_EG_2006_Eg009_DQ988077
4a_EG_2006_Eg007_DQ988076
4a_EG_2006_Eg010_DQ988078
4a_EG_2006_Eg004_DQ988075
4a_JP_1989.75_HEMA51_D45193
4a_AR_2000_AR45_AF308576
4a_EG_2006_Eg012_DQ988079
4a_CA_1995_QC27_U33436
4a_EG_2006_Eg002_DQ988073
4a
4c_CD_1991_Z6_U10238
0.26
4c_GA_2007.042_HO18012_JN642725
4c_GA_2007.042_OI150201_JN642742
4c_GA_2007.042_HO18015_JN642726
4c_GA_2007.042_NY88508_JN642753
4c_GA_2007.042_HO15906_JN642723
4c_GA_2007.042_OI102112_JN642738
4c_GA_2007.042_HO27701_JN642731
4c(4e)_GA_2007.042_OL113824_JN642775
4c_GA_2007.042_MO370196_JN642745
4c_GA_1994_S73421_S73421
4c_GA_1994_S73404_S73404
4c_GA_2007.042_HO15910_JN642724
4c_CD_1991_Z7_U10239
4c_GA_2007.042_OI103007_JN642739
4c_GA_1992_GB358_L29607
4c_GA_2007.042_OL110610_JN642770
4C_CD_2004.917_QC381_FJ462436
4c_CD_2007_DRC3967_KF813094
4c_GA_2007.042_HO21206_JN642728
4c_CD_2007_DRC1400_KJ408432
C2
4c_CD_2007_DRC2239_KF813082
0.8
4c_CD_2007_DRC0742_KJ408431
4c_CD_2007_DRC0327_KC506767
0.96
4c_CD_2007_DRC2429_KF813084
4c_CD_2007_DRC3231_KF813089
C1
4c_CD_2007_DRC2148_KF813081
0.97
4c_CD_2007_DRC2324_KF813083
4c_CD_2007_DRC2141_KC506770
4c_CD_2007_DRC1755_KF813077
4c_CD_2007_DRC2143_KF813080
4c_CD_2007_DRC2518_KC506775
4c_CD_2007_DRC4278_KJ408429
0.53
4c_CD_2007_DRC2639_KF813086
4c_CD_2007_DRC0734_KJ408430
4c
0.04

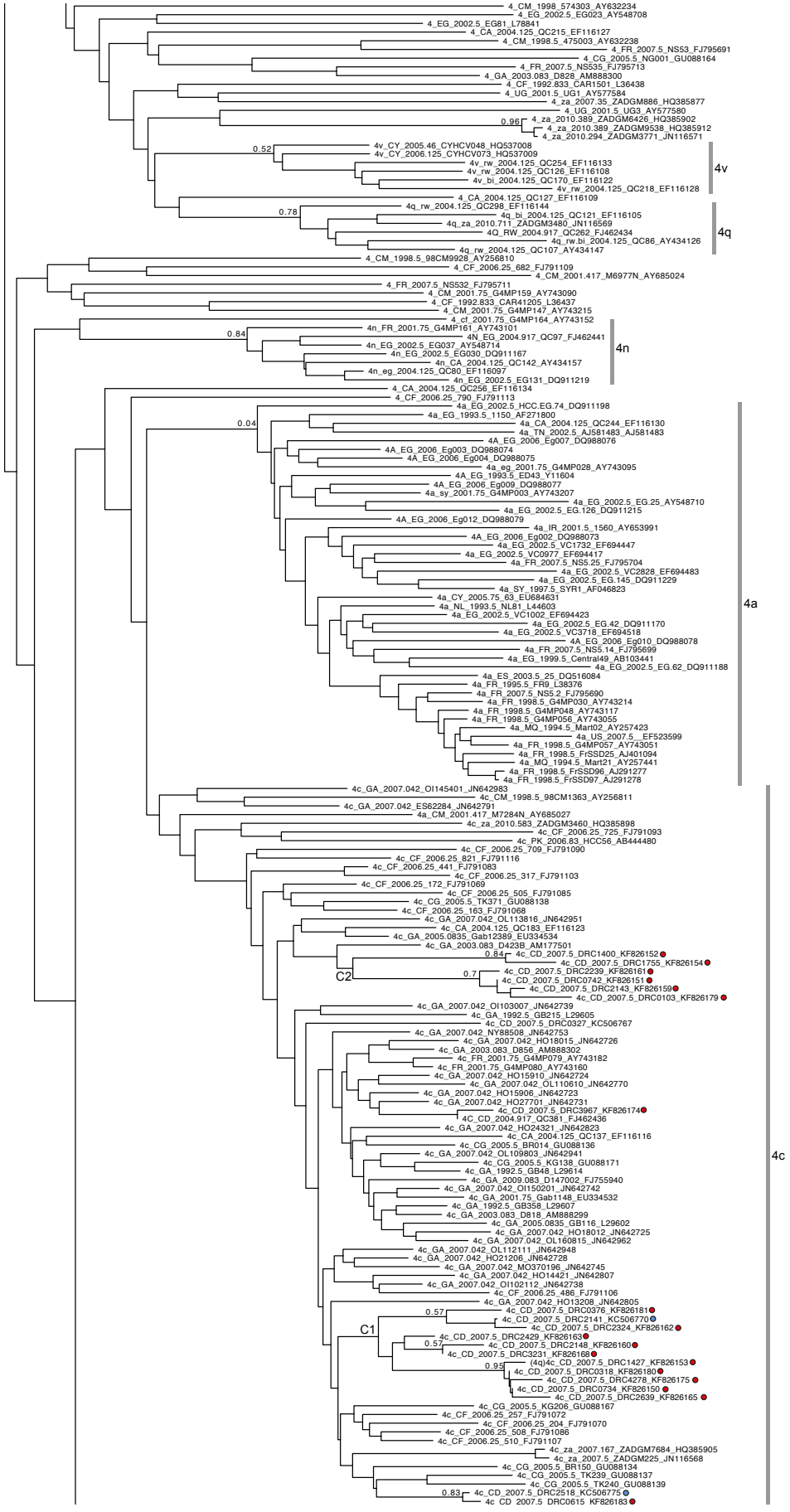
7.3.2 NS5B ML Phylogeny

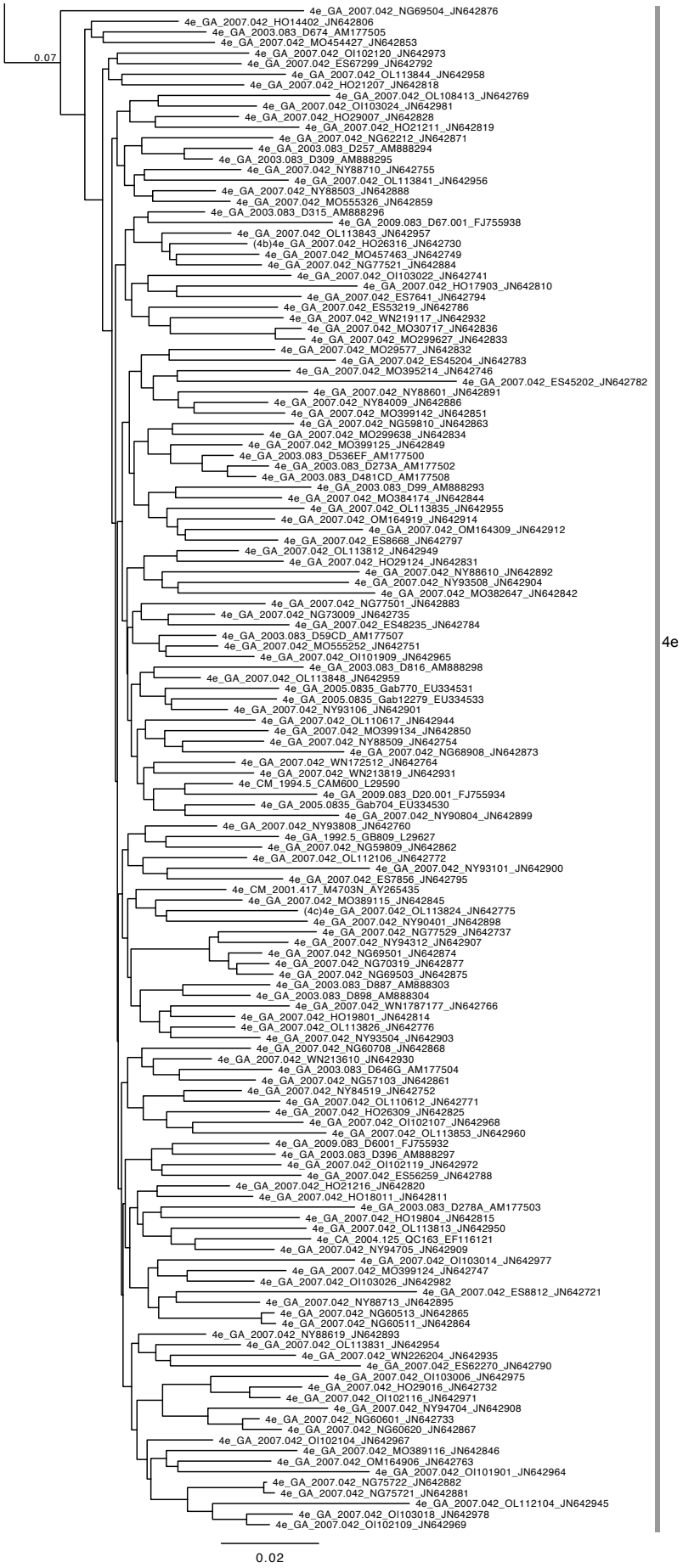
Estimated maximum likelihood phylogeny for the NS5B gene alignment. Subtypes in parentheses denote isolates whose core and NS5B subtypes are discordant. The subtype in parenthesis represents the subtype of the corresponding core sequence. See the legend of Figure 7.3.1 for further figure details.


4r_cd_1998_FrSSD120_AJ291282
4r_CG_2005.5_NG079_GU088185
4r_bi_2004.125_QC92_EF116098
4R_BI_2004.917_QC384_FJ462439
4r_cd.cg_2004.125_QC306_EF116146
4r_CF_2001.75_G4MP073_AY743066
4r_GA_2007.042_HO23307_JN642821
4r_GA_2003.083_D831_AM888301
(4t)4r_GA_2007.042_HO19805_JN642727
4r_FR_2001.75_G4MP078_AY743171
4r_DE_2005.5_S97_EF467648
4r_CG_2005.5_KG024_GU088186
4r_CG_2005.5_KG032_GU088182
4r_FR_1998.5_FRSSD173_AJ291293
4r_cd.cg_2004.125_QC141_EF116118
4r_FR_2001.75_G4MP075_AY743159
4r_cd_1998_FrSSD136_AJ291284
4r_CG_2005.5_NG036_GU088184
4r_cd_2001.75_G4MP074_AY743147
4r_GA_2007.042_OL108801_JN642940
4r_FR_2007.5_NS529_FJ795708
4r_CF_2006.25_26_FJ791063
4r_CF_2006.25_753_FJ791112
4r_za_2009.7_ZADGM8690_HQ385908
4r_FR_2001.75_G4MP076_AY743098
4r_UG_2001.5_UG2_AY577579
4r_CG_2005.5_TK116_GU088155
4r_FR_2001.75_G4MP077_AY743099
4r_CG_2005.5_BR049_GU088181
4r_cd.cg_2004.125_QC304_EF116145
4r_cd_1998.5_FRSSD162_AJ291288
4r_CD_2007.5_DRC4188_KF813076
4r_CD_2007.5_DRC1930_KF813072
4r_CD_2007.5_DRC4005_KF813071
4r_CD_2007.5_DRC2386_KF813075
4r_CD_2007.5_DRC2454_KF813073
4r_CD_2007.5_DRC2410_KF813074
4r_CD_2007.5_DRC0889_KF826187
4r_CD_2007.5_DRC1904_KC506769
4r_CD_2007.5_DRC1423_KF826191
4r_CD_2007.5_DRC1424_KC506776
4h_CA_2004.125_QC128_EF116110
4h_CA_2004.125_QC118_EF116103
4h_CA_2004.125_QC119_EF116104
4h_cd_1998.5_FrSSD35_AJ291249
4h_GA_2001.75_G4MP146_AY743206
4h_CA_2004.125_QC281_EF116141
