HIV-1 in the United Kingdom: population dynamics and genetic evolution

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by Geraldine Marie Foster

January 2014

Contents

Title **Page**

List of Tables and Figures

Title Page

Chapter 1: Introduction

Chapter 3: Contribution of viral recombination to the UK HIV-1 epidemic

Title Page

Chapter 4: Identification of potentially recombinant HIV-1 sequences and development of a near full-length single genome sequencing protocol

Chapter 5: Identification of CRF50_A1D

*** On CD attached to back cover**

Acknowledgements

I would like to express my thanks to the following people who made the writing of this thesis possible:

- My primary supervisor, Professor Anna Maria Geretti, who supported me throughout the project in two different locations;
- Professor Vince Emery, my second supervisor during my time at UCL;
- Professor Neil Hall, my second supervisor during my time at UoL;
- Professor Andrew Leigh Brown and Dr. Stephane Huè for phylogenetics support and advice;
- Colleagues at the UK HIV Drug Resistance Database, particularly David Dunn and Esther Fearnhill, who de-linked aggregated clinic data at source during the process of identifying additional CRF50_A1D cases;
- Professor Caroline Sabin at UK CHIC, who de-linked and provided CD4 cell count data for anonymised CRF50_A1D patients;
- Mr. John Ambrose at the University of Liverpool, who wrote the Perl script used to measure maximum genetic distance over all branches of an approximate maximum likelihood tree.

Abstract

HIV-1 in the United Kingdom: population dynamics and genetic evolution

Background: Phylogenetic characterisation of local HIV-1 epidemics can be used to understand emerging transmission networks. We analysed subtype-unassigned sequences in the UK HIV Drug Resistance Database (UK HIV DRD) to map the emergence and origin of genetically diverse recombinant strains using phylogenetic analyses, near full-length sequencing of plasma HIV-1 and CD4 cell count decline.

Methods: 55,556 genotyped *pol* sequences (protease amino acids 1-99, RT amino acids 1-234) were analysed for evidence of recombination. Near full-length single genome sequencing of plasma HIV-1 was performed on stored specimens from six patients with a putative A1/D novel recombinant strain. Recombination and phylogenetic analyses were performed using RIP, PhyML, jpHMM, Simplot, SCUEAL and FastTree v2.3.1. Evolutionary analysis of putative novel recombinants was performed using Bayesian Evolutionary Analysis by Sampling Trees (BEAST). Geographic screening for additional instances of recombinant structures was performed using HIV BLAST and the Los Alamos National HIV database. Demographic data were available for 50/72 patients. CD4 cell count decline was assessed using linear regression.

Results: The proportion of subtype-unassigned HIV-1 strains in the UK HIV DRD increased significantly during 1999-2008 (p=<0.01). 2,030 putative B-recombinant sequences were analysed for evidence of transmission of novel recombinant strains; 15 novel recombinant clusters comprising 94 individuals were identified. The proportion of intravenous drug users (IVDU), males and people of white ethnicity was significantly higher among novel recombinant clusters than among people infected with pure subtypes and recognised CRFs. Geographic screening showed co-circulation of novel strains in seven countries over three continents and import and export of novel strains from the UK was identified.

Near full-length sequencing of six plasma specimens showed five patients sharing an identical A1/D strain; this was registered as CRF50_A1D and 67 further instances in the UK HIV DRD were identified. Geographic analysis showed close relation of component subtype A1 and D regions to East African strains; monophyletic clustering indicated a single introduction of this variant into the UK. Time scaled analysis showed a time to most recent common ancestor of approximately 1992. Demographic data showed the earliest CRF50 transmissions were confined to men who have sex with men (MSM), with subsequent transmissions to heterosexuals and IVDUs. Analysis of the sixth, complex, sequence demonstrated onward recombination of CRF50 with a subtype B strain (median divergence year 2000). CD4 cell count analysis suggested infections with CRF50 progressed in a similar manner to subtype B infections.

Conclusions: Significant increasing genetic diversity of HIV-1 in the UK is linked to both UK and international transmissions among multiple exposure routes. These findings highlight the changing dynamics of HIV transmission in the UK and the converging of the two previously distinct MSM and heterosexual epidemics.

List of abbreviations

AIC	Akaike Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
APOBEC3	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3
ART	Antiretroviral Therapy
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
BP	Base Pairs
CAR	Central African Republic
cDNA	Complementary DNA
CCR5	C-C Chemokine Receptor Type 5
CHAVI	Centre for HIV/AIDS Vaccine Immunology
CI	Confidence Interval
CRF	Circulating Recombinant Form
CXCR4	C-X-C Chemokine Receptor Type 4
DRC	Democratic Republic of Congo
ESS	Effective Sample Size
GTR	Generalised Time Reversible
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
HKY	Hasegawa, Kishino and Yano
HLA	Human Leukocyte Antigen
HPA	Health Protection Agency
HPD	Highest Posterior Density
HTA	Heteroduplex Assay
IVDU	Intravenous Drug User
jpHMM	Jumping Profile Hidden Markov Model
LANL	Los Alamos National HIV Database
LTR	Long Terminal Repeat
MCMC	Markov Chain Monte Carlo
MRC-CTU	Medical Research Council Clinical Trials Unit
MSM	Men who have Sex with Men
NFL-SGS	Near Full-Length Single Genome Sequencing
NRTI	Nucleos(t)ide Reverse Transcriptase Inhibitor
NNRTI	Non-nucleoside Reverse Transcriptase Inhibitor
NTC	No Template Control
PBMC	Peripheral Blood Mononuclear Cell
PCR	Polymerase Chain Reaction
PhyML	Phylogenetic Estimation Using Maximum Likelihood
PI	Protease Inhibitor
PR	Protease
Ptt	*Pan troglodytes troglodytes*
RDT	Recombinant HIV-1 Drawing Tool
RIP	Recombinant Identification Program
RT	Reverse Transcriptase
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SCUEAL	Subtype Classification Using Evolutionary Algorithms
SGS	Single Genome Sequencing

SIV	Simian Immunodeficiency Virus
SIV_{CPZ}	Simian Immunodeficiency Virus (Chimpanzee)
SIV_{GOR}	Simian Immunodeficiency Virus (Gorilla)
tMRCA	Time to Most Recent Common Ancestor
TRIM5α	Tripartite Motif 5 Alpha
UK	United Kingdom
UK CHIC	UK Collaborative HIV Cohort Study
UK HIV DRD	UK HIV Drug Resistance Database
URF	Unique Recombinant Form
vRNA	Viral RNA

List of papers and presentations associated with this project

Publications

Foster, G.M., Ambrose, J.C., Hué, S., Delpech, V.C., Fearnhill, E., Abecasis, A.B., Leigh Brown, A.J., Geretti, A.M., and on behalf of the UK HIV Drug Resistance Database (2014). Novel HIV-1 Recombinants Spreading across Multiple Risk Groups in the United Kingdom: The Identification and Phylogeography of Circulating Recombinant Form (CRF) 50_A1D. PLoS ONE *9*, e83337.

Presentations

Foster, G.M., Transmission Dynamics of Novel HIV-1 Recombinants in the United Kingdom. Oral Presentation: ARV/HIV Case Discussion Merseyside BASHH Meeting, September 2013.

Foster, G.M., Novel HIV-1 recombinants spreading among risk groups in the United Kingdom: the identification and phylogeography of circulating recombinant form (CRF)50_A1D. Oral presentation: Institute of Infection and Global Health Day, University of Liverpool, 2012.

Foster, G.M., Danger looming on the horizon: the changing face of HIV. Oral Presentation: Institute of Infection and Global Health Staff Retreat, 2012.

Foster, G.M., CRF50_A1D is spreading from the men who have sex with men community. Oral presentation: 19th International HIV Dynamics and Evolution, 2012.

Foster, G.M., Ambrose, J.C., Conibear, T.C.R, Fearnhill, E., Abecasis, A.B., Asboe, D., Mackie, N E., Ustinowski, A., Leigh Brown, A.J, Geretti, A.M on behalf of the UK HIV Drug Resistance Database. The origin and distribution of CRF50_A1D. Poster presentation: 25th International Workshop on HIV & Hepatitis Virus: Drug Resistance and Curative Strategies, 2011.

Foster, G.M., Ambrose, J.C., Conibear, T.C.R, Fearnhill, E., Abecasis, A.B., Asboe, D., Mackie, N.E., Leigh Brown, A.J, Geretti, A.M on behalf of the UK HIV Drug Resistance Database. First description of a novel HIV A1/D recombinant circulating in men who have sex with men in the UK. Poster presentation: 18th International HIV Dynamics & Evolution, 2011.

Foster , G.M., Ambrose, J.C, Hué, S., Conibear, T.C.R, Fearnhill, E., Abecasis, A. B, Leigh Brown, A.J, Geretti, A.M on behalf of the UK HIV Drug Resistance Database. Mapping the Circulation of HIV-1 Recombinants among Men who have Sex with Men. Poster Presentation: CROI 2011.

Foster G.M., Mapping the emergence of novel HIV-1 recombinants by full genome analysis. Oral presentation:14th Annual Resistance Meeting, 2010.

Foster, G.M., Tracing the HIV-1 epidemic through near full-length genome sequencing. Oral Presentation: Divisional Postgraduate Colloquium, UCL, 2010

Chapter 1: Introduction

1.1 The HIV virus and how it causes infection

Human Immunodeficiency Virus (HIV), a lentivirus from the *Retroviridae* family, is a diploid, positive sense, single stranded RNA virus (Baltimore classification VI) (Baltimore, 1971; Jetzt et al., 2000). HIV is divided into two types: HIV-1 and HIV-2. Of these, HIV-1 is more widespread globally than HIV-2, which is mainly found in West Africa, and causes a less severe disease (Marlink et al., 1994). HIV-1 is further subdivided into four groups: M, N, O and P (Keele et al., 2006; Plantier et al., 2009; Sharp et al., 2005). This review and the following investigation is primarily concerned with HIV-1 group M viruses, although other virus types within HIV are occasionally relevant and mentioned.

The RNA genome of HIV consists of 9749 nucleotides and comprises three structural genes – *gag, pol* and *env* – and several non-structural genes including *tat, rev, nef, vpr* and *vpu* (*vpx* in HIV-2). Of the structural genes, *gag* codes for the *gag* polyprotein which is involved in viral encapsidation, assembly and budding of viral particles, and the organisation of the envelope protein on the virion surface (reviewed in Hill et al., 1996). The *pol*, or polymerase, gene codes for the viral enzymes reverse transcriptase (RT), RNase H, integrase and protease; the *env* gene codes for the precursor protein gp160 which undergoes cleavage into two proteins which are active in viral entry: gp120, which binds to the CD4 receptor, and gp41, which is the fusion fragment (Allan et al., 1985; Frey et al., 2010).

Of the non-structural genes, *vif* enhances the infectivity of viral particles and is essential for productive HIV-1 infection of natural target cells, and *vpu* is active in viral release (Gupta et al., 2009; Hinz et al., 2010; Wieland et al., 1997). *Tat* and *rev* are accessory genes which are unique to lentiviruses and are believed to be responsible for these viruses' ability to maintain a chronic persistent infection even in the presence of a pronounced host immune response (Ranki et al., 1994). *Nef* plays a role in virus infectivity. The long terminal repeat regions (LTR) located at each end of the genome have regulatory functions (Kent et al., 2001).

Entry of HIV-1 into target cells is mediated through binding of the gp120 envelope glycoprotein to the CD4 receptor and a chemokine co-receptor, which is commonly CCR5 on macrophages and a subset of memory CD4 T-cells, and CXCR4 on many cell types including CD4 T-cells and macrophages (Esté and Telenti, 2007). Sequential binding of the gp120 to the CD4 receptor and co-receptor (CCR5 or

CXCR4) induces conformational changes which trigger the dissociation of gp120 and a refolding of the gp41 protein (Frey et al., 2010; Harrison, 2008). This irreversible refolding of gp41 triggers the fusion of viral and target cell membranes (Frey et al., 2010).

Following fusion, the viral core enters the cytoplasm and the viral RNA (vRNA) acts as the template for double stranded proviral DNA synthesis by the virus-encoded reverse transcriptase (RT) enzyme. The DNA migrates to the nucleus, where the viral integrase mediates integration into the host cell genome. The initial transcription and translation of mRNA and proteins within activated cells is then governed by host cell functions (Hinz et al., 2010).

The process of viral entry, expression, assembly and budding can be disrupted by a number of host restriction factors, notably the apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3 (APOBEC3) family of proteins that delaminate C to U residues during transcription; tripartite motif 5 alpha (TRIM5α), which interferes with viral uncoating and blocks reverse transcription and subsequent transport to the nucleus, and tetherin, an antiviral protein which blocks virus release (Gupta and Towers, 2009; Huthoff et al., 2009; Jern et al., 2009; McNatt et al., 2009; Perez-Caballero et al., 2005; Sauter et al., 2009; Stremlau et al., 2005). These selected restriction factors highlight the complexity of host-virus interactions in HIV infections, however, many other such factors exist that govern this process in individuals.

Individual host genetics also affect how readily HIV can cause and prolong infection, particularly Human Leukocyte Antigen (HLA). Individuals with HLA type B-27 or B-57 tend to have infections that progress more slowly than individuals with different HLA profiles. An association has been described between protective HLA alleles and lower viral replication capacities, suggesting that the targeting of HLA responses at functionally important epitopes in *gag* results in escape variants that have undergone a fitness cost (Wright et al., 2010). The TL9 epitope in *gag* (p24) is an immunodominant epitope which is restricted by HLA B-7 supertype alleles, of which HLA type B*81:01 appears to exert the greatest selective pressure on the virus, at least in the first 18 months post-infection (Ntale et al., 2012). Other host genetic factors that can act to restrict HIV-1 are CCR5 deletions and KIR receptors (Mothe et al., 2009).

Finally, the HIV-1 infection and replication process can also be disrupted by antiretroviral agents, of which the main classes are: reverse transcriptase inhibitors (nucleos(t)ide reverse transcriptase inhibitors (NRTI) and non-nucleoside reverse transcriptase inhibitors (NNRTI)), which prevent synthesis of double stranded viral DNA; protease inhibitors (PI), which prevent the cleavage of protein precursors; entry inhibitors, which target the 'prehairpin intermediate' conformation of gp41; integrase inhibitors, which block the integration of linear viral cDNA, and CCR5 antagonists which inhibit viral binding to CCR5-expressing cells (Buzón et al., 2010; Esté and Telenti, 2007; Frey et al., 2010; Harrison, 2008). The widespread use of highly active antiretroviral therapy (HAART) in the treatment of HIV-1 has led to the emergence of antiretroviral resistance to classes of antiretroviral agents (Tang and Pillay, 2004), which has had consequences for both individuals and populations.

1.2 In-host viral diversity in HIV infections

Transmission of HIV-1 usually results from virus exposure at mucosal surfaces followed by viral replication in submucosal and loco-regional lymphoid tissues subsequent to overt systemic infection (Salazar-Gonzalez et al., 2009). In its early phase after transmission, an HIV-1 infection is characterised by extreme levels of viral replication and a high peak viraemia, coupled with massive depletion of CD4 T-cells in the gastrointestinal tract. As the infection moves from the acute into chronic phase, viraemia is lowered by 2-4 logs, which is accompanied by progressive loss of CD4 cells and eventual progression to AIDS (reviewed in (Fischer et al., 2012)).

Both humoral and cell-mediated immune responses are mounted to HIV-1 infection; however, these are eventually evaded by HIV-1 replication and adaptation to the host environment (Jost and Altfeld, 2012; Lemey et al., 2007; Wei et al., 2003). The lowering of viraemia by 2-4 logs during the move from the acute to the chronic stage of infection is primarily due to host immune responses such as the production of neutralising antibodies and the cytotoxic T cell response (Karlsson et al., 2007; Wei et al., 2003). Wei noted that the viral inhibitory activity of neutralising antibodies in early HIV-1 infection resulted in a complete replacement of neutralising antibody-sensitive virus with populations of neutralising antibody-resistant virus (2003).

Although neutralising antibodies impose a strong selective pressure on the *env* gene of HIV-1, they do not control viral replication; partial control of HIV-1 replication (and therefore a lowering of HIV-1 viral load) is associated with the appearance of HIV-1-specific cytotoxic T cell responses (Koup et al., 1994; Lemey et al., 2007). Natural killer cells expressing the KIR_3DS_1 receptor can also inhibit HIV-1 replication in vitro; however, HIV-1 accessory proteins seem to have evolved in particular to counteract NK cell responses (reviewed in Jost and Altfeld, 2012). The various host immune responses to HIV-1 infection mean that during the nonsymptomatic phase preceding progression to AIDS, there is great variation in the level of circulating virus in the plasma (Fellay et al., 2007).

Viral population fluctuations are recurrent in HIV-1 natural infections, and they have important consequences for viral evolution (Lorenzo-Redondo et al., 2011). When only one variant establishes productive infection (the founder virus), this produces a population bottleneck. During the massive population expansion that subsequently occurs, a swarm of mutants is generated. This swarm has been described as comprising a viral quasispecies, a concept first introduced by Eigen and Schuster in 1979 (Eigen and Schuster, 1979; Lauring and Andino, 2010; Lorenzo-Redondo et al., 2011); however, it should be acknowledged that this definition is not universally accepted, and that observation of intra-patient genetic variation is not necessarily sufficient evidence to demonstrate quasispecies behaviour (reviewed in Holmes, 2010).

The extreme levels of genetic diversity displayed in HIV infections are in keeping with other lentiviruses (Artenstein et al., 1995; Buonaguro et al., 2007; Geretti, 2006; Jetzt et al., 2000; Keele and Derdeyn, 2009; Robertson et al., 1995; Shriner et al., 2004; Vessière et al., 2010). This in-host viral diversity is due to: a) the error-prone nature of the RT enzyme, which lacks proofreading ability; this leads to point mutations, insertions and deletions in progeny genomes, with an estimated error rate of ~$3x10^{-5}$ per nucleotide base per cycle of replication; b) high rates of replication, with ~10^{10} virus particles produced daily; c) the mutagenic activity of APOBEC3 and other host factors; d) a high tendency to genetic recombination due to the ability of RT to switch templates during reverse transcription; e) immune pressure; f) repair mechanisms (Allen et al., 2000; Archer et al., 2008; Huthoff et al., 2009; Jetzt et al., 2000; Karlsson et al., 2007; Lemey et al., 2007; Salazar-Gonzalez et al., 2009).

a) The error prone nature of the RT enzyme

The error-prone nature of the HIV-1 RT enzyme has been described several times; in addition, there are also several other sources of retroelement mutation: RNA polymerase II errors, RNA editing, and spontaneous decay of RNA or DNA (Bebenek et al., 1989; Holland et al., 1991; Preston et al., 1988; reviewed in Preston and Dougherty, 1996). Since then, Lauring and Andino proposed that the intrinsic error rate of the viral replicase determines the mutation rate for the virus and therefore the range of genetic variation on which natural selection can act (2010). Of the errors that occur in progeny genomes, G-to-A mutations (not APOBEC-mediated) are the most common RT-mediated error (Jern et al., 2009).

b) High rates of viral replication

HIV-1 has highly productive replication *in vivo*, with, on average, a turnover of half of the plasma virions every two days (Ho et al., 1995) and a mean replication rate of 0.68+/- 0.13x10^9 virions per day. Nearly the entirety of the circulating plasma virus comes from recently infected cells (Ho et al., 1995).

c) Mutagenic activity of host factors

In vivo viral evolution can also be mediated by host factors. APOBEC proteins can induce lethal mutagenesis of HIV through widespread deamination of the HIV genome during reverse transcription (Jern et al., 2009; Lauring and Andino, 2010). The fact that HIV replicates close to the error threshold makes it particularly sensitive to slight increases in mutational load (Lauring and Andino, 2010); however, it should be noted that this may be solely a function of the high replication rates of HIV-1, rather than HIV-1 replicating closer to the error threshold than other organisms.

Low levels of APOBEC3-mediated mutagenesis have also affected long-term, inter-host HIV-1 evolution, leading to enhanced rates of variation at sites (Jern et al., 2009).

d) Genetic recombination due to RT template switching

Genetic recombination is a source of viral diversity for many RNA viruses (Beemon et al., 1974; Coffin, 1979; Hu and Temin, 1990). As a retrovirus, HIV contains two copies of its RNA genome in each virion (Hu and Temin, 1990). Viral diversity in RNA viruses as a result of homologous recombination as opposed to reassortment of genome segments has been described since the early 1970s (Beemon et al., 1974), however, the precise method of retroviral recombination was initially unknown. The two main theories concerning mechanisms of retroviral recombination were the copy choice model and the strand displacement model (Coffin, 1979; Skalka et al., 1982). The copy choice model proposed that only one provirus would result from each cellular infection event, whilst the strand displacement model postulated that two copies of DNA would be produced from each virion, and that recombination would occur during plus-strand DNA synthesis due to displacement of DNA and assimilation into the DNA synthesised by the second template (Coffin, 1979; Junghans et al., 1982). Hu and Temin showed in 1990 that single recombinant proviruses were the progeny of heterozygous virions, supporting the copy choice model.

The copy-choice model originally proposed that the RT enzyme switched templates when it encountered breaks in RNA, however this was broadened to include recombination occurring during minus-strand DNA synthesis without the requirement of breaks in viral RNA. The low processivity of RT causes the enzyme to dissociate from the template, allowing the short DNA-RNA hybrid to be disrupted so that the growing DNA strand is displaced to the other RNA template (Huber et al., 1989; Jetzt et al., 2000).

The estimated mean rate of HIV-1 recombination *in vivo* is 1.38 x 10^{-4} recombination events/adjacent sites/generation, which is 5.5 fold greater than the reported point mutation rate of 2.5x10^{-5}/site/generation (Shriner et al., 2004). This recombination between RNA strands mainly occurs during minus-strand DNA synthesis (Jetzt et al., 2000; Zhuang et al., 2002). Unlike other aspects of HIV-1 replication, recombination is not error-prone (Zhang and Temin, 1994), and the combination of polymorphisms into a new genome in a single round of replication enables viruses to rapidly access a greater sequence space than is possible by the stepwise accumulation of point mutations (Simon-Loriere et al., 2010). The net effect of this is to facilitate the combination of advantageous mutations within highly fit genomes and the removal of deleterious mutations from the viral population (Galli et al., 2010; Simon-Loriere et al., 2010). The effect of recombination on viral evolution within an HIV-1 infected individual is of the same order of magnitude as point mutational change and is therefore a major evolutionary force (Shriner et al., 2004); however, it is important to note that, unlike point mutational change, recombination does not necessarily result in increased genetic diversity within an infected individual.

e) Immune pressure

Early cytotoxic T cell responses have been shown to influence genetic variation in HIV-1 infections (Allen et al., 2000; Karlsson et al., 2007; Salazar-Gonzalez et al., 2009). Karlsson found that HIV-1 undergoes a continuous dynamic development of cytotoxic T cell responses that was associated with viral escape and increased in-host viral diversity (2007). Cytotoxic T cell responses tend to target epitopes in the viral *gag* and *nef* genes (Lemey et al., 2007). HIV-1 is also able to evade neutralising antibody responses by accumulating multiple amino acid changes, particularly in the hypervariable regions of *env* (Lemey et al., 2007).

f) Repair mechanisms

In vitro, HIV-1 has been shown to undergo fitness recovery of debilitated clones following multiple large-population passages (Lorenzo-Redondo et al., 2010). Following 20 passages of debilitated viral clones, Lorenzo-Redondo observed that 24 mutations were fixed at passage 11 (with a dominance of synonymous changes), but 79 mutations were fixed after passage 21 (with a dominance of non-synonymous changes), indicating that fitness recovery is a multistage process in which non-synonymous mutations are indicative of the influence of selective pressures and generate increasing viral heterogeneity (2010).

HIV-1 also exploits host cellular DNA repair mechanisms, triggering the DNA repair machinery of infected cells during the integration process (Smith and Daniel, 2006).

1.3 HIV recombination and consequences for population-level genetic diversity

A critical feature of recombination as an evolutionary force is that of allowing reshuffling of genetic information carried by distantly related viral strains (Simon-Loriere et al., 2010). The existence of an HIV infection is not protective against super- or co-infection with another HIV strain (Artenstein et al., 1995), and it is possible for a cell to become infected with two different viruses and for a plus-strand RNA from each virus to be packaged into a new virion (Harris et al., 2002; Jetzt et al., 2000). Although the rate of recombination between two partially identical molecules is less than that between two identical RNA molecules (Zhang and Temin, 1994), recombination between highly divergent viral strains is possible (Robertson et al. 1995). Therefore, should the above scenario occur, the strand switching activity of the RT can create intra- or intersubtype or inter-group recombinants (Harris et al., 2002; Jetzt et al., 2000; Simon-Loriere et al., 2009).

In order to understand how the ready generation of intra- and inter-subtype recombinants effects both the properties that lead to one virus variant predominating over others, and overall trends in the HIV-1 epidemic, an understanding of the emergence of HIV subtypes and recombinants is essential (Balotta et al., 2001; Holguín et al., 2008).

HIV-1 infections are now accepted to have arisen from at least four separate cross-species transmissions of Simian Immunodeficiency Virus (SIV) from primates to humans: two involving transmission from *Pan troglodytes troglodytes (Ptt)* of SIV chimpanzee (SIV$_{cpz}$Ptt) in the early 20th Century, creating HIV groups M and N (Keele et al., 2006; Sharp et al., 2005; Vidal et al., 2000), and one, HIV group P, from a cross-species transmission involving gorillas (SIV$_{gor}$) (Plantier et al., 2009); although HIV group O was traditionally considered to have arisen following transmission of SIV$_{cpz}$, emerging evidence suggests this group is more closely related to SIV$_{gor}$ (D'arc et al., CROI 2014, Abstract 51; Figure 1_1). The origins of HIV groups M and N have been defined using phylogeography, which identified two regions in Cameroon that appear to contain the wild *Ptt* populations infected with the source SIV viruses (Keele et al., 2006). Groups O and N have remained largely confined to Cameroon and Central Africa, but Group M has not (Lemey et al., 2004). Although the precise reasons why HIV-1 group M is solely responsible for the global pandemic have not been elucidated, it has been proposed that it is the ability, restricted to group M viruses, to use the *vpu* gene to antagonise the host cellular antiviral factor tetherin that has allowed this (Gupta and Towers, 2009).

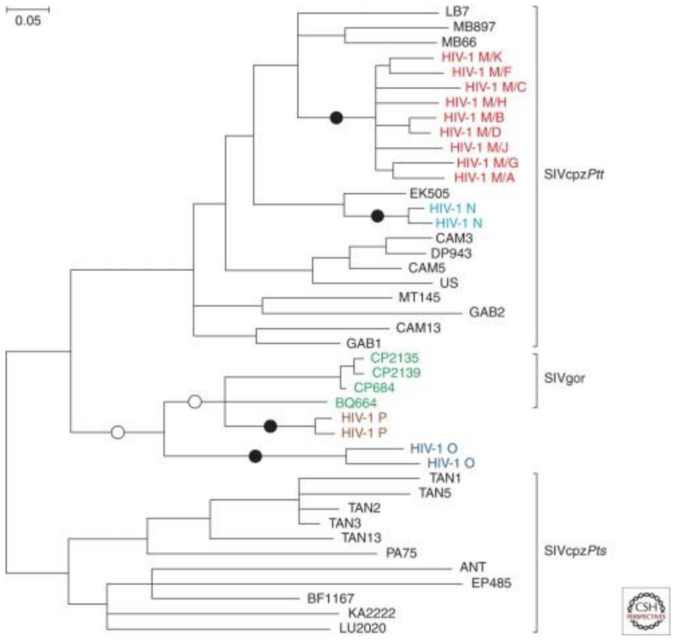


Figure 1_1. HIV origins. Maximum likelihood tree displaying the phylogenetic relationships of SIV$_{cpz}$, HIV-1, and SIV$_{gor}$ strains for a region of the *pol* gene (HXB2 coordinates 3887–4778). SIV$_{cpz}$ sequences are shown in black, SIV$_{gor}$ sequences are shown in green, HIV-1 group M is shown in red, HIV-1 group N in light blue, HIV-1 group O in dark blue and HIV-1 group P in brown, respectively. Black circles indicate the four branches where cross-species transmission-to-humans has occurred; white circles indicate two possible alternative branches on which chimpanzee-to-gorilla transmission occurred. The close relationship of HIV-1 group O sequences to HIV-1 group P sequences adds weight to recent evidence that HIV-1 group O also resulted from a cross-species transmission from gorillas. The scale bar shows 0.05 nucleotide substitutions per site. Figure taken from Sharp and Hahn, 2011.

From Cameroon, HIV-1 group M (via infected individuals) travelled south to Kinshasa in the Democratic Republic of Congo (DRC), where the growth of sub-Saharan African cities allowed for the rapid expansion of HIV-1 infections and the birth of the modern HIV-1 pandemic (Keele et al., 2006; Worobey et al., 2008). Worldwide dissemination of the virus was initiated by multiple ‘founder events’, whereby individual HIV-1 lineages moved to new regions and established epidemics, sometimes recombining in the process, and creating a globally diverse epidemic comprised of multiple subtypes and recombinant forms (Pybus and Rambaut, 2009).

HIV-1 group M variants can be subdivided into phylogenetically distinct genetic subtypes and recombinant forms (Artenstein et al., 1995; Buonaguro et al., 2007; Geretti, 2006; Jetzt et al., 2000; Keele and Derdeyn, 2009; Robertson et al., 2000; Vessière et al., 2010). Currently, there are nine "pure" subtypes (A-D, F-H, J and K); some subtypes can be further divided into sub-subtypes, for example A1-A5 (Peeters, 2001; Vidal et al., 2009). The idea of subtypes was originally proposed because many sequences of Group M fall into discrete clades that are approximately equidistantly related to each other across the entire genome (Peeters, 2001). This distance varies both within and between genes: for example, whilst *gag* and *pol* are relatively conserved, *env* is less so, with an average distance of 10-13% from a common ancestral node; within *env* itself, the hypervariable regions have less than 25% conservation of amino acids (Leitner et al., 1995; Simmonds et al., 1990). Importantly, the subtype classification system reflects only the order in which strains were classified, rather than the evolutionary history of the virus (Vidal et al., 2009). Equally, some subtypes, most notably D and B, are so similar that they would be more accurately classified as sub-subtypes of each other; that they are not is due to reasons for wishing to preserve the historical link with previous literature and classification (Geretti, 2006; Peeters, 2001).

The mechanisms of viral recombination described earlier have also led to the identification of at least 54 circulating recombinant forms (CRFs) of HIV-1, and one of HIV-2 (http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html), alongside several unique recombinant forms (URFs). When the same inter-subtype recombinant is transmitted between multiple individuals (typically at least three subjects not immediately epidemiologically related), and therefore has the potential to be of epidemiological significance, it is termed a CRF. Members of a CRF should resemble each other over the entire genome and share identical breakpoints reflecting common ancestry from the same recombination event (Peeters, 2001; Robertson et al., 2000). Unique recombinant forms, or URFs, have a more limited distribution, and are found in either fewer individuals or in epidemiologically-linked circumstances such as a single household.

Intersubtype recombination may have occurred in approximately 20% of lineages evolving over a period of 30 years in early group M sequences from Kinshasa, DRC, confirming that intersubtype recombination was a substantial force in generating HIV-1 group M diversity (Ward et al., 2013). Since then, CRFs have spread globally alongside the pure subtypes of the group M viruses, and in some areas of the world, form the predominant variant (Easterbrook et al., 2010; Simon-Loriere et al., 2009; Vidal et al., 2009).

The most widely noted impact of recombination on the genetic diversification of HIV is the frequent natural occurrence of inter-subtype recombinants in parts of the world where multiple subtypes co-circulate (Simon-Loriere et al., 2009). Dual infection was posited to affect the global HIV-1 epidemic as early as 1995, and continues to play a major role in the creation of new HIV-1 variants, particularly in Africa and other regions that experience high levels of co-circulation (Artenstein et al., 1995; Robertson et al., 1995; Vidal et al., 2013). For example, a study performed in Cameroon found a dual infection rate of approximately 22% (Vidal et al., 2013), and a recent study among men who have sex with men (MSM) in San Diego found an intrasubtype B dual infection rate of 14.4% (Wagner et al., 2013); these levels of dual infection vastly increase the chances of inter-strain in-host recombination, and therefore onward transmission of new strains to the newly infected.

1.4 Distribution of HIV-1 subtypes

Whilst all group HIV-1 group M subtypes can be found in Central Africa, consistent with this area being the source of the pandemic, the remainder of the world has an unequal distribution of subtypes (Arnold et al., 1995; Vidal et al., 2000). Globally the main variant is subtype C, which predominates in south and east Africa, followed by subtype A and CRF02_AG in west and west-central Africa (Easterbrook et al., 2010). Subtype A is also predominant in central and eastern Africa and in Eastern European countries of the former Soviet Union (Buonaguro et al., 2007).

In contrast to Africa, infections with subtype B have dominated the HIV-1 epidemic in Western Europe and North America (Balotta et al., 2001; Geretti, 2006). Although this strain of HIV-1 has often dominated the headlines, on a global scale, infections with subtype B cause approximately 12% of infections, compared with approximately 48% caused by subtype C (Geretti et al., 2009). However, given its predominance in so many countries, it can be argued that subtype B is the most widespread of the HIV-1 variants (Gilbert et al., 2007).

Historically, subtype B likely moved from Africa to Haiti in approximately 1966, where it spread for some years before dispersing elsewhere (Gilbert et al., 2007). A pandemic clade encompassing the vast majority of subtype B infections in the United States and elsewhere around the world subsequently emerged after a single foundation event out of Haiti in or around 1969, creating the basis of the Western epidemic (Gilbert et al., 2007). However, an increase in co-circulation of multiple HIV-1 strains in Western countries occurred following immigration from sub-Saharan Africa, Asia and Eastern Europe, leading to parallel epidemics comprised of different strains segregated according to ethnicity and risk group (Geretti et al., 2009). In Western countries, subtype B is predominantly observed in white MSM while non-B subtypes circulate among heterosexual man and women of other ethnic groups (Geretti et al., 2009). Although increasing migration toward Europe is changing the predominantly B pattern and increasing non-B infections, and the overall HIV-1 epidemic is becoming more homogenous, highly compartmentalised epidemics are still in evidence (Abecasis et al., 2013; Ciccozzi et al., 2012; Gifford et al., 2007).

A conservative estimate places the global prevalence of intersubtype recombinant forms at approximately 20% of the total infections worldwide (Galli et al., 2010). A high proportion of HIV-1 strains in both Uganda and Kenya are subtype A/D recombinants (Dowling et al., 2002; Harris et al., 2002). In Uganda, 24% of the HIV-1 strains analysed in Rakai were unique A/D recombinants (Harris et al., 2002). Interestingly, the proportion of each subtype within recombinants differs between, and within, each country: whereas subtype D is the most important component of the HIV-1 strains in Rakai (Harris et al., 2002), 50% of *vif* genes in a different region of Uganda were subtype A (Wieland et al., 1997), and subtype A also predominates in Kenya (Dowling et al., 2002), demonstrating that regional dependency is also exhibited by local recombinant networks.

1.5 Risk group and geography vs. recombination

Along with the inherent genetic plasticity of the virus, the range of different risk factors for HIV-1 infection contributes differing distribution pressures across geographic regions; the introduction of a new variant into a high risk population can lead to rapid spread and the establishment of that variant as the dominant subtype, e.g. subtype B in MSM in Europe and North America, and CRF01_AE in intravenous drug users (IVDU) in Asia (Arnold et al., 1995; Sanders-Buell et al., 2007; Vidal et al., 2009).

The rate at which HIV-1 spreads in a host population also affects the evolutionary rate of HIV-1 in that population (Berry et al., 2007). Differences in genetic diversity of HIV-1 strains exist both between and within risk groups, for example, subtype A spreading heterosexually in Africa had an evolutionary rate that was 8.4 times higher than the spread of subtype A among IVDUs in the former Soviet Union (Berry et al., 2007). IVDUs have been shown to both do and do not have differing intra-host HIV evolution rates (Berry et al., 2007). In South and Southeast Asia, rapid growth of HIV-1 infections was first observed among Thai IVDUs in 1988, mainly involving subtypes B and B' (Berry et al., 2007). Spread of CRF01_AE among female sex workers introduced this recombinant to the IVDU risk group, resulting in a mixed epidemic of CRF01_AE and B' infections in a complicated transmission network between sex workers and IVDUs (Berry et al., 2007).

In a heterosexual epidemic, where HIV-1 is transmitted from person to person as new contacts are formed, there is an evolutionary rate at the population level that corresponds to the intra-host evolutionary rate or (due to convergent bottleneck effects) a rate slightly slower than this. However, in an IVDU epidemic, HIV-1 spreads from person to person in the first stage of infection, leading to a fast spread of very similar viruses (Berry et al., 2007). Although some research has found that transmitted variants have less diversity than currently circulating variants in the infecting partner and are more closely related to ancestral strains (Sagar et al., 2009), more recent research indicates that sexually transmitted founder viruses cannot be directly predicted through analysis of donor quasispecies (Frange et al., 2013). At this stage, further research to determine the viral traits that accord the capacity to establish infection is needed.

Once the dominance of a variant is established in a population, the homogeneity of the epidemic increases the difficulty in measuring the degree of contribution of recombination to genetic diversity (Harris et al., 2002). However, in situations where there is either a genetically diverse epidemic, a well-defined high-risk population, or where a new variant is in the process of becoming established, the frequency of viral recombination and the effect on viral pathogenesis can be assessed (Geretti, 2006; Harris et al., 2002; Sanders-Buell et al., 2007; Vidal et al., 2009).

1.6 The UK HIV-1 epidemic

The UK epidemic, although historically dominated by subtype B infections in white MSM, has always displayed a diverse range of subtypes in non-MSM infected individuals (Arnold et al., 1995; Balotta et al., 2001; Geretti, 2006). The subtype B epidemic has been phylogenetically characterised and is thought to have arisen from at least six separate introductions into the UK population, which argues for the existence of distinct, non-overlapping sexual networks within the predominant MSM group (Hué et al., 2005). Similarly, most non-B infections in the UK have been characterised as resulting from separate introductions and at least 36 geographically distinct HIV-1 strains have been identified in the UK epidemic, including a Ugandan subtype D and subtype A strains associated with East and West Africa (Gifford et al., 2007).

Although parallel epidemics exist in the UK (MSM = subtype B; heterosexual = non-B), HIV-1 distribution and diversity is a highly dynamic process (Geretti, 2006; Gifford et al., 2007; Peeters, 2001; Vidal et al., 2009), and non-B infections have been appearing in the UK MSM community for a number of years. There has also been a change in the manner of acquisition of HIV-1 in the UK (HPA, 2012). Previously, whilst MSM infections were largely acquired in-country, heterosexual infections were mainly imported (Geretti et al., 2009; HPA, 2012). However, this has now changed, such that infections acquired in-country now outnumber imported infections, indicating increasing onwards transmission within the UK (Easterbrook et al., 2010; Geretti et al., 2009; HPA, 2012; Hughes et al., 2009). Additionally, in both the UK and in Europe, MSM and heterosexual epidemics in the post-HAART era have begun to increase again, following a period showing a decrease in the number of new infections (Easterbrook et al., 2010; HPA, 2012; van Sighem et al., 2012). Taken together, this indicates that indigenously-acquired infections are likely to increase in importance in the coming years.

HIV-1 recombinant forms have not traditionally been predominant in either the MSM or heterosexual UK HIV-1 epidemics. However, a study which investigated the genetic diversity of HIV-1 in the UK identified a number of sequences that could not be definitively assigned to a subtype or recognised HIV-1 CRF; further investigation of these showed that many were recombinant in origin, indicating that the number of these strains in the UK is increasing (Gifford et al., 2006). As HIV-1 circulates in localised epidemics, and as new variants have the potential to be rapidly established in high risk populations, investigation of unrecognised recombinant forms is likely to offer insights into current population dynamics and routes of infection.

1.7 Using phylogenetic analyses to classify novel HIV-1 CRFs and reconstruct HIV-1 epidemics

As noted previously, the generation of novel CRFs and URFs is especially likely in areas or communities where there is co-circulation of strains within the same transmission networks, particularly in high-risk populations, where multiple exposures may be more frequent (Artenstein et al., 1995; Balotta et al., 2001; Buonaguro et al., 2007; Peeters, 2001; Sanders-Buell et al., 2007). The origin and temporal spread of these strains can be reconstructed using phylogenetic methods (An et al., 2012; Drummond et al., 2006), which, at their most accurate, use full-length HIV-1 genomes.

1.7.1 Classification of HIV-1 based on full genome sequencing

The original designation of HIV-1 subtypes was based on the sub-genomic characterisation of a few individual genes (Buonaguro et al., 2007; Harris et al., 2002; Robertson et al., 2000). This is also the case for the majority of the work performed regarding the spread and distribution of subtypes, which is most commonly measured using partial *pol* sequencing (Balotta et al., 2001). A consequence of this is that the presence of recombination, and, by extension, CRFs, may have been underestimated (Balotta et al., 2001; Harris et al., 2002). However, advancements in sequencing methods have meant that HIV-1 phylogenetic classifications are currently based on either sequences derived from multiple sub-genomic regions or full-length genomic sequence analysis (Buonaguro et al., 2007; Harris et al., 2002).