4h_CA_2004.125_QC122_EF116106
4h_CG_2005.5_BR053_GU088141
4h_CD_2007.5_DRC2016_KF826157
4h_CD_2007.5_DRC2045_KF826158
4h_CA_2004.125_QC136_EF116115
4h_GA_1992.5_GB438_L29613
4h_cg_1998.5_FrSSD61_AJ291264
4h_CG_2005.5_NG123_GU088173
4h_CG_2005.5_NG144_GU088175
4_CG_2005.5_NG053_GU088160
4g_GA_2007.042_OM137904_JN642762
4g_GA_2007.042_NY89802_JN642756
4g_GA_2009.083_D106002_FJ755937
4G_AFRICA_2004.917_QC193_FJ462432
4g_GA_2007.042_NY89801_JN642896
4g_GA_1992.5_GB549_L29617
4g_CG_2005.5_BR013_GU088172
4_CF_2006.25_145_FJ791100
4_CF_2006.25_296_FJ791102
4_CF_2006.25_794_FJ791095
4_CF_2006.25_133_FJ791065
4_CF_2006.25_355_FJ791077
4_CF_2006.25_217_FJ791071
4_CF_2006.25_471_FJ791084
4_CF_2006.25_165_FJ791101
4_CF_2006.25_435_FJ791082
4_CF_2006.25_444_FJ791105
4_CF_2006.25_273_FJ791073
4_CF_2006.25_394_FJ791078
4_CF_2006.25_392_FJ791104
4_CF_2006.25_844_FJ791117
4_CF_2006.25_293_FJ791074
4_CF_2006.25_425_FJ791080
4_CF_2006.25_650_FJ791108
4_CF_2006.25_27_FJ791098
4_CF_2006.25_155_FJ791066
4_CF_2006.25_430_FJ791081
4k_CD_2007.5_DRC0418_KC506768
4k_CA_2004.125_QC124_EF116107
4k_CF_2006.25_326_FJ791076
4k_za_2006.95_ZADGM1903_JN116563
4k_CD_2007.5_DRC0095_KC506766
4k_CD_2007.5_DRC1243_KF826189
4k_cg_2001.75_G4MP155_AY743201
4k_CF_2001.75_G4MP149_AY743210
4k_CF_2006.25_795_FJ791114
4k_BE_1993.5_B14_L44597
4k_FR_2001.75_G4MP148_AY743068
4k_CF_2001.75_G4MP156_AY743202
4k_CG_2005.5_KG095_GU088176
4k_CG_2005.5_KG103_GU088177
4k_GB_2008.5_PS3_EU392173
4k_CM_2001.75_G4MP151_AY743211
4k_CD_2007.5_DRC2849_KF813087
4k_CD_2007.5_DRC1878_KF813078
4K_CD_2004.917_QC383_FJ462438
4k_CA_2004.125_QC237_EF116129
4k_FR_2007.5_NS5.36_FJ795714
4k_CD_2007.5_DRC3308_KF826169
4k_CG_2005.5_BR042_GU088144
4k_CD_2007.5_DRC4616_KF826177
4k_CD_2001.75_G4MP153_AY743203
4k_CA_2004.125_QC130_EF116112
4k_RW_1999.5_DON91_DQ238700
4k_CG_2005.5_NG110_GU088180
4k_GA_2007.042_ES56263_JN642720
4k_RW_2001.75_G4MP150_AY743167
4k_GA_2007.042_NY93515_JN642759
4k_GA_2007.042_MO542492_JN642750
4k_GA_2007.042_ES53223_JN642719
4k_GA_2007.042_MO541513_JN642855
4k_GA_2007.042_NY93109_JN642758
4k_CG_2005.5_BR044_GU088145
4k_za_2007.417_ZADGM655_HQ385891
4k_GA_2007.042_ES39295_JN642781
4k_bi_1998.5_FrSSD174_AJ291294
4k_CF_2006.25_843_FJ791097
4k_CD_2007.5_DRC3089_KF826167
4k_CD_2007.5_DRC0870_KF826186
4k_GA_2007.042_NG62210_JN642734
4k_GA_2007.042_MO30723_JN642744
4k_CM_1998.5_1398603_AY632237
4k_cg_2001.75_G4MP152_AY743204
4k_RW_2001.75_G4MP154_AY743205
4k_CD_2007.5_DRC1129_KF826188
4k_CD_2007.5_DRC3481_KF813091
4k_CD_2007.5_DRC0017_KF826178
4k_CD_2007.5_DRC1300_KF826190
4k_TN_2002.5_AJ581496_AJ581496
4k_GB_2008.5_PB65185_EU392171
4k_CA_2004.125_QC129_EF116111
4k_CD_2007.5_DRC2497_KC506773
4k_CD_2007.5_DRC0660_KF826184
4k_TN_2002.5_AJ581495_AJ581495
4k_CD_2007.5_DRC2492_KC506772
4k_CD_2007.5_DRC0728_KF826185
4k_CD_2007.5_DRC1947_KF826156
4k_CD_2007.5_DRC4299_KF826176
4k_CD_2007.5_DRC3680_KF826172
4k_CD_2007.5_DRC2452_KF826164
4k_CD_2007.5_DRC3417_KF826170
4k_CD_2007.5_DRC3851_KF826173
4k_TN_2002.75_C7Tun_DQ508482
4k_TN_2002.5_AJ581510_AJ581510
4k_TN_2002.5_AJ581502_AJ581502
4k_TN_2002.5_AJ581509_AJ581509
4k_TN_2002.5_AJ581504_AJ581504
4k_TN_2002.5_AJ581501_AJ581501
4k_TN_2002.5_AJ581500_AJ581500
4k_TN_2002.5_AJ581508_AJ581508
4k_TN_2002.5_AJ581499_AJ581499
4k_TN_2002.5_AJ581497_AJ581497
4k_TN_2002.5_AJ581498_AJ581498
4k_TN_2002.5_AJ581506_AJ581506
4k_TN_2002.5_AJ581507_AJ581507
0.83
0.79
C4
0.25
0.99
0.82
0.86
0.37
0.58
0.9
0.97
0.54
C3
0.68
0.74
0.87
0.85
4r
4h
4g
4car
4k


4f_CF_2006.25_9_FJ791062
4f_cm_1998.5_FrSSD159_AJ291286
4f_CM_2001.75_G4MP140_AY743091
4f_FR_2007.5_NS5.31_FJ795710
4f_CM_1998.5_1460203_AY632214
4f_CM_1998.5_480503_AY632215
4f_CF_2006.25_799_FJ791096
4f_CM_1998.5_3294803_AY632213
4f_CM_2001.417_M5177N_AY265440
4f_CM_2001.417_M7335N_AY685028
4f_GB_2008.5_PS6_EU392175
4f_dz_1998.5_FrSSD160_AJ291287
4f_CM_1998.5_1574303_AY632212
4f_GA_2007.042_WN173209_JN642920
4f_CM_1992.5_G22_L29595
4f_CM_1998.5_1476103_AY632211
4f_FR_2007.5_NS5.21_FJ795701
4f_CM_2001.417_M5225N_AY265441
4f_CA_2004.125_QC135_EF116114
4f_GA_2007.042_MO382666_JN642843
4f_GA_2007.042_WN197103_JN642928
4f_MQ_1994.5_Mart19_AY257439
4f_CM_2001.75_G4MP144_AY743145
4F_FR_2001.75_IFBT88_EF589161
4f_GA_1992_G4MP139_AY743063
4F_FR_2002.04_IFBT84_EF589160
4f_CM_2001.75_G4MP143_AY743125
4f_CM_2001.417_M182E_AY265434
4f_GA_2007.042_NG75714_JN642879
4f_GA_2007.042_ES8671_JN642798
4f_GA_2007.042_WN172507_JN642918
4f_GA_2007.042_WN1787181_JN642924
4f_CM_2001.417_M6638N_AY685021
4f_GA_2007.042_WN1787178_JN642923
4f_CM_1998.5_913703_AY632218
4f_CM_1998.5_326403_AY632217
4f_CM_2001.417_M7516N_AY685032
4f_CM_1998.5_491203_AY632208
4f_CM_1998.5_125003_AY632207
4f_CA_2004.125_QC276_EF116140
4f_CM_1998.5_868803_AY632203
4f_CM_2001.417_M3645N_AY685012
4f_GA_2007.042_WN190918_JN642926
4f_CM_1998.5_251203_AY632201
4f_CM_1998.5_1055303_AY632219
4f_GA_2007.042_ES9002_JN642801
4f_GA_2007.042_ES7864_JN642796
4f_GA_1994_ES20149_JN642778
4f_GA_2007.042_MO454415_JN642852
4f_CM_1998.5_992303_AY632210
4f_GA_2007.042_ES9004_JN642802
4f_MQ_1994.5_Mart18_AY257438
4f_CM_2001.417_M6872N_AY685023
4f_CM_1998.5_A13150125C_AY265421
4f_GA_2007.042_ES7638_JN642793
4f_CM_2001.417_M6310N_AY685020
4f_CM_1998.5_158603_AY632197
4f_CM_2001.417_M7858N_AY685038
4f_GA_2007.042_OL153006_JN642961
4f_CM_1998.5_18503_AY632191
4f_CM_2001.417_M7231N_AY685025
4f_CM_2001.417_M1527E_AY685052
4f_CM_1998.5_1057103_AY632204
4f_CM_1998.5_2998702_AY632206
4f_CM_2001.417_M5807N_AY685014
4f_CM_2001.417_M7480N_AY685031
4f_CM_2001.417_M7863N_AY685039
4f_CM_1998.5_3016202_AY632199
4f_GA_2007.042_OI99214_JN642963
4f_CM_1998.5_280403_AY632202
4f_CM_1998.5_193503_AY632205
4f_CM_1998.5_3355502_AY632192
4f_CM_2001.417_M718E_AY685045
4f_CM_1998.5_665603_AY632195
4f_CM_2001.417_M5647N_AY685013
4f_GB_2008.5_PS4_EU392174
4f_CM_1998.5_323403_AY632198
4f_CM_2001.417_M6560N_AY685016
4f_CM_2001.75_G4MP141_AY743092
4f_CM_1998.5_98CM1431_AY256815
4f_CM_1998.5_A07300246C_AY265428
4f_CM_1998.5_1257003_AY632193
4f_GA_2007.042_ES48246_JN642785
4f_CM_2008.5_SP1578_EU392170
4f_CM_1998.5_98CM1578_AY257075
4f_CM_1993.5_DAV9905_EU392169
4f_GA_2007.042_WN183905_JN642925
4f_GA_2007.042_ES20156_JN642779
4f_CM_2001.75_G4MP145_AY743056
4f_cm_1998.5_FrSSD125_AJ291283
4f_CM_1998.5_3173603_AY632216
4f_CM_2001.75_G4MP142_AY743148