Although the advent of full-length analysis was a dramatic step forward in characterising HIV-1 genomes, the methods used are not without inherent problems, especially with respect to detecting and classifying recombination. Firstly, the majority of reported sequences are obtained from either PBMC-derived proviral

DNA or from plasma RNA with reverse-transcription PCR (RT-PCR) followed by cloning (Nadai et al., 2008). PBMC proviral DNA is easier to work with than plasma RNA, but it does not reflect the actively replicating virus in the same way as sequences captured from plasma RNA, as the half life of plasma virions is only a few hours (Ho et al., 1995; Nadai et al., 2008). Also, bulk PCR methods and cloning are prone to introducing methodological artefacts in the amplified virus (Jansen and Ledley, 1990; Marton et al., 1991; Meyerhans et al., 1990; Salazar-Gonzalez et al., 2008). Another method, the heteroduplex assay (HTA), only allows for qualitative assessments of subgenomic structure (reviewed in Salazar-Gonzalez et al., 2008). To circumvent these problems, the single genome sequencing (SGS) approach was developed by Palmer and colleagues, and was extended by Salazar-Gonzalez and colleagues to encompass full-length viral genome analysis (Palmer et al., 2005; Salazar-Gonzalez et al., 2008). This method allows the obtainment of DNA sequences derived from many single viral genomes in a sample, and has low reported error rate of 0.03% (Palmer et al., 2005; Salazar-Gonzalez et al., 2008). The theory is based on the statistical premise of Poisson's distribution, which states that a limiting dilution of cDNA that results in 30% of reactions being positive will have had a reaction input of one cDNA molecule 80% of the time (Palmer et al., 2005; Nadai et al., 2008). The lack of *in vitro* recombination when using SGS techniques makes this approach ideally suited to obtaining sequences that correspond to the *in vivo* circulating plasma virus, which is essential when investigating viral pathogenesis and evolution (Nadai et al., 2008; Salazar-Gonzalez et al., 2008).

1.7.2 Recombination analyses

Recombination events can be identified most clearly in the context of phylogenetic analyses, and are indicated when phylogenetic relationships for different parts of the genome are discordant (Robertson et al., 1995).

Analysis of recombinant HIV sequences is traditionally performed using sliding window analyses, followed by phylogenetic classification of putative non-recombinant fragments of the genome (Kosakovsky Pond et al., 2009; Salminen et al., 1995). This method involves bootscanning, which involves aligning the sequence of interest with related sequences and computing phylogenies locally over the alignment, usually over a sequence length of 200-500 nucleotides (Salminen et al., 1995). When combined with informative sites analysis, a statistical assessment regarding the placement of a breakpoint (the genomic location where a recombination event occurred) at a particular position can be made (Robertson et al., 1995).

A number of automated and semi-automated algorithms based on statistical probability models, profile hidden markov models, likelihood-based models and comparison to standardised reference alignments have been developed, but these can give contradictory results (Holguín et al., 2008; Kosakovsky Pond et al., 2009; Lole et al., 1999; Oliveira et al., 2005; Pond et al., 2006; Schultz et al., 2006; Truszkowski and Brown, 2011). Additionally, difficulties exist in locating exact breakpoints between subtypes, since different subtypes are often highly conserved in some sequence regions (Gifford et al., 2006; Truszkowski and Brown, 2011).

Given that discrepancies between automated subtyping methods are common, Pond proposes that a phylogeny-based method is adopted for accurate subtyping. The programme Subtype Classification Using Evolutionary Algorithms (SCUEAL), which uses a phylogeny-based method, has a reported breakpoint recovery of 88.3% for sequences with 5% or greater divergence between parental strains and a length of 200 nucleotides or longer (Kosakovsky Pond et al., 2009).

1.7.3 Epidemic reconstruction

Phylogenetic analyses can be used as a molecular epidemiological strategy to characterise epidemics on the basis of the genetic interrelatedness of DNA sequences (Brenner et al., 2013; Drummond et al., 2006; Hué et al., 2004; Korber et al., 2000). It can be used to discern the introduction and dissemination of HIV-1 viral subtypes in different regional settings, including the transmission patterns of heterosexual, MSM and IVDU epidemics, the role of disease stage in transmission dynamics and underlying trends in regional epidemics; these can all be important in the selection of control interventions to limit HIV-1 transmission (Brenner et al., 2013).

Phylodynamics describes infectious disease behaviour that arises from a combination of evolutionary and ecological processes, and commonly use molecular clock models to represent the relationship between genetic distance and time (Pybus and Rambaut, 2009). Early models assumed a constant rate of genetic variation, but later, relaxed, models contain the ability to incorporate rate variation either between strains or through time (Pybus and Rambaut, 2009).

A common program used in phylodynamic analyses is BEAST (Bayesian Evolutionary Analysis by Sampling Trees), which uses Markov chain Monte Carlo (MCMC) probability distributions to provide a framework for analysing molecular sequence data through parameter estimation and hypothesis testing of evolutionary models (Drummond and Rambaut, 2007). In this program, an evolutionary model (nucleotide substitution, rate model among nucleotide sites and tree branches), the phylogenetic tree (which models the phylogenic relationships of the sequences) and the tree prior (a distribution for the node heights and tree topology) are used to infer a time scaled phylogeny from nucleotide sequence data, enabling epidemic reconstruction (Drummond and Rambaut, 2007). Different evolutionary models can be compared by calculating the Bayes Factor, which is a ratio of the marginal likelihoods between the models (Drummond and Rambaut, 2007). To ensure statistical sufficiency of results from these methods, it is advised to use multiple long BEAST runs and combine them (Ho and Phillips, 2009).

The process of epidemic reconstruction can be enhanced by using methods such as the coalescent, which links patterns of genetic diversity to processes such as changing population size and population structure, and is typically used to infer past rates of population growth. This tool has been successfully used to reconstruct both HIV-1 and hepatitis C populations (Pybus and Rambaut, 2009; Pybus et al., 2001). Evolutionary and spatial change can also be considered by applying phylogeography tools, which use time scaled phylogenies and geographic data to infer the geographic origin of emerging infections, the route of transmission and the rate of geographic spread (Afonso et al., 2012; Lemey et al., 2009, 2010; Mbisa et al., 2012; Pybus and Rambaut, 2009). This can be particularly interesting in the context of the emergence of recombinants, such as in Rwanda, where it was found that the HIV-1 epidemic in Kigali was characterised by the emergence of A1/C recombinants and was phylogenetically distinct from the HIV-1 epidemic in neighbouring countries (Paraskevis et al., 2004; Rusine et al., 2012; Tee et al., 2009). Local infections with non-B subtypes in B-dominated regions have also been identified by phylogenetic linkage of individuals diagnosed with primary HIV-1 infection (Brenner et al., 2007).

In order to reconstruct epidemic histories, nucleotide sequence data is required. Full-length HIV-1 genomes are preferred for this, as undetected recombination could contribute to errors made in dating HIV-1 phylogenies by increasing the apparent variation in rates among nucleotide sites and reducing the genetic distances between sequences (Rambaut et al., 2004; Schierup and Hein, 2000;

Worobey and Holmes, 2001). However, drug resistance programmes, introduced in the 2000s, have provided large *pol* (protease/RT) sequence datasets which can be used for analysis of transmission trends in regional epidemics; using *pol* only has been found to be sufficient for epidemic reconstruction, notwithstanding the presence of undetected recombination that may exist in the unsequenced regions of the genome (Hué et al., 2004; Lewis et al., 2008; van Sighem et al., 2012).

1.8 Individual-level consequences of HIV-1 genetic diversity

There is great variation in the clinical outcome following HIV-1 infection (Fellay et al., 2007; Langford et al., 2007). As such, individual-level effects of genetic diversity, such as disease progression associated with different subtypes and strains, exist.

One of the strongest predictors of disease progression is T cell activation (Ormsby et al., 2012). During the steady state (when CD4 lymphocyte production and destruction are balanced), the population of CD4 lymphocytes turns over every 15 days (Ho et al., 1995). As HIV-1 infection advances and escapes immune pressures from the host, the immunological state of the infected individual deteriorates (Lemey et al., 2007; Ormsby et al., 2012). Progression to AIDS in humans is characterised by high levels of immune activation associated with accelerated T cell turnover rates and apoptotic death (Langford et al., 2007; Schindler et al., 2006). As such, CD4 cell decline can be used as a proxy for disease progression in individuals (Badri et al., 2008; Phillips et al., 2010).

Other predictors of HIV-1 disease progression include plasma HIV-1 RNA viral load, which has been classed as the best single predictor of HIV-1 disease progression (Mellors et al., 1997). Plasma viral load is predictive of subsequent CD4 cell decline and can discriminate risk at different levels of CD4 count, but the most accurate predictor of HIV-1 disease progression comes from the use of plasma HIV-1 viral load and CD4 cell count in combination (Mellors et al., 1997). HIV-1 proviral DNA viral load is also a predictor of disease progression (independent of plasma RNA viral load and CD4 cell count), particularly in the first six months of infection (Rouzioux et al., 2005). Although HIV-1 proviral DNA viral loads are not commonly measured during routine clinical monitoring, it may be that a combination of all three measurements would provide the best overall monitoring of HIV-1 disease progression (Rouzioux et al., 2005).

There is some evidence to indicate that the rate of HIV-1 disease progression varies depending on the subtype of infection, and whether the infection was caused by a multi-subtype recombinant. In both Kenya and Uganda, infections with viruses classed as subtype D based upon envelope sequences were found to carry a significantly higher increase in the risk of death, and a faster rate of CD4 cell decline and disease progression than subtype A infections, despite similar HIV-1 plasma viral loads (Baeten et al., 2007; Kaleebu et al., 2002; Kiwanuka et al., 2010). This was also seen in a study conducted in London, where subtype D infections were associated with significantly faster rate of CD4 decline when compared with subtypes B, A and C (Easterbrook et al., 2010). An earlier switch to using the CXCR4 co-receptor for viral entry may explain subtype D infections' faster progression (Easterbrook et al., 2010). In other studies, subtype A and C have been shown to have slower disease progressions relative to subtype B infections (Klein et al., 2011).

A further study in the Rakai district of Uganda assessed A/D recombinants, and found subtype A and A/D infections were transmitted at a higher rate than subtype D infections, although the A/D recombinant figure was not statistically significant (Kiwanuka et al., 2009). This suggests that over time the rate of subtype D infections may decrease relative to the rate of subtype A infections, and is perhaps reflected in the lesser worldwide distribution of subtype D (Abecasis et al., 2007; Kiwanuka et al., 2010). This has been seen in Rakai, where an 8-year study found a significant decrease in subtype D viruses from 71% to 63% and an increase in A viruses from 15% to 20% (Kiwanuka et al., 2009).

1.9 Aims and Objectives

The aim of this PhD project was to conduct a comprehensive investigation into HIV-1 recombination in the UK, focusing on the following research questions:

1. Is the composition of the UK HIV-1 landscape changing on a genetic level?
2. Are there novel HIV-1 CRFs circulating in the UK?
3. What are the origin and likely pathogenesis of these novel strains?

Question 1 (Chapter 3) was addressed by assessing the degree and nature of subtype unclassified recombination present in the UK as captured by the UK HIV Drug Resistance Database (UK HIV DRD), followed by phylogenetic identification of clusters of strains likely to be novel HIV-1 CRFs. Clusters comprised of strains likely to be of epidemiological importance were subjected to phylodynamic analysis, including geographic investigations.

Although Question 2 (Chapters 4 and 5) was partially answered through the investigation performed for Question 1, confirmation of one strain identified through this process as a novel HIV-1 CRF was sought. In order to perform detailed characterisation and analyses of one strain likely to be a novel HIV-1 CRF, a method for near full-length, single genome sequencing of HIV-1 RNA was developed and optimised. Stored plasma samples from six patients sharing a similar breakpoint in the *pol* gene were retrieved and characterised, and the recombination breakpoints identified. Identification of both the recombinant breakpoint locations and the classification of parental subtypes was performed using both a traditional sliding window analysis with phylogenetic reconstruction of non-recombinant fragments and a probabilistic jumping profile hidden markov model (jpHMM) (Lole et al., 1999; Schultz et al., 2006, 2009). Following characterisation, the registration of CRF50_A1D and one B/CRF50_A1D URF was reported.

Question 3 (Chapters 6 and 7) was answered through a reconstruction of the epidemic history of CRF50_A1D in the UK using phylogeography and Bayesian skyline analysis, an investigation into the global origin of the component parental A1 and D subtypes, and a preliminary analysis of the likely pathogenesis of CRF50_A1D infections conducted using CD4 decline slopes from patients not yet on antiretroviral treatment.

Chapter 2: Methods

2.1 Access to HIV drug resistance testing sequences and associated clinical data

The UK Collaborative Group on HIV Drug Resistance (UK CHIC) (http://www.ukchic.org.uk/) and the UK HIV Drug Resistance Database (UK HIV DRD) (http://www.hivrdb.org.uk/) provided access to stored HIV reverse transcriptase and protease sequences produced at the time of drug resistance testing in routine practice and, where available, relevant clinical and demographic data. These bodies are coordinated by the Medical Research Council Clinical Trials Unit (MRC-CTU) (http://www.ctu.mrc.ac.uk/). Additional demographic data for some patients were obtained through an agreement between the Health Protection Agency (HPA) and the UK HIV DRD.

2.1.1 UK CHIC

The UK collaborative group on HIV Drug Resistance was established in 2001 and collects clinical (including treatment) data from patients undergoing care at 15 collaborating centres across the UK (http://www.ukchic.org.uk/). The earliest available data are from 1996, and, as of 2012, the CHIC database contained over 45,000 records. The data retrieved from UK CHIC for this project were date of birth, gender, risk group (route of HIV exposure), country of origin, ethnicity, date of HIV-1 diagnosis, CD4 cell counts with dates, plasma HIV-1 RNA loads with dates, and antiretroviral treatment (ART) history.

2.1.2 The UK HIV DRD

The UK HIV Drug Resistance database is a repository of *pol* gene sequences (protease and reverse transcriptase) obtained by Sanger population sequencing in a large number of HIV-infected patients undergoing drug resistance testing in the UK (http://www.hivrdb.org.uk/). Sequences submitted by contributing centres are genotyped by the database steering committee using the Rega and SCUEAL methodologies at approximately 2-year intervals (Kosakovsky Pond et al., 2009; Oliveira et al., 2005). Both these methods classify HIV-1 sequences into pure or CRF HIV-1 subtypes. Sequences that are not recognised by Rega as a pure subtype or a known CRF are designated 'unassigned', whereas SCUEAL suggests a putative recombinant structure for these sequences, given as both a 'simplified subtype' and a 'detailed subtype'. The ‘detailed subtype’ field includes both intra-

and inter-subtype recombination and is often quite complex (e.g. B,B,B,B,CRF14,D,G inter-subtype recombinant), whereas the simplified subtype field shows a basic recombinant structure only (e.g. B,G recombinant).

Following genotyping, the accessible database download is referred to by the year of genotyping e.g. 2007, 2010 or 2012. This project used both the 2007 and 2010 database downloads; the database download used in each project section is clearly indicated within the text. Although SCUEAL results are usually reported using the nucleotide position from the beginning of the sequence length, e.g. 212, for ease of understanding, all results in this report were converted to standard HXB2 numbering.

The 2007 UK HIV DRD database download contained 34,469 sequences, and the 2010 download contained 55,556 sequences. As sequences were collected from both treatment-naive and treatment-experienced patients, the database contained multiple sequences from some patients (baseline and failure resistance test/s). Accordingly, the 2007 download contained 25,631 unique patient records, and the 2010 download contained 43,002 unique records (Andrew Leigh Brown, personal communication, 2010; Hughes et al., 2009). For patients captured only in the UK HIV DRD and not by UK CHIC or HPA surveillance, the demographic information available for each sequence comprised the sample date and a geographical identifier for an aggregated group of HIV clinics.

2.1.3 HPA data

The Health Protection Agency is an independent body that collects epidemiological and surveillance data on UK HIV infections (http://www.hpa.org.uk/). As part of an agreement with the UK HIV DRD, access to some demographic data was available for those patients not captured by UK CHIC. As of 2012, the HPA database held approximately 19,500 records containing date of birth, gender, risk group, country of origin and date of HIV-1 diagnosis.

2.1.4 Project proposals and ethical approval

Access to data contained in UK CHIC and the UK HIV DRD is decided by the steering committee on a project-by-project basis. Two proposals were submitted and accepted for this project.

2.1.4.1 UK HIV DRD project proposal for the analysis of unassigned sequences

This proposal requested access to sequence data stored in the UK HIV DRD for the purposes of analysing subtype-unassigned sequences and identifying potential recombinant specimens of interest (Appendix 2_1). Ethics approval was gained from the Royal Free Hospital Ethics Committee for the anonymised study of up to 15 recombinant specimens. The MRC-CTU assisted with the anonymised retrieval of six specimens from three HIV centres in London and Northwest England. Plasma specimens were retrieved from routine storage and transferred to the virology department of the Royal Free Hospital, London, where the laboratory work took place.

2.1.4.2 UK CHIC proposal for CD4 cell slope analysis

A proposal was submitted to UK CHIC to access HIV-1 RNA viral loads, CD4 cell counts and ART data for 19 patients with recombinant HIV-1 infections (Appendix 2_2). These data were used to analyse the rate of HIV-1 disease progression as estimated from CD4 cell decline.

2.2 Near full-length amplification and sequencing of HIV-1 from plasma RNA

Following specimen retrieval, characterisation of patient specimens was performed using single genome, near full-length amplification and sequencing of HIV-1. This process involved: extraction of HIV-1 RNA (vRNA) from plasma; reverse transcription of vRNA into cDNA; limiting dilution of cDNA; amplification of near full-length HIV-1 DNA using nested PCR and Sanger sequencing of resulting PCR products. Each stage of the process was extensively optimised.

2.2.1 Sample preparation

140µl of plasma was normalised to contain 20,000 copies of vRNA. In specimens with low HIV viral loads, the appropriate volume of plasma was centrifuged for 90 minutes, 20,000xg, 4°C, and the supernatant removed to obtain a final volume of 140µl. The viral pellet was resuspended in the remaining plasma supernatant before proceeding to vRNA extraction. In cases where the combination of HIV-1 viral load and the volume of plasma received was too low to contain 20,000 virus copies, the entire available plasma volume was concentrated to 140µl using the above centrifugation protocol to obtain the maximum number of viral copies available. This pre-extraction step was used prior to extraction using the QiAmp Viral RNA Mini Kit.

2.2.2 Extraction and reverse transcription of viral RNA from plasma

2.2.2.1 Viroseq extractions

Viroseq extractions were used during the optimisation of the Nadai et al. protocol for near full-length amplification of HIV-1. These were performed according to the manufacturer’s instructions. Briefly, 500µl thawed plasma was placed in an 1.5ml Eppendorf and centrifuged for 1 hour, 22,000xg, 4°C to pellet virions, after which the supernatant was removed. The pellet was resuspended in 600µl lysis buffer and incubated at room temperature for 10 minutes. 600µl isopropanol was added and the mixture centrifuged at 13,000xg, 5 minutes. Samples were aspirated to dryness; 1ml chilled 70% ethanol was added, and the mixture spun at 13,000xg, 5 minutes. After a final aspiration, the pellets were air-dried and 50µl cold RNA diluent added to resuspend the extracted RNA.

2.2.2.2 EasyMAG extractions

Automated nucleic acid extractions were performed using the NucliSENS EasyMAG (Biomerieux, Marcy l'Etoile, France). 1ml of thawed plasma was used in the generic on board extraction protocol, according to the manufacturer's instructions. Following incubation of plasma with 2ml of Lysis Buffer for 10 minutes at room temperature, 140µl of magnetic silica solution was added to the sample cartridge. Nucleic acids were eluted in 25µl Elution Buffer.

2.2.2.3 QiAmp Viral RNA Mini kit

To extract vRNA, the QiAmp Viral RNA Mini kit (Qiagen, Hilden, Germany) was used according to the manufacturer’s instructions, with a modified final elution volume of 65µl rather than 60µl. Briefly, 140µl of prepared sample and 560µl Buffer AVL with carrier RNA were combined and incubated at room temperature for 10 minutes. 560µl absolute ethanol was added, and the entire mixture was applied to a spin column in two 630µl steps, using two spins at 6,000xg, 1 minute. 500µl of Buffer AW1 was added, and the mixture centrifuged at 6,000xg, 1 minute, after which 500µl Buffer AW2 was added and the column centrifuged at 14,000xg, 3 minutes. An extra spin was performed using an empty spin column to ensure complete removal of buffer. RNA was eluted using 65µl AVE buffer, followed by incubation at room temperature for 1 minute, and centrifugation at 6,000xg, 1 minute.

An identical extraction process using nuclease-free water as the input was performed simultaneously to provide a contamination control in subsequent downstream analyses. This is henceforth referred to as the negative control.

2.2.3 Reverse Transcription

All extracted RNA was immediately transcribed in four separate reactions using the Superscript III First Strand Synthesis Supermix Kit (Life Technologies, Paisley, UK) using the following protocol per reaction: 0.63µl of a 20µM solution (0.25µM final concentration) of reverse primer 1.R3.B3R 5'-ACTACTTGAAGCACTCAAGGCAAGCTTTATTG (CHAVI-MBSC 2009, unpublished) was combined with 1.87µl nuclease free water and 2.5µl Annealing Buffer in a nucleic acid free environment on ice. 15µl (5,000 copies) RNA template was added, and the resulting mixture denatured at 65°C for 5 minutes. PCR tubes were removed and placed on ice for at least 1 minute before adding 25µl of 2xReaction Mixture and 5µl of Enzyme Mixture (Superscript III and M-MLV-RT). Reactions were heated to 50°C for 90 minutes, paused, and 2µl of Superscript III RT Enzyme added, after which reactions were incubated at 55°C for a further 90 minutes, then 85°C for 5 minutes. The resulting cDNA was combined, and either frozen at -80°C or used immediately as input into a nested PCR reaction.

2.2.4 cDNA dilution and nested PCR protocol

A two step cDNA dilution protocol was used to ensure single genome amplification. The first step was a limiting dilution to find the dilution at which 30% of the PCR reactions were positive. When a Poisson distribution is assumed, this 30% positive level is understood to contain a single copy of amplified DNA in 80% of cases. 96 nested PCR reactions were set up, of which there was one no template control (NTC), one negative control, six reactions using undiluted cDNA, 24 wells at a 1:10 cDNA dilution and 64 wells at 1:100 cDNA dilution. This plate was primarily a control to assess that the cDNA was amplifying at rates suggesting near-ideal extraction and reverse transcription conditions, i.e. that the undiluted cDNA contained 100 copies/µl. Following acceptable results from the limiting dilution plate, a further 96 nested PCR reactions were performed. These reactions comprised one NTC reaction and 95 template reactions at a 1:200 cDNA dilution (a theoretical input of 1 copy/reaction). In circumstances where there was a reduced viral load, or the limiting dilution plate suggested that the sample was amplifying suboptimally, the dilutions were adjusted accordingly. The PCR amplicons from this second amplification reaction were those used in subsequent Sanger sequencing.

All PCR reactions were set up on ice. Both first round and nested PCR reactions used the Platinum PCR Supermix High Fidelity Kit (Life Technologies, Paisley, UK). The first round per reaction volumes were 45µl PCR supermix, 1.25µl each (0.25µM each final concentration) of forward primer 1.U5.B1F 5'CCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGT and reverse primer 1.R3.B3R, 0.5µl nuclease free water and 2µl cDNA. Cycling conditions were 94°C 2 minutes, followed by 40 cycles of 94°C 15s, 60°C 30s, 68°C 9.5 m, and a final extension at 68°C 20 minutes.

The nested PCR reaction was identical to the first round PCR, except that the primers used were 2.U5.B4F 5'-AGTAGTGTGTGCCCGTCTGTTGTGTGACTC (forward) and 2.R3.B6R 5'-TGAAGCACTCAAGGCAAGCTTTATTGAGGC (reverse). The template input was 2µl of the first-round reaction. Cycling conditions were 94°C 2 minutes, followed by 45 cycles of 94°C 15s, 60°C 30s, 68°C 9.5 m, and a final extension at 68°C 20 minutes. The resulting ~9kb product spanned HXB2 nucleotides 552-9636.

2.2.5 Agarose gel electrophoresis and preparation of DNA for Sanger sequencing

Positive nested PCR reactions were identified using agarose gel electrophoresis on a 1% agarose gel with a 5µl DNA input. PCR products were prepared for Sanger sequencing using Milipore filters (Watford, UK). Post-filtration, cleaned PCR products were quantified using agarose gel electrophoresis, whereby the intensity of the DNA band was compared to that of a quantified DNA ladder (Hyperladder I, Bioline, London, UK) and diluted with molecular biology grade water such that the subsequent sequencing reactions produced signal intensity within the parameters of the ABI 3730xl sequencing instrument (Life Technologies). Diluted PCR products were pooled for sequencing analysis as a quality control measure.

2.2.6 DNA sequencing

Prepared PCR products were directly sequenced using fluorescently labelled dideoxy chain terminators (BigDye Terminator v3.1 Cycle Sequencing Assay, Life Technologies) and an automated 3730xl sequencer in a 20µl sequencing reaction containing 1.7µl Big Dye, 9.3µl nuclease free water, 1µl of primer (0.5µM final concentration) and 8µl template. Sequencing primers were sourced from in-house protocols obtained from the Molecular biology and Sequencing core at the Centre for HIV/AIDS Vaccine Immunology (CHAVI), published protocols in Van Laethem et al., 2005, 2006; Nadai et al., 2008 and designed using Primer3 version 4.0 (available at http://frodo.wi.mit.edu/primer3). A total of 100 primers were used to sequence the full HIV genome (Appendix 2_3). Sequencing reactions were repeated until near full bi-directional coverage was obtained.

2.2.7 Sequence assembly and quality control

Sequences were assembled using SeqScape version 2.6 software (Life Technologies). During sequence assembly, the fragment sequences from individual sequencing primers were examined for the presence of mixed bases, which could potentially indicate the amplification of >1 target molecule. The electropherograms and raw data from all positions showing evidence of this were examined; cases where the mixed base was due to low sequencing intensity were repeated using the individual primer; for cases in which the mixed base appeared genuine the specimen was re-amplified from scratch. There were isolated cases for which there was consistently low sequencing signal at a particular base, even after repeat sequencing. In these cases, the mixed base was allowed to stand rather than manually edit the sequence.

2.3 Phylogenetic analyses

Phylogenetic analyses were used in the following tasks: assessing the number of subtype-unassigned and complex sequences present in the UK HIV DRD; cluster analyses of unassigned and complex sequences; recombination analyses of unassigned and complex sequences and the amplified full-length patient specimens; identification of further individuals carrying the novel CRF50_A1D strain; phylogeographic determination of the emergence and spread of CRF50_A1D in the UK and an investigation into the likely geographic region from which the CRF50_A1D parental strains originated.

A variety of different methods were used to perform the phylogenetic analyses in this thesis. For sequence alignment, MUSCLE was chosen in preference to more widely used alignment programmes such as Clustal W due to the faster alignment time, especially when aligning full-length sequences (Edgar, 2004; Nuin et al., 2006). Bio-Edit was selected for manual adjustments to automated sequence alignments due to the flexibility of the program and the ability to convert files into a variety of formats (Hall, 1999).

To construct phylogenetic trees, three programmes were used: FastTree, PhyML and BEAST (Drummond and Rambaut, 2007; Guindon et al., 2005; Price et al., 2009). FastTree was used to analyse alignments with large numbers of sequences (>100) due to the speed of the programme; PhyML was used to perform maximum likelihood analysis when classifying the putative pure regions of full-length sequences, and BEAST was used for more complex analyses such as time-scaled phylogenies and phylogeographic analysis. PhyML was used in preference to BEAST during subtyping classification because it was considered that maximum likelihood analysis was a sufficiently robust methodology to perform subtyping classifications and that the high computational requirements of Bayesian analysis were best saved for intensive, time-scaled analyses.

Recombination analyses were performed using four methods: RIP, SCUEAL, jpHMM and Simplot (Kosakovsky Pond et al., 2009; Lole 1999, Schultz et al., 2006; Siepel et al., 1995). RIP was used as a 'quick look' programme to gain a rough understanding of the likely recombinant structure of a sequence and Simplot was used to run manual screens of sequences; this was especially useful when classifying novel recombinants as the programme allowed the use of a user-defined reference sequence when performing a similarity scan. jpHMM and SCUEAL were used for the bulk of recombinant classification. SCUEAL was chosen in preference to Rega (another widely used subtyping method) due to its incorporation of evolutionary analyses into its algorithm through the measurement of branch lengths for MRCA and the use of statistical criteria (the Bayesian Information Criterion) to calculate the fitness of each proposed structural model; in contrast, Rega tests its subtyping classification using a sliding window approach and a measurement of phylogenetic signal only. jpHMM was chosen because of the probabilistic approach to assigning subtype classifications and recombinant breakpoints: local segments of the query are aligned to the segments of the reference alignment that are the most similar to them. It was considered that concordant recombination results from jpHMM and SCUEAL had a high possibility of accuracy, given the differences in the construction of the algorithms used.

In terms of workflow, from a starting point of unaligned sequences, MUSCLE and Bio-Edit were used to create an alignment. For downstream analyses involving cluster identification, FastTree was used to build an approximate maximum likelihood tree, followed by measurement of genetic distances and subtyping classification using SCUEAL and jpHMM; following screening, BEAST was used to build trees and time-scaled phylogenies of the identified clusters. For downstream analyses involving recombinant classification, RIP, Simplot and jpHMM were used to predict the likely recombinant structure; after slicing the alignment into regions of putative pure subtypes, PhyML was used for maximum likelihood classification of these regions. Subsequently, BEAST was used for evolutionary analyses such as time-scaled phylogenies, prediction of tMRCA and phylogeographic analysis.

2.3.1 Creation of a near full-length HIV-1 reference alignment

A near full-length HIV-1 reference alignment was constructed for use in recombination and phylogenetic analyses. Sequences covering the HXB2 region 552 - 9636 (the region amplified using the near full-length sequencing protocol) were downloaded from the Los Alamos HIV Sequence Database (LANL) (http://www.hiv.lanl.gov/components/sequence/HIV/search/search.html). Sequences were downloaded in subtype- and sub-subtype-specific groups of A1, A2, B, C, D, F1, F2 G, H, J and K. To reduce computational requirements, the implementation of MUSCLE housed at the European Bioinformatics Institute (EBI) website (http://www.ebi.ac.uk/Tools/muscle; Edgar, 2004) was used to make subtype-specific alignments, which were then manually edited using BioEdit version 7.0.5.3 (Hall, 1999). The genetic distance between the sequences in each subtype alignment was measured using a Tamura Nei 93 (TN93) pairwise distance matrix implemented in HyPhy (distancematrix.bf), and sequences with less than 6% genetic distance were discarded (Poon et al., 2009; Tamura and Nei, 1993). The 6% genetic distance was chosen for two reasons: firstly, because a distance of 6% would allow the capture of the range of genetic variation that is present within individual subtype groupings; second, to keep consistency with the conditions that were used when the original SCUEAL reference alignment was created (Kosakovsky Pond et al., 2009).

The final subtype-specific alignments were combined with the Los Alamos 2008 subtype reference alignment, re-aligned using MUSCLE, edited with Bio-Edit, and genetic distances measured using the above parameters. At this stage, the V1 and V2 hypervariable regions of *env* were stripped from the alignment owing to excessive sequence variability across subtypes. Finally, the remaining sequences were tested for recombination using the Recombinant Identification Program (RIP) at the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov/content/sequence/RIP/RIP.html) with a window size of 400 base pairs (bp), gap stripped, and a confidence level of 90%; any sequences showing evidence of recombination were discarded. As the resulting alignment was heavily weighted with subtype B and C sequences, several of these were stripped to provide a more manageable and relevant alignment. The final reference alignment contained 78 sequences comprising 10 A1, 1 A2, 19 B, 24 C, 4 D, 6 F1, 4 F2, 5 G, 3 H, 1 J and 1 K sequences. This alignment was used for all subsequent analyses where the program implementation did not provide its own reference alignment, and is henceforth referred to as the 'reference alignment'.

2.3.2 Contribution of recombination to the UK HIV-1 epidemic

The 2010 download of the UK HIV DRD was used to investigate the contribution of recombinant HIV-1 strains to the UK HIV-1 epidemic. This investigation had two parts: a) an assessment of the proportion of recombinant HIV-1 strains present in the UK HIV DRD (both as CRFs and as non-CRF recombinants), and b) whether any of the non-CRF strains present had been transmitted to sufficient people to constitute novel CRF strains.

The SCUEAL subtype classification from the 2010 UK HIV DRD download was used to classify sequences as pure or recombinant strains of HIV-1 for part a) of this investigation. To perform the work for part b), however, both the subtype classification and the location of any recombination breakpoints was essential data. Owing to restrictions in storage capacity, the stored subtype classification in the UK HIV DRD did not include recombinant breakpoint locations. Therefore, for part b) of this investigation, the sequences studied for the cluster analysis had recombination breakpoints determined using two methods: the SCUEAL implementation housed on the Datamonkey server (http://www.datamonkey.org), and the jpHMM implementation at the GOBICS server (http://jphmm.gobics.de/). The jpHMM program provided a schematic of the likely recombinant structure, subtype designations for each genomic region, breakpoint locations with 95% confidence intervals, and an indication of regions of subtyping uncertainty. Large 95% confidence intervals at breakpoint locations indicated uncertainty within the model, and suggested that further analysis was required to designate a precise location (Schultz et al., 2006, 2009).

The results from these analyses were used in all downstream analyses. To differentiate between the two sets of SCUEAL results, the terms 'SCUEAL 2010' and 'SCUEAL 2012' were used throughout.

2.3.2.1 Definition of ‘non-CRF recombinant’, ‘unassigned’ and ‘complex’ sequences, and ‘clusters’

Both ‘non-CRF recombinant’ and ‘unassigned’ sequences were defined as those sequences that had been classified as recombinant strains of HIV-1 but were not a recognised CRF HIV-1 strain. ‘Complex’ sequences were defined as sequences comprising four or more subtype strains that did not share the defined recombinant structure of a recognised HIV-1 CRF. For the purposes of this document, all three sequence types have been referred to using the term 'unassigned sequence', which was defined to include any recombinant HIV-1 strain not included in the Los Alamos HIV CRF guide (http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html#CRF55) as of December 2012.

It should be noted that this project did not seek to determine transmission direction among individuals sharing similar recombinant HIV-1 strains. The term 'cluster' is used throughout to refer to groups of related HIV-1 strains that have met particular phylogenetic criteria.

2.3.2.2 Number of unassigned/recombinant sequences present in the UK HIV DRD by year

To assess the degree of recombination in the UK HIV-1 epidemic (as captured by the UK HIV DRD) the 55,556 sequences in the UK HIV DRD were stratified by year of sampling; further stratification used the SCUEAL subtype classification to classify sequences into pure subtypes, recognised CRFs, and unassigned. Duplicate sequences from the same patient were excluded. The total number of unassigned sequences and the total number of all recombinant sequences (unassigned + recognised CRFs) per year were compared to the total number of sequences per year. The SCUEAL simplified subtype field was used for sequence classification. Statistical significance was assessed using the Chi squared test and the Chi squared trend. The years 1996 and 2009 were excluded from statistical analysis due to the low number of sequences captured in 1996 and incomplete data from 2009 (6 months only).

2.3.3 Identification of clusters of unassigned sequences

Following the investigation into the degree of recombination present in the UK HIV-1 epidemic, the specific recombinant structures present in the 'subtype-unassigned' fraction of the UK HIV DRD were characterised and clusters of identical structures were identified. The purpose of this was to identify potential new CRFs that were circulating in Britain, rather than specifically examining transmission clusters. Subtype B-containing recombinants were selected for this analysis as this subtype was traditionally restricted almost exclusively to the MSM risk group, and it was hypothesised that any emerging novel population dynamics would be identified more readily from this higher risk group as opposed to the lower risk heterosexual community.

2.3.3.1 Subtyping classification of sequences in UK HIV DRD.

The 2010 detailed SCUEAL subtype was used to capture all potential B-recombinant sequences in the UK HIV DRD. There were two reasons for choosing the detailed subtype field rather than the simplified subtype field as the starting point for this analysis. Firstly, the speed of HIV-1 evolution makes the analysis of complex recombinant sequences notoriously difficult (Oliveira et al., 2005), and secondly, criteria with a wide latitude were desirable in order to capture all possible recombinants in the database. The sequences were analysed for breakpoints and subtype classification using SCUEAL and jphMM as detailed previously. The results were examined for concordance using the SCUEAL detailed subtype, SCUEAL simplified subtype and the jpHMM results. Statistical significance was assessed using Fisher’s exact test (two tailed), and a p value of 0.05 was the cut-off for statistical significance.

2.3.3.2 Approximate maximum likelihood analyses

Unassigned sequences belonging to subtype B were selected and aligned with the reference alignment. FastTree v2.1.3 was used to perform approximate maximum likelihood analyses using the Generalised Time Reversible (GTR) +CAT model (available from http://meta.microbesonline.org/fasttree/).

2.3.3.3 Cluster screening

The topology of the approximate maximum likelihood tree (but not the approximate likelihood values) was used to determine potential clusters of unassigned recombinants. Jukes-Cantor maximum genetic distances were calculated over the entire tree using a Perl script written by J. Ambrose in the HIV, HBV and HCV pathogenesis group at the University of Liverpool. The usual cluster criteria when working with data from UK HIV DRD was a genetic distance of 4.5% or less and a bootstrap support of 95% (internal criteria set by the steering committee of the UK HIV DRD). However, these criteria were created to identify transmission clusters rather than novel CRFs; consequently, relaxed criteria were used for this analysis to ensure that all potentially related recombinant structures were captured.

A maximum genetic distance of 10% was used to screen for clusters within the approximate maximum likelihood tree. The distance of 10% was chosen in order to capture groups of related recombinant structures e.g. groups of sequences with the same overall recombinant structure but that may have slightly different breakpoint locations. The distribution of the genetic distances of the clusters was measured using the D'Agostino-Pearson omnibus normality test. Clusters comprised of pairs and clusters containing pure subtype reference sequences were excluded. Following this, the SCUEAL 2012 and jpHMM subtype classifications were used to exclude clusters containing any sequences that had been previously classified as B-recombinant (SCUEAL 2010) but were now classified as pure subtype B. Some sequences were now classified as belonging to pure subtypes other than B; clusters containing these sequences were excluded.

Clusters remaining following this screening were screened using recombinant breakpoint locations. Owing to the variability inherent in the analysis of recombinant sequences, both the SCUEAL 2012 and the jpHMM results were used in this screening. If a cluster contained sequences showing similar breakpoint locations and subtype classifications in either SCUEAL 2012 or jpHMM then it was taken forward for further analysis. Finally, as neither SCUEAL nor jpHMM included all recognised CRFs in their algorithms, the Los Alamos CRF guide was consulted to ensure the remaining clusters were not comprised of sequences from a previously described CRF.

2.3.3.4 BEAST analyses

To be considered a potential novel CRF, sequences in each of the identified clusters were required to also exhibit clustering when broken down into component subtype regions. This was investigated using Bayesian Evolution using Sampling Trees (BEAST) version 1.8.0.

In order to use the full range of subtype diversity for these analyses, the reference alignment was expanded to include selected sequences from the Los Alamos 2010 HIV subtype reference alignment (http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html#ref). To reduce computational requirements, only sequences relevant to the putative recombinant structure were included; due to the close relation between subtypes B and D, subtype D sequences were also included. The final alignment for each cluster comprised: the sequences under investigation; relevant pure subtype reference sequences; reference sequences from all recognised CRFs with similar subtype compositions, and an appropriate outgroup for the putative recombinant structure.

Following the completion of each reference alignment, sequences were sliced into component subtype regions according to the identified SCUEAL and jpHMM breakpoints. Where conflict existed among the identified breakpoints for the cluster, the cluster was split into segments reflecting each conflicting set of breakpoints and analysed multiple times. 3xMCMC runs of 1x10^8 states were run for each alignment segment using a strict molecular clock, no tip dates, the Hasegawa, Kishino and Yano (HKY) nucleotide model, and a Bayesian Skyline coalescent with 10 groups. Trees were visualised using FigTree version 3.1 (available from http://tree.bio.ed.ac.uk/software/figtree/). If the sequences under investigation clustered monophyletically in each component tree, with a posterior prior of at least 0.7, they were considered to be a potential novel CRF.

2.3.3.5 Demographic analysis

Risk group analysis was conducted on the clusters that passed all the above criteria using the recorded risk groups (where available) in the UK HIV DRD and HPA data. The geographic location of each sequence was investigated using the aggregated centre data captured by the UK HIV DRD (this gave a general geographic region only, e.g. London and Southeast England or Northwest England).

HIV BLAST at the Los Alamos National Database (http://www.hiv.lanl.gov/content/sequence/BASIC_BLAST/basic_blast.html) was used to conduct an initial investigation into further instances of the recombinant structures beyond the UK HIV DRD, using every sequence in a putative cluster as a reference sequence for a BLAST search. To be included in downstream analyses, a BLAST match was required to be one of the 10 most related sequences for at least 75% of the sequences in a cluster.

Clusters with interesting demographic properties or geographic links were re-analysed using time-scaled phylogenies in BEAST v1.8.0 using the conditions above (with tip dates) in order to elucidate further linkages.

2.3.4 Analysis of near full-length sequences

Following near full-length sequencing of specimens of interest, the following phylogenetic analyses were used:

2.3.4.1 Recombination analyses using RIP

Following sequence assembly of a query specimen, the sequence was submitted to the online RIP program using window size of 400bp, gap stripped, and a confidence level of 90% to have a 'quick look' at the likely recombination profile. Sequences were first submitted against the entire reference alignment of pure subtypes, and then subsequently submitted using a reduced profile containing only those subtypes showing the highest similarity scores. As this program does not identify recombination breakpoints, it was used to gain an overall idea of the likely recombinant structure. Specimens showing a recombination profile that warranted further breakpoint identification were analysed using two different methods: a) jpHMM, and b) sliding window and bootscanning analysis using Simplot.