4f_CM_1998.5_1272703_AY632200
4f_FR_2007.5_NS5.28_FJ795707
4f_CM_2001.417_M376E_AY265448
4f_FR_2007.5_NS5.38_FJ795716
4f_CM_1998.5_811203_AY632209
4f_CM_1998.5_1025903_AY632194
4f_CM_1998.5_350303_AY632196
4f_CM_2001.417_M5574N_AY265445
4f_ga_1998.5_FrSSD36_AJ291250
4f_CM_2001.417_M7920N_AY685042
4f_CM_1998.5_820403_AY632153
4f_GA_2007.042_OI151911_JN642987
4f_CM_1998.5_A02210237C_AY265424
4f_CM_2001.417_M4962N_AY265438
4f
4m_EG_2002.5_EG050_DQ911177
4M_EG_2004.917_QC249_FJ462433
4m_DE_2005.5_S103_EF467653
4m_EG_2002.5_EG159_DQ911238
4m_SA_2008.5_SD035_D86543
4m_EG_1993.5_1797_AF271813
4m_EG_2002.5_EG027_AY548712
4m_EG_2002.5_EG.26_AY548711
4m_EG_2002.5_EG.44_DQ911172
4m_EG_2002.5_EG.154_DQ911235
4m_EG_2002.5_VC0996_EF694334
4m_EG_2002.5_HCC.EG.148_AY548731
4m_EG_2002.5_LYM.EG.152_DQ911233
4m_EG_2002.5_EG162_DQ911240
4m
4_CF_2006.25_820_FJ791115
4_CM_1998.5_164003_AY632236
4P_CM_2004.917_QC139_FJ462431
4p_GA_2003.083_D71_AM888292
4p_FR_2007.5_NS542_FJ795720
4p_CM_1998.5_A07321152C_AY265429
4p_CM_1998.5_994403_AY632233
4p_CM_1998.5_575603_AY632232
4p_CM_1998.5_847103_AY632231
4p_CM_2001.75_G4MP157_AY743124
4p_CM_2001.417_M6732N_AY685017
4p_DE_2005.5_S69_EF467646
4p_cm_1998.5_FrSSD158_AJ291285
4p_CM_1998.5_1184503_AY632235
4p_CM_2002.5415_M944E_AY685047
4p
0.61
0.52
0.62


4t_GA_2007.042_WN226216_JN642936
4t_GA_2007.042_MO542493_JN642857
4t_GA_2007.042_WN219120_JN642933
4t_CM_1998.5_3039802_AY632228
4t_CM_1998.5_A13020210C_AY265430
4t_CM_2001.75_G4MP165_AY743163
4t_GA_2007.042_WN227102_JN642768
4t_CM_2001.417_M329E_AY265446
4T_AFRICA_2003.89_QC155_FJ839869
4t_GA_2007.042_MO555350_JN642860
4t_CM_1998.5_A02240258C_AY265425
4t_CM_1998.5_769703_AY632222
4t_CM_1998.5_3109102_AY632220
4t_CM_1998.5_920603_AY632223
4t_GA_2007.042_OI150324_JN642743
4t_CM_1998.5_857703_AY632224
4t_GA_2007.042_WN219123_JN642767
4t_CM_1998.5_1339403_AY632229
4t_CM_1998.5_98CM1458_AY256808
4t_GA_2007.042_WN178304_JN642765
4t_GA_2007.042_OL112108_JN642773
4t_CM_2001.75_G4MP160_AY743212
4t_GA_2007.042_ES9007_JN642722
4t_GA_2007.042_ES9008_JN642804
4t_CM_1998.5_891003_AY632221
4t_CM_2001.417_M7395N_AY685030
4t_FR_2001.75_G4MP158_AY743208
4t_CA_2004.125_QC85_AY706997
4t_CM_1998.5_1006603_AY632225
4t_CM_1998.5_554703_AY632226
4t_CM_1998.5_777703_AY632227
0.83
4t
4D_africa_2005.5_3.18_DQ418786
4d_SA_1995.5_SD008_D86538
4d_SA_1995.5_SD006_D86537
4d_IT_1996.5_S16026_AY944677
4d_IR_2001.5_1447_AY653994
4d_IR_2001.5_1460_AY653995
4d_IR_2001.5_930_AY653992
4d_IR_2001.5_943_AY653993
4d_IR_2001.5_918_AY653996
4d_IR_2001.5_1457_AY653997
4d_IR_2001.5_1454_AY653998
0.58
4d_FR_2001.75_G4MP137_AY743154
4d_DE_2005.5_S99_EF467650
4d_FR_2001.75_G4MP088_AY743076
4d_FR_2001.75_G4MP114_AY743114
4d_FR_2001.75_G4MP132_AY743199
4d_FR_2001.75_G4MP085_AY743054
4d_FR_2001.75_G4MP082_AY743169
4D_CA_2004.917_QC382_FJ462437
4d_FR_2007.5_NS5.1_FJ795689
4d_FR_2001.75_G4MP111_AY743110
4d_FR_2001.75_G4MP117_AY743119
4d_FR_2001.75_G4MP097_AY743190
4d_FR_1998_FrSSD163_AJ291289
4D_ES_2002.5_24_DQ516083
4d_FR_2001.75_G4MP081_AY743065
4d_pt_1998_FrSSD164_AJ291290
4d_FR_2007.5_NS5.39_FJ795717
4d_FR_2001.75_G4MP089_AY743078