2.3.4.2 Recombination analyses using jpHMM

Query sequences were submitted to the online jpHMM implementation at the GOBICS web server (http://jphmm.gobics.de). The identified breakpoints and subtype classifications were compared to the breakpoints and subtype classifications from the Simplot and maximum likelihood analyses.

2.3.4.3 Recombination analyses using Simplot

To perform sliding window and bootscan analyses using Simplot, the query sequence(s) were added to the full-length reference alignment. Query sequence(s) were added to the reference FASTA file, and this file was aligned using MUSCLE with manual editing using BioEdit as described previously. Sliding window analyses were performed using Simplot version 3.5.1 (Lole et al., 1999). A 400bp window was set with a step size of 20bp. Sequences were gap stripped, and Kimura 2-parameter genetic distances were used to find the 50% consensus. The transition/transversion ratio was set at 2.0. The results from the similarity screening were used to select the subtype groups for bootscan analysis.

Bootscanning of the query sequence was performed using subtypes A1, B, D, and F2 with informative sites analysis. The analysis parameters were as for the similarity screening, and 100 bootscan replicates were performed using neighbour-joining tree building. Recombination breakpoints were set using the highest statistically significant x^2 value around the 50% crossover point between subtypes. The statistical significance of the identified breakpoints was assessed using Fisher's exact test. Following breakpoint assignment, slices of the alignment corresponding to the putative pure subtype regions between each breakpoint were created and saved for downstream analyses.

2.3.4.4 Likelihood mapping and maximum likelihood genotyping of putative non-recombinant fragments

Prior to maximum likelihood analysis, each slice of the alignment was assessed using likelihood mapping to determine whether sufficient phylogenetic signal was present for robust phylogenetic analysis. Likelihood mapping was performed using Tree-puzzle (Schmidt et al., 2002). Ungrouped sequences were analysed for 10,000 neighbour-joining quartets, using the HKY model of nucleotide substitution. Transition/transversion and nucleotide frequency parameters were estimated from each dataset, and uniform rate heterogeneity was assumed. Results were viewed using EPS Viewer (available from http://epsviewer.org/), and the percentage of un- and partly-resolved quartets counted.

The Entropy tool at the Los Alamos database (available from http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html) was used to investigate the variation between subtypes B and D over the genomic region covered by alignment slice 4 (HXB2 2101-2503, *pol* gene). The subtype B and D alignments used in the full-length reference alignment were submitted to the tool, and the consensus sequence for each subtype produced from this was submitted to the EMBOSS pairwise alignment at the European Bioinformatics Institute website (http://www.ebi.ac.uk/Tools/emboss/align). Following this, the number of informative sites was counted over the length of the region in question.

Following likelihood mapping, the likelihood parameters for each putatively pure subtype region of the HIV-1 genome were estimated using PAUP version 4.0 (Sinauer Associates, Massachusetts, USA). HKY85 distance and neighbour-joining trees were used to estimate the transition/transversion ratio, nucleotide base frequencies, and among-site rate variation under a gamma distribution. These parameter estimates were used for maximum likelihood analysis using the

Phylogenetic Estimation Using Maximum Likelihood (PhyML) implementation housed at the ATGC server (http://www.atgc-montpellier.fr/phyml/; Guindon et al., 2005), with 1000 bootstrapping replicates, with the exception of alignment slice 5, which was too long for 1000 replicates to be performed. In this case, 100 bootstrap replicates were considered sufficient. To restrict computational requirements, the alignment used for the PhyML analysis was restricted to subtypes A1, B, D and K; subtypes A1 and D were chosen as the closest matches to the query sequences, subtype B due to its genomic similarity to subtype D, and subtype K as an outgroup. Regions of the genome not showing clear identification with one subtype after this process (paraphyletic clustering or bootstrap support <70%) were resubmitted to PhyML using a greater range of subtypes from the reference alignment (A1, A2, B, C, D, F1, F2 and K).

The phylogenetic trees produced from the maximum likelihood analysis were visualised using Dendroscope version 2.3 (available from http://ab.inf.uni-tuebingen.de/data/software/dendroscope3/download/welcome.html). Schematics of finalised recombinant structures were drawn using the Recombinant HIV-1 Drawing Tool (RDT) from the Los Alamos website (available from http://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html).

2.3.5 Confirmation and registration of a novel HIV-1 CRF

2.3.5.1 Observation of clustering across the entire genome

Full-length sequencing was performed on six specimens in total. These specimens were selected following a two-stage process. Originally, a cluster analysis was performed on the 2007 download of the UK HIV DRD by colleagues at the Rega Institute in Belgium. This process identified five sequences that could not be assigned to a definite subtype. At this stage, ethics approval for full-length sequencing of up to 15 specimens was obtained, and the Rega Institute transferred the project to University College London and the Royal Free Hospital. Following this, 4/5 specimens were obtained from three different clinics in London. These four specimens (33365, 8179, 40534 and 34567) were subjected to full-length sequencing and the *pol* regions of the sequences were used to identify further instances of the same recombinant structure in the UK HIV DRD. Additional specimens were sought for further full-length sequencing, but due to reasons of preserving patient anonymity, it was decided to only attempt to obtain specimens from clinics with greater than 10 patients sharing the novel recombinant structure. Following record de-linking performed at the UK HIV DRD headquarters, it became apparent that only one clinic, located in Northwest England, had sufficient numbers of these patients. This clinic was contacted, and stored samples from two patients were available, bringing the total number of specimens obtained for full-length sequencing to six.

To confirm clustering of the A1/D recombinant specimens across the entire HIV-1 genome, the sequences from all six recombinant specimens were added to the full-length reference alignment, and sliced into the pure subtype regions indicated by the breakpoint analysis. Each alignment slice was analysed using FastTree version 2.1. Approximate maximum likelihood trees were generated using the GTR+CAT model on nucleotide alignments. Trees were visualised using FigTree version 3.1. Each tree was examined for evidence that the A1/D recombinant sequences clustered with each other within the pure subtype group across the entire HIV-1 genome.

2.3.5.2 Comparison with the Los Alamos HIV CRF register

Following breakpoint identification and subtype assignment of the HIV-1 infections of interesting patients, the identified structures were compared with the list of registered CRFs on the Los Alamos website (http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html). CRFs containing the same subtypes as identified recombinants were compared to the breakpoints of the novel recombinant to confirm the novel recombinant structure had not been previously identified as a CRF.

2.3.5.3 Global BLAST analysis of full-length and individual gene regions

The full-length structure of recombinants and individual genetic regions of the *gag, pol* and *env* genes were submitted to Basic Local Alignment Search Tool (BLAST) at NCBI (http://blast.ncbi.nlm.nih.gov/) in order to identify any submitted HIV-1 sequences with similar subtype and breakpoint patterns. Sequences returned as matches were scanned for evidence of recombination using RIP and jpHMM.

2.3.6 Identification of additional cases of CRF50_A1D in the UK HIV DRD

2.3.6.1 BLAST, SCUEAL and jpHMM analysis

To capture additional cases of CRF50_A1D that were not identified during the original analysis, a three stage process was used. Firstly, the *pol* sequences of three specimens were used as reference sequences for a local BLAST search of the UK HIV DRD. Following this, the top 500 closest matches to each sequence were subtyped using SCUEAL and jpHMM, and the breakpoint locations (and statistical confidence intervals) used to identify possible matches.

Possible matches were confirmed using the following criteria: a) a matching SCUEAL breakpoint location, or a breakpoint that included the SCUEAL breakpoint in the 95% confidence interval; b) a matching jpHMM breakpoint location, or a nearby breakpoint location that included the jpHMM breakpoint in the 95% confidence interval.

In addition to matching breakpoint locations, potential additional cases were required to show a matching recombinant structure. Both B/A1 and D/A1 structures were admitted as possible matches owing to the similarity of subtypes D and B in *pol* (see Shannon Entropy results below).

Following this initial determination, sequences were confirmed as additional cases by building an approximate maximum likelihood tree using FastTree to confirm monophyletic clustering. Multiple sequences from the same patient were excluded using UK CHIC ID and demographic information.

2.3.6.1.2 Entropy analysis

Shannon entropy testing was performed using the Shannon Entropy tool located at the LANL HIV database (http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html). Subtypes B and D were compared between HXB2 positions 2253 and 3572.

2.3.7 Emergence and distribution of CRF50_A1D in the UK

The time to Most Recent Common Ancestor (tMRCA) of CRF50_A1D in the UK was determined using time scaled phylogenies in BEAST v1.6.1. All CRF50_A1D sequences (both full-length and *pol* only) were aligned with the following reference sequences: 4 full-length subtype A sequences; 4 full-length subtype D sequences; the 3 closest full-length sequence matches to CRF50_A1D in BLAST; 4 subtype C sequences (outgroup – chosen from the full-length reference alignment). The alignment was performed using HIV align (http://www.hiv.lanl.gov/content/sequence/VIRALIGN/viralign.html) and manually edited using Bioedit. The month of sampling (sequence date) was converted into decimal to assume sampling in the middle of the month using the formula:

$$decimal\ month = year + (\frac{month\ -\ 0.5}{12})$$

2.3.7.1 Model selection

Find Model at the Los Alamos National database was used to predict the best nucleotide substitution model for use (http://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html). In each case, the initial tree was constructed using PAUP Jukes Cantor/Neighbour Joining and the reduced set of possible models was used. The model with the lowest Akaike Information Criterion (AIC) score was considered the best model.

Strict and relaxed molecular clocks and GTR and HKY nucleotide substitution models were tested. A Bayesian Skyline coalescent was used throughout. Three runs of 1x10^{8} states were performed for each set of model conditions. To ease computational requirements the multiple tree and log files were combined using Log Combiner (available from http://beast.bio.ed.ac.uk/LogCombiner) with a resampling rate of 10,000 and a burn-in of 10%. Maximum clade credibility trees were annotated using Tree Annotator (available from http://beast.bio.ed.ac.uk/TreeAnnotator*)* and viewed using FigTree. Bayes factors were calculated using Tracer (available from http://beast.bio.ed.ac.uk/Tracer).

2.3.7.2 Time scaled phylogenies

Time scaled phylogenies were performed to investigate the likely emergence date of CRF50 using all identified CRF50 sequences, the subtype A region, subtype D region and the concatenated A/D region of *pol*.

Time scaled phylogenies in BEAST v1.6.1 were performed using three runs of 1x10^{8} states, a relaxed uncorrelated log normal molecular clock, the GTR nucleotide substitution model, tip dates and a Bayesian Skyline coalescent with 10 groups and a constant expansion. No other priors were set. Multiple runs were combined using Log Combiner with a resampling rate of 10,000 and a burn-in of 10%. All Effective Sample Sizes (ESS) were required to be greater than 200 for run acceptance.

To confirm the tMRCA, the five full-length sequences were used to analyse subtype A1 and D regions of *gag, pol* and *env.* Due to the small number of sequences, a simpler model and a longer run length were used for this analysis. Six runs of

2.5x10^{8} states, a relaxed uncorrelated log normal molecular clock, the HKY nucleotide substitution model and a Bayesian skyline coalescent with two groups and a constant expansion were used. A gamma distribution prior was set on the covariance. Multiple runs were combined using Log Combiner as described above.

The regions of each gene used for analysis were dictated by the recombinant structure of CRF50, and were as follows (numbers represent HXB2 nucleotide positions):

Subtype A:

- *gag:* 552-1274
- *pol:* 2504-3256
- *env:* 6612-7248

Subtype D:

- *gag:* 1274-1884
- *pol:* 2253-2504
- *env:* 8568-7248

The final characterised full-length sequence was a CRF50/B URF. In order to ascertain whether this recombinant was a product of onward recombination, or whether this sequence was the source of CRF50, the time scaled phylogenies were performed both with and without this sequence.

2.3.7.2.1 Bayesian skyline reconstruction

The demographic history of CRF50_A1D infections in the UK was reconstructed using a Bayesian skyline coalescent. A stepwise (constant) skyline variant was used to model the skyline population and group sizes to estimate the number of effective CRF50_A1D infections.

2.3.7.3 Phylogeographic analysis

2.3.7.3.1 Geographic locations of CRF50_A1D patients

Geographic locations of patients were determined using aggregated centre data as detailed above. A midpoint longitude/latitude point was used to create an arbitrary geographic location within each group of aggregated clinics.

2.3.7.3.2 Discrete phylogeographic analysis using BEAST

The combined tree file from the time-scaled emergence analysis was resampled to contain 1,000 trees and used as the input tree for discrete phylogeographic analysis in BEAST v1.6.1. The .xml files from the emergence analysis were edited according to the instructions at http://beast.bio.ed.ac.uk/Discrete_Phylogeographic_Analysis in order to estimate the path of CRF50 geographic spread across the UK. 2xMCMC runs of 2.5x10^{8} states were performed and combined using Log Combiner as detailed above. Trees were annotated and visualised using Tree Annotator and FigTree as detailed above.

2.3.7.3.3 Google Earth analysis

Following the BEAST phylogeographic analysis, the tree files were converted to .kml format and converted for visualisation in Google Earth. CamStudio (http://camstudio.org/) was used to capture the Google Earth rendering on video.

2.3.8 Global origin of CRF50_A1D

The likely global origin of the A1 and D strains comprising CRF50_A1D was investigated. *Gag, pol* and *env* A1 and D regions of CRF50 were investigated separately.

2.3.8.1 Construction of global alignments

Alignments of A1 and D sequences were constructed using publicly available sequences downloaded from the LANL database, using the geography search. Sequences were initially downloaded from East African countries, as this region traditionally contains the highest number of subtype A, subtype D and A/D recombinant infections. Additional sequences were obtained from the LANL database by selecting all countries that had a prevalence in the LANL database of A1 or D sequences of 10% or more. UK sequences were included wherever possible; in practice this was only in the *pol* region. All pure A1 and D UK sequences lodged in the UK HIV DRD were included in the *pol* analysis. The list of countries searched was: Angola, Belarus, Botswana, Burundi, Cameroon, Central African Republic (CAR), Chad, DRC, Ethiopia, Gabon, Georgia, Ghana, India, Kenya, Latvia, Tanzania, Russian Federation, Rwanda, Ukraine, Uganda, and the UK.

Potential reference sequences were screened by year and genetic distance. Sequences predating the earliest sequence in the UK HIV DRD were used in order to try and ascertain whether the recombinant was imported into the UK as a recombinant or whether the recombination event occurred in the UK.

Genetic distance screening was used to reduce the number of sequences in each alignment, and therefore ease computational requirements. Due to the difference in genetic variation present in different HIV-1 genes, a different genetic distance cut-off was used for each gene. Different cut-offs were also used for subtype A and subtype D in the *pol* gene due to the number of sequences available. The cut-offs used were: *gag:* 6%; *pol:* 14% (subtype A1) and 6% (subtype D), *env:* 20%.

The number of sequences contained in the trees were too many to allow tip dates to be used in the analysis, or to list each reference sequence used. Therefore, the phylogenetic trees from this analysis have been presented with branches coloured by either country of origin or geographic region; the relevant accession numbers of reference sequences have been included where appropriate.

2.3.8.1.1 Selection of A1 and D regions

The sequence regions used for analysis were those used in the time scaled phylogenies detailed in section 2.3.7.2 above.

2.3.8.2 Approximate maximum likelihood analysis

Approximate maximum likelihood analysis was performed using the FastTree conditions detailed previously. The following criteria were used to assess the origin of the CRF50 A1 and D strains:

- Clustering with of both A1 and D CRF50 strains with exclusively UK A1 and D sequences indicated that CRF50 arose from a recombination event in the UK (possible for the *pol* gene trees only);
- Clustering with strains from other countries in the presence of UK A1 and D strains indicated likely recombination in another country followed by importation into the UK as a recombinant structure.

2.4 Statistical analyses

An estimation of the likely speed of progression of CRF50_A1D infections was performed using CD4 decline slopes as a proxy for disease progression.

2.4.1 ID of CRF50_A1D patients in the CHIC database

CRF50_A1D patients who had clinical and demographic data captured as part of UK CHIC were identified. The anonymised CHIC ID reference numbers were transferred to the MRC-CTU, de-anonymised at source, and the new reference IDs were transferred to Professor Caroline Sabin at CHIC. Professor Sabin extracted the data and linked it to the original anonymised reference identifiers. No linkable patient identifiable information was transferred outside UK CHIC.

The CHIC inclusion criteria were: age over 16; entered the cohort between January 1, 1998 and June 2010; known HIV-1 subtype and at least two CD4+ T-lymphocyte counts within one year while ART naive. CD4 cell counts, HIV-1 RNA viral loads and ART history were extracted from the CHIC database for each patient that met the criteria.

Once data were received, post-treatment HIV-1 RNA viral loads and CD4 cell measurements were excluded. Patients were also excluded from further analysis if there were less than three pre-treatment CD4 counts available.

2.4.2 Transformation of data and statistical analysis over time

To calculate CD4 cell count decline per year, CD4 cell measurement dates were normalised by subtracting the first date of measurement for each individual (baseline) from each subsequent measurement date for that individual. This associated each CD4 cell measurement with a 'days from baseline' value; dividing the 'days from baseline' by 365.25 scaled these values to 'years from baseline'. Following normalisation, the square root of each CD4 cell measurement was taken and simple linear regression was performed for each patient. Due to the low number of patients involved the slope for each patient was studied individually.

2.4.3 Comparison to subtype B slopes

The CD4 count slopes for patients were compared to the CD4 count slopes for subtype B patients that were estimated by Klein in 2010 (CROI 2011, Abstract B-131).

Chapter 3: Contribution of viral recombination to the UK HIV-1 epidemic.

3.1 Proportion of recombinant HIV sequences in the UK HIV DRD

3.1.1 Genotyping the UK HIV DRD: Rega and SCUEAL results

The 2010 UK HIV DRD SCUEAL and Rega genotyping results were obtained from the UK HIV DRD steering committee. Sequences were initially categorised according to whether a subtype could be assigned. Categorisation of the results into sequences that were subtype–assigned and those that were subtype-unassigned showed that SCUEAL returned significantly more unassigned results than Rega (6,076 vs. 3,801, p=<0.001; Table 3_1). Both methods identified the same number of Group O sequences in the database; Rega identified one HIV-2 sequence, whereas SCUEAL categorised this sequence as unknown. The SCUEAL method had 73 subtyping failures. No failures were recorded with the Rega method.

Classification	**Rega**	**SCUEAL**
Assigned	51,733	49,405
Unassigned	3,821	6,076
HIV-2	1	0
Group O	1	1
Unknown	0	1
Failed	0	73
Total	**55,556**	**55,556**

Table 3_1. Classification of UK HIV DRD records using two different genotyping methods. The 55,556 sequences in the 2010 download of the UK HIV DRD were stratified into subtype-assigned, subtype-unassigned, failed and unknown using Rega and SCUEAL genotyping methods.

3.1.2 Proportion of unassigned/recombinant sequences in the UK HIV DRD

After removing any duplication, stratifying the UK HIV DRD sequences by year of sample showed that the number of unassigned sequences in the database increased from 0 in 1996 to 1140 (12.6%) in 2008, the last complete year of sampling (p = <0.001; Table 3_2). The chi squared test for trend over the same time period was also significant (p = <0.001). One year (2004 - 2005) showed a statistically significant year-on-year increase (p = <0.01). The total number of recombinant records (unassigned sequences + recognised CRF) increased from 0 in 1996 to 1,377 in 2008 (p = <0.001; chi squared test for trend p = <0.001); two years showed statistically significant year-on-year increases: 2004 - 2005 (p = <0.01) and 2006 - 2007 (p = 0.02).

During this same period, the proportion of recognised CRFs increased from 0% in 1996 to 2.6% in 2008 (Figure 3_1). Unassigned sequences increased more rapidly than CRFs, demonstrating that the increase in recombinants as captured by the UK HIV DRD was independent of the increase in CRF circulation.

Year	Total records	Unassigned records	% Unassigned	p	Recombinant records	% Recombinant	p
1996	7	0	0.0		0	0.0	
1997	334	17	5.1		21	6.3	
1998	464	19	4.1	0.64	21	4.5	0.38
1999	588	26	4.4	0.92	30	5.1	0.79
2000	855	50	5.8	0.31	61	7.1	0.18
2001	1326	89	6.7	0.50	109	8.2	0.44
2002	1795	128	7.1	0.72	163	9.1	0.48
2003	2396	184	7.7	0.57	225	9.4	0.80
2004	3886	332	8.5	0.29	420	10.8	0.11
2005	5975	642	10.7	<0.01	801	13.4	<0.01
2006	7474	849	11.4	0.33	1038	13.9	0.50
2007	7938	987	12.4	0.07	1226	15.4	0.02
2008	9053	1140	12.6	0.80	1377	15.2	0.73
2009	906	124	13.7		153	16.9	

Table 3_2. Unassigned and total recombinant sequences stratified by year. Unassigned records include non-CRF recombinants; total recombinant records include non-CRF + CRF recombinants. P values were calculated using Chi-squared with Yates' correction and represent year-on-year increases.

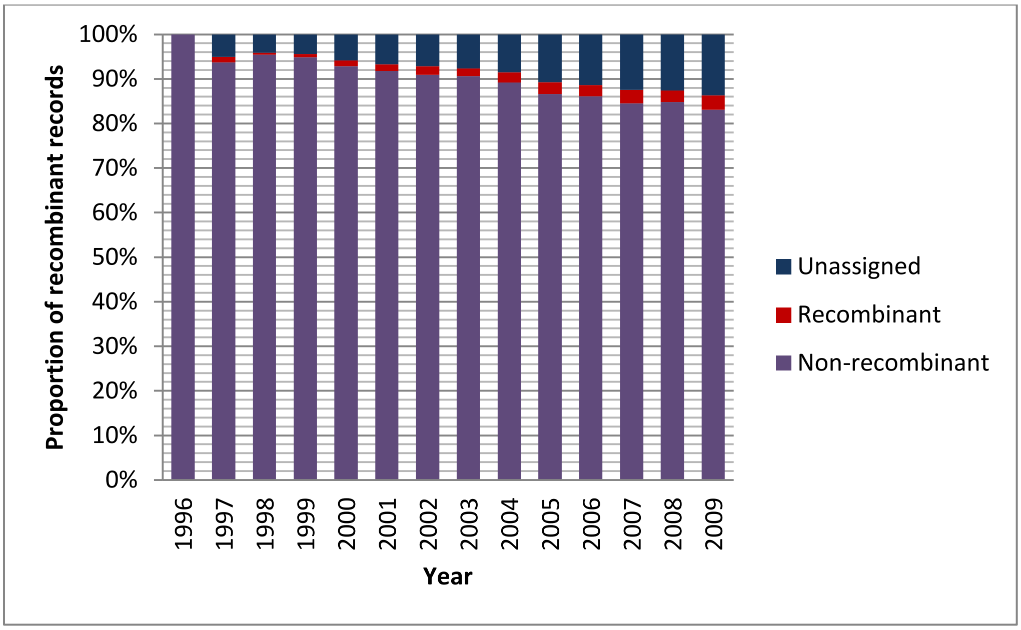


Figure 3_1. Unassigned and recombinant records as a proportion of total records in the UK HIV DRD. Non-recombinant, CRF recombinant and unassigned recombinant sequences as a proportion of the total UK HIV DRD records. Non-recombinant sequences are coloured purple, CRF-recombinant records are coloured red and unassigned-

recombinant records are coloured blue. The proportion of unassigned-recombinant records increased more than the proportion of CRF-recombinant records.

3.2 Identification of clusters of complex sequences

Following confirmation that the contribution of subtype-unassigned sequences to the UK HIV-1 epidemic was increasing, the nature of these unassigned sequences was investigated. As indigenous HIV transmission in the UK has been traditionally composed of subtype B infections transmitted among MSMs, we focused the investigation on unassigned sequences that contained subtype B in their mosaic structure. This was in order to draw inferences regarding whether population mixing between different risk groups was driving the increase in unassigned recombinant sequences.

3.2.1 Subtyping classification of sequences in UK HIV DRD

2,030 sequences were classified as unassigned subtype B-recombinant in the 2010 SCUEAL genotyping results. All 2,030 sequences were submitted to SCUEAL 2012 and jpHMM for breakpoint analysis and the results between all three methods were compared.

3.2.1.1 Comparison of SCUEAL 2010 and SCUEAL 2012

The comparison first compared the SCUEAL 2010 and SCUEAL 2012 results. Using the SCUEAL detailed subtype field, 268/2,030 (13.2%) of subtyping results were concordant. The proportion of concordant results increased when using the SCUEAL simplified subtype field (1,174/2,030, 57.8%), but were still less than would generally be expected for two sets of results from the same method.

Of the 858 non-concordant sequences (when using the simplified subtype field), 409/858 (47.8%) were sequences that had been classified as subtype B-recombinant by SCUEAL 2010 but were classified as pure subtype B by SCUEAL 2012. To investigate these 409 sequences the jpHMM result was used to try and gain a consensus. Overall, 365/409 (89.2%) sequences were also classified as pure subtype B by jpHMM. Of the remaining 44/409 discrepant sequences, 27/44 (61.2%) were sequences that contained gaps; and 26/27 (96.3%) contained an erroneous A2 classification by jpHMM (Figure 3_2). If the misclassified gap sequences were discarded, the concordance between jpHMM and SCUEAL 2012 for those sequences classified as subtype B-recombinant by SCUEAL 2010 but pure subtype B by SCUEAL 2012 was 95.3% (365/383).

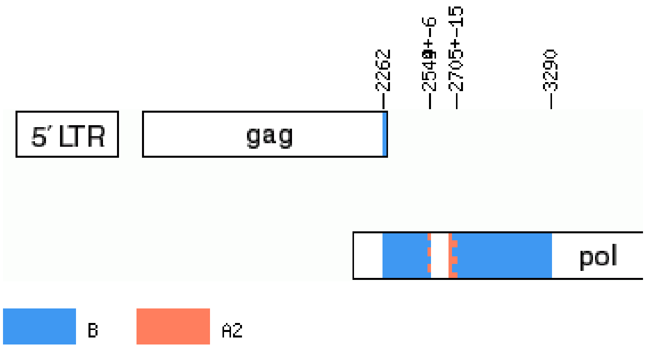


Figure 3_2. Erroneous jpHMM A2 classification in sequences with gaps. Example of a jpHMM result for a sequence with a gap between HXB2 2549 and 2705. A gap (i.e. a region for which no nucleotides could be determined) in this position is typical of sequencing results from the Trugene genotyping assay. Small regions on either side of the sequence gap have been misclassified by jpHMM as subtype A2, suggesting that sequence with gaps should be genotyped using an alternative method.

Of the remaining 449 sequences that showed discrepant results from the two SCUEAL runs, 183/449 (40.1%) were not truly discrepant but showed minor typographic differences in the displayed results displayed, 130/449 (29.0%) were sequences that were recombinant (e.g. G/D/B/K) by one set of results but classified as complex by the other, 29/449 (6.5%) were sequences originally classified as pure subtype B by SCUEAL 2010 but as subtype B-recombinant by SCUEAL 2012, and 31/449 (6.9%) were sequences that were subtype B-recombinant by SCUEAL 2010 but pure non-B subtypes by SCUEAL 2012. The remaining 76/449 (16.9%) sequences showed discrepant recombinant structures.

If the 183 sequences showing typographical differences were classified as concordant, then the proportion of concordant results increased to 1,357/2,030 (66.8%) for the simplified subtype results. If those sequences classified as recombinant by one genotyping run and complex by another were also included as concordant, the proportion further increased to 1,487/2,030 (73.2%).

3.2.1.2 Comparison of SCUEAL 2010, SCUEAL 2012 and jpHMM

The breakdown of the SCUEAL 2010 results was as such:

- 2,030 B-recombinant sequences (detailed subtype field)
    - 1176 B-recombinant (simplified subtype field)
    - 854 pure subtype B (simplified subtype field)

Owing to the variability observed in the detailed subtyping results in this previous section, the simplified subtype field was used to investigate the concordance between all three sets of genotyping results. Firstly, all sequences that were pure subtype B in the simplified subtype classification were removed. This left 1,176 SCUEAL 2010 subtype B-recombinant sequences. The jpHMM results for these sequences showed 609/1176 (51.8%) pure subtype B, 120/1176 (10.2%) non-B subtypes, 444/1176 (37.8%) subtype B-recombinant, and 3/1176 (0.26%) were subtyping failures.

The SCUEAL 2012 results for the 1,176 SCUEAL 2010 subtype B-recombinant sequences showed 409/1176 (34.8%) pure subtype B, 68/1176 (57.8%) non-B subtypes, and 699/1176 (59.4%) subtype B-recombinants. Sequences that were classified as subtype B-recombinant by jpHMM were not necessarily classified as subtype B-recombinant by SCUEAL 2012. jpHMM showed a significantly higher proportion of pure subtype B results than SCUEAL 2012 (51.8% vs. 34.8%, p=<0.01) and a significantly higher proportion of non-B results (10.2% vs. 5.8%, p=<0.01). Overall, 379 sequences were subtype B-recombinant sequences by all three methods (Table 3_3).

Method	B	Non-B	B-recombinant	Failures	Total
jpHMM	609	120	444	3	1176
SCUEAL 2012	409	68	699	0	1176

Table 3_3. Comparison of jpHMM and SCUEAL 2012 results to SCUEAL 2010 results using the SCUEAL simplified subtype classification. 1176 SCUEAL 2010 B-recombinant sequences were genotyped using SCUEAL 2012 and jpHMM. jpHMM showed the lowest number of B-recombinant results and the highest number of pure subtype B results. jpHMM also showed the highest number of results that did not include subtype B.

3.2.2 Approximate maximum likelihood analyses

The comparison between the three subtyping methods showed that, even when using the simplified subtype classification, a high degree of non-concordance was observed between the three subtyping methods. Therefore, all 2,030 sequences identified by the detailed SCUEAL 2010 results as subtype B-recombinant were carried over to the approximate maximum likelihood analysis that was the first stage of cluster identification (Figure 3_3).

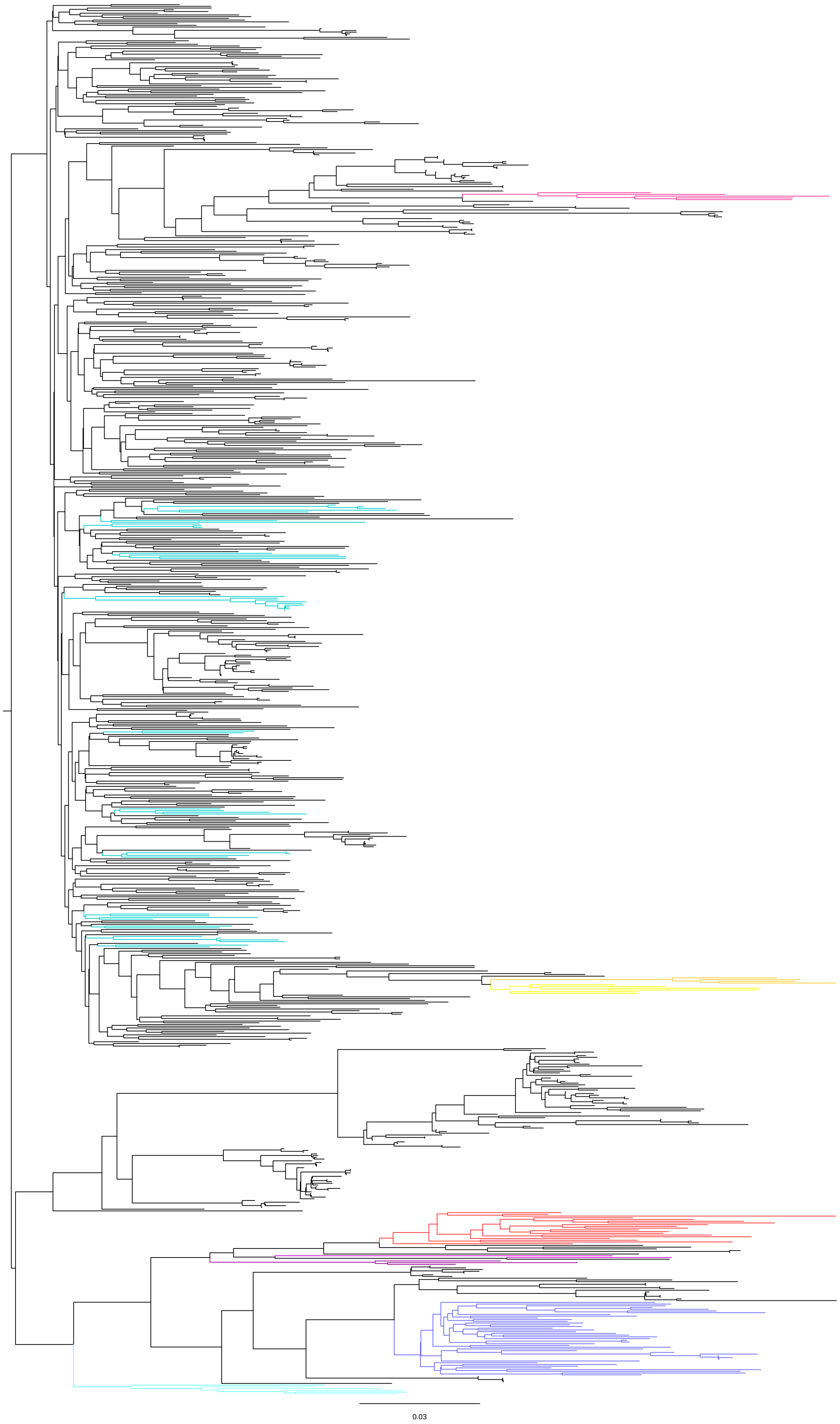


Figure 3_3. Approximate maximum likelihood analysis of 2,030 putative B-recombinant sequences. 2,030 sequences classified as B-recombinant by SCUEAL 2010 were aligned with 78 pure subtype reference sequences and analysed using FastTree v2.1. The tree is mid-point rooted. Pure subtype sequences were coloured as follows: A = red; B =

teal, C = blue, D = turquoise, F1 = yellow, F2 = mustard, G = pink, H/J/K = purple. The 2,030 query sequences are coloured black. A high proportion of sequences were located in branches containing subtype B reference sequences.

The tree showed the majority of the sequences distributed among the subtype B reference sequences, which was expected given the high number of pure subtype B subtyping results in the SCUEAL 2012 and jpHMM analyses. The tree topology indicated potential novel CRFs in three regions. The first of these was near the subtype G reference sequences, where several branches were located between the subtype B and subtype G reference sequences (Figure 3_4). These branches indicated the presence of B/G recombinant sequences; six of these branches contained greater than three sequences, suggesting novel CRFs.

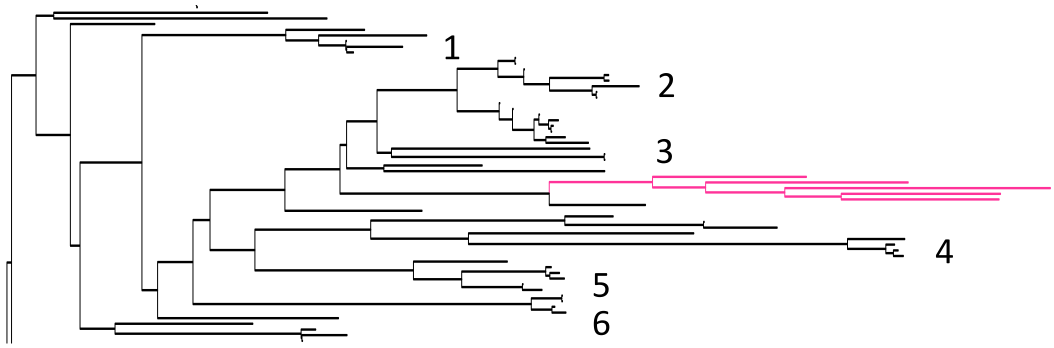


Figure 3_4. Putative subtype B/G recombinant CRF sequences. Enlarged section of Figure 3_3 showing the region surrounding the subtype G reference sequences (in pink). Several branches were located between the subtype B and subtype G references, indicating the presence of B/G recombinant sequences. Numbers 1-6 indicate branches with sufficient sequences to be potential novel CRFs.

The second region of interest was located near the subtype F reference sequences (Figure 3_5), where the tree showed five branches with sufficient sequences to comprise novel CRFs.

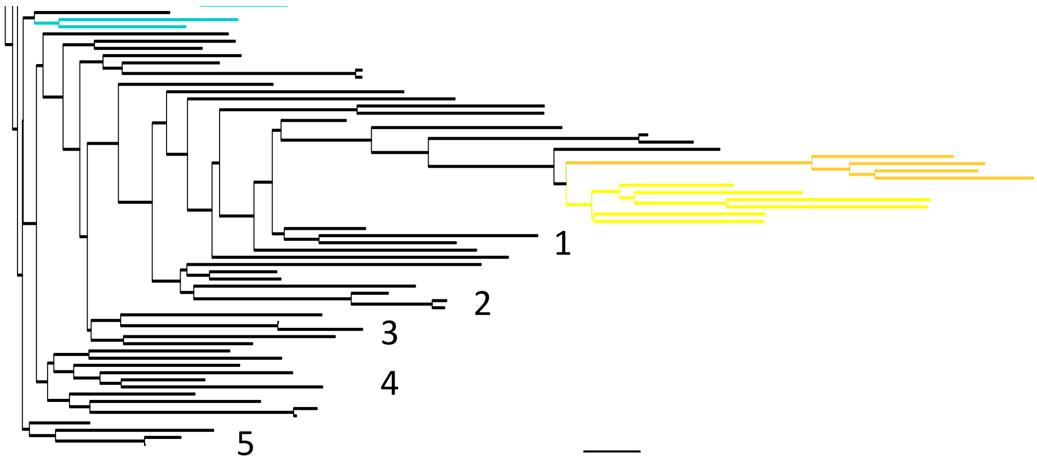


Figure 3_5. Putative subtype B/F recombinant CRF sequences. Enlarged section of Figure 3_3 showing the branches surrounding the subtype F reference sequences (F1 = yellow, F2 = mustard). Several branches are located between the subtype B and subtype F references, indicating the presence of B/F recombinant sequences. Numbers 1-5 indicate branches with sufficient sequences to be potential novel CRFs.

The final region was located at the base of the tree (Figure 3_6). This region showed seven branches with sufficient sequences to be novel CRFs. Three regions were located between the subtype B reference sequences and the remainder of the reference sequences, one region was located close to the subtype A sequences, and the remaining three regions were located close to the subtype C sequences. Regions one, two and six contained sizeable numbers of sequences (58, 27 and 14, respectively), which, if confirmed as novel CRFs, could indicate a significant population impact.

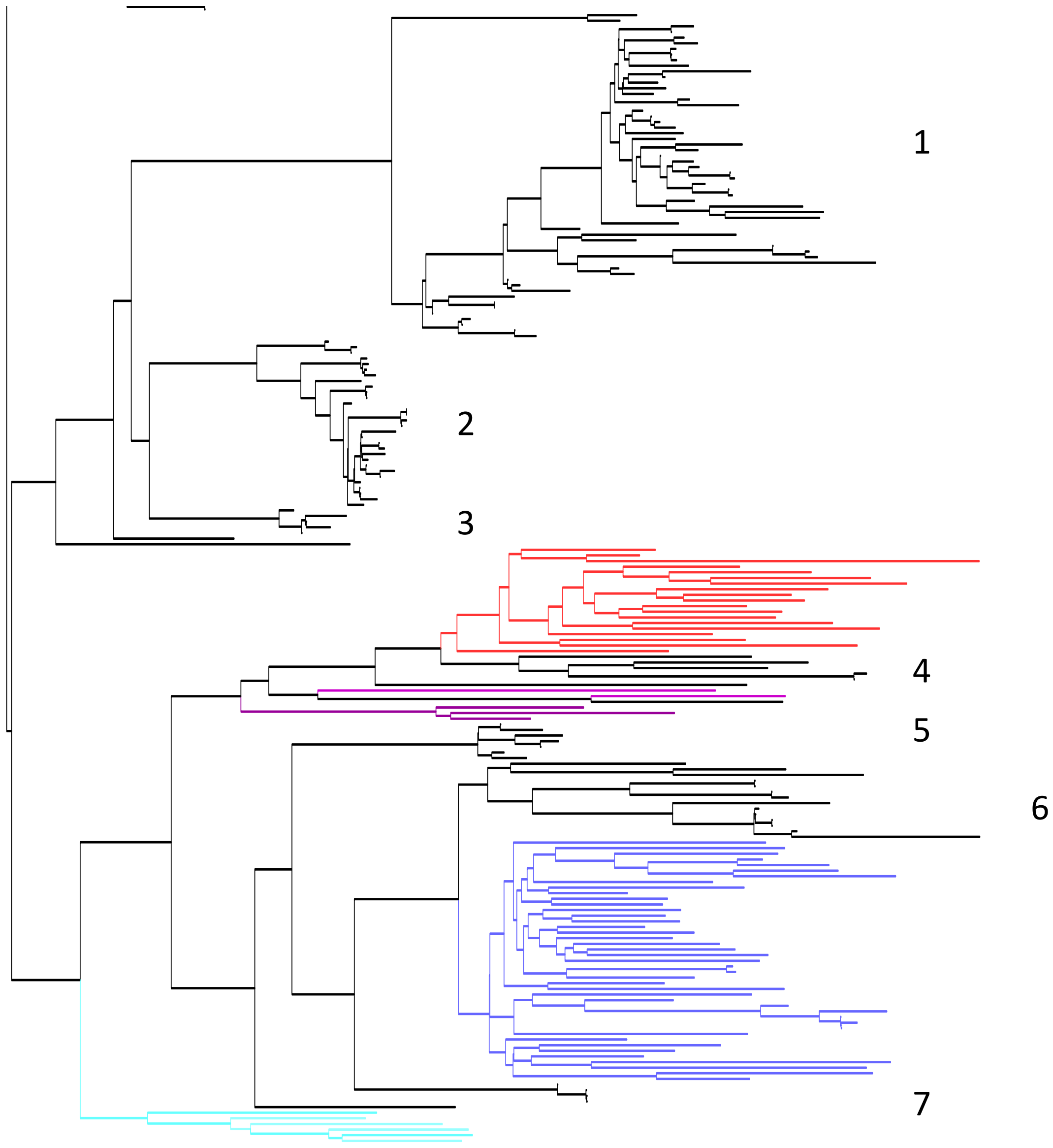


Figure 3_6. Putative novel CRF branches. Enlarged section of Figure 3_3 showing the region below the subtype B reference sequences. Seven branches indicated potential novel CRF clusters. Branches 1-3 were located between the subtype B sequences and the remainder of the reference sequences, branch 4 was located close to the subtype A sequences (in red), branches 5, 6 and 7 were located close to the subtype C sequences (in blue). H/J/K reference sequences were coloured purple, and subtype D sequences were coloured turquoise.

3.2.3 Cluster screening

Cluster screening was performed on the whole tree and was reviewed alongside the SCUEAL 2012 and jpHMM genotyping results in order to identify groups of closely related sequences with identical recombinant structures. Maximum genetic distance measurements of the approximate maximum likelihood tree showed 2088 clusters with GDs ranging from 0 to 15.9%. The D'Agostino-Pearson K2 value was 182.0, indicating that the GD of the clusters were not normally distributed. Overall, 1959 of these had a maximum GD of 10% or less; 658 clusters were comprised of two sequences and 139 clusters contained a pure subtype reference sequence. After further screening using the SCUEAL 2012 and jpHMM results to exclude those clusters containing sequences now classified as pure subtype B strains and sequences now classified as other pure subtype strains, 281 clusters remained. These 281 clusters were screened according to recombinant breakpoint location and subtype classification; following this process 28 clusters remained. The maximum genetic distance in any cluster was 9.0% (Table 3_4). Some clusters displayed recombinant structures similar (but not identical) to recognised CRF structures, notably clusters 3 and 4 (CRF 24-like), 8 (CRF17, 38, 39, 44-like), 10 (CRF03), and 19 and 20 (CRF31-like), respectively. The sequences identified as part of Cluster 9 contained some sequences that belonged to an already-classified CRF (CRF50; see subsequent chapters for details of the identification of CRF50_A1D). In order to confirm that Cluster 9 did not contain sequences from two very closely related (but separate) CRFs, the CRF50 sequences were excluded from Cluster 9 and the remaining sequences were subjected to the same analysis process as the other clusters.

Cluster number	Recombinant Structure	Max GD (%)	No. of members
1	G/B	8.6	26
2	G/B	7.4	15
3	G/B	7.4	12
4	G/B	3.0	6
5	G/B	1.2	4
6	G/F/B	0.5	3
7	F/B	4.9	4
8	F/B	5.9	14
9	D/A	6.8	60
10	D/A	5.2	12
11	D/A/B	1.6	3
12	01/B	6.5	5
13	A/B	6.5	8
14	B/A/J/G	2.8	6
15	B/C	0.3	5
16	B/C	5.0	4
17	B/G/B	3.0	7
18	B/01	2.1	3
19	C/B/C	3.2	7
20	B/C/B	1.5	5
21	B/U	1.1	4
22	B/A	4.4	4
23	B/A	3.6	3
24	B/A	1.5	5
25	B/A/B	8.1	21
26	C/B/C	6.4	5
27	C/B/C	1.2	4
28	G_complex	9.0	23

Table 3_4. Potential CRF clusters identified through recombinant structure and maximum genetic distance screening. 28 potential clusters were identified and selected for phylogenetic analysis of putative pure subtype regions using BEAST. Genetic distance measurements and recombinant structure were based on partial *pol* sequences captured by the UK HIV DRD.

3.2.4 Phylogenetic and demographic analyses

Phylogenetic screening was performed using the putative pure subtype regions of each cluster suggested by the jpHMM and SCUEAL genotyping results. Please note that in the following cluster descriptions, the full breakpoint results refer to sequences definitively identified as belonging to a cluster, not the complete list of putative cluster members listed in Table 3_4.

Following phylogenetic analysis, BEAST analysis identified 15 confirmed clusters comprised of sequences with novel recombinant structures, comprising 94 individuals in total (Table 3_5). Relating the confirmed clusters back to the original approximate maximum likelihood tree (Figure 3_3) showed that Cluster 1 was group 2 in Figure 3_4, Cluster 2 was group 6 in Figure 3_4, Cluster 3 was group 1 in Figure 3_4, Cluster 5 was group 4 in Figure 3_6, Cluster 6 was group 1 in Figure 3_5, Cluster 15 was group 7 in Figure 3_6, Cluster 19 was group 6 in Figure 3_6, Cluster 25a was group 3 in Figure 3_6, Cluster 25b was group 2 in Figure 3_6 and Cluster 25c was located in group 1 in Figure 3_6. Cluster 27 was group 5 in Figure 3_6. The remaining identified clusters were located in un-enlarged branches.

The largest cluster contained 23 members; 1/15 (6.7%) clusters contained 10 and eight members, respectively, 2/15 (13.3%) contained six members, 4/15 (26.6%) contained five members, 3/15 (20%) contained four members and 3/15 (20%) clusters contained three members. The genetic distance range was 0.2-5%.

Overall, 58/94 (61.7%) of cluster members were male, 14/94 (14.9%) were female, 16/94 (17.0%) were IVDU, 36/94 (38.3%) were MSM and 15/94 (16.0%) were heterosexual. 58/94 (61.7%) cluster members were white, 10/94 (10.6%) were Black-African and 1/94 (1.1%) was Black-Caribbean. When compared to the proportions seen among individuals infected with pure subtypes and recognised CRFs, the proportion of males in clusters was significantly higher (61.7% vs. 39.5%, p= <0.01), as was the proportion of IVDUs (17.0% vs. 1.5%, p= <0.01) and people of white ethnicity (61.7% vs. 31.4%, p= <0.01).

The proportion of females in clusters was not significantly different compared to individuals infected with pure subtypes and recognised CRFs (14.9% vs. 15.3%, p=0.93), as was the proportion of MSMs (38.5% vs. 28.7%, p= 0.17), the proportion of heterosexuals (16.0% vs. 23.0%, p= 0.23), the proportion of Black-Africans (10.6% vs. 16.3%, p= 0.25) or the proportion of Black-Caribbeans (1.1% vs. 1.9%, p= 0.85).

Overall, 3/15 (20%) of clusters showed exposure profiles composed of >1 exposure route. One cluster was mixed IVDU/heterosexual, composed exclusively of white ethnicity and contained both males and females (Cluster 3). One cluster (6.7%) was mixed Heterosexual/MSM and was composed exclusively of white men (Cluster 15).

Two clusters were comprised exclusively of sequences from heterosexuals (Cluster 2 and Cluster 14); both clusters contained both male and female members. Cluster

2 was comprised exclusively of people of white ethnicity, and Cluster 14 showed Black-African members only. Cluster 14 had a complex recombinant structure that did not allow for evolutionary analysis prior to full-length sequencing.

Overall, 2/4 sequences in Cluster 2 belonged to white heterosexual men and 2/4 sequences belonged to white heterosexual women. The geographic region identified with these sequences was exclusively Southwest England/Wales.

Eight clusters were comprised exclusively of MSM sequences, of which 5/8 were exclusively white, 2/8 were mixed white/Black-African and one was mixed white/Black-African/Black-Caribbean; 1/5 white MSM cluster, comprising 10 members, was located in London/Southeast England and Northwest England (cluster 25a). Another white MSM cluster containing three members was located in London/SE England and Scotland; 1/5 white MSM clusters were located in Northwest England (Cluster 5), and a further 1/5 white MSM cluster was located in both Northwest England and Scotland (Cluster 25b). The remaining white MSM cluster was located in East Anglia, a region of the UK immediately north of London.

Of the three MSM clusters showing mixed ethnicity, 2/3 were located exclusively in London (Clusters 13 and 19); the remaining MSM cluster with mixed ethnicity was located in SW England and Wales (Cluster 18).

Finally, extra-UK geographic links were found in 5/15 (33.4%) (Clusters 1, 2, 15, 25, 19).

Please see below for the detailed cluster analysis. Following this analysis, full-genome sequencing of the identified clusters was required to confirm the recombinant structures as novel CRF strains of HIV-1.

Cluster	No.	Max GD (%)	Structure	Breakpoint location (HXB2)	Risk group	Gender	Ethnicity	Location
1	23	3.5	G/B	2985	IVDU, MSM, Het	M/F	White/ mixed	London/ SE England
2	4	0.2	G/B	2406	Het	M/F	White	SW England/ Wales
3	5	2.8	CRF14/B	2563	IVDU, Het	M/F	White	SW England/ Wales
5	4	1.2	G/?B	3049	MSM	M	White	NW England
6	3	0.5	G(or related)/F/ B	2500/2698	IVDU	M/F	White	East Anglia
13	6	0.87	A1/B	2716	MSM	M	White/ Black African/ Black Caribbean	London/ SE England
14	6	2.8	B/AE/J/ CRF18	2354/2815/ 3093	Het	M/F	Black-African	London/ SE England
15	5	0.3	B/C	2589	Het, MSM	M	White	London/ SE England and NW England
16	4	5.0	CRF07/C	3079	MSM	M	White	East Anglia
18	3	2.1	B/AE	3231	MSM	M	White/ Black-African	SW England/ Wales
19	5	1.8	C/B/C	2726/3089	MSM	M	White/ Black-African	London/ SE England
25a	10	2.1	B/CRF50/ B	2555/3048	MSM	M	White	London/ SE England
25b	8	1.5	B/CRF50/ B	2545/2941	MSM	M	White	Scotland/ NW England
25c	3	0.4	CRF50/B	2632/2943	MSM	M	White	London/ SE England
27	5	4	C/B/C	2377/2719	U	U	U	Mixed

Table 3_5. Summary table of likely novel CRF clusters identified through BEAST phylogenetic analysis. 15 clusters were identified that may comprise novel CRFs. Further confirmation using full-genome screening is necessary before cluster sequences can be registered as novel CRFs.

3.2.4.1 Cluster 1 (G/B)

Cluster 1 had 26 members and a maximum genetic distance of 8.6% (Table 3_4). The alignment slice was placed at HXB2 2985, the most commonly predicted breakpoint (jpHMM, 19/26 sequences).

The tree for slice 1 showed 24/26 sequences clustering monophyletically with a posterior probability of 1 (Figure 3_7a). The sequences clustered closest to subtype G/CRF14. Slice 2 showed 24/26 sequences clustering monophyletically with a posterior probability of 0.85. One of the non-clustering sequences was the same sequence as in slice 1 and one was different; when these were excluded from the cluster, the posterior probability decreased to 0.73 (Figure 3_7b). None of the three non-clustering sequences had a predicted breakpoint at HXB2 2985. The 24 sequences clustered closest to subtype B with moderate support (0.63). The three sequences that did not exhibit consistent monophyletic clustering were excluded from the cluster; this reduced the maximum genetic distance to 3.5%.

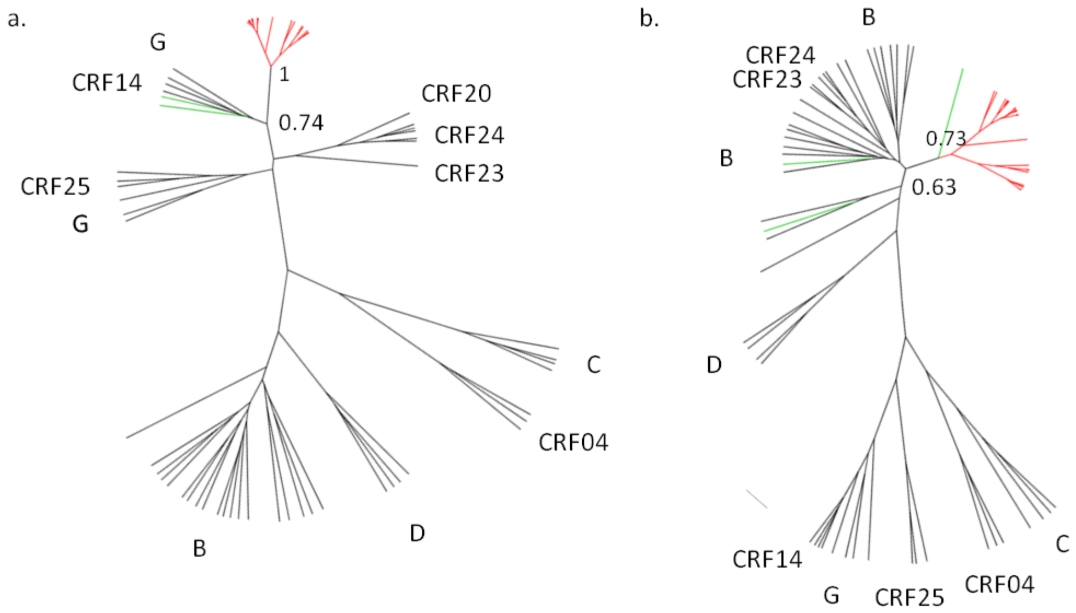


Figure 3_7. Cluster 1: G and B fragments. Maximum clade credibility trees of putative pure subtype regions. The breakpoint was placed at HXB2 2985. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. Sequences that do not exhibit monophyletic clustering are highlighted in green. a) Slice 1; b) Slice 2.

The full breakpoint results showed that jpHMM identified 22/23 sequences as G/B recombinant and 1/23 as B/G/B recombinant (Table 3_6). The breakpoint predictions were consistent and had an interval of 40 nucleotides (HXB2 2956 –

HXB2 2996). SCUEAL 2012 consistently predicted 3 breakpoints: HXB2 2585-2630; HXB2 2705-2827 and HXB2 2956-2959.

No.	jpHMM GT	jpHMM breakpoint (95% CI)	SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)			
1	G/B	2985 (2942-3028)	G/B	CRF14/CRF24/CRF14/ B/B	2585 (2584-2586)	2719 (2718-2720)	2956 (2955-2957)	3226 (3224-3228)
2	G/B	2968 (2941-2995)	G/B	CRF14/B/G/B	2605 (2591-2619)	2723 (2722-2724)	2959 (2957-2961)	
3	B/G/B	2717 (2669-2765) 2979 (2942-3016)	G/B	G/G/B		2759 (2747-2771)	2956 (2955-2957)	
4	G/B	2996 (2943-3049)	G/B	CRF14/G/CRF14/B/ CRF20/B	2605 (2604-2606)	2705 (2704-2706)	2959 (2958-2960)	3223 (3222-3224) 3381 (3380-3382)
5	G/B	2983 (2941-3024)	G/B	CRF14/B/G/B/B	2605 (2588-2622)	2759 (2758-2760)	2959 (2958-2960)	3382 (3381-3383)
6	G/B	2985 (2941-3029)	G/B	G/B/G/B	2611 (2598-2624)	2729 (2727-2730)	2959 (2958-2960)	
7	G/B	2985 (2942-3028)	G/B	G/G/B	2630 (2610-2650)	2830 (2829-2831)		
8	G/B	2985 (2942-3028)	G/B	CRF14/G/CRF14/B	2605 (2598-2600)	2705 (2699-2711)	2959 (2958-2960)	
9	G/B	2957 (2940-2974)	G/B	G/CRF28/G/B	2605 (2593-2617)	2723 (2722-2724)	2959 (2958-2960)	
10	G/B	2985 (2942-3028)	G/B	CRF14/B/G/B/CRF20/B	2605 (2604-2606)	2723 (2722-2724)	2959 (2958-2960)	3205 (3203-3207) 3381 (3375-3387)
11	G/B	2985 (2942-3028)	G/B	CRF14/CRF28/G/B/B	2605 (2586-2624)	2723 (2722-2724)	2959 (2958-2960)	3382 (3381-1130)
12	G/B	2965 (2940-2989)	G/B	G/CRF28/G/B	2558 (2551-2565)	2723 (2721-2725)	2959 (2958-2960)	
13	G/B	2956 (2941-2971)	G/B	G/B/G/B	2605 (2601-2609)	2759 (2758-2760)	2959 (2958-2960)	
14	G/B	2985 (2942-3028)	G/B	CRF14/B/G/B	2520 (2519-2521)	2723 (2722-2724)		3004 (2999-3009)
15	G/B	2985 (2942-3028)	G/B	G/B/G/B/B/B	2558 (2557-2559)	2729 (2728-2730)	2959 (2958-2960)	3170 (3149-3193) 3381 (1127-3382)
16	G/B	2971 (2927-3015)	G/B	G/G/CRF14/B	2605 (2600-2610)	2705 (2704-2706)	2959 (2956-2962)	
17	G/B	2985 (2942-3028)	G/B	CRF14/B/G/B	2520 (2516-2524)	2723 (2721-2725)	2959 (2958-2960)	
18	G/B	2985 (2942-3028)	CRF23-like	CRF23/G/B		2720 (2696-2744)	2959 (2957-2961)	
19	G/B	2985 (2942-3028)	G/B	G/B/B		2827 (2826-2828)		3064 (3047-3081)
20	G/B	2985 (2942-3028)	G/B	G/B			2959 (2955-2963)	
21	G/B	2985 (2942-3028)	G/B	CRF14/G/CRF14/B/CRF20	2605 (2602-2608)	2705 (2704-2706)	2959 (2955-2963)	3226 (3225-3227)
22	G/B	2967 (2922-3012)	G/B	CRF14/CRF24/G/B/CRF20	2585 (2584-2586)	2720 (2719-2721)	2959	3145 (3144-3146)
23	G/B	2980 (2942-3018)	G/B	G/B/G/B	2597 (2587-2607)	2729 (2728-2730)	2959 (2958-2960)	

Table 3_6. Full breakpoint results for Cluster 1. jpHMM and SCUEAL 2012 breakpoint results for Cluster 1. 20/23 and 23/23 sequences, respectively, had a SCUEAL or jpHMM predicted breakpoint in the region of HXB2 2956-2985.

The 23 remaining members of the cluster showed a mixed exposure profile, comprising 15/23 (65.2%) IVDU (11/15 male, 4/15 female), 2/23 (8.7%) MSM, and 2/23 (8.7%) female heterosexuals. Overall, 20/23 (87.0%) members were white and 1/23 (4.3%) showed a mixed ethnicity. Available geographic information showed the cluster was concentrated in London and Southeast England (21/23, 87.0%), with one member in Scotland and one in Southwest England/Wales.

Six additional sequences with identical recombination structures and a breakpoint at HXB2 2985 were identified using HIV BLAST; 5/6 were from Portugal and 1/6 was from Spain. MCMC analysis showed a time to most recent common ancestor (tMRCA) in the UK of 1999.72 (95% highest posterior density (HPD) = 1997.70 - 2002.05); branches with other G/B recombinant sequences from Portugal and Spain showed tMRCA of 1989.14 (95% HPD = 1984.14 - 1994.92) and possible import of this recombinant structure into the UK from this region (Figure 3_8).

The results suggested a novel recombinant spreading primarily in white IVDU with MSM and female heterosexual involvement, which was imported into the UK from Spain or Portugal.

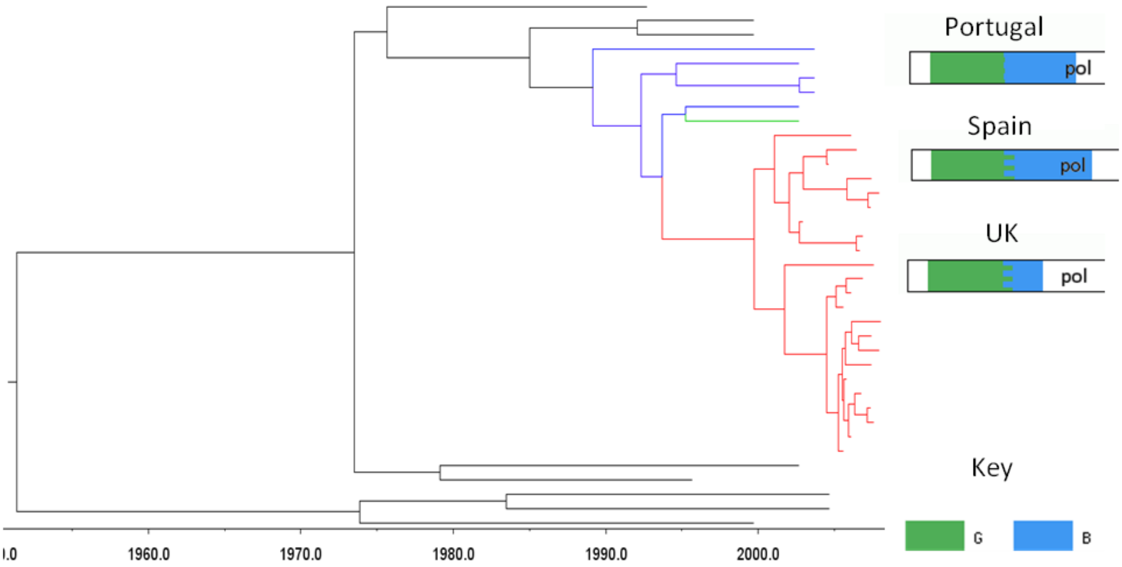


Figure 3_8. Time scaled phylogeny of Cluster 1, a mixed MSM, IVDU and heterosexual cluster. MCMC analysis of a UK-based novel subtype G/B recombinant cluster, plus strains showing identical recombination profiles from Portugal and Spain. Sequences from the UK are shown in red, sequences from Portugal are shown in blue, and sequences from Spain are shown in green. Subtype reference B, G and CRF14 sequences are shown in black. Later tMRCAs of the UK sequences relative to the Spanish and Portuguese sequences indicate possible import into the UK from this region.

3.2.4.2 Cluster 2 (G/B)

Cluster 2 had 15 members and a maximum genetic distance of 7.4%. 9/15 sequences had a breakpoint at HXB2 2406 and the alignment slice was placed at this position. The trees of slices 1 and 2 showed 4/15 sequences that consistently clustered together with a posterior probability of 1 in each tree (Figure 3_9a and 3_9b). In slice 1 this was closest to subtype G (posterior probability = 0.62) and in slice 2 this was closest to subtype B (posterior probability = 0.81). Restricting the cluster to these four specimens reduced the genetic distance of the cluster to 0.2%

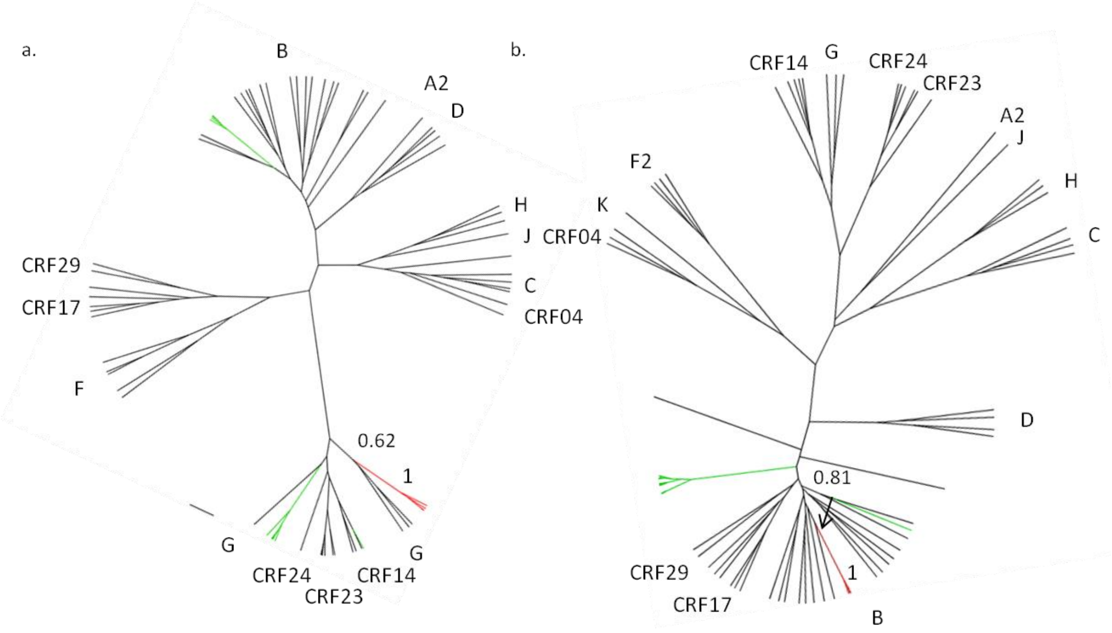


Figure 3_9. Cluster 2: G and B fragments. Maximum clade credibility trees of putative pure subtype regions. The breakpoint was placed at HXB2 2406. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. Sequences that do not exhibit monophyletic clustering are highlighted in green. a) Slice 1; b) Slice 2.

The full breakpoint results showed no clear consensus in the jpHMM results, with 2/4 sequences showing an uncertain subtype classification, and no clear consensus regarding breakpoint location (Table 3_7). The SCUEAL 2012 results showed closely related breakpoints in 3/4 sequences, and matching simplified subtypes in 3/4 cases.

No.	jpHMM GT	jpHMM breakpoint (95% CI)	SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)	
1	U/B	2334 (2253-2415)	G/B	G/B	2427 (2395-2459)	
2	G/B	2396 (2369-2423)	G/B	G/B	2377 (2322-2432)	
3	G/B	2401 (2368-2434)	G/B	G/B/CRF29	2425 (2423-2427)	2582 (2566-2598)
4	UB	2344 (2253-2435)	F2/B	F2/B/CRF29	2437 (2426-2448)	2564 (2257-2571)

Table 3_7. Full breakpoint results for Cluster 2. The SCUEAL 2012 breakpoints show a consistent breakpoint close to position HXB2 2427 in 3/4 sequences. This is not reflected in the jpHMM results, which show uncertain classification in 2/4 sequences and no clear consensus on breakpoint location.

2/4 sequences in Cluster 2 belonged to white heterosexual men and 2/4 sequences belonged to white heterosexual women. The geographic region identified with these sequences was exclusively Southwest England/Wales.

The HIV BLAST search identified four sequences sharing a recombinant structure and breakpoint location; 3/4 were from Portugal and 1/4 from Luxembourg. The time-scaled phylogeny showed a tMRCA for the UK sequences of 2004.78 (95% HPD = 2003.97-2011.41) (Figure 3_10). The tMRCA for the Portuguese and Luxembourgish sequences was 2004.60 (95% HPD = 2004.56-2011.08); the almost identical tMRCAs suggested co-circulation of this potential novel CRF in the UK, Portugal and Luxembourg.

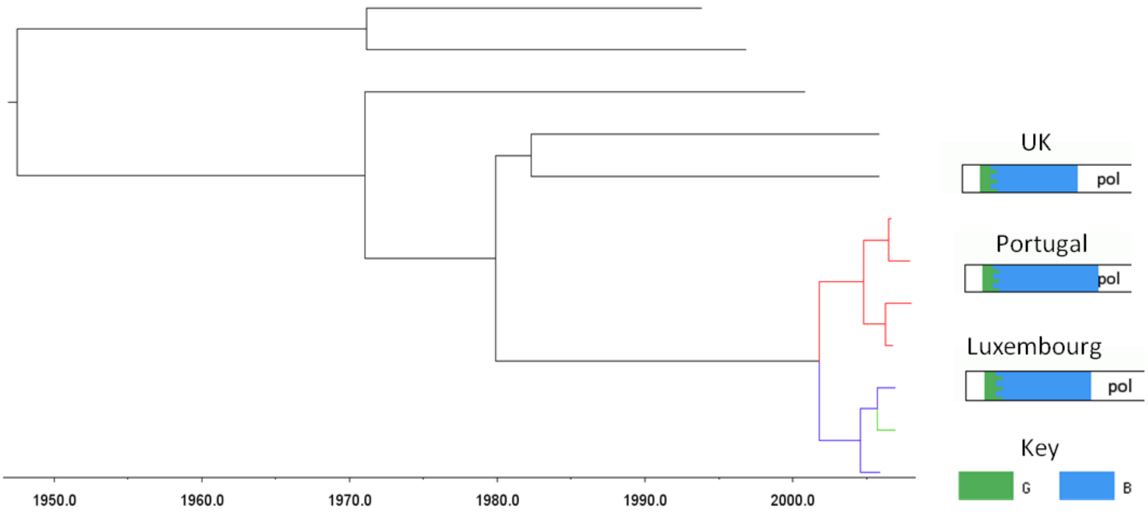


Figure 3_10. Time scaled phylogeny of Cluster 2, a heterosexual cluster. MCMC analysis of a novel subtype G/B recombinant cluster, plus strains showing identical recombination profiles from Portugal and Luxembourg. Sequences from the UK are shown in red, sequences from Portugal are shown in blue, sequences from Luxembourg are shown in green. Subtype reference B and G sequences are shown in black. Branches showing almost identical tMRCAs of this recombinant in the UK and Portugal indicates co-circulation of this recombinant strain in three distinct geographic locations, with probable import from an unknown third location.

3.2.4.3 Cluster 3 (G/B)

Cluster 3 had 12 members and a maximum genetic distance of 7.4%. The alignment slice was placed at position HXB2 2563. 5/12 sequences clustered together consistently in slices 1 and 2 (Figure 3_11a and 3_11b). In slice 1 this was closest to CRF14 (posterior probability = 0.71) and in slice 2 it was closest to subtype B (posterior probability = 0.86). The posterior probability for the clustered sequences was 1 in both trees. Restricting the cluster to these sequences reduced the GD to 2.8%.

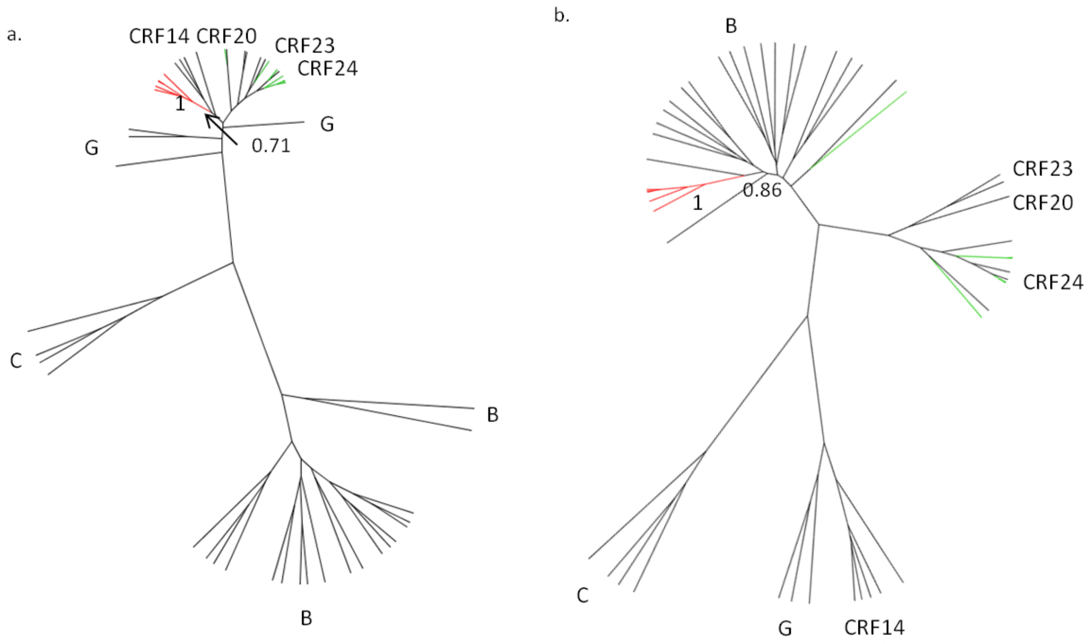


Figure 3_11. Cluster 3: G and B fragments. Maximum clade credibility trees of putative pure subtype regions. The breakpoint was placed at HXB2 2563. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. Sequences that do not exhibit monophyletic clustering are highlighted in green. a) Slice 1; b) Slice 2.

Full breakpoint results showed consistent predicted recombinant structure and breakpoint locations in both the jpHMM and SCUEAL 2012 results, with a breakpoint located between HXB2 2504 and 2552 (Table 3_8).

No.	jpHMM GT	jpHMM breakpoint (95% CI)	SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)	
1	G/B	2516 (2503-2529)	G/B	CRF14/B	2552 (2540-2564)	
2	G/B	2528 (2504-2552)	G/B	G/B	2505 (2483-2527)	
3	G/B	2504 (2435-2573)	G/B	G/B/B	2530 (2525-2535)	3154 (3151-3157)
4	G/B	2528 (2504-2552)	G/B	G/B	2505 (2484-2526)	
5	G/B	2524 (2503-2545)	G/B	CRF14/B/B/B	2539 (2537-2541)	2843 (2833-2853) 3364 (3363-3365)

Table 3_8. Full breakpoint results for Cluster 3. A clear consensus on recombinant structure is shown by both SCUEAL 2012 and jpHMM results. The SCUEAL breakpoints are located at positions HXB2 2505, HXB2 2530, HXB2 2539 and HXB2 2552, consistent with the jpHMM breakpoints.

There were 3/5 sequences with demographic information available; 1/3 sequences was from a white heterosexual woman and 2/3 were from white male IVDU. Available geographical information showed the cluster was concentrated in Southwest England and Wales.

The HIV BLAST search did not locate any matching recombinant structures. These results suggested that this cluster formed a novel recombinant spreading primarily in white male IVDU with white female heterosexual involvement, with no evidence of extra-UK circulation.

3.2.4.4 Cluster 4 (G/B)

Cluster 4 was composed entirely of sequences that were analysed as part of cluster 3. These sequences did not exhibit consistent clustering.

3.2.4.5 Cluster 5 (G/B)

Cluster 5 had four members and a maximum genetic distance of 1.2%. Overall, 3/4 sequences had an identical SCUEAL breakpoint at HXB2 3049; accordingly, the slice position was placed there. All four sequences clustered consistently in slices 1 and 2 with a posterior probability of 1 (Figure 3_12a and 3_12b). In slice 1 the sequences clustered closest to subtype G and related CRFs (posterior probability = 1), but the tree of slice 2 showed weak support for clustering with subtype B with a posterior probability of 0.55. This may have been related to the short fragment length of slice 2 of 240 nucleotides.

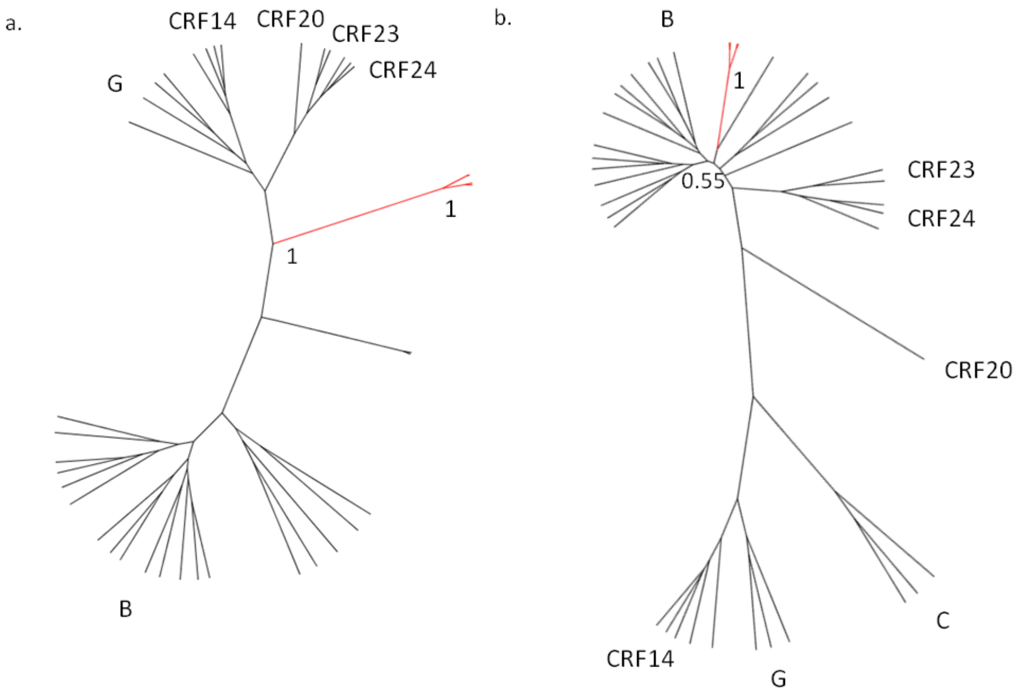


Figure 3_12. Cluster 5: G and B fragments. Maximum clade credibility trees of putative pure subtype regions. The breakpoint was placed at HXB2 3049. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. a) Slice 1; b) Slice 2.

Full breakpoint results showed consistent predicted recombinant structure between the two genotyping methods, but little agreement on breakpoint location (Table 3_9). The SCUEAL breakpoint, which was used for the cluster analysis, was approximately 250 nucleotides distant from the jpHMM breakpoint, indicating that sequencing of a longer region additional was necessary to enable precise breakpoint identification.

Overall, 2/4 sequences had associated demographic data; both were white male MSM. The geographic region associated with these sequences was Northwest England. The HIV BLAST search did not identify any similar recombinant structures.

These results suggested a novel G/B recombinant circulating among white MSM; however, longer sequence lengths were required to confirm the genotyping of the putative subtype B region.

No.	jpHMM GT	jpHMM breakpoint (95% CI)	SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)		
1	G/B	2759	G/B	G/B		3049 (2998-3100)	
2	G/B	2772 (2760-2784)	G/B	G/CRF03		3049 (3033-3045)	
3	G/B	2778	G/B	G/B/B	2885 (2875-2895)	3001 (2996-3006)	
4	G/B	2886	G/B	G/B/B		3049 (3044-3054)	3378 (3368-3388)

Table 3_9. Full breakpoint results for Cluster 5. Both jpHMM and SCUEAL 2012 results show consistent predicted recombinant structures. The SCUEAL 2012 breakpoint at 3049 was approximately 250 nucleotides distant from the jpHMM breakpoints.

3.2.4.6 Cluster 6 (G/F/B)

Cluster 6 had three members and a maximum genetic distance of 0.5%. Alignment slices were placed at HXB2 2500 and 2968. All three sequences clustered together consistently in each slice with posterior probabilities of 1 (Figure 3_13a, 3_13b, 3_13c). Slice 1 clustered with subtype G and related CRFs (posterior probability = 1), slice 2 clustered with subtype F1 (posterior probability = 0.99) and slice 3 clustered with subtype B (posterior probability = 1).

The full breakpoint results showed consistent recombinant structure and breakpoint locations between both genotyping methods, which were consistent with the results shown in the trees (Table 3_10).

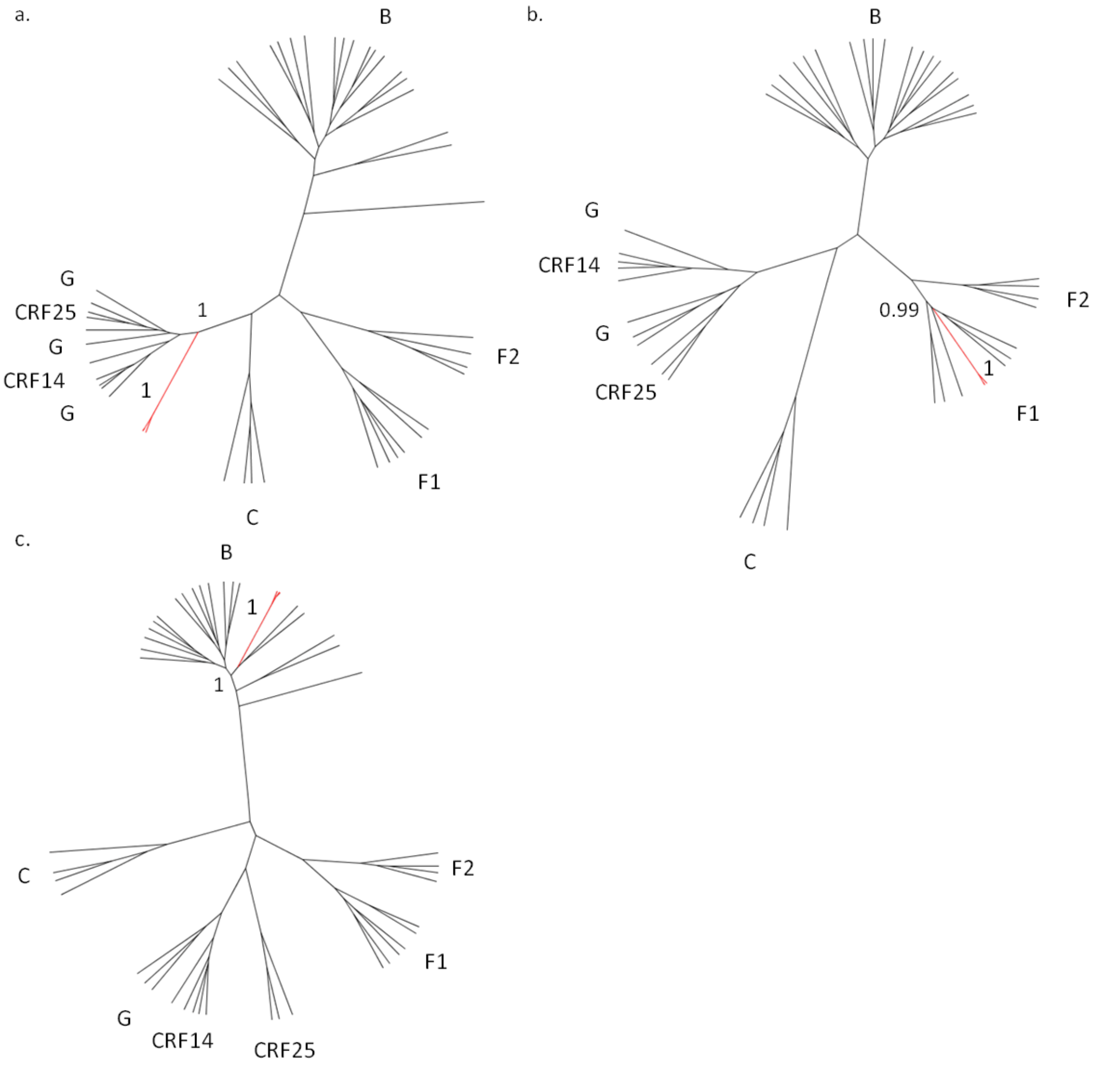


Figure 3_13. Cluster 6: G, F and B fragments. Maximum clade credibility trees of putative pure subtype regions. The breakpoints were placed at HXB2 2500 and 2968. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. a) Slice 1; b) Slice 2; c) Slice 3.