4d_FR_2001.75_G4MP108_AY743168
4d_FR_2001.75_G4MP129_AY743198
4d_MQ_1994.5_Mart50_AY257467
4d_FR_1997.875_MRS91_AF516054
4d_GB_2008_PS2_EU392172
4d_DE_2005.5_S98_EF467649
4d_RU_1999.41_624_AF388477
4d_DE_2005.5_S101_EF467652
4d_DE_2005.5_S7_EF205238
4d_FR_2007.5_NS5.8_FJ795696
4d_MQ_1994.5_Mart01_AY257422
4d_FR_1997.875_P9Don_AF515974
4d_FR_1998_FrSSD37_AJ291251
4d_FR_1998_FrSSD85_AJ291271
4d_AR_2001_AR47_AF308575
4d_FR_1997.875_MRS95_AF516057
4d
4_CG_2005.5_NG008_GU088161
4b_CG_2005.5_KG107_GU088163
4b_CG_2005.5_KG083_GU088162
4b_CA_2004.125_QC58_EF116096
0.64
4_PT_2005.5_P026_FJ025854
0.33
4w_PT_2005.5_P245_FJ025856
4w_PT_2005.5_P212_FJ025855
4w
4B_CD_2004.917_QC264_FJ462435
4b_CA_2004.125_QC253_EF116132
4b_cg_2004.125_QC108_EF116101
4b_cg_2004.125_QC109_EF116102
4b_CM_2001.75_G4MP166_AY743139
(4e)4b_GA_2007.042_HO24316_JN642729
4b_GA_2007.042_HO24322_JN642824
4b
4_CM_1998.5_98CM9774_AY256812
4_CF_2006.25_740_FJ791111
4_TN_2002.5_AJ581493_AJ581493
0.72
4_TN_2002.5_AJ581490_AJ581490
4_TN_2002.5_AJ581492_AJ581492
4_TN_2002.5_AJ581491_AJ581491
4_TN_2002.5_AJ581505_AJ581505
4_TN_2006.5_I16_DQ508478
4o_CM_2001.75_G4MP162_AY743105
4o_TN_2002.5_AJ581494_AJ581494
0.85
4o_EG_2002.5_HCCEG011_AY548723
4o_DE_2005.5_S100_EF467651
4o_EG_2002.5_EG052_DQ911179
4o_EG_2002.5_HCC.EG.7_DQ911163
4o_EG_2002.5_LYM.EG.22_AY548734
4o_EG_2002.5_VC0492_EF694328
4o_TN_2002.5_AJ581489_AJ581489
4o_EG_2002.5_LYM.EG.32_AY548736
4o_EG_2002.5_VC1001_EF694335
4o_EG_2002.5_HCC.EG.17_DQ911166
4O_EG_2004.917_QC93_FJ462440
4o_EG_2002.5_EG.65_AY548716
4o_EG_2002.5_HCC.EG.20_AY548729
4o_EG_2002.5_LYM.EG.153_DQ911234
4o_EG_1993.5_2113_AF271814
4o_EG_2002.5_HCp.EG.12_DQ911165
4o_EG_2002.5_EG160_DQ911239
4o_EG_1993.5_2153_AF271815
4o_EG_2002.5_HCp.EG.4_DQ911162
4o_FR_2001.75_G4MP163_AY743113
4o_EG_2002.5_HCC.EG.8_AY548721
4o_EG_2002.5_HCC.EG.13_AY548724
4o_EG_2002.5_HCC.EG.18_AY548728
4o
4_CF_2006.25_6_FJ791061
4_CD_2007.5_DRC0387_KF826182
0.45
4_CD_2007.5_DRC2431_KC506771
4_CD_2007.5_DRC2450_KC506774
4drc
4_CA_2004.125_QC132_EF116113
4_CA_2004.125_QC147_EF116119
4_CA_2004.125_QC282_EF116142
4L_CA_2004.917_QC274_FJ839870
0.2
4l_EG_1993.5_2116_AF271816
4l_EG_2002.5_HCC.EG.147_AY548730
4l_CA_2004.125_QC95_EF116100
4l_EG_2002.5_EG038_AY548715
4l_SA_2008.5_SD002_D86534
4l_CA_2004.125_QC263_EF116137
4l_SA_2008.5_SD015_D86539
4l_EG_2002.5_EG089_AY548717
4l_EG_2002.5_EG139_DQ911224
4l
4u_EG_2002.5_VC2790_EF694481
0.93
4u_EG_2002.5_VC1139_EF694336
4u_EG_2002.5_VC1961_EF694461
4u_EG_2002.5_VC3183_EF694369
4u_EG_2002.5_VC1258_EF694435
4u_EG_2002.5_VC2064_EF694466
4u_EG_2002.5_VC3023_EF694488
4u_EG_2002.5_VC1904_EF694348
4u_EG_2002.5_VC3672_EF694516
4u_EG_2002.5_VC2587_EF694474
4u_EG_2002.5_VC3168_EF694494
4u


4_CM_1998_574303_AY632234
4_EG_2002.5_EG023_AY548708
4_EG_2002.5_EG81_L78841
4_CA_2004.125_QC215_EF116127
4_CM_1998.5_475003_AY632238
4_FR_2007.5_NS53_FJ795691
4_CG_2005.5_NG001_GU088164
4_FR_2007.5_NS535_FJ795713
4_GA_2003.083_D828_AM888300
4_CF_1992.833_CAR1501_L36438
4_UG_2001.5_UG1_AY577584
4_za_2007.35_ZADGM886_HQ385877
4_UG_2001.5_UG3_AY577580
0.96
4_za_2010.389_ZADGM6426_HQ385902
4_za_2010.389_ZADGM9538_HQ385912
4_za_2010.294_ZADGM3771_JN116571
0.52
4v_CY_2005.46_CYHCV048_HQ537008
4v_CY_2006.125_CYHCV073_HQ537009
4v_rw_2004.125_QC254_EF116133
4v_rw_2004.125_QC126_EF116108
4v_bi_2004.125_QC170_EF116122
4v_rw_2004.125_QC218_EF116128
4v
4_CA_2004.125_QC127_EF116109
0.78
4q_rw_2004.125_QC298_EF116144
4q_bi_2004.125_QC121_EF116105
4q_za_2010.711_ZADGM3480_JN116569
4Q_RW_2004.917_QC262_FJ462434
4q_rw.bi_2004.125_QC86_AY434126
4q_rw_2004.125_QC107_AY434147
4q
4_CM_1998.5_98CM9928_AY256810
4_CF_2006.25_682_FJ791109
4_CM_2001.417_M6977N_AY685024
4_FR_2007.5_NS532_FJ795711
4_CM_2001.75_G4MP159_AY743090
4_CF_1992.833_CAR41205_L36437
4_CM_2001.75_G4MP147_AY743215
4_cf_2001.75_G4MP164_AY743152
0.84
4n_FR_2001.75_G4MP161_AY743101
4N_EG_2004.917_QC97_FJ462441
4n_EG_2002.5_EG037_AY548714
4n_EG_2002.5_EG030_DQ911167
4n_CA_2004.125_QC142_AY434157
4n_eg_2004.125_QC80_EF116097
4n_EG_2002.5_EG131_DQ911219
4n
4_CA_2004.125_QC258_EF116134
4_CF_2006.25_790_FJ791113
4a_EG_2002.5_HCC.EG.74_DQ911198
4a_EG_1993.5_1150_AF271800
0.04
4a_CA_2004.125_QC244_EF116130
4a_TN_2002.5_AJ581483_AJ581483
4A_EG_2006_Eg007_DQ988076
4A_EG_2006_Eg003_DQ988074
4A_EG_2006_Eg004_DQ988075
4a_eg_2001.75_G4MP028_AY743095
4A_EG_1993.5_ED43_Y11604
4A_EG_2006_Eg009_DQ988077
4a_sy_2001.75_G4MP003_AY743207
4a_EG_2002.5_EG.25_AY548710
4a_EG_2002.5_EG.126_DQ911215
4A_EG_2006_Eg012_DQ988079
4a_IR_2001.5_1560_AY653991
4A_EG_2006_Eg002_DQ988073
4a_EG_2002.5_VC1732_EF694447
4a_EG_2002.5_VC0977_EF694417
4a_FR_2007.5_NS5.25_FJ795704
4a_EG_2002.5_VC2828_EF694483
4a_EG_2002.5_EG.145_DQ911229
4a_SY_1997.5_SYR1_AF046823
4a_CY_2005.75_63_EU884631
4a_NL_1993.5_NL81_L44603
4a_EG_2002.5_VC1002_EF694423
4a_EG_2002.5_EG.42_DQ911170
4a_EG_2002.5_VC3718_EF694518
4A_EG_2006_Eg010_DQ988078
4a_FR_2007.5_NS5.14_FJ795699
4a_EG_1999.5_Central49_AB103441
4a_EG_2002.5_EG.62_DQ911188
4a_ES_2003.5_25_DQ516084
4a_FR_1995.5_FR9_L38376
4a_FR_2007.5_NS5.2_FJ795690
4a_FR_1998.5_G4MP030_AY743214
4a_FR_1998.5_G4MP048_AY743117
4a_FR_1998.5_G4MP056_AY743055
4a_MQ_1994.5_Mart02_AY257423
4a_US_2007.5__EF523599
4a_FR_1998.5_G4MP057_AY743051
4a_FR_1998.5_FrSSD25_AJ401094
4a_MQ_1994.5_Mart21_AY257441
4a_FR_1998.5_FrSSD96_AJ291277
4a_FR_1998.5_FrSSD97_AJ291278
4a
4c_GA_2007.042_OI145401_JN642983
4c_CM_1998.5_98CM1363_AY256811
4c_GA_2007.042_ES62284_JN642791
4a_CM_2001.417_M7284N_AY685027
4c_za_2010.583_ZADGM3460_HQ385898
4c_CF_2006.25_725_FJ791093
4c_PK_2006.83_HCC56_AB444480
4c_CF_2006.25_709_FJ791090
4c_CF_2006.25_821_FJ791116
4c_CF_2006.25_441_FJ791083
4c_CF_2006.25_317_FJ791103
4c_CF_2006.25_172_FJ791069
4c_CF_2006.25_505_FJ791085
4c_CG_2005.5_TK371_GU088138
4c_CF_2006.25_163_FJ791068
4c_GA_2007.042_OL113816_JN642951
4c_CA_2004.125_QC183_EF116123
4c_GA_2005.0835_Gab12389_EU334534
4c_GA_2003.083_D423B_AM177501
0.84
4c_CD_2007.5_DRC1400_KF826152
4c_CD_2007.5_DRC1755_KF826154
C2
0.7
4c_CD_2007.5_DRC2239_KF826161
4c_CD_2007.5_DRC0742_KF826151
4c_CD_2007.5_DRC2143_KF826159
4c_CD_2007.5_DRC0103_KF826179
4c_GA_2007.042_OI103007_JN642739
4c_GA_1992.5_GB215_L29605
4c_CD_2007.5_DRC0327_KC506767
4c_GA_2007.042_NY88508_JN642753
4c_GA_2007.042_HO18015_JN642726