No.	jpHMM GT	jpHMM breakpoint (95% CI)		SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)	
1	G/F1/B	2495 (2426-2564)	2968 (2955-2981)	Complex	CRF25/F1/B	2369 (2352-2386)	2975 (2974-2976)
2	G/F1/B	2503 (2426-2580)	2968 2955-2981)	Complex	CRF25/F1/B	2444 (2443-2445)	2975 (2973-2976)
3	G/F1/B	2500 (2426-2574)	2967 (2955-2979)	Complex	CRF14/F1/B	2450 (2438-2462)	2977 (2976-2978)

Table 3_10. Full breakpoint results for Cluster 6. Both jpHMM and SCUEAL results show consistent predicted recombinant structures and breakpoint locations. The highest sequence position for the first SCUEAL is HXB2 2450; the second predicted breakpoint is at HXB2 2975.

The HIV BLAST search identified nine sequences that matched to all three cluster members. Two sequences were G/F/B recombinants from Spain and Macau, and seven sequences were F1/B recombinants from Argentina. jpHMM analysis of the breakpoints showed one G/F/B sequence with breakpoints at HXB2 2813 and 2968, and eight F/B sequences with 5/8 breakpoints at HXB2 2992, two breakpoints at HXB2 2990, one breakpoint at 2979 and one breakpoint at 3034. A time-scaled phylogeny including these sequences was attempted; inadequate mixing between the chains and subsequent low ESS were observed. This indicated that further analysis without full-length sequencing of the genome was unlikely to yield satisfactory results.

There were 2/3 sequences with demographic information available; 1/3 sequences was from a white woman and 1/3 was from a white male. Both were IVDU. The geographic location for all three sequences was East Anglia.

These results suggested a novel CRF circulating in IVDU in East Anglia. Some evidence was identified that suggests these sequences may be related to sequences isolated in Spain and Argentina, but sequencing of a longer genomic region is required for confirmation.

3.2.4.7 Cluster 7 (F/B)

Cluster 7 had four members and a maximum genetic distance of 4.9%. The alignment slice was placed at position HXB2 2963. When split into component subtype regions, none of the four sequences clustered together, indicating that these sequences did not share a common evolutionary path, and therefore did not comprise a novel CRF.

3.2.4.8 Cluster 8 (F/B)

Cluster 8 had 14 members and a maximum genetic distance of 5.9%. Of these, 7/14 sequences had a breakpoint located at HXB2 2486 and the alignment slice was placed at this position; 4/7 sequences clustered consistently with CRF44 in both trees; 1/7 sequences did not cluster with any other sequences, and 2/7 sequences clustered as a pair in both trees.

There were 3/14 sequences (Cluster 8a) that clustered in both trees with a posterior probability of 1. These sequences clustered weakly with subtype F1 in tree 1 (posterior probability = 0.44) and subtype CRF29 in tree 2 (posterior probability = 1). Demographic information was available for 2/3 sequences; both sequences belonged to white MSM. The geographic location was the south coast of England and Wales for all three sequences.

There were 4/14 sequences (Cluster 8b) that clustered in both trees with posterior probabilities of 0.83 and 1, respectively. These sequences clustered closest to subtype CRF29 in tree 1 and subtype B/CRF28/CRF29 in tree 2 (with weak support).

Both trees were very weakly supported in the subtype F and CRF29 branches of tree one and the subtype B and CRF29 branches of tree two. Combined with the position of the likely breakpoint at HXB2 2486, only 86 nucleotides distant from the F/B breakpoint of CRF29, the most likely explanation was that these strains were closely related to CRF29 rather than a true novel CRF.

3.2.4.9 Cluster 9 (D/A)

After excluding obvious CRF50 sequences, Cluster 9 had eight members and a maximum genetic distance of 6.0%. The alignment slice was placed at HXB2 2545. All eight sequences clustered with CRF50 sequences. These results did not suggest a further novel CRF, but instead sequences related to CRF50.

3.2.4.10 Cluster 10 (D/A)

Cluster 10 had 12 members and a maximum genetic distance of 5.2%. Overall, 6/12 sequences had a jpHMM breakpoint at HXB2 2610 and the alignment slice was placed here. All 12 sequences clustered consistently with CRF50 sequences in both slices.

3.2.4.11 Cluster 11 (D/A/B)

Cluster 11 had three members and a maximum genetic distance of 1.6%. The sequences had jpHMM breakpoints at HXB2 2553 and 3298; the second breakpoint would have resulted in a too-short fragment for analysis so the end of the alignment was trimmed to HXB2 3298 and the alignment slice was placed at HXB2 2553. All four sequences clustered with CRF50 in both trees.

3.2.4.12 Cluster 12 (CRF01/B)

Cluster 12 had five members and a maximum genetic distance of 6.5%. The alignment slice was placed at position HXB2 2945. The sequences did not cluster monophyletically in either tree.

3.2.4.13 Cluster 13 (A/B)

Cluster 13 had eight members and a maximum genetic distance of 6.5%. Of these, 4/8 sequences had a breakpoint at HXB2 2716 and the alignment slice was placed here; 6/8 sequences clustered monophyletically in trees 1 and 2 with posterior probabilities of 1 (Figure 3_14a and 3_14b). In tree 1 the sequences clustered strongly with subtype A1 (posterior probability = 0.97) and in tree 2 the sequences clustered with subtype B (posterior probability = 0.87). Restricting the cluster to these sequences reduced the maximum genetic distance of the cluster to 0.87%.

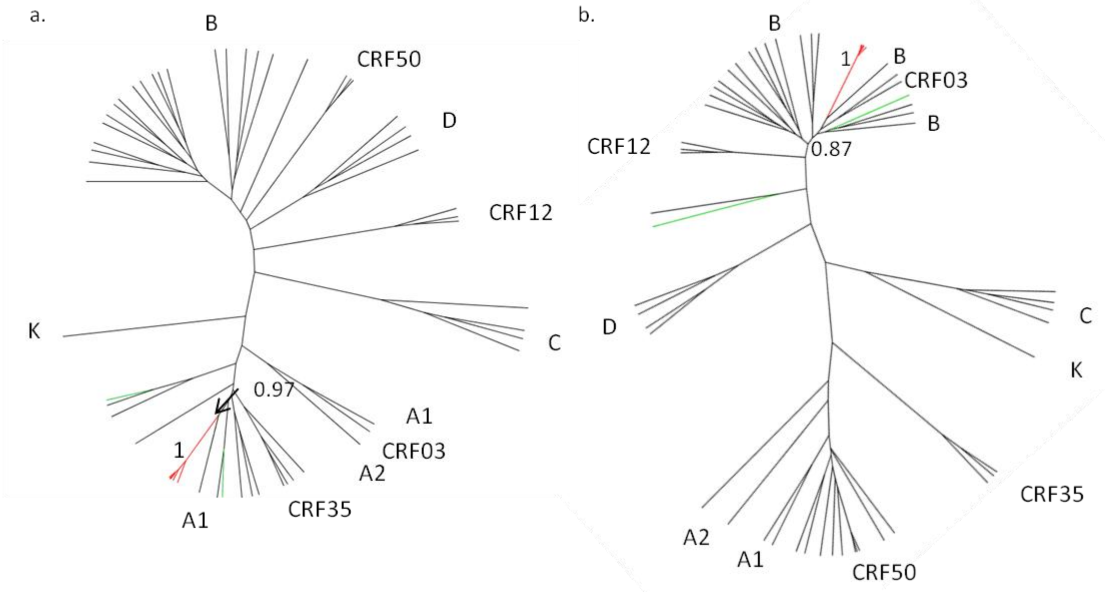


Figure 3_14. Cluster 13: A and B fragments. Maximum clade credibility trees of putative pure subtype regions. The breakpoint was placed at HXB2 2716. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red; sequences that do not cluster monophyletically are shown in green. a) Slice 1; b) Slice 2

The full breakpoint results showed consistent predictions in jpHMM and SCUEAL results (Table 3_11); 6/6 recombinant structures were classified as A1/B by jpHMM, as were 4/6 sequences by SCUEAL. The remaining 2/6 sequences were classified by SCUEAL as complex, due to the prediction of a subtype D fragment in 1/6 sequences and subtype K and CRF12 fragments in 1/6 sequences. The SCUEAL-predicted breakpoint at HXB2 2738 was consistent with the jpHMM-predicted breakpoint of 2716 in 4/6 sequences.

Overall, 5/6 sequences were MSM and 1/6 was of an unknown risk group; 3/6 were white, 2/6 were Black-African and 1/6 was Black-Caribbean. The geographic location of all six was London and Southeast England. The time scaled phylogeny showed a tMRCA of 2005.52 in the UK, indicating a novel recombinant located primarily in white and Black-African MSM (Figure 3_15). No evidence was found to indicate a geographic origin of the recombinant outside the UK.

No.	jpHMM GT	jpHMM breakpoint (95% CI)	SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)		
1	A1/B	2689 (2617-2761)	Complex	A1/B/D		2738 (2729-2747)	3457 (3456-3458)
2	A1/B	2716 (2654-2778)	Complex	K/A1/B/CRF12	2354 (2351-2357)	2738 (2737-2739)	3428 (3425-3431)
3	A1/B	2705 (2624-2786)	A1/B	A1/B/CRF12		2771 (2749-2793)	3428 (3420-3436)
4	A1/B	2716 (2654-2778)	A1/B	A1/B/B		2738 (2730-2746)	3452 (3436-3468)
5	A1/B	2716 (2654-2778)	A1/B	A1/B/B		2738 (2737-2739)	3337 (3314-3360)
6	A1/B	2716 (2654-2778)	A1/B	A1/A1/B/CRF12	2389 (2388-2390)	2752 (2747-2756)	3428 (3427-3429)

Table 3_11. Full breakpoint results for Cluster 13. 4/6 predicted recombinant structures were identical in both jpHMM and SCUEAL 2012 results. The jpHMM results showed consistent breakpoint predictions across all six sequences; the SCUEAL breakpoint in 4/6 sequences was HXB2 2738.

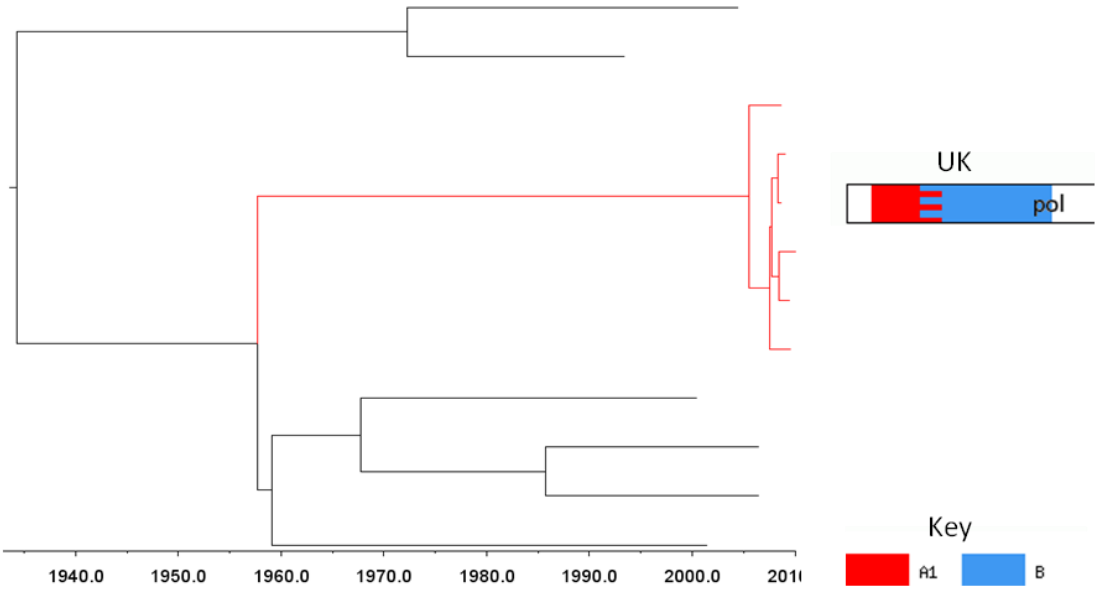


Figure 3_15. Time scaled phylogeny of Cluster 13, a MSM cluster of mixed ethnicity. MCMC analysis of a novel subtype A1/B recombinant cluster. Sequences from the UK are shown in red. Subtype reference A1 and B sequences are shown in black.

3.2.4.14 Cluster 14 (B/A/J/G)

Cluster 14 had six members and a maximum genetic distance of 2.3%. All six sequences were designated 'complex' by SCUEAL 2012 and the suggested recombinant structure contained three breakpoints in the 1301bp *pol* fragment analysed. Overall, 5/6 sequences had a first breakpoint at HXB2 2354; subsequent breakpoints were located at HXB2 2815 and HXB2 3093, respectively. Owing to the short length of fragment 1, the first two sequence regions were analysed as one fragment; slices 3 and 4 were analysed separately.

In all three phylogenetic trees (slices 1+2, slice 3, slice 4, Figure 3_16a, 3_16b and 3_16c, respectively), all six sequences clustered together monophyletically with a posterior probability of 1. In the tree of slices 1 and 2 the sequences were clustered closest to subtypes CRF01_AE and CRF22 (posterior probability = 0.93), in the tree of slice 3 the sequences were clustered closest to subtype J (posterior probability = 1), and in the tree of slice 4 the sequences were clustered closest to subtypes K and CRF18 (posterior probability = 0.72).

The full breakpoint results showed relatively consistent breakpoint locations in both the jpHMM and SCUEAL 2012 results, but no clear consensus as to the predicted recombinant structure (Table 3_12). Although the simplified SCUEAL results designated all six sequences as complex, the detailed subtype results showed no consensus either within the other SCUEAL results, or between the SCUEAL and jpHMM results.

Demographic information was available for 6/6 sequences. All six individuals were heterosexual Black-African; 4/6 were female and 2/6 were male. The geographic location was London and Southeast England for 5/6 sequences and Southwest England and Wales for 1/6 sequences. The HIV BLAST search returned two sequences that were present in 6/6 closest sequence matches: DQ886038, an A/K/U recombinant from the USA and GQ241034, a CRF45 sequence from Spain.

These results suggest these sequences form a novel recombinant, however due to the complexity of the putative recombinant structure further work to confirm this is needed. No clear evidence was present to suggest that this recombinant is circulating in regions other than the UK.

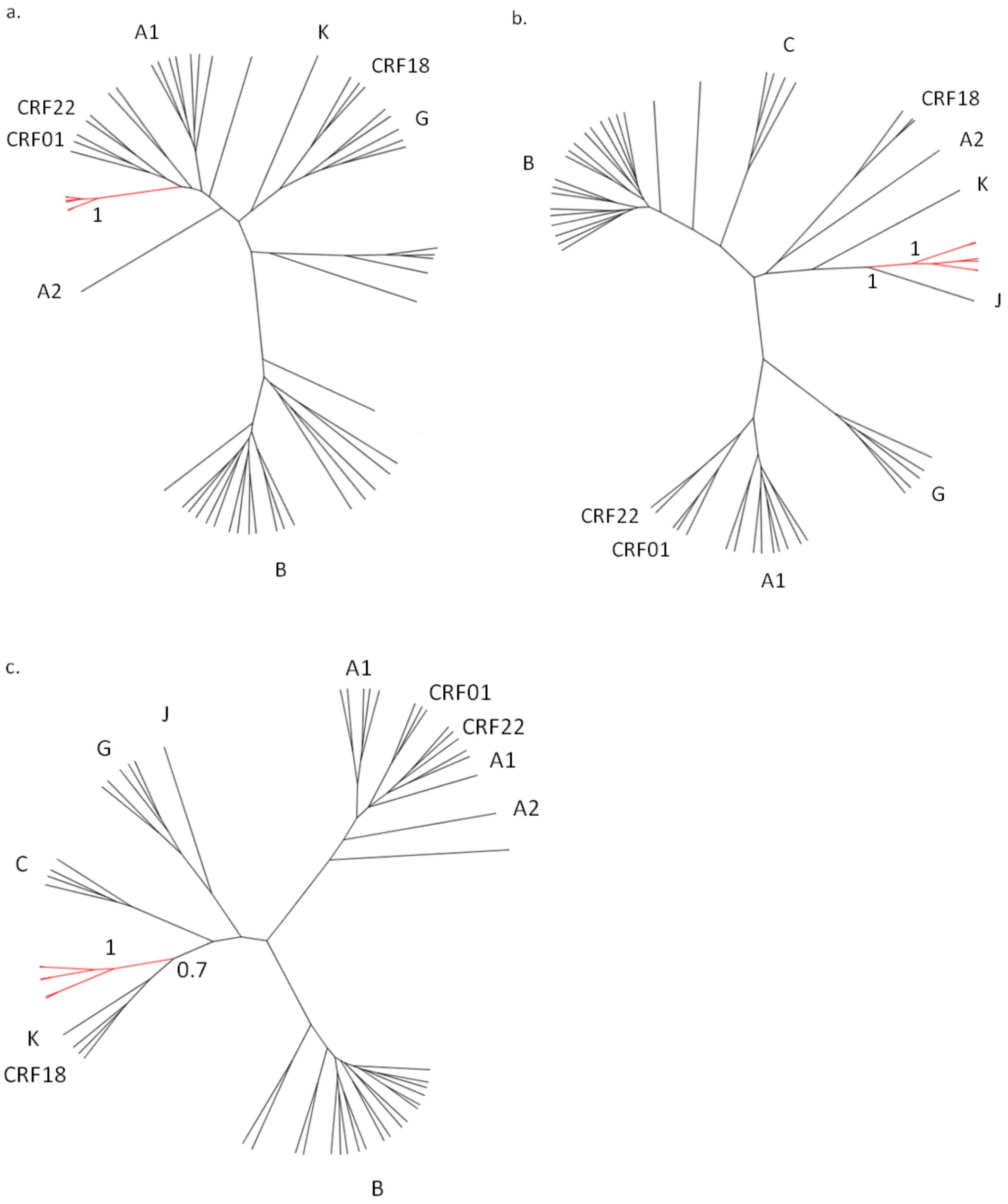


Figure 3_16. Cluster 14: B, A, B, J, G recombinant fragments. Maximum clade credibility trees of putative pure subtype regions. Breakpoints were placed at HXB2 2354, 2815 and 3093, respectively. Owing to the short length of the first fragment, slices 1 and 2 were analysed in a single tree. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. a) Slices 1 and 2; b) Slice 3; c) Slice 4.

No.	jpHMM GT	jpHMM breakpoint (95% CI)			SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)			
1	U/A1/J/G	2305 (2253-2357)	2815 (2790-2840)	3110	Complex	U/AE/A1/J/B/J		2420 (2419-2421)	2669 (2668-2670) 2791 (2790-2792)	3118 (3117-3119) 3217 (3216-3218)
2	J/A1/J/G	2356 (2326-2386)	2808 (2782-2834)	3092 (3074-3110)	Complex	B/AE/J/CRF18	2354 (2353-2355)		2806 (2790-2832)	3118 (3117-3119)
3	B/A1/J/B	2370 (2313-2427)	2815 (2790-2840)	3093 (3070-3116)	Complex	B,A,AE,CRF22,J,CRF18	2354 (2353-2355)	2512 (2508-2516) 2613 (2604-2622)	2791 (2790-2792)	3113 (3112-3114)
4	B/A1/J/G	2369 (2313-2425)	2810 (2783-2837)	3110	Complex	B/AE/J/K/J	2354 (2353-2355)		2806 (2804-2808)	3118 (3117-3119) 3223 (3222-3224)
5	01/J/G	2503 (2253-2753)	2795 (2754-2836)	3110	Complex	B/AE/J/J/J	2354 (2353-2355)		2756 (2755-2757)	3056 (3055-3057) 3167 (3055-3279)
6	01/J/G		2782 (2754-2810)	3092 (3070-3114)	Complex	B/A/AE/J/B/J	2354 (2353-2355)	2485 (2484-2486)	2789 (2788-2790)	3124 (3121-3125) 3223 (3222-3224)

Table 3_12. Full breakpoint results for Cluster 14. The predicted results show a complex recombinant structure, with no clear consensus as to the constituent subtypes. The breakpoint locations in jpHMM are relatively consistent. Breakpoints appear consistently in SCUEAL at positions 2354 (5/6 sequences), 2789-2806 (5/6 sequences) and 3113-3124 (5/6 sequences).

3.2.4.15 Cluster 15 (B/C)

Cluster 15 had five members and a maximum genetic distance of 0.3%. All five recombinant structures had a jpHMM breakpoint at HXB2 2589. The MCMC tree of slice 1 showed monophyletic clustering, closest to subtype B, with a posterior probability of 1; the support for subtype B was moderate, with a posterior probability of 0.54 (Figure 3_17a). The tree of slice 2 also showed monophyletic clustering with a posterior probability of 1; this was closest to subtype C (Figure 3_17b).

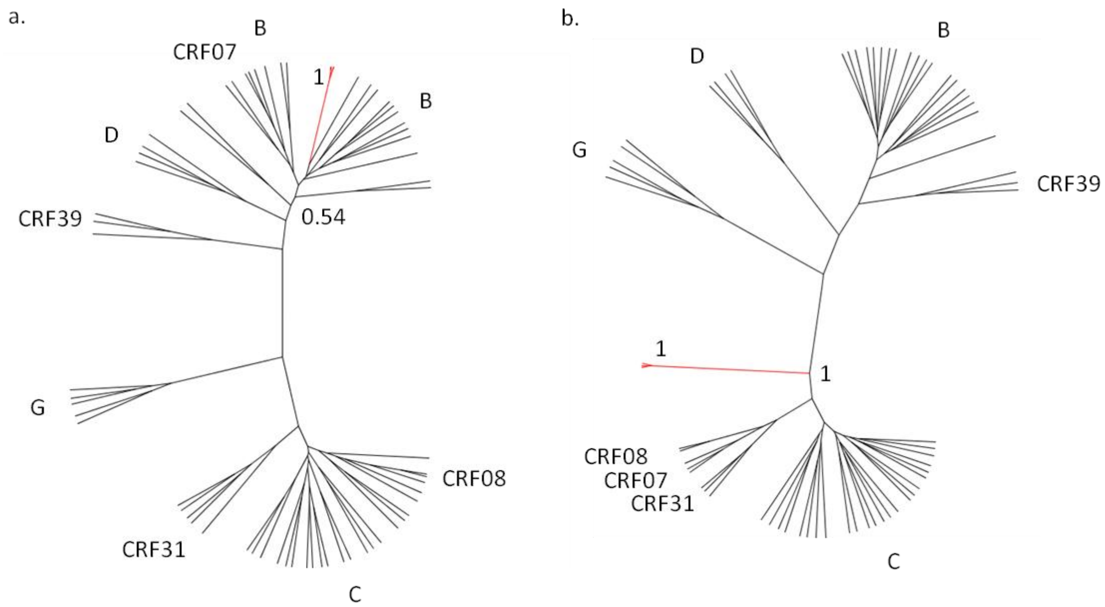


Figure 3_17. Cluster 15: B and C fragments. Maximum clade credibility trees of putative pure subtype regions. The breakpoint was placed at HXB2 2589. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. a) Slice 1; b) Slice 2.

The full breakpoint results showed identical recombinant structure predictions from both jpHMM and SCUEAL 2012 (Table 3_13). The SCUEAL breakpoint was placed at HXB2 2523, only 66 nucleotides distant from the jpHMM predicted breakpoint of HXB2 2589.

Demographic information was available for 4/5 sequences and showed 4/4 white men; 3/4 were heterosexual and 1/4 was MSM. Geographic data was available for 5/5 sequences and showed 3/5 from London and Southeast England and 2/5 from Northwest England.

The time scaled phylogeny showed a cluster with a tMRCA of 2004.86 (95%HPD = 2003.36-2007.34) (Figure 3_18). This cluster contained sequences from the UK, Spain, and Paraguay; all sequences shared identical breakpoints and recombinant structures. No close relationship seen with the sequences from Brazil. The branch containing sequences from the UK contained sequences from Spain and Paraguay, suggesting co-circulation of this recombinant structure in three countries.

No.	jpHMM GT	jpHMM breakpoint (95% CI)	SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)		
1	B/C	2589 (2464-2714)	B/C	B/C		2523 (2512-2534)	
2	B/C	2589 (2464-2714)	B/C	B/C/CRF31		2522 (2521-2523)	3432 (3424-3440)
3	B/C	2589 (2464-2714)	B/C	B/B/C	2376 (2366-2386)	2522 (2521-2523)	
4	B/C	2589 (2464-2714)	B/C	CRF39/B/C/CRF31	2418 (2417-2419)	2522 (2521-2523)	3432 (3424-3440)
5	B/C	2589 (2464-2714)	B/C	B/C		2520 (2505-2535)	

Table 3_13. Full breakpoint results for Cluster 15. Both jpHMM and SCUEAL show a B/C predicted recombinant structure; the breakpoint was at 2589 in all jpHMM results and between HXB2 2520 and 2523 in all SCUEAL results.

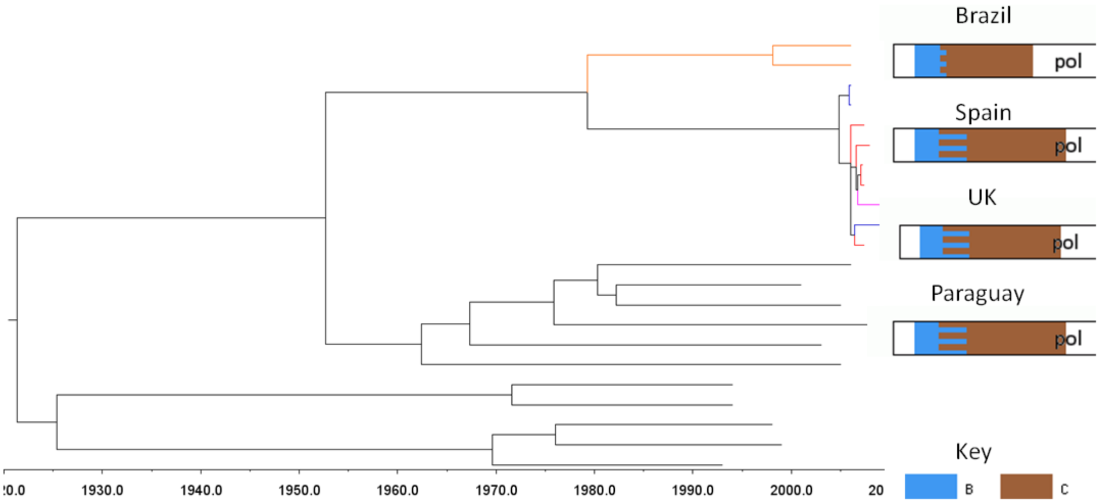


Figure 3_18. Time scaled phylogeny of Cluster 15, a mixed MSM/heterosexual cluster. MCMC analysis of a novel subtype B/C recombinant cluster, plus strains showing identical recombination profiles from Spain, Paraguay and Brazil. Sequences from the UK are shown in red, sequences from Spain are shown in blue, sequences from Brazil are shown in orange and sequences from Paraguay are shown in lavender. Subtype reference B, C, and F1

sequences are shown in black. Clustering of the Spanish, UK and Portuguese sequences indicates co-circulation of this recombinant strain in three distinct geographic locations.

3.2.4.16 Cluster 16 (B/C)

Cluster 16 had four members and a maximum genetic distance of 5.0%. The breakpoint was placed at HXB2 3079 (shared by 2/4 sequences). The tree of slice 1 showed monophyletic clustering with a posterior probability of 0.97, closest to subtype CRF07 (posterior probability = 0.99, Figure 3_19a). The tree of slice 2 also showed monophyletic clustering with a posterior probability of 0.93; this was in subtype C (posterior probability = 0.99, Figure 3_19b).

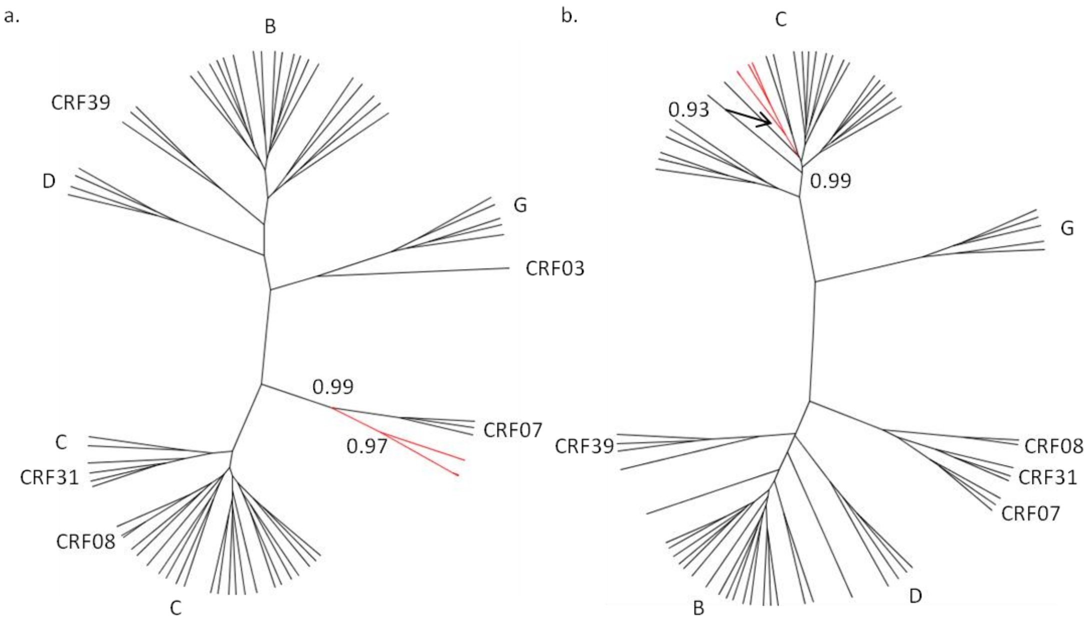


Figure 3_19. Cluster 16: CRF07 and B recombinant fragments. Maximum clade credibility trees of putative pure subtype regions. The breakpoint was placed at HXB2 3079. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. a) Slice 1; b) Slice 2.

The full breakpoint results showed a predicted B/C structure in 3/4 jpHMM results and 3/4 SCUEAL 2012 results (Table 3_14). There was a jpHMM-predicted breakpoint at HXB2 3077-3079 in 3/4 sequences; these sequences all had a SCUEAL-predicted breakpoint at HXB2 3077-3083.

No	jpHMM GT	jpHMM breakpoint (95% CI)	SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)		
1	B/C	3079 (3068-3090)	B/C	B/B/C/ CRF03	2611 (2597-2625)	3083 (3082-3084)	3245 (3244-3246)
2	B/C	3079 (3068-3090)	B/C	B/C/B/C/ CRF03	2611 (2597-2625) 2710 (2709-2711)	3083 (3082-3084)	3259 (3258-3260)
3	B/C	3077 (3068-3086)	Complex	B/C/B/D	2564 (2563-2565) 2719 (2718-2720) 2918 (2917-2919)	3077 (3076-3079)	
4	B/C/B	2607 (2560-2654) 3404 (3390-3421)	B/C	B/C/B	2611 (2597-2625)		3394 (3386-3402)

Table 3_14. Full breakpoint results for Cluster 16. Both the jpHMM and the SCUEAL results showed consistent breakpoint prediction in 3/4 sequences. The SCUEAL-predicted breakpoint at HXB2 3077-3083 is consistent with the jpHMM-predicted breakpoints.

Demographic information was available for 3/4 sequences; 4/4 were white MSM. Geographic location was available for 4/4 sequences; this showed 3/4 in East Anglia and 1/4 in London and Southeast England.

The HIV BLAST search showed seven sequences that appeared in at least 75% of the closest sequence match lists. They were: JN223163, a B/C recombinant from Myanmar; GQ303845 and HM468715, subtype B sequences from Canada; JN944923, a subtype B sequence from the USA; JX299665, a subtype B sequence from Germany; GQ399355, a subtype B sequence from Portugal, and EU611261, a subtype B sequence from an unrecorded location. jpHMM analysis of the BC recombinant sequence showed a pure subtype B sequence.

These results suggested these sequences form a novel recombinant located in white MSM. No evidence indicated this recombinant had extra-UK circulation.

3.2.4.17 Cluster 17 (B/G/B)

Cluster 17 had seven members and a maximum genetic distance of 3.0%. Both jpHMM and SCUEAL showed relatively consistent breakpoint location results (jpHMM: HXB2 2409/2952/3110; SCUEAL 2377/2578/3212). Owing to the short fragment length that would result from slicing the sequences at HXB2 3212, a single slice was made at HXB2 2954. Both trees showed all seven sequences clustering together consistently, however in both instances the clustering was with CRF02_AG sequences.

3.2.4.18 Cluster 18 (B/CRF01)

Cluster 18 had three members and a maximum genetic distance of 2.1%. The jpHMM and SCUEAL analyses showed conflicting results regarding these sequences: jpHMM returned results of B, B/U and A1/B, respectively, whilst SCUEAL gave 3/3 results of B/AE with a breakpoint at HXB2 3145 (Table 3_15). The slice was placed here for the phylogenetic analysis.

No.	jpHMM GT	jpHMM breakpoint (95% CI)	SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)
1	B/U	3203 (3155-3251)	B/AE	B/AE	3145 (3128-3162)
2	B		B/AE	B/AE	3145 (3129-3161)
3	A1/B	3122	B/AE	B/AE	3145 (3096-3200)

Table 3_15. Full breakpoint results for Cluster 18. No clear agreement existed between the jpHMM and SCUEAL-predicted results. The SCUEAL results show clear concordance.

The tree from slice 1 showed all three sequences clustering together with a posterior probability of 1 (Figure 3_20a). The sequences clustered with subtype B with a posterior probability of 0.64; the posterior probability for the B branches was 0.97. The tree from slice 2 also showed clustering with a posterior probability of 1; in this fragment the sequences clustered with subtype A1 (posterior probability = 0.52, Figure 3_20b). The query sequences did not cluster near the CRF01 sequences. The region of the slice 2 tree containing the A1, CRF01, CRF15, CRF33 and CRF34 sequences was not well-supported; a longer fragment length than 146 nucleotides may have shown a clearer relationship between the query sequences and the reference sequences.

Demographic information was available for 3/3 sequences. This showed 3/3 sequences belonging to MSM; 2/3 were white and 1/3 was Black-African. The geographic region was Southwest England and Wales for 2/3 sequences and the Midlands for 1/3 sequences.

The HIV BLAST search showed four sequences that were present in at least 75% of the closest sequences matches. Three of these were subtype B submissions from the UK HIV DRD (JN101126, JN101698, JN101327) and one was a subtype B sequence from the Netherlands (JQ650838).

These results indicated that these sequences comprise a novel CRF circulating in the UK only.

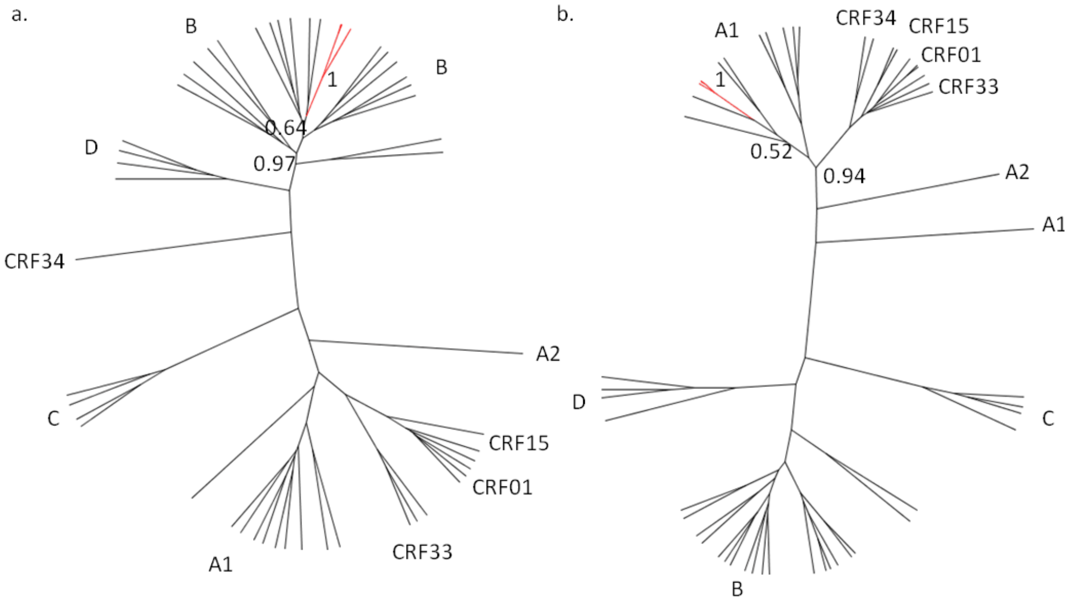


Figure 3_20. Cluster 18: B and A1 recombinant fragments. Maximum clade credibility trees of putative pure subtype regions. The breakpoint was placed at HXB2 3145. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. a) Slice 1; b) Slice 2.

3.2.4.19 Cluster 19 (C/B/C)

Cluster 19 had seven members and a maximum genetic distance of 3.2%. Breakpoints were placed at HXB2 2726 and 3089, which was consistent with both the jpHMM and SCUEAL subtyping results (Table 3_16). The tree for slice 1 showed monophyletic clustering of all seven sequences closest to subtype C with a posterior probability of 1 (Figure 3_21a); the posterior probability of the clustering with subtype C was 0.73. The tree for slice 2 showed 5/7 sequences clustering monophyletically with subtype B (Figure 3_21b, posterior probability = 1, posterior probability of clustering with subtype B = 0.25), and the remaining 2 sequences continuing to cluster closer to subtype CRF31 (posterior probability = 0.99, posterior probability of clustering with CRF31 = 0.23). This tree was very poorly supported, with only the subtype C, D, F1 and G branches showing strong support for the tree topology.

The tree for slice three showed all seven sequences clustering monophyletically with subtype C (Figure 3_21c, posterior probability = 1).

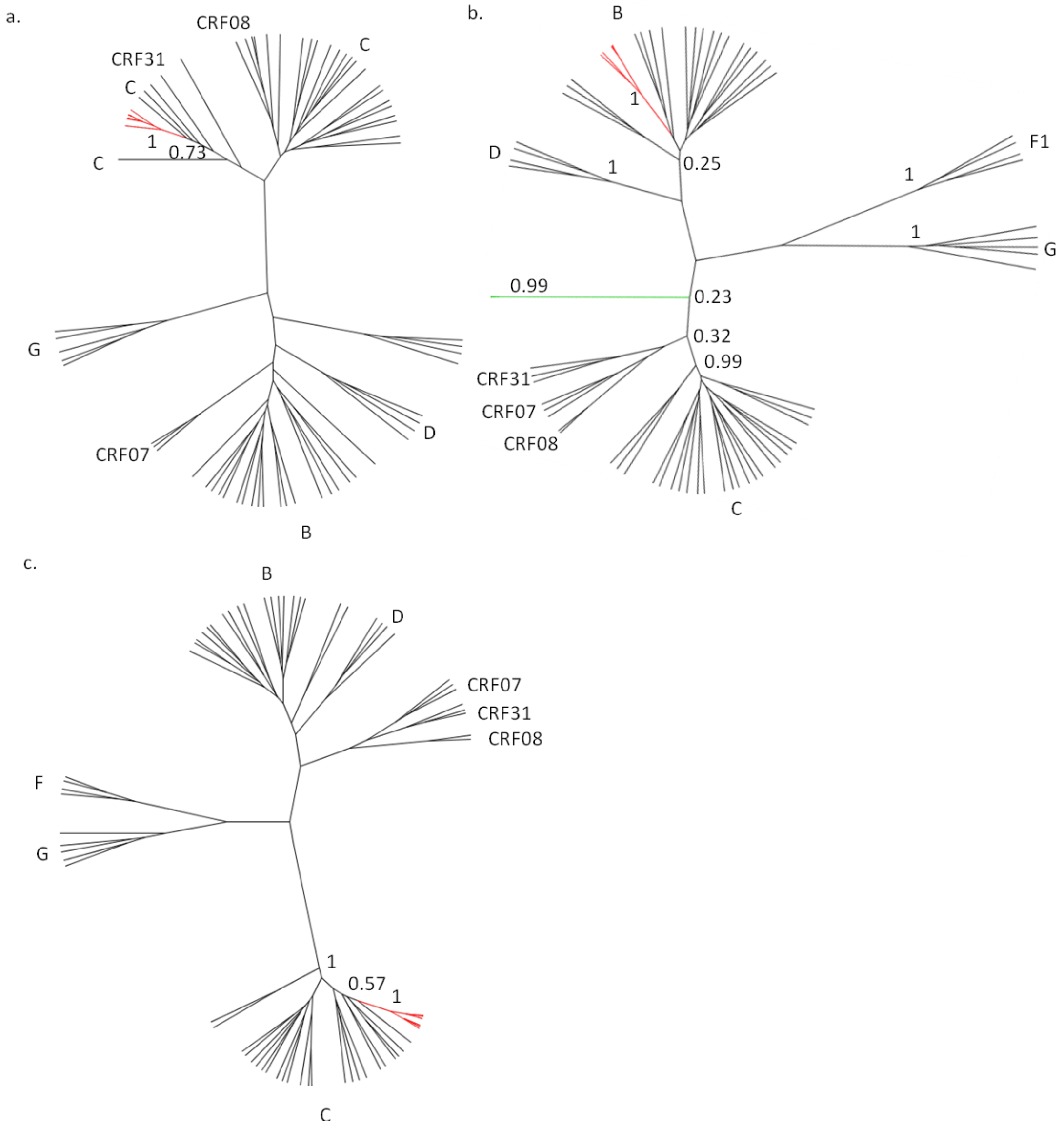


Figure 3_21. Cluster 19: C and B recombinant fragments. Maximum clade credibility trees of putative pure subtype regions. Breakpoints were placed at HXB2 2726 and 3089. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. a) Slice 1; b) Slice 2; c) Slice 3.