4c_GA_2003.083_D856_AM888302
4c_FR_2001.75_G4MP079_AY743182
4c_FR_2001.75_G4MP080_AY743160
4c_GA_2007.042_HO15910_JN642724
4c_GA_2007.042_OL110610_JN642770
4c_GA_2007.042_HO15906_JN642723
4c_GA_2007.042_HO27701_JN642731
4c_CD_2007.5_DRC3967_KF826174
4C_CD_2004.917_QC381_FJ462436
4c_GA_2007.042_HO24321_JN642823
4c_CA_2004.125_QC137_EF116116
4c_CG_2005.5_BR014_GU088136
4c_GA_2007.042_OL109803_JN642941
4c_CG_2005.5_KG138_GU088171
4c_GA_1992.5_GB48_L29614
4c_GA_2009.083_D147002_FJ755940
4c_GA_2007.042_OI150201_JN642742
4c_GA_2001.75_Gab1148_EU334532
4c_GA_1992.5_GB358_L29607
4c_GA_2003.083_D818_AM888299
4c_GA_2005.0835_GB116_L29602
4c_GA_2007.042_HO18012_JN642725
4c_GA_2007.042_OL160815_JN642962
4c_GA_2007.042_OL112111_JN642948
4c_GA_2007.042_HO21206_JN642728
4c_GA_2007.042_MO370196_JN642745
4c_GA_2007.042_HO14421_JN642807
4c_GA_2007.042_OI102112_JN642738
4c_CF_2006.25_486_FJ791106
4c_GA_2007.042_HO13208_JN642805
4c_CD_2007.5_DRC0376_KF826181
0.57
4c_CD_2007.5_DRC2141_KC506770
4c_CD_2007.5_DRC2324_KF826162
C1
4c_CD_2007.5_DRC2429_KF826163
0.57
4c_CD_2007.5_DRC2148_KF826160
4c_CD_2007.5_DRC3231_KF826166
(4x)4c_CD_2007.5_DRC1427_KF826153
0.95
4c_CD_2007.5_DRC0318_KF826180
4c_CD_2007.5_DRC4278_KF826175
4c_CD_2007.5_DRC0734_KF826150
4c_CD_2007.5_DRC2639_KF826165
4c_CG_2005.5_KG206_GU088167
4c_CF_2006.25_257_FJ791072
4c_CF_2006.25_204_FJ791070
4c_CF_2006.25_508_FJ791086
4c_CF_2006.25_510_FJ791107
4c_za_2007.167_ZADGM7684_HQ385905
4c_za_2007.5_ZADGM225_JN116568
4c_CG_2005.5_BR150_GU088134
4c_CG_2005.5_TK239_GU088137
4c_CG_2005.5_TK240_GU088139
0.83
4c_CD_2007.5_DRC2518_KC506775
4c_CD_2007.5_DRC0615_KF826183
4c


4e_GA_2007.042_NG69504_JN642876
4e_GA_2007.042_HO14402_JN642806
4e_GA_2003.083_D674_AM177505
4e_GA_2007.042_MO454427_JN642853
0.07
4e_GA_2007.042_OI102120_JN642973
4e_GA_2007.042_ES67299_JN642792
4e_GA_2007.042_OL113844_JN642958
4e_GA_2007.042_HO21207_JN642818
4e_GA_2007.042_OL108413_JN642769
4e_GA_2007.042_OI103024_JN642981
4e_GA_2007.042_HO29007_JN642828
4e_GA_2007.042_HO21211_JN642819
4e_GA_2007.042_NG62212_JN642871
4e_GA_2003.083_D257_AM888294
4e_GA_2003.083_D309_AM888295
4e_GA_2007.042_NY88710_JN642755
4e_GA_2007.042_OL113841_JN642956
4e_GA_2007.042_NY88503_JN642888
4e_GA_2007.042_MO555326_JN642859
4e_GA_2003.083_D315_AM888296
4e_GA_2009.083_D67.001_FJ755938
4e_GA_2007.042_OL113843_JN642957
(4b)4e_GA_2007.042_HO26316_JN642730
4e_GA_2007.042_MO457463_JN642749
4e_GA_2007.042_NG77521_JN642884
4e_GA_2007.042_OI103022_JN642741
4e_GA_2007.042_HO17903_JN642810
4e_GA_2007.042_ES7641_JN642794
4e_GA_2007.042_ES53219_JN642786
4e_GA_2007.042_WN219117_JN642932
4e_GA_2007.042_MO30717_JN642836
4e_GA_2007.042_MO299627_JN642833
4e_GA_2007.042_MO29577_JN642832
4e_GA_2007.042_ES45204_JN642783
4e_GA_2007.042_MO395214_JN642746
4e_GA_2007.042_ES45202_JN642782
4e_GA_2007.042_NY88601_JN642891
4e_GA_2007.042_NY84009_JN642886
4e_GA_2007.042_MO399142_JN642851
4e_GA_2007.042_NG59810_JN642863
4e_GA_2007.042_MO299638_JN642834
4e_GA_2007.042_MO399125_JN642849
4e_GA_2003.083_D536EF_AM177500
4e_GA_2003.083_D273A_AM177502
4e_GA_2003.083_D481CD_AM177508
4e_GA_2003.083_D99_AM888293
4e_GA_2007.042_MO384174_JN642844