No.	jpHMM GT	jpHMM breakpoint (95% CI)		SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)			
1	C/B/C	2716 (2692-2740)	3089 (3072-3106)	B/C	CRF31/CRF31/B/C	2573 (2528-2618)	2723 (2722-2724)	3100 (3099-3101)	
2	C/D/C	2726 (2712-2740)	3095 (3087-3103)	C/U	CRF31/CRF31/U/C/C	2573 (2571-2575)	2743 (2742-2744)	3100 (3095-3105)	3209 (3208-3210)
3	C/B/C	2716 (2692-2740)	3089 (3072-3106)	B/C	CRF31/CRF31/B/C	2573 (2540-2606)	2723 (2720-2726)	3100 (3098-3101)	
4	C/B/C	2720	3099 (3092-3106)	B/C	CRF31/CRF31/B/C/CRF31	2573 (2561-2585)	2723 (2720-2726)	3100 (3099-3101)	3208 (3207-3209)
5	C/B/C	2723	3100 (3092-3108)	B/C	CRF31/B/C/CRF31		2743 (2742-2744)	3087 (3086-3088)	3373 (3343-3403)

Table 3_16. Full breakpoint results for Cluster 19. The jpHMM results show consistent predictions for breakpoint locations within a 10 nucleotide interval. The SCUEAL breakpoints at HXB2 2723-2743 and HXB2 3087-3100, which are present in 5/5 sequences, are consistent with the jpHMM results.

Demographic information was available for 4/5 sequences; 4/4 were from men, 3/4 were MSM, 3/4 were white, 1/3 were from UK, 1/3 from Denmark, 1/4 Black-African from Nigeria. The geographic location of the 5 sequences showed 4/5 in London and Southeast England and 1/5 in Southwest England.

The time scaled phylogeny included seven sequences with identical breakpoints and recombinant structures, all from Italy (Figure 3_22). The tMRCA for the branch containing the UK sequences was 2003.19 (95% HPD = 1998.98-2009.40). The Italian sequences were interspersed with the UK sequences, indicating co-circulation of this recombinant in the UK and Italy, and possible export from the UK to Italy of this recombinant.

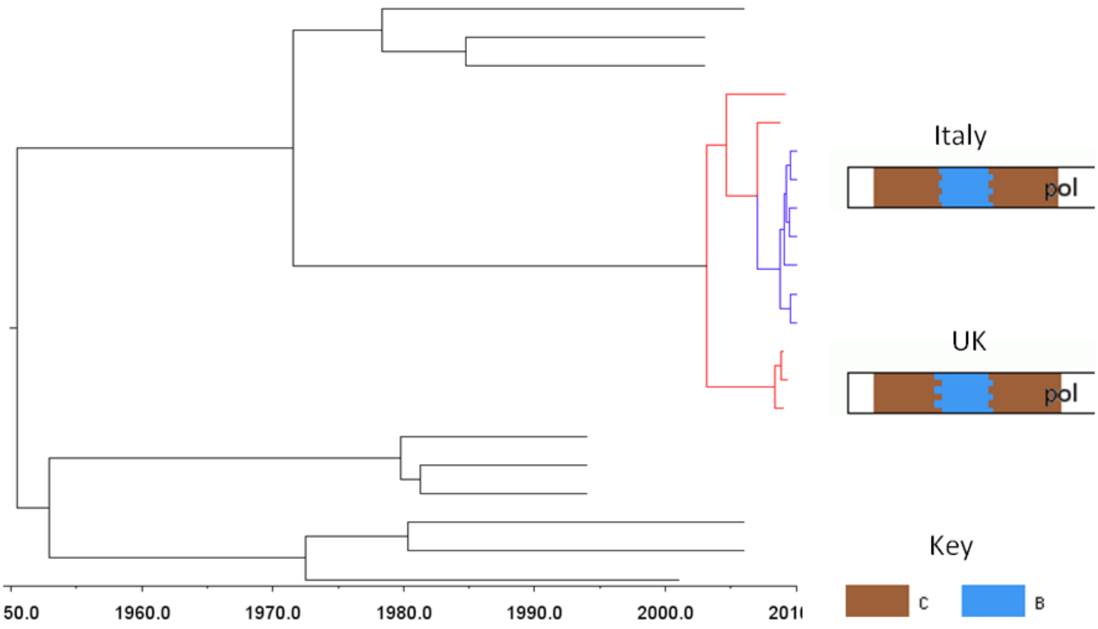


Figure 3_22. Time scaled phylogeny of Cluster 19, a MSM cluster of mixed ethnicity. MCMC analysis of a recombinant C/B/C cluster. Sequences from the UK are shown in red. Sequences from Italy are shown in blue. Later tMRCAs of the Italian sequences relative to the UK sequences indicate possible export from the UK to Italy, and subsequent co-circulation of this strain.

3.2.4.20 Cluster 20 (B/C/B)

Cluster 20 had five members and a maximum genetic distance of 1.5%. Breakpoints were placed at HXB2 2386 and 2704. The tree for slice 1 showed 4/5 sequences clustering weakly in subtype C (posterior probability of 0.51) and 1/5 sequences clustering elsewhere in subtype C. The tree for slice 2 showed all five sequences clustering together in subtype B with a posterior probability of 1, and the tree for slice 3 also showed monophyletic clustering with a posterior probability of 1, in this case closest to subtype C. The fragment length of slice 1 was very short (133 bp), which may account for the weak clustering behaviour. However, the jpHMM result for these sequences showed 2/5 as having B/C/B structures, 2/5 as C/B structures and 1/5 as B only. It is more likely that these differences are the source of the weak clustering in slice 1 and that these sequences are somewhat related in terms of transmission, but not representative of a novel CRF.

3.2.4.21 Cluster 21 (B/U)

Cluster 21 had four members and a maximum genetic distance of 1.1%. 3/4 sequences showed an identical SCUEAL breakpoint at sequence position 212 (HXB2 2465) and the breakpoint was placed here to slice the alignment. The tree for slice 1 showed monophyletic clustering with a posterior probability of 1; this was weakly associated with subtype D (posterior probability of 0.40). The tree for slice 2 also showed 4/4 sequences clustering monophyletically, in this case closest to subtype B. All four sequences were from MSM located in southeast England. Although the sequences exhibited monophyletic clustering throughout, the tree for slice 1 showed the sequences located in a mixed subtype D, B, CRF21 region of the tree. It seems more likely that these sequences may be related in terms of transmission, but not a novel CRF.

3.2.4.22 Cluster 22 (B/A)

Cluster 22 had four members and a maximum genetic distance of 4.4%. The breakpoint was placed at HXB2 2558. The sequences did not exhibit monophyletic clustering in either tree, making it unlikely that they comprise a novel CRF.

3.2.4.23 Cluster 23 (B/A)

Cluster 23 had three members and a maximum genetic distance of 3.6%. The breakpoint position was placed at HXB2 3099. The sequences did not exhibit monophyletic clustering in either tree, making it unlikely that they comprise a novel CRF.

3.2.4.24 Cluster 24 (B/A)

Cluster 24 had five members and a maximum genetic distance of 1.5%. The breakpoint was placed at HXB2 2615. All five sequences exhibited clustering closest to CRF50_A1D sequences in both trees, making it likely that these sequences are related to CRF50 rather than a novel CRF.

3.2.4.25 Cluster 25 (B/A/B)

Cluster 25 had 21 members and a maximum genetic distance of 8.1%. Owing to conflicting breakpoint analysis results, the alignment was sliced three different ways for analysis. Method 1 placed breakpoints at HXB2 positions 2555 and 3048 (Figure 3_23), method 2 placed breakpoints at 2545 and 2941 (Figure 3_24), and method 3 placed breakpoints at 2632 and 2943 (Figure 3_25; Table 3_17). The clustering displayed in the three groups of trees showed that, rather than one large cluster, there were three distinct groups of sequences, each potentially comprising a novel CRF. Cluster 25a contained 10 sequences and had breakpoints at HXB2 2555 and 3048, Cluster 25b contained seven sequences and had breakpoints at HXB2 2545 and 2941, and Cluster 25c contained three sequences and had breakpoints at HXB2 2632 and 2943. One sequence did not cluster with any other query sequence in any of the phylogenetic analyses.

Tree 1 for Cluster 25a showed all 10 sequences clustered together with a posterior probability of 1 (Figure 3_23a). The sequences clustered closest to subtype B (posterior probability = 0.95). Tree 2 (Figure 3_23b) showed well supported posterior probabilities. All ten sequences clustered together (posterior probability = 1), closest to CRF50 sequences (posterior probability = 1). Tree three (Figure 3_23c) showed all ten sequences clustered together (posterior probability = 1) closest to subtype B (posterior probability = 0.89).

Restricting the cluster to these 10 sequences gave a maximum genetic distance of 2.1%. The geographic location of Cluster 25a showed 9/10 sequences in London and Southeast England and 1/10 sequences in Northeast England.

Demographic information was available for 10/10 sequences; this showed that all sequences came from white MSM.

No.	jpHMM GT	jpHMM breakpoint (95% CI)		SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)			
Cluster 25a									
1	B/A1/B	2555 (2519-2591)	3048 (3035-3061)	A1/B	CRF05/B/A1/CRF17	2438 (2437-2439)	2540 (2539-2541)		3034 (3033-3035)
2	B/A1/B	2555 (2519-2591)	3048 (3035-3061)	A1/B	CRF05/B/A1/B	2438 (2437-2439)	2540 (2539-2541)		3034 (3033-3035)
3	B/A1/B	2555 (2519-2591)	3048 (3035-3061)	A1/B	CRF05/B/A1/B	2360 (2359-2361)	2540 (2539-2541)		3034 (3033-3035)
4	B/A1/B	2555 (2518-2592)	3048 (3035-3061)	Complex	D/A1/B		2540 (2539-2541)		3034 (3024-3046)
5	B/A1/B	2555 (2518-2592)	3048 (3035-3061)	A1/B	B/A1/B		2540 (2539-2541)		3034 (3024-3046)
6	B/A1/B	2555 (2518-2592)	3048 (3035-3061)	Complex	B/G/A1/A1/B		2552 (2551-2553)	2722 (2712-2732)	2890 (2888-2891) 3034
7	B/A1/B	2568 (2540-2596)	3048 (3035-3061)	A1/B	CRF05/B/A1/B	2362 (2361-2363)	2540 (2539-2541)		3034 (3033-3035)
8	B/A1/B	2562 (2519-2605)	3048 (3035-3061)	A1/B	B/A1/B		2540 (2497-2583)		3034 (3033-3035)
9	B/A1/B	2557 (2519-2595)	3048 (3035-3061)	A1/B	CRF39/B/A1/B	2418 (2411-2425)	2540 (2539-2541)		3034 (3033-3035)
10	B/A1/B	2567 (2540-2594)	3048 (3035-3061)	Complex	CRF05/U/A1/B	2437 (2425-2449)	2540 (2539-2541)		3034 (3033-3035)
Cluster 25b									
1	B/A1/B	2545 (2538-2552)	2941 (2911-2971)	B	B				
2	B/A1/B	2545 (2538-2552)	2941 (2911-2971)	Complex	B/B/A1/U/B	2354 (2353-2355)	2540 (2539-2541)	2671 (2647-2695)	3016 (3013-3020)
3	B/A1/B	2545 (2538-2552)	2941 (2910-2972)	B/U	B/U/B		2540 (2539-2541)		3016 (3004-3028)
4	B/U/B	2530	2974	B/U	U/CRF03				3033 (2951-3117)
5	B/U/B	2554	2932	Complex	B/A1/F1/A1/B		2555	2671	2918

No.	jpHMM GT	jpHMM breakpoint (95% CI)		SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)			
							(2556-2558)	(2670-2772) 2819 (2818-2820)	(2917-2919)
6	B/A1/B	2592 (2553-2631)	2922 (2907-2937)	B	B				
7	B/U/B	2520	2924	A1/B	B/A1/B			2818 (2816-2820)	2917 (2916-2918)
Cluster 25c									
1	B/A1/B	2653 (2612-2694)	2923 (2909-2937)	B	B	2434 (2430-2438)	2540 (2539-2541)		
2	B/A1/B	2632 (2612-2652)	2923 (2909-2937)	A1/B	B/A1/B/B			2729 (2727-2731)	2924 (2923-2925) 3184 (3166-3202)
3	B/A1/B	2648 (2612-2684)	2923 (2909-2937)	A1/B	B/B/A1/CRF03	2371 (2367-2373)	2612 (2601-2623)		2924 (2923-2925)

Table 3_17. Full breakpoint results for Cluster 25a,b,c. In Cluster 25a, the jpHMM results show near-identical breakpoint predictions for all ten sequences. The most commonly SCUEAL-predicted breakpoints are at HXB2 2540 and HXB2 3035. In Cluster 25b, the first jpHMM breakpoint has a similar placement to cluster 25a, but the second breakpoint is approximately 100 nucleotides distant from the second breakpoint identified for that cluster. The SCUEAL results show little consistency. In Cluster 25c, the jpHMM results indicate a B/A1/B structure for these sequences, but the phylogenetic results indicate a CRF50/B recombinant. CRF50 is not included at present in the jpHMM algorithm, which accounts for the seemingly discrepant results. The only SCUEAL breakpoint that is present in 2/3 sequences, HXB2 2924, is consistent with the second predicted jpHMM breakpoint.

The HIV BLAST search showed 10 sequences that appeared in at least 75% of the closest sequence matches. Four of these sequences were submissions from the UK HIV DRD, and the remaining six were A1/B sequences from Canada. jpHMM analysis of the six Canadian sequences showed that all sequences had a single breakpoint only; 3/7 were B/A1 recombinants and 4/7 were D/A1 recombinants.

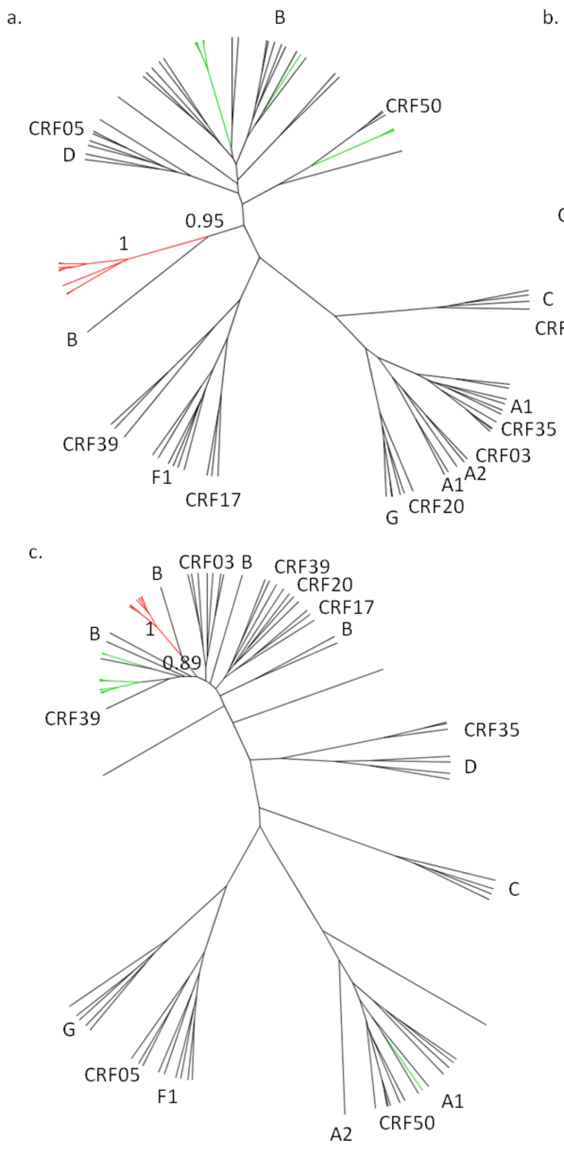


Figure 3_23. Cluster 25a: B and CRF50 recombinant fragments. Maximum clade credibility trees of putative pure subtype regions. Breakpoints were placed at HXB2 2555 and 3048 (Method 1). Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. Sequences that do not exhibit monophyletic clustering are highlighted in green. a) Slice 1; b) Slice 2; c) Slice 3.

Tree 1 for cluster 25b showed all seven sequences clustering monophyletically, with a posterior probability of 1 (Figure 3_24a). The sequences clustered with subtype B, however the support for this was weak (posterior probability = 0.38).

Tree 2 showed all seven sequences clustered closest to CRF50 sequences, with moderate support (posterior probability = 0.60), and strong support once subtype A1 was included (posterior probability=1; Figure 3_24b). Tree 3 showed all seven sequences clustered close to sequences from cluster 25c, and closest to subtype B (posterior probability = 0.83, Figure 3_24c). Restricting the cluster to these seven sequences reduced the maximum genetic distance to 1.5%.

The predicted structure for cluster 25b was B/CRF50/B, with breakpoints at breakpoints at 2545 and 2941. The geographic location showed 6/7 sequences in Scotland and 1/7 sequences in Northwest England. Demographic information was available for 1/7 sequences, which was a white MSM.

The HIV BLAST search showed eight sequences that appeared in at least 75% of the closest sequence matches. 5/8 were sequences that were submissions from the UK HIV DRD, 2/8 were subtype B sequences from Sweden and 1/8 was a subtype B sequences from the USA.

Trees 1 and 2 from cluster 25c showed the three sequences clustering with CRF50 sequences, with relatively strong support in both trees (Figure 3_25a, 3_25b). The tree from slice 3 showed all three sequences clustering with subtype B sequences (posterior probability = 0.82, Figure 3_25c). Restricting the cluster to these sequences gave a maximum genetic distance of 0.4%. The geographic location showed that 2/3 sequences came from London and Southeast England and 1/3 sequences came from Scotland. Demographic information was available for 2/3 sequences; both sequences came from white MSM. The country of origin for one sequence was the USA.

The HIV BLAST search showed seven sequences that were present in at least 75% of the closest sequence matches. Two of these were subtype B sequences from the UK (not submitted from the UK HIV DRD), two were subtype B sequences from Sweden, and three were D/A1 recombinant sequences from Canada. jpHMM analysis of the recombinant sequences showed a breakpoint located between HXB2 2550 and 2561, similar to the D/A1 breakpoint in CRF50.

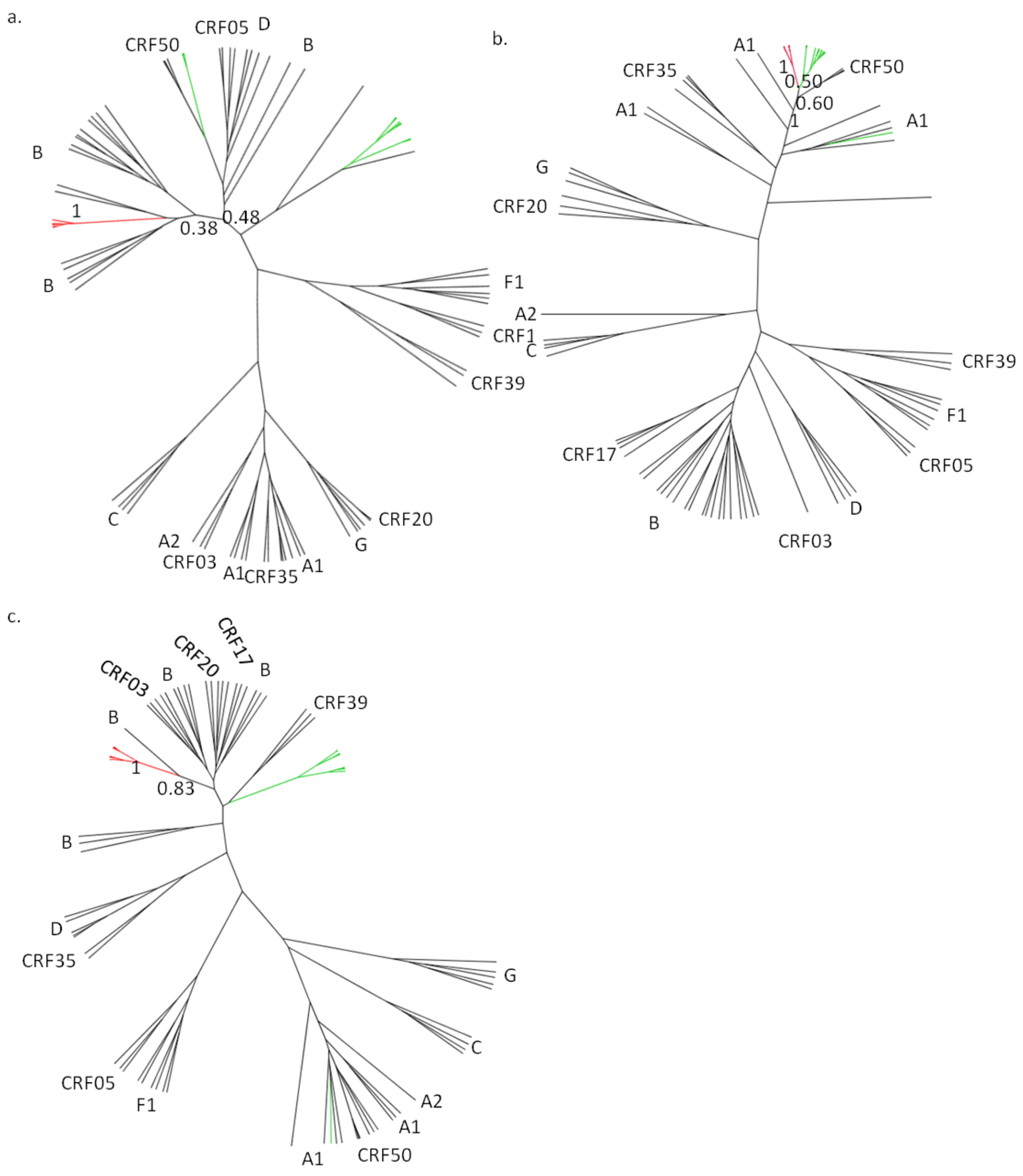


Figure 3_24. Cluster 25b: B and CRF50 recombinant fragments. Maximum clade credibility trees of putative pure subtype regions. Breakpoints were placed at HXB2 2545 and 2941 (Method 2).Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red; sequences that do not exhibit monophyletic clustering are highlighted in green. a) Slice 1; b) Slice 2; c) Slice 3.

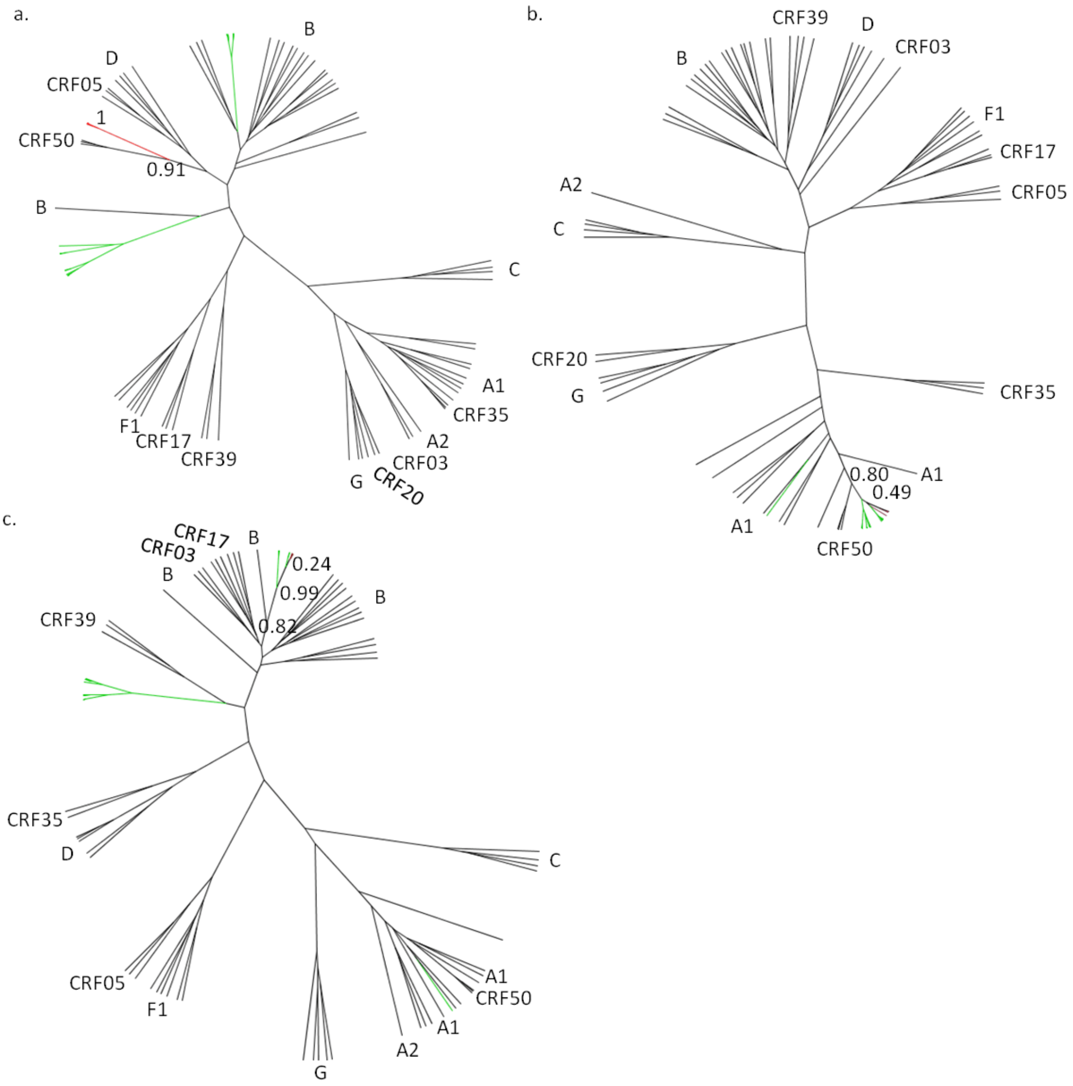


Figure 3_25. Cluster 25c: CRF50 and B recombinant fragments. Maximum clade credibility trees of putative pure subtype regions. Breakpoints were placed at HXB2 2632 and 2943 (Method 3). Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red; sequences that do not exhibit monophyletic clustering are highlighted in green. a) Slice 1; b) Slice 2; c) Slice 3.

A time-scaled phylogeny including all 72 identified CRF50_A1D strains showed that Cluster 25a had a tMRCA of 1997.60 (95% HPD = 1994.26-2001.14) in the UK (Figure 26). MCMC analysis showed a tMRCA for Cluster 25b of 1999.14 (95% HPD = 1995.95 – 2002.48). Three of the sequences were in a sub-branch containing subtype B sequences from the UK and Sweden. The remaining four sequences formed their own branch with a tMRCA of 2005.07 (95% HPD = 2004.59-2005.94). The time scaled phylogeny for Cluster 25c showed a tMRCA of

1999.55 (95% HPD = 1997.54-2001.82) and contained a sequence from the UK showing identical recombinant structure and breakpoints.

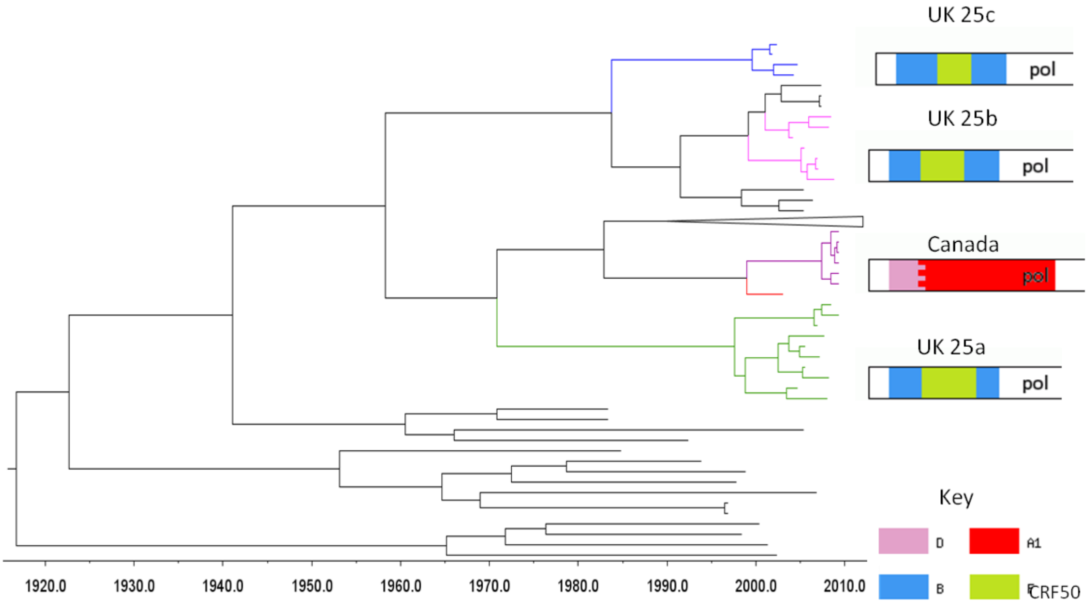


Figure 3_26. Time scaled phylogeny of Clusters 25a, b, and c, white MSM clusters. MCMC analysis of three CRF50_A1D/B recombinant clusters. CRF50_A1D sequences are contained in the collapsed node. Canadian sequences showing almost identical recombination breakpoints as CRF50 are shown in purple; the later tMRCA of this cluster indicates possible export of CRF50 from the UK to Canada. Cluster 25a is shown in green, cluster 25b is shown in lavender and cluster 25c is shown in blue.

3.2.4.26 Cluster 26 (C/B/C)

Cluster 26 had four members and a maximum genetic distance of 6.4%. Breakpoints were placed at HXB2 2572 and 3026. The sequences did not cluster together in any of the three trees, indicating that these sequences were not a novel CRF.

3.2.4.27 Cluster 27 (C/B/C)

Cluster 27 comprised five members and had a maximum genetic distance of 4%. Overall, 3/5 sequences had identical SCUEAL breakpoints at HXB2 equivalent 2377 and 2719, respectively (Table 3_18), accordingly, the alignment was sliced in these positions. 5/5 sequences clustered together in all three trees (posterior probabilities = 0.79, 1 and 1, respectively). Slice 1 showed the sequences clustering closest to subtype C sequences (Figure 3_27a), slice 2 showed the sequences clustered with subtype B (Figure 3_27b) and slice 3 showed clustering closest to subtype C

(Figure 3_27c). No demographic information was available. The geographic location showed 2/5 in Southwest England and Wales, 1/5 in Northeast England, 1/5 in Northwest England and 1/5 in Scotland.

The HIV BLAST search showed two sequences that were consistently present in the closest sequence matches. Both were subtype C sequences; one from Zimbabwe and one from South Africa.

The trees indicate that these sequences may comprise a novel CRF that is circulating in an unknown population.

No.	jpHMM GT	jpHMM breakpoint (95% CI)		SCUEAL GT (S)	SCUEAL GT (D)	SCUEAL breakpoints (95% CI)		
1	C/B/C	2379 (2323-2435)	2703 (2688-2719)	Complex	C/B/C/U	2377 (2376-2378)	2716 (2718-2720)	3142 (3140-3148)
2	C/B/C	2381 (2322-2440)	2703 (2688-2719)	B/C	C/B/B/C/C	2398 (2397-2399) 2501 (2500-2502)	2710 (2709-2711)	3058 (3051-3065)
3	C/B/C	2389 (2346-2432)	2704 (2688-2720)	Complex	C/B/C/G	2377 (2376-2378)	2719 (2714-2725)	3142 (3141-3143)
4	C/B/C	2385 (2335-2435)	2712 (2689-2735)	Complex	C/B/C/U	2377 (2376-2378)	2719 (2718-2720)	3142 (3138-3146)
5	C/D/A2/C	2349 (2340-2358) 2443 (2358-2528)	2706 (2685-2727)	C	C			

Table 3_18. Full breakpoint results for Cluster 27. The jpHMM results show consistent breakpoint predictions in 4/5 sequences; the SCUEAL breakpoints at positions 124 and 466 are equivalent to HXB2 2377 and 2719, respectively. The predictions for the fifth sequence are different between jpHMM and SCUEAL and different from the predictions for the other four sequences.

3.2.4.28 Cluster 28 (G/complex)

Cluster 28 had 23 members and a maximum genetic distance of 9%. The predicted recombinant structures were comprised of complex structures based around subtype G. Although the sequences appeared related, there was no consensus regarding breakpoint locations in either of the two genotyping results; consequently the alignment was sliced using four different sets of breakpoints and analysed multiple times. The first set of breakpoints was placed at HXB2 2369, 2737 and 3297, respectively, the second at 2429, 2542, 3033, the third at 2655, 2768, 3094 and the final set at 2671 and 3192. Consistent clustering was not observed in any of the trees.

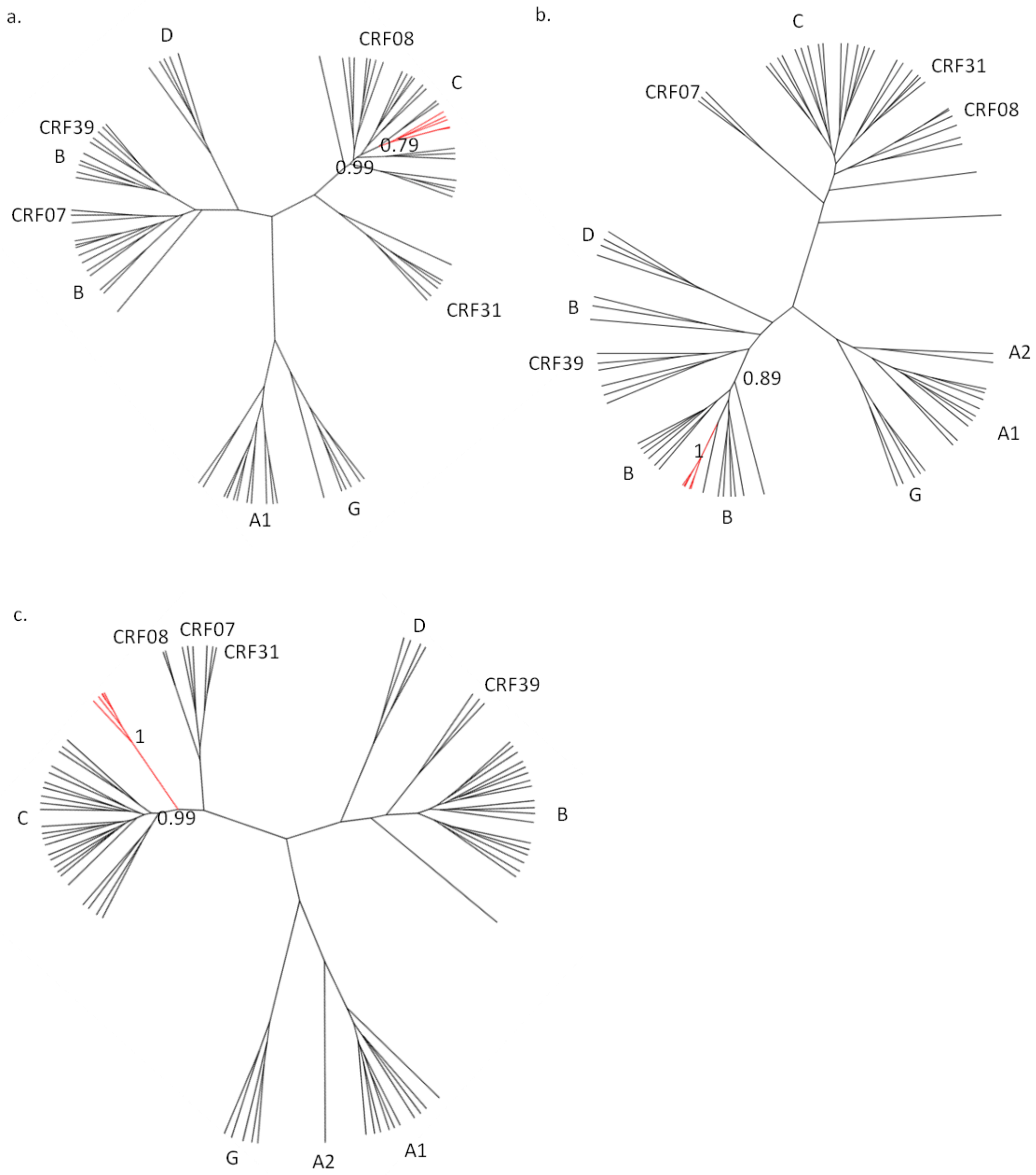


Figure 3_28. Cluster 27: C and B recombinant fragments. Maximum clade credibility trees of putative pure subtype regions. Breakpoints were placed at HXB2 2377 and 2719. Sequences exhibiting monophyletic clustering with a posterior probability of >0.7 are highlighted in red. a) Slice 1; b) Slice 2.

Chapter 4: Identification of potentially recombinant HIV-1 sequences and development of a near full-length single genome sequencing protocol

4.1 Phylogenetic screening of *pol* gene sequences in the UK HIV DRD (2007 and 2010 downloads)

A screen of the *pol* gene (RT and PR) population sequences present in the 2007 anonymised download of the UK HIV DRD was performed using the REGA tool developed by the REGA Institute in Leuven, Belgium. This process identified five mosaic sequences that were unable to be assigned to a definite subtype. All five sequences shared a similar recombination profile, which was a single breakpoint between subtype D and A1 between nucleotides HXB2 2603-2703. Alignments of subtype D and sub-subtype A1 sequences of the same genomic region were generated and used to confirm the subtype classifications, using the 100 closest BLAST matches in the UK HIV DRD.

The five putative A1/D recombinant specimens were all from white MSM diagnosed with HIV infection between 2000 and 2003. This part of the work was done in collaboration with our collaborators at the REGA Institute. Following the REGA investigation, stored samples from four of the subjects were subsequently retrieved and transferred to the Royal Free Hospital for near full-length, single genome sequencing (NFL-SGS).

In 2011, following sequencing of the four specimens, an additional screen of the 2010 download of the database was performed to expand the previous analysis. The single genome sequences were used as reference sequences for a BLAST search of the 2010 UK HIV DRD, and the top 400 matches were genotyped using SCUEAL to identify sequences sharing the same D/A1 breakpoint. 42 sequences sharing the same *pol* D/A1 breakpoint were identified, of which 16/42 were located in clinics in London and southeast England and 26/42 were located in clinics in northwest England. As the original four specimens that were sequenced were from clinics in London and Southeast England, two additional samples from clinics in Northwest England were obtained and sequenced.

4.1.1 Sample retrieval

In order to preserve patient anonymity, we elected to characterise samples from clinics with large numbers of patients. Furthermore, when retrieving samples for testing we used the MRC-CTU (who held the linked database and clinic patient identifiers) as an intermediate between the laboratory (holding database identifiers only) and the clinic, thereby ensuring we remained fully blinded to the patients' personal details. Finally, the aliquots of plasma were anonymised before transfer to the Royal Free Hospital.

4.2 Development and validation of a near full-length, single genome sequencing (NFL-SGS) protocol for plasma HIV-1 RNA

Although protocols for HIV-1 NFL-SGS were available in published literature, the majority of these were for amplification of viral DNA extracted from peripheral blood cells (PBMC); few published data were available for the amplification of near full-length HIV-1 genomes from plasma HIV-1 RNA. Additionally, there were particular challenges related to the development of this technique. Firstly, although two of the samples retrieved from storage were less than one year old, four samples were between six and ten years old; all samples had been stored at -80^{o}C under routine conditions and the viral RNA may have undergone a degree of deterioration. Secondly, the volume of plasma available varied according to the storage protocols at each centre; sample volumes received were between 270µl and 1.5ml. Finally, the plasma HIV-1 RNA ("viral") load ranged between 9,148 and 500,000 copies/ml, meaning that the number of viral copies available for amplification was generally low (see Table 5_1 in the next chapter for further details of viral loads). Obtaining extra stored samples was not possible under the terms of the ethics agreement. Therefore, a robust protocol was optimised for low sample volumes and viral load. Additionally, because the ultimate aim was single genome amplification of recombinant specimens, a two-step, nested protocol using a single set of primers per PCR round was required. The two-step approach allowed for dilution of cDNA following reverse transcription, and therefore amplification of single genomes, and the single set of primers per PCR round eliminated the possibility that multiple sets of primers were amplifying a mixed, rather than recombinant, viral population.