4e_GA_2007.042_OL113835_JN642955
4e_GA_2007.042_OM164919_JN642914
4e_GA_2007.042_OM164309_JN642912
4e_GA_2007.042_ES8668_JN642797
4e_GA_2007.042_OL113812_JN642949
4e_GA_2007.042_HO29124_JN642831
4e_GA_2007.042_NY88610_JN642892
4e_GA_2007.042_NY93508_JN642904
4e_GA_2007.042_MO382647_JN642842
4e_GA_2007.042_NG77501_JN642883
4e_GA_2007.042_NG73009_JN642735
4e_GA_2007.042_ES48235_JN642784
4e_GA_2003.083_D59CD_AM177507
4e_GA_2007.042_MO555252_JN642751
4e_GA_2007.042_OI101909_JN642965
4e_GA_2003.083_D816_AM888298
4e_GA_2007.042_OL113848_JN642959
4e_GA_2005.0835_Gab770_EU334531
4e_GA_2005.0835_Gab12279_EU334533
4e_GA_2007.042_NY93106_JN642901
4e_GA_2007.042_OL110617_JN642944
4e_GA_2007.042_MO399134_JN642850
4e_GA_2007.042_NY88509_JN642754
4e_GA_2007.042_NG68908_JN642873
4e_GA_2007.042_WN172512_JN642764
4e_GA_2007.042_WN213819_JN642931
4e_CM_1994.5_CAM600_L29590
4e_GA_2009.083_D20.001_FJ755934
4e_GA_2005.0835_Gab704_EU334530
4e_GA_2007.042_NY90804_JN642899
4e_GA_2007.042_NY93808_JN642760
4e_GA_1992.5_GB809_L29627
4e_GA_2007.042_NG59809_JN642862
4e_GA_2007.042_OL112106_JN642772
4e_GA_2007.042_NY93101_JN642900
4e_GA_2007.042_ES7856_JN642795
4e_CM_2001.417_M4703N_AY265435
4e_GA_2007.042_MO389115_JN642845
(4c)4e_GA_2007.042_OL113824_JN642775
4e_GA_2007.042_NY90401_JN642898
4e_GA_2007.042_NG77529_JN642737
4e_GA_2007.042_NY94312_JN642907
4e_GA_2007.042_NG69501_JN642874
4e_GA_2007.042_NG70319_JN642877
4e_GA_2007.042_NG69503_JN642875
4e_GA_2003.083_D887_AM888303
4e_GA_2003.083_D898_AM888304
4e_GA_2007.042_WN1787177_JN642766
4e_GA_2007.042_HO19801_JN642814
4e_GA_2007.042_OL113826_JN642776
4e_GA_2007.042_NY93504_JN642903
4e_GA_2007.042_NG60708_JN642868
4e_GA_2007.042_WN213610_JN642930
4e_GA_2003.083_D646G_AM177504
4e_GA_2007.042_NG57103_JN642861
4e_GA_2007.042_NY84519_JN642752
4e_GA_2007.042_OL110612_JN642771
4e_GA_2007.042_HO26309_JN642825
4e_GA_2007.042_OI102107_JN642968
4e_GA_2007.042_OL113853_JN642960
4e_GA_2009.083_D6001_FJ755932
4e_GA_2003.083_D396_AM888297
4e_GA_2007.042_OI102119_JN642972
4e_GA_2007.042_ES56259_JN642788
4e_GA_2007.042_HO21216_JN642820
4e_GA_2007.042_HO18011_JN642811
4e_GA_2003.083_D278A_AM177503
4e_GA_2007.042_HO19804_JN642815
4e_GA_2007.042_OL113813_JN642950
4e_CA_2004.125_QC163_EF116121
4e_GA_2007.042_NY94705_JN642909
4e_GA_2007.042_OI103014_JN642977
4e_GA_2007.042_MO399124_JN642747
4e_GA_2007.042_OI103026_JN642982
4e_GA_2007.042_ES8812_JN642721
4e_GA_2007.042_NY88713_JN642895
4e_GA_2007.042_NG60513_JN642865
4e_GA_2007.042_NG60511_JN642864
4e_GA_2007.042_NY88619_JN642893
4e_GA_2007.042_OL113831_JN642954
4e_GA_2007.042_WN226204_JN642935
4e_GA_2007.042_ES62270_JN642790
4e_GA_2007.042_OI103006_JN642975
4e_GA_2007.042_HO29016_JN642732
4e_GA_2007.042_OI102116_JN642971
4e_GA_2007.042_NY94704_JN642908
4e_GA_2007.042_NG60601_JN642733
4e_GA_2007.042_NG60620_JN642867
4e_GA_2007.042_OI102104_JN642967
4e_GA_2007.042_MO389116_JN642846
4e_GA_2007.042_OM164906_JN642763
4e_GA_2007.042_OI101901_JN642964
4e_GA_2007.042_NG75722_JN642882
4e_GA_2007.042_NG75721_JN642881
4e_GA_2007.042_OL112104_JN642945
4e_GA_2007.042_OI103018_JN642978
4e_GA_2007.042_OI102109_JN642969
4e
0.02