The plasma specimens used for the validation process were left over samples from the virology diagnostic department at the Royal Free Hospital. These samples had been previously genotyped to determine the presence of drug resistance, as part of routine care. Once retrieved from storage, samples were identified by laboratory number only and not linked to patient identifiable information.

4.2.1 Preliminary protocols

Development of an NFL-SGS system commenced with an attempt to optimise a protocol published by Nadai et al. in 2008. The protocol involved amplification of the HIV-1 genome between HXB2 nucleotides 769 and 9181 in either two or three overlapping fragments of sizes 2.6kb (HXB2 769 - 3477), 3.3kb (HXB2 5861 - 9181), 3.7kb (HXB2 2483 - 6352), and 7.0kb (HXB2 2166 - 9181), respectively. The reported success rate for the 2.6kb fragment (which was used in both 2- and 3-fragment strategies) was 100% for viral loads greater than 750,000 copies/ml, 75% for viral loads between 100,000 and 200,000, 40% for viral loads between 10,000 and 100,000 copies/ml and 0% for viral loads less than 10,000 copies/ml. Therefore, an attempt was made to optimise the protocol for specimens with low viral loads. The protocol was designed for the amplification of subtype B specimens, and therefore the initial optimisation attempts were made using clinical subtype B plasma specimens.

4.2.1.1 Amplification of the 2.6kb fragment

The amplification of the 2.6kb fragment was the first step of the optimisation process. HIV-1 RNA was extracted using the Biomerieux easyMAG automated extractor using 1ml of plasma as the input in a generic on-board lysis protocol, with a final elution of 25µl. cDNA synthesis was trialled using both gene-specific and Oligo dT priming during reverse transcription (RT) with Superscript III Reverse Transcription reagents (Life Technologies). A total of 3µl RNA, 500µM dNTPs, and 2.5µM of either Oligo dT or primer UNINEF 7’ 5’-GCACTCAAGGCAAGCTTTATTGAGGCTT-3’ were combined and heated at 65°C for 5 minutes. Subsequently, PCR reactions were removed from the thermal cycler, and a mastermix of 2µl RT Buffer, 25mM $MgCl_2$, 0.1M DTT, 1µl RNaseOUT, 2µl Superscript III and 3µl nuclease-free H_2O added for a combined reaction volume of 20µl. Reverse transcription was performed at 50°C 2 hours, 85°C 5 minutes, followed by addition of 1µl RNase H, and then 37°C 20 minutes, 70°C 15 minutes. These conditions (except for priming with Oligo dT) were identical to the conditions detailed in Nadai et al.

To amplify the cDNA, the Expand Long Template PCR Kit (Roche Diagnostics, Sussex, UK) was used. The 50µl first-round PCR reaction mixture comprised 5µl

10x PCR Buffer, 350µM dNTPs, 0.4µM each of forward primer msf12b 5'-AAATCTCTAGCAGTGGCGCCCGAACAG and reverse primer RT347R 5'-GAATCTCTCTGTTTTCTGCCAGTTC, 5U enzyme, 29.3µl nuclease-free H_2O and 10µl cDNA. Cycling conditions were 94°C 2 minutes, followed by 10 cycles of 94° 10 s, 60°C 30 s, 68°C 3 m, then 20 cycles of 94°C 10s, 55°C 30s, 68°C 3 m; the final extension was 68° C 10 minutes. The second round of PCR used forward primer f2nst 5'-GCGGAGGCTAGAAGGAGAGAGATGG and reverse primer proRT 5'-TTTCCCCACTAACTTCTGTATGTCATTGACA. 5µl of first-round PCR product was used as the reaction template, and the volume of nuclease-free H_2O was 34.3µl. All other reaction components and conditions remained the same. The PCR was performed without the physical wax barrier used in the Nadai et al. paper. This approach failed to yield an amplified product with both gene-specific and Oligo dT approaches.

Subsequently, the PCR protocol was attempted using a physical wax barrier (AmpliWax PCR Gem 50, Life Technologies) as described in the Nadai et al. paper. Briefly, 350µM dNTPs, 0.4µM of each primer (msf12b and RT3473R in the first round and f2nst and proRT in the nested round) and 20µl of nuclease-free H_2O were combined and placed in a thin-walled PCR reaction tube with a wax bead. PCR reactions were heated to 85°C for 5 minutes, and then placed on ice to re-solidify the wax and create a physical barrier that would prevent premature primer hybridisation during the enzyme activation stage. Following this, 5µl 10x PCR Buffer, 5U of enzyme, 9.3µl of nuclease-free H_2O (14.3 µl in the nested round) and either 10 (first round) or 5 (nested round) µl template was combined and added to the PCR reaction. The cycling conditions were identical to those detailed above. Unfortunately, the addition of the wax barrier also failed to yield any amplified PCR product when using both gene-specific and Oligo dT priming.

At this stage, three subtype B clinical specimens were extracted in parallel using the easyMAG system and Viroseq manual extractions, and tested using an already-validated HIV-1 *pol* PCR. Only the Viroseq extracts amplified, and so the easyMAG extraction protocol was discarded. All subsequent testing was performed using Viroseq extractions.

A new RT approach using random hexamers (Qiagen) with Promega reagents (London, UK) was adopted. The reaction mixture comprised 4µl 5xRT Buffer, 1µl 25mM dNTPs, 1.25µl random hexamers, 400U RNAseIn, 50U MMLV RT, 1.75µl nuclease-free H_2O and 10µl RNA. The RNA, dNTPs and random hexamers were combined and heated to 65°C for 30 seconds, followed by 42°C for five minutes. Following this, the remaining reagents were added, and reverse transcription was performed at 42°C 60 minutes, followed by 99°C for 5 minutes. The subsequent PCR amplification was performed with the wax barrier as detailed above. This approach resulted in the successful amplification of the 2.6kb fragment of three subtype B clinical specimens (Figure 4_1a).

Although the fragment amplification was successful, the gel showed smears of DNA in 2/3 specimens, indicating an unoptimised methodology. To optimise the reaction, the annealing temperature was increased by 2°C, to 57°C, in the second round of cycles in the nested PCR, and amplification was repeated using the same three clinical specimens. This resulted in both smears of DNA and evidence of mispriming (smaller, incorrectly-sized products) (Figure 4_1b), and so the annealing temperature was returned to 55°C, and a further subtype B specimen was amplified in triplicate. When this showed clean bands in 2/3 replicates (Figure 4_1c), the decision was made to delay final optimisation of each fragment until successful amplification was achieved across the entire HIV genome.

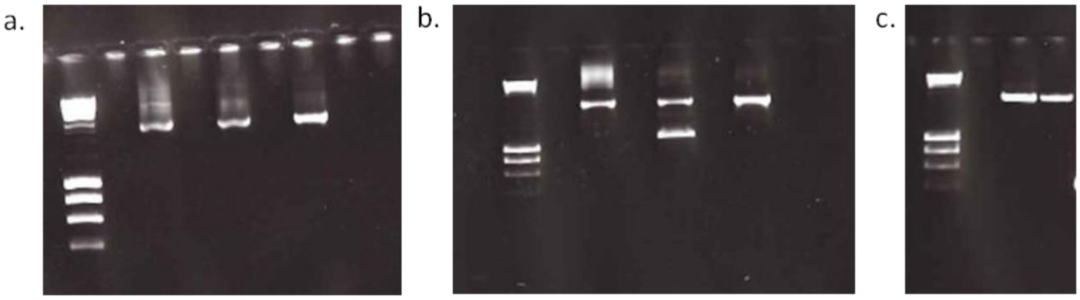


Figure 4_1. Amplification of the 2.6kb fragment spanning *gag-pol* (HXB2 769 - 3477) from the Nadai et al. protocol for near full-length sequencing of HIV-1 from plasma.
a) Amplification of the 2.6kb fragment using three clinical HIV-1 subtype B stored plasma specimens. Lane 1 contained Hyperladder I, lanes 3, 5 and 7 contained one replicate each of a clinical HIV-1 subtype B specimen. Lane 9 contained the negative control. Although the DNA bands were of the correct size, DNA smears were present, indicating an unoptimised methodology. **b)** Optimisation of the second round PCR annealing temperature. The same clinical specimens as in Figure 4_1a were amplified using an annealing temperature of 57°C in the second round of cycles in the nested PCR. Lane 1 showed Hyperladder I, and lanes 3, 5 and 7 contained one replicate each of a clinical HIV-1 subtype B specimen. Lane 9 contained the negative control. Increasing the annealing temperature in this step did not optimise the method, as DNA smears were still present, and mispriming occurred in the second specimen. **c)** Triplicate amplification of a single subtype B specimen at the original annealing temperature of 55°C. Hyperladder I is in lane 1; lanes 2 - 4 show triplicate replicates of a single specimen. Although 1 replicate failed to amplify, the other two replicates showed clean DNA bands.

4.2.1.2 Amplification of the 3.3kb and 3.7kb fragments

Following the successful amplification of the 2.6kb fragment, amplification of the remaining two fragments in the three fragment protocol (3.3kb (HXB2 5861-9181), 3.7kb (HXB2 2483-6352)) was attempted. Owing to the consistent failure of the extraction and RT steps from the Nadai et al. paper, the Viroseq extraction and random hexamer RT were retained. The PCR reactions comprised the same components as for the 2.6kb amplicon; primers for the 3.3kb fragment were ENVoutF1 5'-AGARGAYAGATGGAACAAGCCCCAG and UNINEF 7' in the first round, and ENVinF1 5,-TGGAAGCATCCRGGAAGTCAGCCT and nefyn05 5'-GTGTGTAGTTCTGCCAATCAGGGAA in the nested round; primers for the 3.7kb amplification were POLoutF1 5'-CCTCAAATCACTCTTTGGCARCGAC and VIF-VPUoutR1 5'-GGTACCCCATAATAGACTGTRACCCACAA in the first round reaction, and POLinF1 5'-AGGACCTACRCCTGTCAACATAATTGG and VIF-VPUinR1 5'-CTCTCATTGCCACTGTCTTCTGCTC in the nested round. The cycling conditions for each were the same as for the 2.6kb fragment, except that the extension time in each stage was increased from 3 minutes to 4 minutes. These conditions were identical to those in Nadai et al.

Although the new RT strategy resulted in consistently successful amplification of the 2.6kb fragment, when this approach was applied to amplification of the 3.3 and 3.7kb fragments, the results showed either products of incorrect size (for the 3.7kb fragment reaction), or complete failure of amplification (for the 3.3kb reaction) (Figure 4_2 a). This was most likely due to the use of random hexamers at the RT stage, which produces many shorter fragments of cDNA rather than the single fragment gained from using gene-specific or Oligo dT priming, but, as this was the only RT strategy that had been successful, further attempts at optimising the reactions for these fragments were made.

The Expand Long Template kit came equipped with three different reaction buffers containing differing concentrations of $MgCl_2$. Amplification of both fragments was subsequently attempted using both Buffer 2 and Buffer 3, using a single clinical specimen from which the 2.6kb fragment was successfully amplified. The reactions with Buffer 2 resulted in failed amplification for the 3.3kb fragment, and amplification at the incorrect size for the 3.7kb fragment (Figure 4_2b). One faint band of the correct size was amplified for the 3.7kb fragment, however this was of insufficient intensity for use in any downstream analyses. The reactions using Buffer 3 also resulted in failure of amplification of the 3.3kb fragment and amplification of incorrectly-sized products for the 3.7kb fragment (Figure 4_2c).

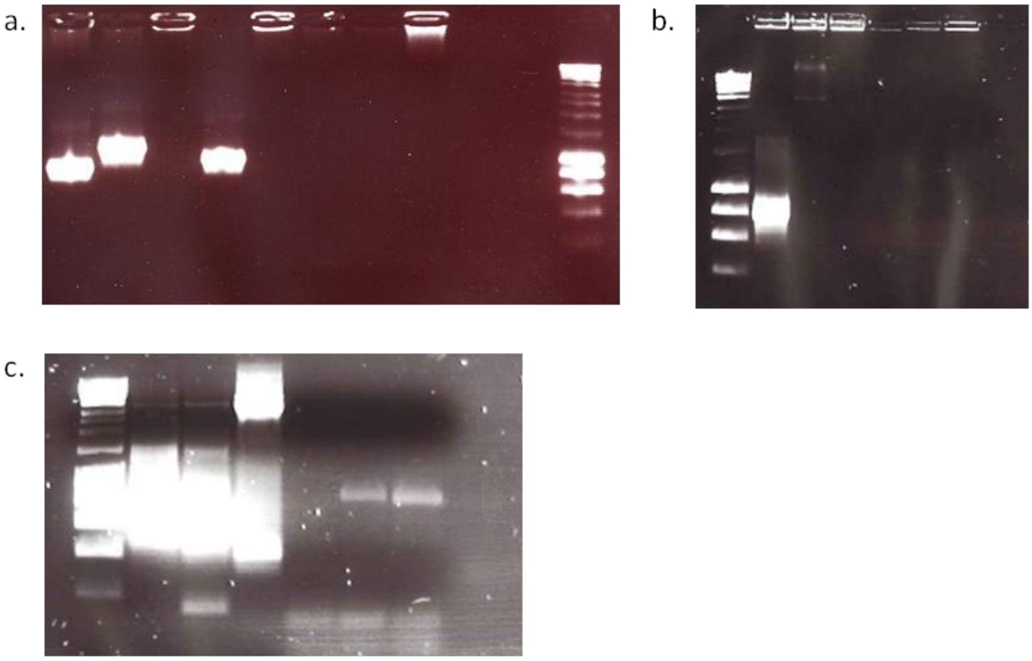


Figure 4_2. Optimisation of the 3.3kb and 3.7kb fragments from the Nadai et al. protocol.

a) Typical results from attempted subtype B amplification of the 3.3kb and 3.7kb fragments of the Nadai et al protocol. Lanes 1-2 and 3-4 contained duplicate reactions of two clinical subtype B specimens using the 3.7kb amplification protocol. Incorrectly sized product was present in 3/4 reactions; the remaining reaction failed to amplify. Lanes 5-6 and 7-8 contained duplicate reactions of the same two subtype B clinical specimens, amplified using the 3.3kb protocol. All reactions failed to amplify. Lanes 9 and 10 contained the negative control reactions, and lane 11 contained Hyperladder I. **b)** Results from attempted amplification of the 3.3kb and 3.7kb fragments using Buffer 2. Lane 1 contained Hyperladder I, lanes 2-4 contained triplicates of a subtype B clinical specimen amplified using the 3.7kb fragment protocol, lanes 5-7 contained triplicates of the same subtype B specimen amplified using the 3.3kb fragment protocol, and lane 8 contained the negative control. There was one faint band of the correct size for the 3.7kb fragment (lane 3); all other replicates for both fragments failed. **c)** Results from attempted amplification of the 3.3kb and 3.7kb fragments using Buffer 3 and the same subtype B clinical specimen used in 4-b. Lane 1 contained Hyperladder I, lanes 2-4 contained triplicates of the subtype B clinical specimen amplified using the 3.7kb protocol, lanes 5-7 contained triplicates of the same subtype B specimen amplified using the 3.3kb fragment protocol, and lanes 8-9 contained the negative controls. Some very faint products of the correct size were observed for the 3.7kb fragment; no products were present for the 3.3kb amplicon. DNA smears and evidence of mispriming was widespread across all reactions.

Given the consistent failure of the 3.3kb fragment, it was decided to focus on amplifying this region of the genome successfully before proceeding to further optimisation of other genomic regions. A new strategy of splitting the 3.3kb fragment into two smaller, overlapping fragments (HXB2 6813-8817 and 5514-7374) was adopted, with a view to extending this strategy to the 3.7kb fragment if successful. The primers for this fragment were sourced from the list of sequencing primers in Nadai et al. (see Appendix 2_4 for the full list). The HXB2 co-ordinates and melting temperatures of primers were compared, and a nested PCR was designed using first round primers ZFF 5'-GGGATCAAAGCCTAAAGCCATGTGTAA and JL89 5'-TCCAGTCCCCCCTTTTCTTTTAAAAA, and second round primers 793SEQ1 5'-AACACCTCAGTCATTACACAGGCC and JL71 5'-TTTTGACCACTTGCCACCCAT. The HXB2 co-ordinates of this fragment were 6813-8817.

Amplification of the new fragment was attempted using triplicates of a single subtype B clinical specimen that had previously successfully amplified. Owing to the size of the fragment (2.0kb), the PCR mastermix components and cycling conditions used were as for the 2.6kb fragment. Each of the three buffers in the Expand Long Template PCR kit was trialled simultaneously, and the best results were achieved using Buffer 3 (Figure 4_3 a). Following this, a further fragment (fragment 4) was designed using the first-round primers ACC1 5'-TTCAGAAGTATACATCCCACTAGG and JL102 5'-GATGGGAGGGGCATACAT, and second round primers VIFC 5'-GAYAAAGCCACCTTTGCCTAGTGTT and JL98 5'-AGAAAAATTCCCCTCCACAATTAA (HXB2 co-ordinates 5514 - 7374; fragment size 1.8kb). This was amplified using the same conditions as fragment 3, but using Buffer 3 only (Figure 4_3b).

Given the success of splitting the 3.3kb fragment into two, smaller, fragments, the same approach was now attempted for the 3.7kb fragment. Accordingly, the fragment was split into regions amplified by the following primers: fragment 2a used first round PoloutF1 5'-CCTCAAATCACTCTTTGGCARCGAC and SP1AS 5'-GGATGAATACTGCCATTTGTACTGC, and PolinF1 5'-AGGACCTACRCCTGTCAACATAATTGG and POLT- 5'-GCAGTCTACTTGTCCATGCATGGC in the nested round (HXB2 co-ordinates 2483-4397; fragment 2b used first round POLP 5'-GGATGGGATATGAACTCCATCC and VIF-VPUinF1 5'-CTCTCATTGCCACTGTCTTCTGCTC, and POLU 5'-ACTTTCTATGTAGATGGGGCAGC and ACC8R 5'-TCTCCGCTTCTTCCTGCCATAG in the nested round (HXB2 co-ordinates 3864 -

5989). These new fragments were trialled using Buffer 3, and the same PCR reaction components and cycling conditions as the 2.6kb fragment. Although fragment 2a amplified (Figure 4_3 c), fragment 2b failed to amplify.

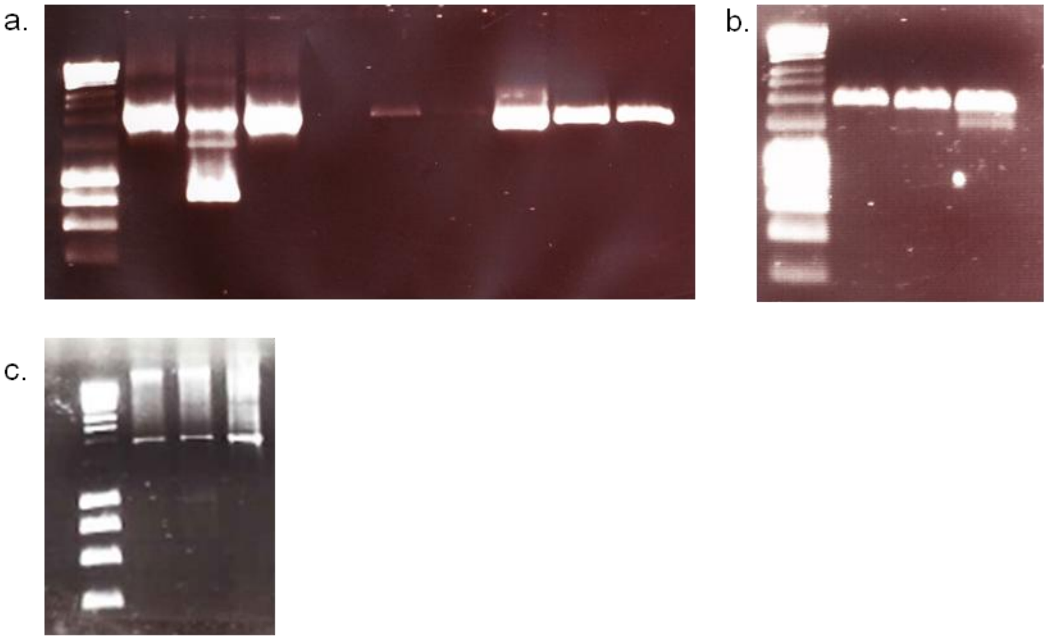


Figure 4_3. Optimisation of newly designed fragments.
a) Amplification of triplicates of a clinical subtype B specimen for the newly designed fragment 3, spanning HXB2 6813 - 8817. Each of the three buffers in the Expand Long Template PCR kit was trialled simultaneously. Lane 1 contained Hyperladder I and Lanes 2 - 10 contained amplification products. Lanes 2 - 4 contained the fragment using Buffer 1, and lanes 5 - 7 and 8 - 10 contained the products from using Buffers 2 and 3, respectively. Consistent amplification without mispriming was achieved using Buffer 3 only. **b)** Amplification of the subsequently designed fragment 4, spanning HXB2 5514 - 7374, using Buffer 3. Lane 1 contained Hyperladder I. Lanes 2 - 4 contained amplification of a clinical HIV-1 subtype B specimen. **c)** Amplification of fragment 2a using triplicates of a subtype B clinical specimen. Lane 1 contained Hyperladder I, and lanes 2 - 4 contained the subtype B specimen. Amplification of the correct size was achieved in all three replicates.

4.2.1.3 Filling in the gaps

At this stage, a total of 6.9kb of the genome had been amplified successfully using subtype B clinical specimens (Figure 4_4). However, there were two regions of the genome that showed consistent amplification failure. Two sets of primers were designed using Primer3 to cover these regions. The primer sets were designed to cover the missing regions with sufficient overlap to allow for subsequent sequence assembly, and with melting temperatures that should allow for use of the same PCR conditions as the other fragments. The primers for fragment *pol-vif* were Pol2fwd 5’-

AATACAGAAGCAGGGGCAAG and Pol2rev 5'-GCAATGAAAGCAACACTTTTT in the first round (HXB2 3533 - 5932), and Pol2nestedfwd 5'-AGGAAACATGGGAAACATGG and Pol2nestedrev 5'-TCGACACCCAATTCTGAAAA (HXB2 3733 - 5790). Unfortunately, both of the newly-designed fragments failed to amplify.

Following this, a final attempt was made to amplify the *pol-vif* region using a combination of the primers POLP 5'-GGATGGGATATGAACTCCATCC and Pol2rev in the first round, and POLU 5'-ACTTTCTATGTAGATGGGGCAGC and Pol2nestedrev in the nested round. Unfortunately, this combination of primers also failed to amplify. Additionally, there was concern regarding splitting the HIV-1 genome into too many fragments, especially when the identification of a recombinant structure was the ultimate goal of the analysis. Therefore, a renewed attempt to amplify the 3.7kb fragment in a single reaction was made using two new long-range PCR systems.

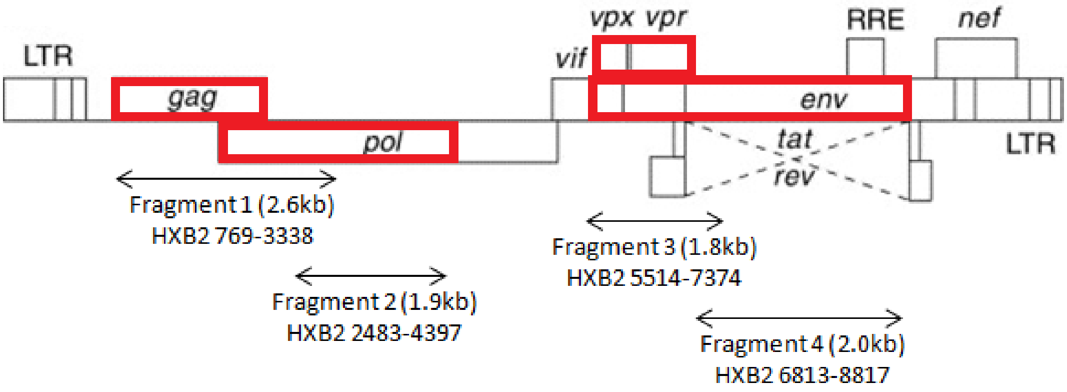


Figure 4_4. Regions of the HIV-1 genome successfully amplified using conditions and primers adapted from the Nadai et al. protocol. A modified schematic of the HIV-1 genome shows regions successfully amplified using modified primers sets and conditions from the Nadai et al. protocol. Genomic regions that had successfully been amplified are shown in red. Although approximately 2/3 of the genome was covered, the two missing regions showed consistent amplification failure. The original genome schematic is available from http://dx.doi.org/10.1016/S0966-842X(00)01816-3.

4.2.1.4 Amplification of the 3.7kb fragment using two long-range PCR systems

Two new long-range PCR systems were used in a final attempt to amplify the 3.7kb fragment: the New England Biolabs Phusion PCR Kit (Hitchen, UK), and the Qiagen Long Range PCR Kit (Hilden, Germany). The Phusion PCR kit strategy was attempted according to the manufacturer's instructions, both with and without the addition of DMSO. Per reaction components without DMSO comprised 25µl Phusion Mastermix, 2.5µl each of forward primer POLoutF1 and reverse primer VIF-VPUoutR1, 15µl nuclease-free H_2O and 5µl template DNA. The per reaction components for the reactions with DMSO comprised 25µl Phusion Mastermix, 2.5µl each of forward primer POLoutF1 and reverse primer VIF-VPUoutR1, 13.5µl nuclease-free H_2O, 1.5µl DMSO and 5µl template DNA. The nested round was identical to the first round except for the use of forward and reverse primers POLinF1 and VIF-VPUinR1. 5µl of first round product was used as the template for the nested PCR. Cycling conditions for both reactions were 98°C 30s, followed by 35 cycles of 98°C 10s, 68°C 30s, 72°C 2m, and a final extension of 72°C 10m. Both these reactions failed to amplify.

Following amplification failure, a gradient PCR using annealing temperatures of 69.0°C, 69.5°C, 70°C, 71.5°C and 72°C was performed to optimise the annealing temperature, and the primer concentration was titrated in concentrations of 0.2µM, 0.25µM, 0.3µM and 0.35µM per reaction. Unfortunately, no amplified product was observed at any stage. An attempt was made to amplify the 2.6kb fragment amplifiable with the Expand Long Template Kit; after this fragment did not amplify in the correct size, the kit was abandoned.

Next, the Qiagen Long Range PCR kit was trialled. This kit contained a proprietary RT step, which was trialled using parallel testing with the random hexamer RT protocol. Per reaction components for the Qiagen RT were: 4 µl RT Buffer, 2µl dNTP mix, 1µl Oligo dT, 0.2µl RNasIn, 1µl RT enzyme, 1.8µl nuclease-free H_2O and 10µl RNA. Cycling conditions were 42°C 90m, followed by 85°C. The per reaction components for the PCR protocol were: 5µl 10xBuffer, 2.5µl dNTPs, 2µl each of forward and reverse primers POLoutF1 and VIF-VPUoutR1, 0.4µl enzyme, 33.1µl nuclease-free H_2O and 5µl cDNA. Cycling conditions were 93°C 3m, followed by 35 cycles of 93°C 15s, 62°C 30s, 68°C 3 m; the final extension was 68°C 10m. The nested PCR protocol was the same as the first round but with the use of primers

POLinF1 and VIF-VPUinR1. The PCR reactions using the proprietary RT and the random hexamer RT failed to amplify.

After amplification failure, a gradient PCR with annealing temperatures of 60°C, 62°C, 66°C and 68°C was performed, followed by experimentation using both touchdown PCR and cycle elongation of 20 successive cycles. After these modifications still failed to yield products of the correct size, the protocol was discarded.

Subsequently, a protocol was successfully adapted from one developed in Beatrice Hahn's laboratory at the Centre for HIV/AIDS Vaccine Immunology (CHAVI) (unpublished, CHAVI-MBSC-2). This protocol involved the amplification of HIV-RNA in a single 9kb fragment.

4.2.2 Optimisation of the CHAVI protocol

Optimisation of the CHAVI protocol was performed at each of the following steps:

4.2.2.1 RNA extraction

To lessen the labour associated with performing manual nucleic acid extractions, RNA extraction was trialled using the QiAmp Viral RNA Mini Kit (Qiagen) as detailed in the CHAVI protocol, and the automated BioMerieux Nuclisens easyMAG system (BioMerieux, Basingstoke, UK). 22/65 (33.8%) reactions using the automated extractor showed evidence of mispriming and PCR products of the incorrect size (~2-3kb) (Figure 4_5a). This was most likely due to the automated extractor shearing the RNA during the extraction process. In contrast, when the QiAmp Viral RNA Mini Kit was used, PCR products of the correct ~9kb size were amplified.

The putative recombinant samples were not only limited in volume, but were potentially of poor quality and may therefore have contained fewer RNA copies than suggested by the viral load result. Therefore, extraction of RNA using theoretical quantities of 20,000 and 40,000 copies was compared (Figure 4_5b). However, RNA extraction using double the amount of theoretical viral copies (40,000 copies instead of 20,000) proved counterproductive, as a higher number of reactions failed when using the increased copy number. Extraction of 20,000 copies was retained for the standard protocol, with the reservation that an extraction of 40,000 copies remained an option in specimens of sufficient volume, but poor quality.

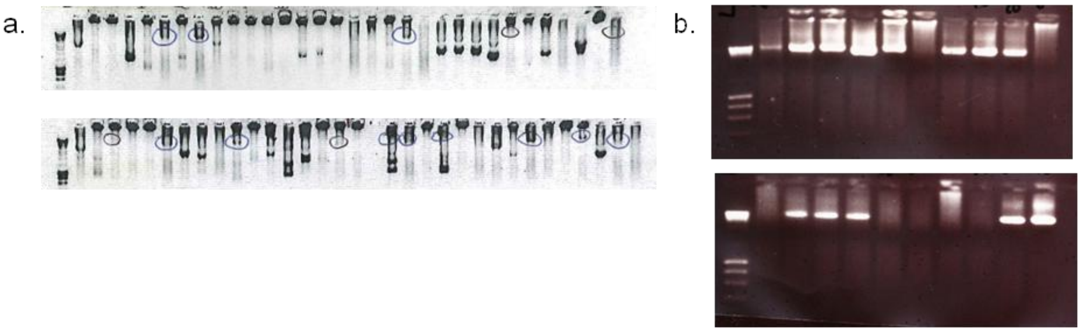


Figure 4_5. **Optimisation of a near full-length, single genome sequencing protocol**. **a)** Optimisation of the RNA extraction method. A clinical subtype B specimen was extracted at 20,000 copies and 40,000 copies using the automated easyMAG system, transcribed, and amplified using neat cDNA and a 1:2 cDNA dilution as the template input. 16 replicates of each extraction/dilution combination were run. Top row: Lane 1 contained Hyperladder I; lanes 2 - 17 contained neat cDNA from the 20,000 copies extraction; lanes 18 - 34 contained the 1:2 cDNA dilution from the 20,000 copies extraction. Bottom row: lanes 1 contained Hyperladder I; lanes 2 - 16 contained neat cDNA from the 40,000 copies extraction; lanes 18 - 34 contained the 1:2 cDNA dilution from the 40,000 copies reaction; lane 35 contained the NTC control. The circled lanes indicate amplified PCR product of the correct size. There was widespread evidence of mispriming across both the 20,000 copies and 40,000 copies extractions and the two cDNA dilutions (22/65 wells). **b)** Optimisation of copy number extraction. Extraction of 20,000 copies and 40,000 copies of RNA was trialled using 5 replicates of two clinical HIV-1 subtype B (top lanes) and C (bottom lanes) plasma samples. Lane 1 in each row contained Hyperladder I, Lanes 2 - 6 contained 5 replicates of a clinical subtype B (row 1) or C (row 2) specimen extracted at 20,000 RNA copies, and lanes 7 - 11 contained 5 replicates of the same subtype B and C specimens extracted at 40,000 RNA copies. Successful amplification was achieved more consistently when 20,000 RNA copies were extracted (lanes 2-6, top and bottom, success rate 5/5 and 3/5, respectively), then when 40,000 copies were extracted (lanes 7 - 11 top and bottom, success rate 3/5 and 2/5, respectively).

4.2.2.2 cDNA synthesis

The CHAVI protocol used Life Technologies' Superscript III Reverse Transcription reagents. However, these reagents had been previously trialled during the optimisation of the Nadai et al. protocol, and were found to be unsatisfactory. Therefore, the Superscript III First-Strand Synthesis Supermix Kit (Life Technologies) was selected for optimisation instead. This kit was chosen for trial due to the reduction in RNase H activity that is provided by the M-MLV-RT enzyme in the enzyme mix. As RNase H activity can promote strand transfer during minus strand DNA synthesis (Shriner et al., 2004), the selection of a system with reduced RNase H activity that could minimise any *in vitro* recombination events was desirable. The initial reaction conditions were identical to the manufacturer's instructions; however, a modified protocol using an additional 2µl of Superscript III RT enzyme (not included in the kit) added to each reaction following the initial 90 minute RT step at 50°C was selected. The extra enzyme circumvented enzyme exhaustion due to both the product size and the length of the RT protocol.

4.2.2.3 Nested PCR protocol

As the primers obtained from the CHAVI protocol were originally designed for use with subtype C viruses, the optimisation process commenced using subtype B and C clinical samples. Subtype B was chosen as the second subtype because the genetic distance of subtype B from subtype C would potentially give a good indication of the likelihood of the primers working with a range of HIV-1 subtypes. Although the CHAVI protocol used the same Expand Long Template PCR Kit as the Nadai et al. protocol, these reagents had already been trialled extensively, and so the Life Technologies Platinum PCR Supermix High Fidelity Kit was trialled alongside the Expand Long Template Kit. Using extracts from subtype B and C clinical samples, parallel PCR reactions were run using the Roche Diagnostics kit and the Life Technologies kit, using the manufacturer's recommendations for each set of reagents. The Life Technologies reaction components and conditions were as follows: 45µl PCR Supermix, 0.2µM each of forward primer 1.U5.B1F and reverse primer 1.R3.B3R, 1µl nuclease-free H_2O and 2µl cDNA were combined and used under the cycling conditions 94°C 2m, followed by 35 cycles of 94°C 15s, 55°C 30s, 68°C 9.5m, with a final extension of 68°C 20m. The nested round used the same components save forward primer 2.U5.B4F and reverse primer 2.R3.B6R were used, and 2µl of the first round reaction was used as the reaction template. The number of cycles was increased to 45 for this round.

The Expand Long Template Kit reactions were run using a wax barrier. The reaction conditions were the same as for the 2.6kb fragment from the Nadai et al. protocol, but with an extension time of four minutes. The specimen input for these reactions was one subtype B and one subtype C clinical specimen that was extracted at both 20,000 and 40,000 copies and was tested in five replicates per extraction for each of the two PCR systems.

All reactions using the Roche kit failed to amplify, whereas the reactions using the Life Technologies kit amplified at the correct size of 9kb in 2/5 (40%) cases using the subtype B and C specimens extracted at 20,000 copies, 4/5 (80%) cases using the subtype B specimen extracted at 40,000 copies, and 1/5 (20%) using the subtype C specimen extracted at 40,000 copies (Figure 4_6). Accordingly, the Life Technologies kit was selected for use.

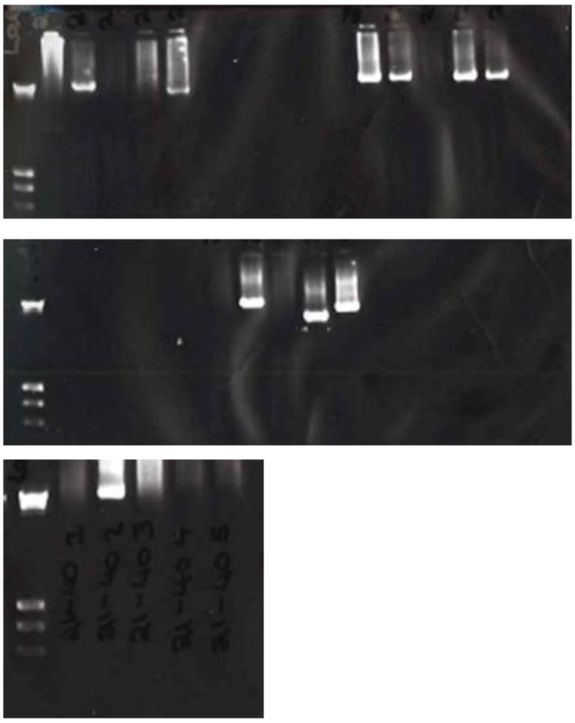


Figure 4_6. Comparison of the Life Technologies Platinum PCR Supermix High Fidelity and the Roche Expand Long Template PCR System. One subtype B clinical specimen and one subtype C clinical specimen were extracted at 20,000 copies and 40,000 copies. Following reverse transcription, cDNA was diluted 1:2 and used as the input into parallel testing of both PCR systems. Each extract was amplified in 5 replicates. Lane 1 of each row contained Hyperladder I. Lanes 2 - 11 of the top row contained the subtype B specimen extracted with 20,000 copies, of which lanes 2 - 6 show the Life Technologies protocol, and lanes 7 - 11 show the Roche protocol. 2/5 replicates of the Life Technologies protocol and no replicates using the Roche protocol successfully amplified. Lanes 12 - 16 of the top row and lanes 2 - 6 of the middle row contain the subtype B specimen extracted at 40,000 copies; lanes 12 - 16 show the Life Technologies protocol and lanes 2-6 show the same Roche protocol. 4/5 replicates of the Life technologies protocol amplified; no replicates of the Roche protocol amplified. Lanes 7 - 16 in the middle row contained the subtype C specimen extracted at 20,000 copies, of which lanes 7 - 11 show the Life Technologies system and lanes 12 - 16 show the Roche system. 2/5 Life Technologies replicates have amplified; none of the Roche replicates were successful. Lanes 2 - 6 of the bottom row contain the subtype C specimen extracted at 40,000 copies, and show the Life Technologies replicates only . 1/5 of the Life Technologies replicates amplified; none of the Roche replicates amplified (not shown).

4.2.2.3.1 Optimisation for subtype A amplification

Having selected the Life Technologies PCR system, the reaction conditions needed optimisation such that subtypes A, B, C and D were amplified with comparable efficiency. The multiple subtype optimisation was begun using a subtype A clinical specimen, using the same reaction conditions used during the Roche/Life Technologies comparison. Although these conditions amplified subtype A, the highest amplification achieved was 31% using the 1:2 cDNA dilution (Figure 4_7a). As this dilution equated to a total input of 100 template molecules per reaction, the reaction efficiency was clearly suboptimal. The reaction conditions were therefore optimised to increase the efficiency of subtype A amplification.

Firstly, the primer concentration was increased to 0.30µM per reaction, and 8 replicates of a clinical specimen extracted at 20,000 and 40,000 copies were run (Figure 4_7b). 7/8 (75%) of the 20,000 copies reactions showed amplification at the correct size, although 1 of these also showed evidence of mispriming. The 40,000 copies reactions showed 5/8 (63%) positive reactions, of which 2/5 (40%) also showed mispriming. 3/8 (38%) of reactions were negative. These conditions clearly improved the subtype A reaction efficiency, however could potentially negatively impact the reactions for other subtypes, so at this stage reactions using subtype A, B, C and D specimens were run at primer concentrations of 0.30 and 0.35µM per reaction.

Figure 4_7c shows the results from the primer titration reactions. Overall, 16 replicates of each of subtype A, B, C and D were performed using a primer concentration of 0.35µM per reaction, and eight replicates of each subtype were performed using a primer concentration of 0.30µM per reaction. At 0.35µM, there were 4/16 (25%) positive reactions for subtype A, 6/16 (38%) positive reactions for subtype B, 12/16 (75%) positive reactions for subtype C and 9/16 (56%) positive reactions for subtype D. The reactions for subtypes A and D showed widespread mispriming (11/16, 69% and 7/16, 44%, respectively). The best results were for subtype C. The reactions using a primer concentration of 0.30µM showed 3/8 (38%) positive reactions for subtypes A and B, 7/8 (88%) positive reactions for subtype C and 5/8 (63%) positive reactions for subtype D. Mispriming was still evident in the subtype A and D reactions (5/8, 62% and 2/8, 25%, respectively), however, this was reduced.

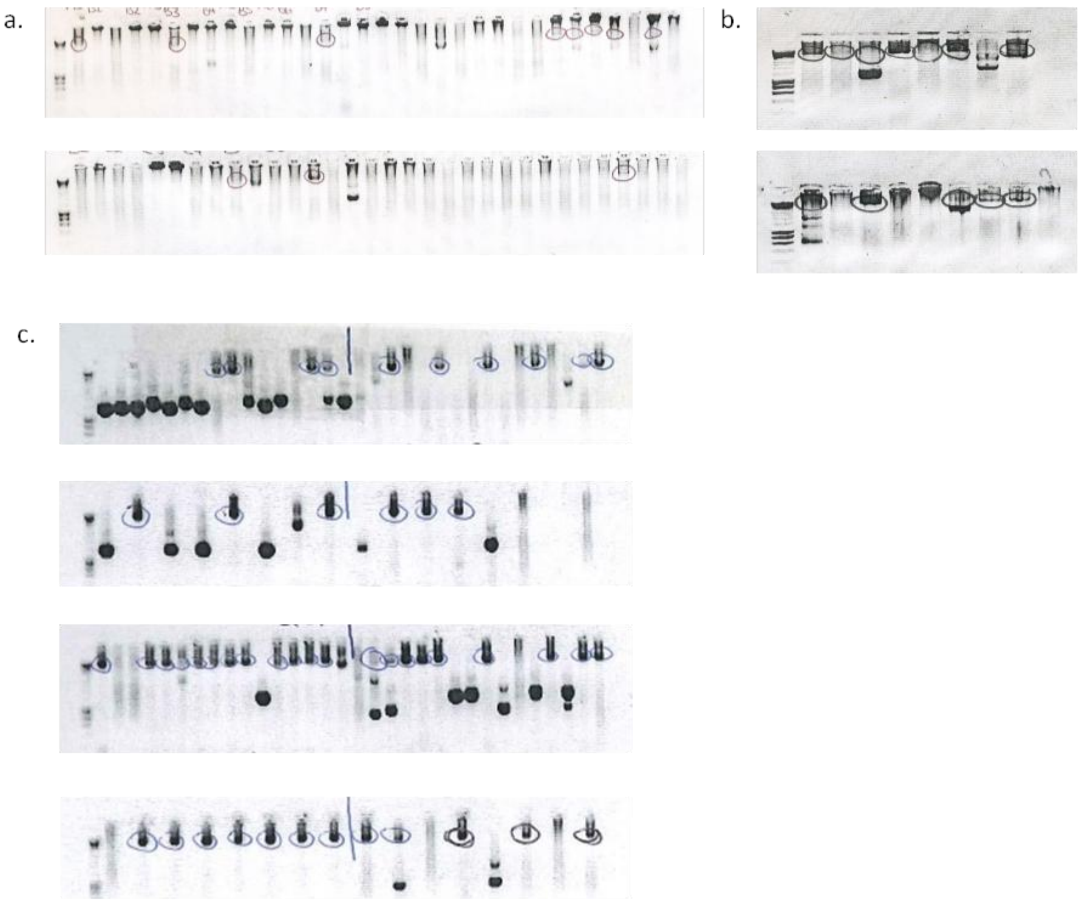


Figure 4_7. **Optimisation of PCR reactions to work successfully with subtypes A, B, C and D. a)** Initial amplification of a subtype A clinical specimen. 16 reactions each using undiluted cDNA and cDNA dilutions of 1:2, 1:4 and 1:10 were performed. Lane 1 of each row contained Hyperladder I. Lanes 2-17 of the top row contained neat cDNA (3/16 positive, 19%); lanes 18-34 contained the1:2 cDNA dilution (5/16, 31% positive). In the middle row, lanes 2-17 contained the 1:4 cDNA dilution (2/16 positive,13%) and lanes 18-34 contained the 1:10 cDNA dilution (1/16, 6%). The highest number of positive reactions (5/16, 31%) was gained from using a 1:2 cDNA dilution. Lane 35 contained the negative control. This low rate of successful reactions indicated suboptimal reaction conditions for the amplification of subtype A specimens. **b)** Optimisation of subtype A reaction conditions. 8 replicates of a subtype A clinical specimen extracted at both 20,000 and 40,000 copies were amplified using an increased primer concentration of 0.35µM per reaction. Lane 1 in each row contained Hyperladder I. Lanes 2-9 in the top row contained the amplified products from the 20,000 copies extraction. Lanes 2-9 in the bottom row contained the products from the 40,000 copies extraction, and lane 10 contained the negative control. 7/8 of the 20,000 copies reactions showed amplification at the correct size; 5/8 of the 40,000 copies extraction showed amplification at the correct size. **c)** Optimisation of primer concentrations across subtypes A, B, C and D. 16 reactions using a primer concentration of 0.35µM each primer per reaction and 8 reactions using a primer concentration of 0.3µM each primer per reaction were performed using extracts from clinical specimens of subtypes A, B, C and D. Lane 1in each row contained Hyperladder I; The top and second rows contained the subtype A and B reactions using 0.35µM (top row) and 0.30µM (second row). The third and bottom rows contained the subtype C and D reactions using 0.35µM (third row) and 0.30µM (bottom row). Top and second rows: lanes 2-17 contained subtype A reactions; lanes 18-34 contained subtype B reactions. Third and bottom rows: lanes 2-17 contained subtype C reactions and

lanes 18-34 contained subtype D reactions. Although the reactions across both concentrations were generally robust when using the subtype C extract, the higher primer concentration caused multiple mispriming events across all four subtypes, particularly subtypes A and D. The lower primer concentration of 0.3µM produced fewer mispriming events.

Further testing showed that a primer concentration at 0.25µM per reaction reduced the degree of mispriming, but also reduced the overall reaction efficiencies for subtypes A and D. The number of PCR cycles in the first round was increased from 35 to 40 to help alleviate this.

The annealing temperature was optimised across subtypes A, B, C, and D. A preliminary gradient PCR was performed using annealing temperatures of 50°C, 53°C, 56°C, 59°C, 62°C, 65°C and four replicates each of clinical specimens with 20,000 RNA copies extracted . The reactions for subtypes B and C viruses were robust over a range of temperatures, but the reactions for subtypes A and D were not. On this basis, the gradient PCR was repeated using eight replicates of subtypes A and D only, in order to perform a more rigorous assessment for these subtypes. The results showed that, whilst mispriming was still occurring with subtype A, the best temperature was between 59°C and 62°C for both subtypes. Based on the results from both gradient PCRs, an annealing temperature of 60°C was chosen.

Two final checks were performed to ensure that the selected reaction conditions were optimal. Firstly, 16 replicates each of a subtype A specimen with 20,000 copies extracted were amplified as neat cDNA, and dilutions of 1:2 and 1:4 at primer concentrations per reaction of both 0.25 and 0.30µM. The results showed comparable amplification success at neat cDNA (10/16, 63%) for both primer concentrations, but better success at the 1:4 cDNA dilution when using the 0.25µM concentration (7/15, 47% vs. 3/14, 21%). In terms of mispriming, the 0.25µM reactions showed generally less mispriming (7/16, 44%, 8/16, 50%, 3/16, 19% for neat cDNA, 1:2 and 1:4, respectively) than the 0.3µM reactions (12/16, 75%, 7/16, 44%, 4/16, 25% for neat cDNA, 1:2 and 1:4, respectively). These results suggested that mispriming was related as much to the template concentration as the primer concentration.

One of the potential recombinant specimens that had a high viral load and a large sample volume available (500,000 copies/ml and 1500µl, respectively) was amplified using eight replicates each at the following cDNA dilutions: neat, 1:4, 1:10, 1:50 and 1:100. The degree of mispriming found in these reactions was 7/8 (88%) using neat cDNA, 3/8 (38%) using the 1:4 dilution, 2/8 (25%) using the 1:10 dilution, 0/8 using 1:50, and 0/8 using 1:100, respectively. Given that the aim of the optimisation was to optimise for low numbers of template molecules, these results suggested that the assay was optimised sufficiently.

The final reaction conditions chosen represented the best available combination that achieved consistent amplification across all four validation subtypes. These conditions differed from the original protocol, in that the optimised assay used a different PCR system, different primer concentrations, a different annealing temperature and a different number of PCR cycles. Table 4_1 summarises the main features of the finalised protocol compared with the Nadai et al. and unadapted CHAVI protocols.

	Nadai et al. (2008)	**CHAVI (2009)**	**Final protocol**
Template	Plasma RNA	Plasma RNA	Plasma RNA
No. of fragments to amplify HIV-1 genome	2 - 3	1	1
Genomic coverage (HXB2)	623 - 9636	552 - 9636	552 - 9636
Optimum HIV-1 RNA copies (input)	50,000 - 375,000	20,000 - 40,000	10,000 - 20,000
Optimised HIV-1 subtypes	B	B, C	A, B, C, D
Method	2-step, nested RT-PCR	2-step, nested RT-PCR	2-step, nested RT-PCR
cDNA synthesis method	Oligo dT or gene-specific priming (UNINEF 7')	Gene-specific priming (1.R3.B3R)	Gene-specific priming (1.R3.B3R)
PCR Reagents	Expand Long Template PCR kit (Roche Diagnostics)	Expand Long Template PCR kit (Roche Diagnostics)	Platinum PCR Supermix High Fidelity (Life Technologies)
Forward and Reverse Primer concentrations (µM)	0.4	0.3	0.25
Number of PCR cycles first round	30	35	40
Number of PCR cycles second round	30	35	45
Annealing Temperature (ºC)	60	55	60

Table 4_1. Comparison of the final optimised RT-PCR protocol for near full-length HIV-1 amplification with the unadapted Nadai et al. and CHAVI protocols.

4.2.2.4 Dilution of cDNA and DNA sequencing for single genome analysis

Following the optimisation of reaction conditions, the final procedural aspect requiring optimisation was the manner in which a single template molecule would be amplified. The limiting dilution of cDNA into a PCR reaction such that it is assumed that only one copy of cDNA is present in each amplified reaction is based on Poisson's distribution. In a situation where cDNA has been diluted such that only 30% of reactions are positive, then Poisson's distribution states that there is only a single amplified copy of DNA in 80% of cases (Simmonds et al., 1990).

However, in long range PCR, the efficiency of a reaction decreases dramatically with the length of the template (Dittmar et al., 1997). Therefore, even in an optimised reaction, the possibility exists that even at cDNA dilutions that produce only 30% positive reactions, there may still be more than one molecule of template DNA present. Additionally, the reaction conditions were a set of 'best fit' conditions chosen to work across a range of subtypes, rather than a set of conditions optimised for a single subtype at maximum efficiency. To test that the Poisson distribution for single template amplification still applied for this particular set of reaction conditions, three specimens were sequenced at a range of dilutions below 30% positivity required, up to a dilution of 1:200, which is the theoretical limit where one molecule of cDNA was input into each reaction (20,000 copies of RNA transcribed into 200µl of cDNA = 100 copies per/µl, input into PCR reaction = 2 µl, therefore, a 1:200 dilution = 0.5 copies/µl, which gives an input of one copy into the reaction) (Table 4_2). During sequence analysis, dilutions for each specimen were analysed in parallel, and assessed for the presence of mixed bases that might indicate a mixed population. Any differences in individual bases between dilutions were assessed by examining the raw electropherogram data for that primer. No differences were found in individual bases that could not be explained by either low quality sequence or a low sequencing signal. Where evidence of low quality signal was found, these primers were repeated to obtain a higher quality sequence. In all cases, this resulted in the difference between nucleotides, or the presence of mixed bases, being resolved. Nevertheless, in order to circumvent the possibility of more than one molecule being amplified in a reaction, the highest dilution that still produced a number of wells that could be sequenced (≥4) was used for single genome analysis. This represented a final departure from the methodology of the CHAVI protocol which used a traditional limiting dilution to produce single molecule amplification.

Study number	cDNA dilution	cDNA copies/µl	cDNA copies/reaction	Positive reactions at this dilution (%)
33365	1:80	1.25	2.5	26
	1:100	1	2	12
	1:200	0.5	1	6
8179	1:10	10	20	12
	1:20	5	10	4
40534	undiluted	22	44	40
	1:4	5.5	11	7

Table 4_2. Comparison of full-length sequences at different cDNA dilutions. Three study specimens were sequenced at 6 dilutions below the 30% positive rate suggested by Poisson's distribution as producing an amplicons from one amplified copy of cDNA 80% of the time, and one dilution above the 30% positive rate, in order to analyse whether the decreased sensitivity of long-range PCR had an impact on the sequences from these specimens. No differences between sequences were found, indicating that decreased sensitivity did not change the statistical distribution of mixed populations. Nevertheless, sequences at the highest dilution that produced ≥4% positive reactions per plate were used for single genome analysis.

Chapter 5: Identification of CRF50_A1D

5.1 Single genome amplification and sequencing of 6 patients with recombinant HIV-1

5.1.1 Normalisation of plasma to contain 20,000 copies of vRNA and amplification of 6 near full-length genomes

All of the study samples obtained were frozen plasma taken between 2000 and 2011. The viral load range was 9,148 - 500,000 copies/ml and the range of sample volumes was 270 - 1500µl (Table 5_1). Owing to low viral loads, 20,000 copies could not be extracted in 3/6 samples.

Relatively consistent amplification was observed across all specimens; differences in reaction efficiency were most likely due to specimen age and routine storage conditions, and the freeze-thaw cycle required to separate aliquots for transfer between centres. Only one genome amplified successfully at the theoretical cDNA dilution of one molecule/reaction; a further specimen amplified at 1.37 molecules/reaction. Each of the remaining four specimens showed amplification lower than the 30% Poisson distribution set point. One specimen (11762) only amplified with an undiluted cDNA input. This specimen was re-extracted, however still only amplified when using undiluted cDNA (Appendix 5_1).

Study number	Sample date	Viral load (copies/ml)	Volume received (µl)	Volume required to extract 20,000 copies (µl)	Copies extracted	Final concentration undiluted cDNA (copies/µl)	Final cDNA concentration used for single genome analysis (copies/µl)	Positive reactions (%)
33365	28/04/2003	500,000	1500	40	20,000	100	0.5	6
8179	25/07/2000	150,380	500	133	20,000	100	5	4
40534*	23/04/2003	31,111	500	643	4,356	22	5.5	7
34567	27/03/2003	11,893	270	1682	3,211	16	16	8
11762†	12/03/2011	74,595	410	268	20,000	100	100	19
12792*	06/07/2010	9,148	610	550	548	2.74	0.685	13

Table 5_1. Characteristics of the study samples received. Four specimens were received from two London centres (33365, 8179, 40534, 34567), and two specimens were received from one Northwest centre (11762, 12792). The amount of plasma received varied depending on storage protocols at each centre. 20,000 RNA copies were extracted where the plasma volume received/viral load allowed, and amplification was performed at the highest dilution resulting in 4% success or higher.

*****These specimens were originally extracted at a volume for 20,000 vRNA copies. However, the amplification of these extractions failed. The value in the 'copies extracted' column reflects subsequent, successful extractions.

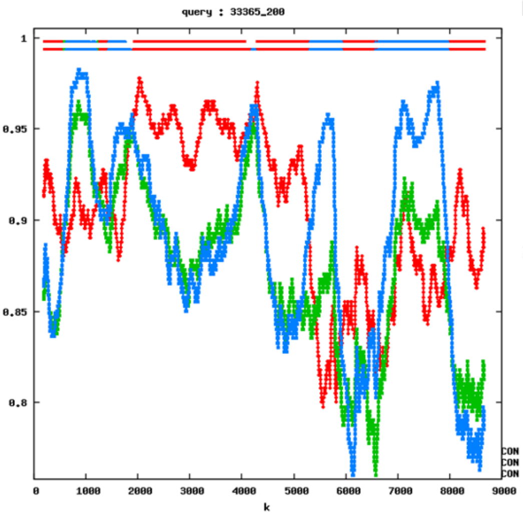
†This sample only amplified as undiluted cDNA. A subsequent extraction of 10,443 extracted vRNA copies also only amplified undiluted. Please see Appendix 5_1 for further details

5.2 Recombination analysis of near full-length sequences

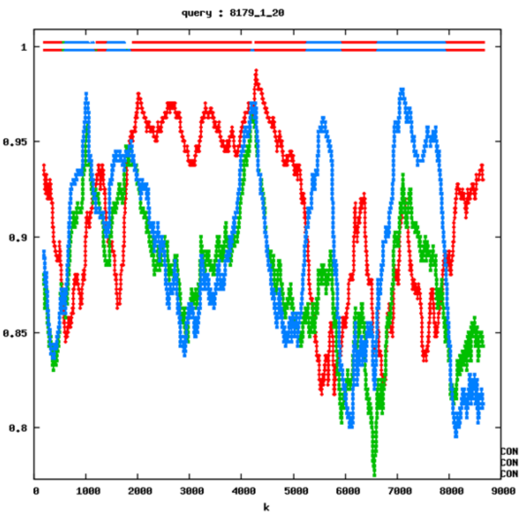
5.2.1 RIP analysis

RIP analysis of the six recombinant specimens showed a putatively identical A1/D structure for 5/6 specimens (33365, 8179, 40534, 11762, 12792), and a complex A1/B/D structure for the remaining specimen (34567) (Figure 5_1). In all six plots the lowest region of similarity was between positions approximately located at 5,900 – 6,700; when translated back to HXB2 numbering, this region corresponds to the hypervariable region of *env* coding for the V1-V3 loops, and is, as such, a region where lower similarity to subtype demarcations is expected. Each of the A1/D plots suggested a recombinant structure with approximately three breakpoints in *gag,* one breakpoint in *pol*, along with a possible region of uncertainty; one breakpoint in the accessory genes, and a further three breakpoints in *env*. The plot for the complex A1/B/D structure switched between subtypes A1, B and D for the first 3000 nucleotides, with no clear regions of identity. This plot showed a clear region of subtype B between positions 3000 and 5000; after this point, the plot showed the same A1/D structure as the other five plots.

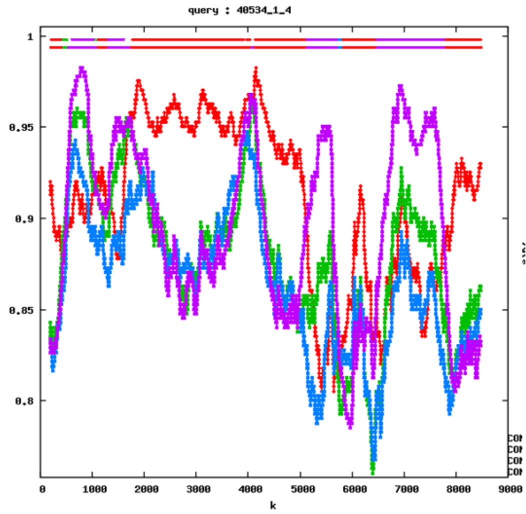
a.


query : 33365_200
1
0.95
0.9
0.85
0.8
s(k)
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
k
CON
CON
CON

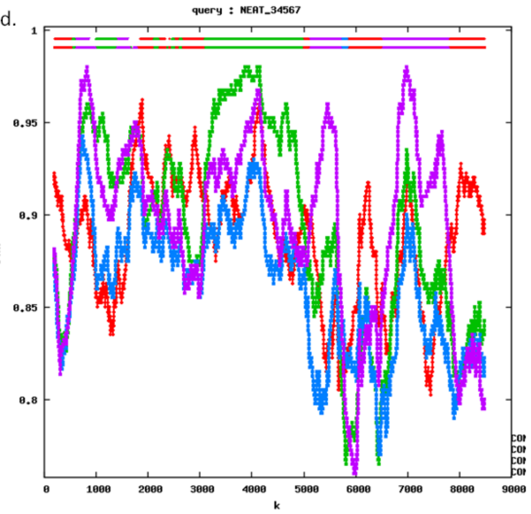
b.


query : 8179_1_20
1
0.95
0.9
0.85
0.8
s(k)
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
k
CON
CON
CON

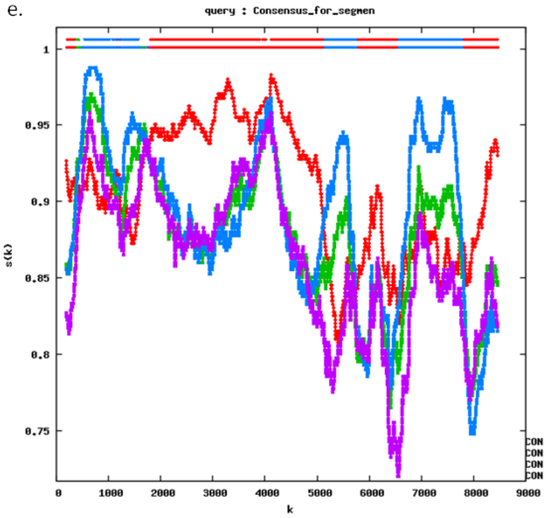
c.


query : 40534_1_4
1
0.95
0.9
0.85
0.8
s(k)
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
k
CON
CON
CON
CON

d.


query : NEAT_34567
1
0.95
0.9
0.85
0.8
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
k
CON
CON
CON
CON

e.


query : Consensus_for_segmen
1
0.95
0.9
0.85
0.8
0.75
s(k)
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
k
CON
CON
CON
CON

f.

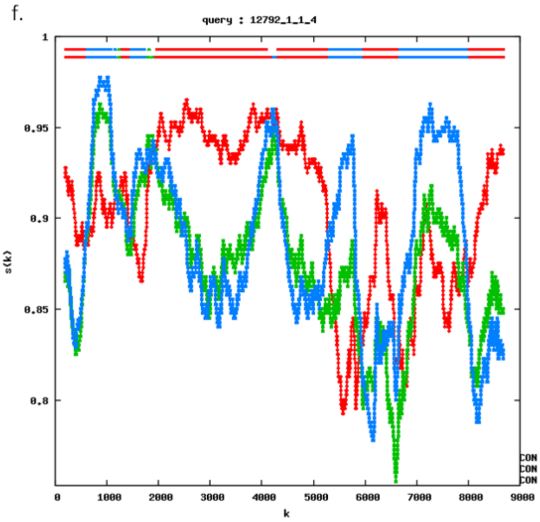

query : 12792_1_1_4
1
0.95
0.9
0.85
0.8
s(k)
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
k
CON
CON
CON

Figure 5-1. Simplified Recombinant Identification Program (RIP) scans of 6 putatively recombinant specimens. The Los Alamos program RIP was used to scan six near full-length recombinant HIV-1 sequences. The settings used were a window size of 400bp, gap stripped and a step size of 20bp. RIP performs a similarity scan and reports results graphically using the closest subtype match from its own set of subtype reference sequences. The x axis is numbered according to the nucleotide position of the query sequence. The bar at the top of the scan shows the most likely subtype for that region of the sequence. In plots, a), b), e) and f), subtype A1 is represented in red, subtype B in green, and subtype D in blue. In plots c) and c) subtype A1 is represented in red, subtype B in green, subtype C in blue and subtype D in purple. In plot e), subtype F2 is represented in purple. Five of the six plots (a - c), e) and f)) showed near-identical A1/D structures; plot d) showed a complex A1/B/D structure. a) 33365, b) 8179, c)40534 d) 34567 e)11762 f)12792.

5.2.2 jpHMM analysis

The RIP analysis gave a basic 'working' structure of the recombinant specimens, but did not designate breakpoint locations; these were initially defined using jpHMM. The jpHMM analysis of the six specimens also showed five specimens with largely identical structures (33365, 8179, 40543, 11762, 12792) and one specimen with a complex A1/A2/D/B/U structure (34567) (Figure 5_2).The breakpoint locations and confidence intervals for the five A1/D specimens are summarised in Table 5_2. Generally, the jpHMM breakpoint locations and subtype classifications showed a good level of consistency among the five A1/D specimens, and the structure suggested by the RIP screening. However, there were three structural differences observed in the jpHMM plots among the five A1/D specimens. Firstly, specimen 11762 (Figure 5_2 e) showed a region of subtype D uncertainty in the p2-p7 regions of the *gag* gene which was classified as subtype A1 in the other four A1/D specimens. However, as the limits of the region of uncertainty corresponded to the confidence intervals of the subtype A1 region (Table 5_2*), this was considered an area for further investigation rather than a true structural difference.

The second structural difference was the location of the *pol* breakpoint in specimen 12792 (Figure 5_2 f), Table 5_2†). The breakpoint location for this specimen was set 109 -111 nucleotides distant from the corresponding breakpoint in the other four specimens. However, the plot indicated that the breakpoint was in the same location as the other four specimens, but there was an area of A1/D uncertainty adjacent to it. In these cases, the jpHMM algorithm sets the 'breakpoint' location as the central point of the uncertainty region; this was therefore considered an area for further analysis rather than a true structural difference.

The final difference in the jpHMM plots concerned specimens 33365 and 12792 (Figure 5_2a and f). These two specimens showed a region of subtype D uncertainty in the *env* gene (confidence intervals 7320 - 7496 and 7314 - 7496, respectively). Again, these were considered areas for further investigation.

The jpHMM analysis of specimen 34567 (Figure 5_2 d) provided a clearer picture of this complex specimen than the RIP analysis. Whereas the RIP plot indicated that the first 3000 nucleotides of the sequence did not clearly correspond to any particular subtype, the jpHMM plot showed two clear regions in the genome with the same structure as the five A1/D specimens. These regions were the very beginning of *gag*, which had an identical A1/D breakpoint (1162 ±8), and from the breakpoint in *tat/rev* (5983 ±23) to the end of the genome. This raised the possibility for the first time that this specimen may be the result of a further recombination event between the A1/D structure seen in the five A1/D specimens, and another infection.

The remaining regions of 34567 were not so clearly defined. While there were clear regions of subtype B, including a B/A1 breakpoint that largely corresponded with the D/A1 breakpoint in *pol* (2535 ±16 c.f. 2489 ±26), there were areas of subtype A uncertainty in *pol,* and a region of *gag* (1479-1625) that could not be assigned to any known subtype.

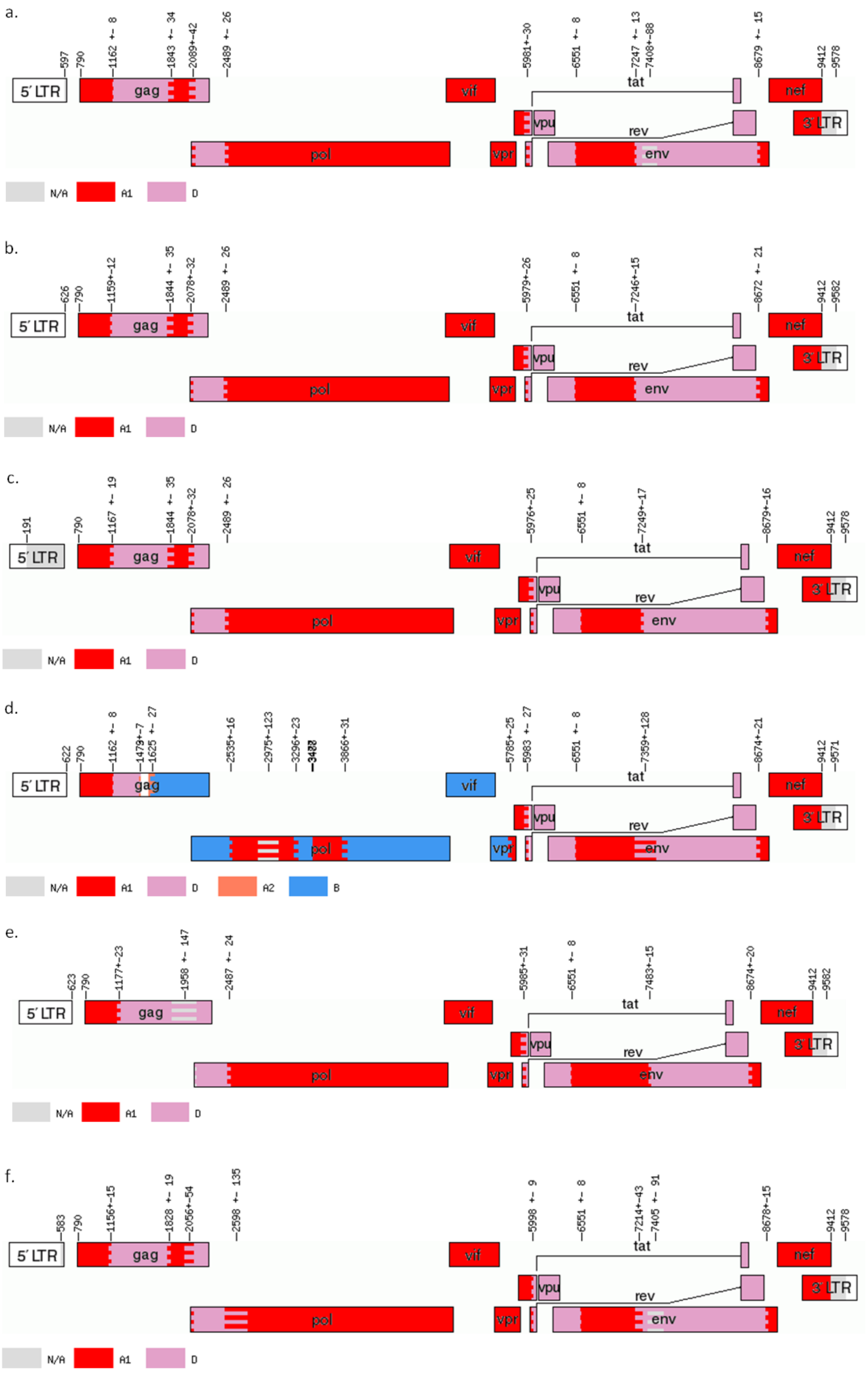

a.
597
790
1162 +- 8
1843 +- 34
2089+-42
2489 +- 26
5981+-30
6551 +- 8
7247 +- 13
7408+-88
8679 +- 15
9412
9578
5′ LTR
gag
pol
vif
vpr
vpu
tat
rev
env
nef
3′ LTR
N/A
A1
D
b.
626
790
1159+-12
1844 +- 35
2078+-32
2489 +- 26
5979+-26
6551 +- 8
7246+-15
8672 +- 21
9412
9582
5′ LTR
gag
pol
vif
vpr
vpu
tat
rev
env
nef
3′ LTR
N/A
A1
D
c.
191
790
1167 +- 19
1844 +- 35
2078+-32
2489 +- 26
5976+-25
6551 +- 8
7249+-17
8679+-16
9412
9578
5′ LTR
gag
pol
vif
vpr
vpu
tat
rev
env
nef
3′ LTR
N/A
A1
D
d.
622
790
1162 +- 8
1479+-7
1625 +- 27
2535+-16
2975+-123
3296+-23
3488
3866+-31
5785+-25
5983 +- 27
6551 +- 8
7359+-128
8674+-21
9412
9571
5′ LTR
gag
pol
vif
vpr
vpu
tat
rev
env
nef
3′ LTR
N/A
A1
D
A2
B
e.
623
790
1177+-23
1958 +- 147
2487 +- 24
5985+-31
6551 +- 8
7483+-15
8674+-20
9412
9582
5′ LTR
gag
pol
vif
vpr
vpu
tat
rev
env
nef
3′ LTR
N/A
A1
D
f.
583
790
1156+-15
1828 +- 19
2056+-54
2598 +- 135
5998 +- 9
6551 +- 8
7214+-43
7405 +- 91
8678+-15
9412
9578
5′ LTR
gag
pol
vif
vpr
vpu
tat
rev
env
nef
3′ LTR
N/A
A1
D

Figure 5_2. jpHMM analysis of six recombinant HIV-1 sequences. Putative recombinant HIV-1 sequences were submitted to the online implementation of jpHMM at the GOBICS server. The program used its own stored reference alignment and statistical algorithm to determine subtype classifications, breakpoint locations and 95% confidence intervals. Breakpoint locations and confidence intervals are marked on each plot and are equivalent to HXB2 numbering. In each plot, subtype A1 is represented in red, subtype A2 in coral, subtype D in lavender and subtype B in blue. Areas of subtype uncertainty are grey. The plots show the same structure as identified by the RIP scanning, that is, five A1/D recombinants (a-c, e and f) and one complex structure (d). The jpHMM result for d) shows a more complex structure than that elucidated by RIP, with regions of subtype A2 and one extra region of subtype B. a)33365. b)8179. c) 40534. d)34567. e)11762. f)12792.

Breakpoint	**Specimen ID**					**Gene/region**	
	33365	**8179**	**40534**	**11762**	**12792**		
1	1162 (1154-1170)	1159 (1147-1171)	1167 (1148-1186)	1177 (1154-1200)	1156 (1141-1171)	*gag*	p24
2	1843 (1809-1877)	1844 (1809-1879)	1844 (1809-1879)	1958* (1811-2105)	1828 (1809-1847)	*gag*	p24
3	2089 (2047-2131)	2078 (2046-2110)	2078 (2046-2110)		2056 (2002-2110)	*gag*	p1
4	2489 (2463-2515)	2489 (2465-2515)	2489 (2465-2515)	2487 (2475-2499)	2598† (2463-2733)	*pol*	Protease
5	5981 (5951-6011)	5979 (5953-6005)	5976 (5951-6001)	5985 (5973-5997)	5998 (5989-6007)	*tat/rev*	-
6	6551 (6543-6559)	6551 (6543-6559)	6551 (6543-6559)	6551 (6539-6563)	6551 (6543-6559)	*env*	gp120
7	7247 (7234-7260)	7246 (7231-7261)	7249 (7232-7266)	7483 (7471-7495)	7214 (7171-7257)	*env*	gp120
8	8679 (8664-8694)	8672 (8651-8693)	8679 (8663-8695)	8674 (8662-8686)	8678 (8663-8693)	*env*	gp41

Table 5_2. jpHMM-assigned breakpoint locations and confidence intervals for the five A1/D recombinant sequences. Breakpoint locations with confidence intervals as determined by jpHMM (HXB2 numbering). Confidence intervals are indicated in parentheses under each breakpoint. The gene/region column indicates the position of each breakpoint in relation to the genetic structure of HIV-1. The breakpoint locations were generally consistent across all five specimens, indicating that the same A1/D recombinant structure is shared.

*****This corresponds to a region of subtype D uncertainty. Refer to Figure 5_2.

†This corresponds to a region of subtype A uncertainty. Refer to Figure 5_2.

Please note that specimens 33365 and 12792 had a region of subtype D uncertainty in *env* (confidence intervals 7320 - 7496 and 7314 - 7496, respectively).

5.2.3 Sliding window and maximum likelihood analysis

5.2.3.1 Sliding window analysis of specimens 33365, 8179, 40534, 11762, 12792

The RIP and jpHMM results from the six sequenced specimens suggested that five of the specimens were of a putatively identical A1/D recombinant structure, and that the remaining specimen was a unique A1/B/D/U structure. However, these results contained regions of structural uncertainty that required a third method to both confirm previous results and refine the genomic areas that were not definitively classified. Accordingly, sliding window analysis was performed using each of the six recombinant specimens as a query sequence.

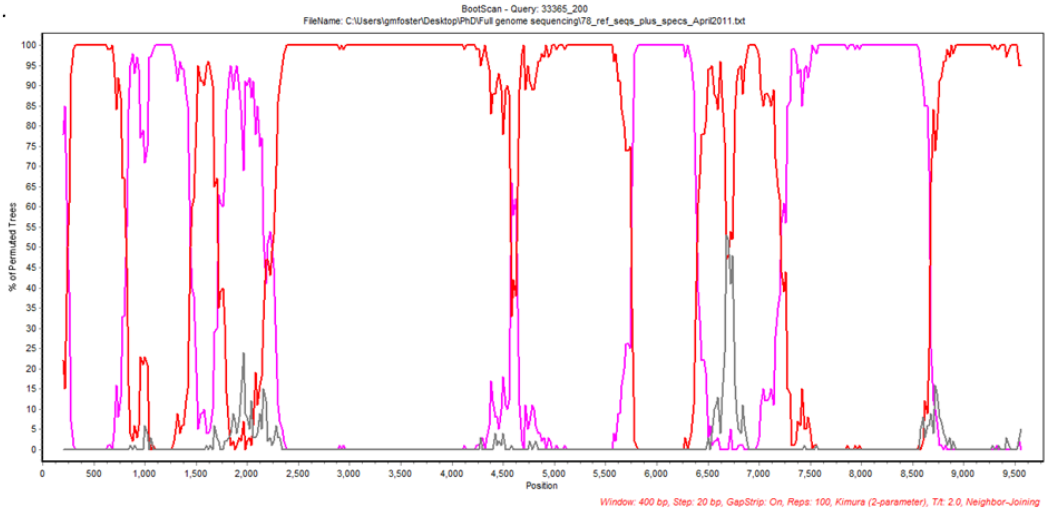
Figure 5_3 shows bootscan plots for specimens 33365, 8179, 40534, 11762 and 12792. The bootscanning further confirmed the A1/D structure indicated by the RIP and jpHMM analyses and resolved areas of structural uncertainty suggested by the other two algorithms. The first area of structural uncertainty indicated by the jpHMM analysis was an uncertain area of subtype classification in the p2-p7 region of the *gag* gene in specimen 11762. The bootscan plot for this specimen (Figure 5_3 e, nucleotide positions 1400 - 1700) classified this region as a clear subtype A1 match with >70% of all permuted trees containing this specimen with subtype A1.

The second region of uncertainty concerned the uncertain region of subtype A1 in specimen 12792 immediately adjacent to the *pol* breakpoint. Figure 5_3 f showed that, firstly, the *pol* breakpoint for this specimen was in the same location as in the other four A1/D specimens (approximately position 2250), and that the region directly following this (nucleotide positions 2250-2600) was a clear subtype A1 region with a near-100% match to subtype A1 across the region.

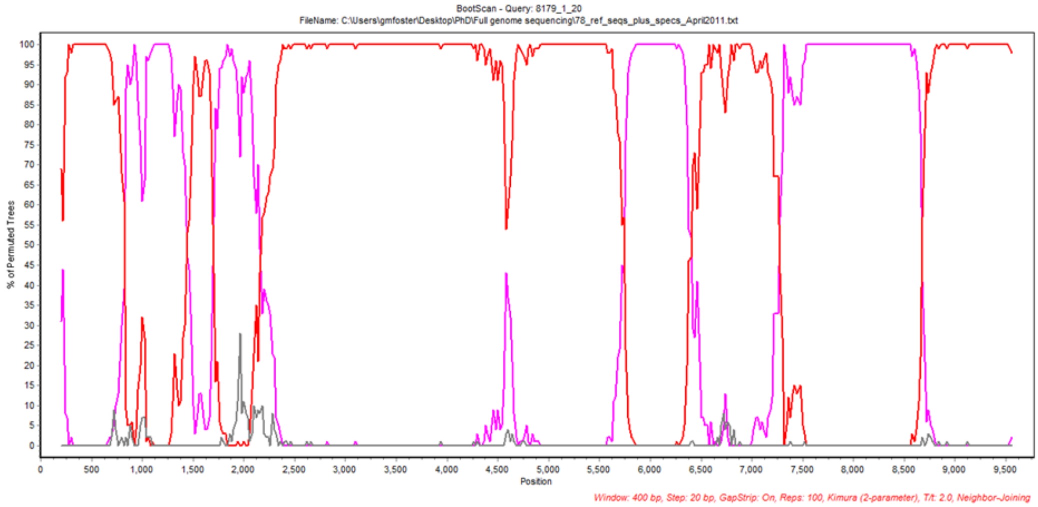
The final structural difference suggested by jpHMM was the area of uncertainty in the *env* gene in specimens 33365 and 12792. Figure 5_3a and f shows this region for both specimens. In both cases, this area (Figure 5_3a and f, nucleotide positions 7,200-7,500) showed a definitive subtype D classification, with >85% of all trees matching this subtype.

The bootscanning analysis also showed some uncertain areas not shown by jpHMM and RIP. These were in specimens 33365 (dip in identity in *env* Figure 5_3a), and 40534 (dip in identity in *env* Figure 5_3c). However, these were not present in the jpHMM or RIP analyses.

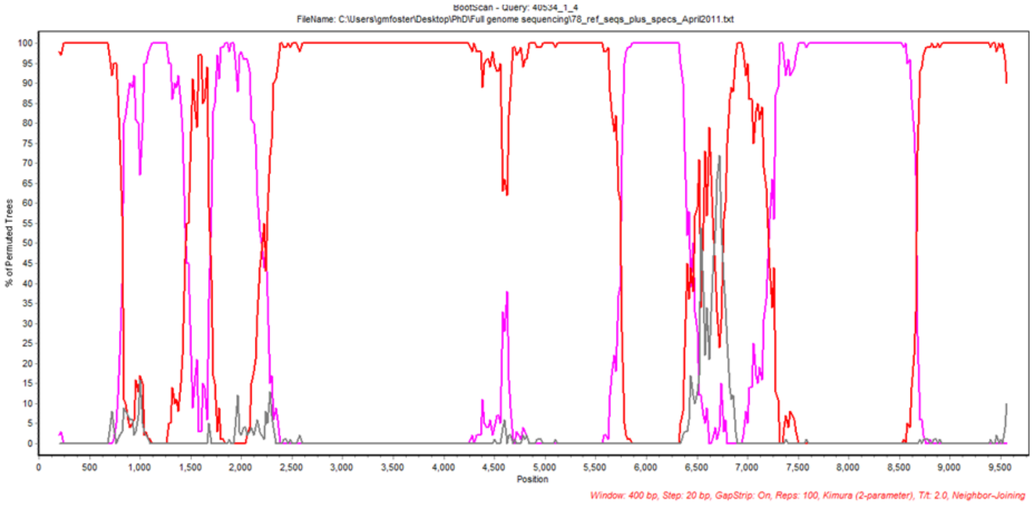
a.


BootScan - Query: 33365_200
FileName: C:\Users\gmfoster\Desktop\PhD\Full genome sequencing\78_ref_seqs_plus_specs_April2011.txt
% of Permuted Trees
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500
6,000
6,500
7,000
7,500
8,000
8,500
9,000
9,500
Position
Window: 400 bp, Step: 20 bp, GapStrip: On, Reps: 100, Kimura (2-parameter), T/t: 2.0, Neighbor-Joining

b.


BootScan - Query: 8179_1_20
FileName: C:\Users\gmfoster\Desktop\PhD\Full genome sequencing\78_ref_seqs_plus_specs_April2011.txt
% of Permuted Trees
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500
6,000
6,500
7,000
7,500
8,000
8,500
9,000
9,500
Position
Window: 400 bp, Step: 20 bp, GapStrip: On, Reps: 100, Kimura (2-parameter), T/t: 2.0, Neighbor-Joining

c.


BootScan - Query: 40534_1_4
FileName: C:\Users\gmfoster\Desktop\PhD\Full genome sequencing\78_ref_seqs_plus_specs_April2011.txt
% of Permuted Trees
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500
6,000
6,500
7,000
7,500
8,000
8,500
9,000
9,500
Position
Window: 400 bp, Step: 20 bp, GapStrip: On, Reps: 100, Kimura (2-parameter), T/t: 2.0, Neighbor-Joining

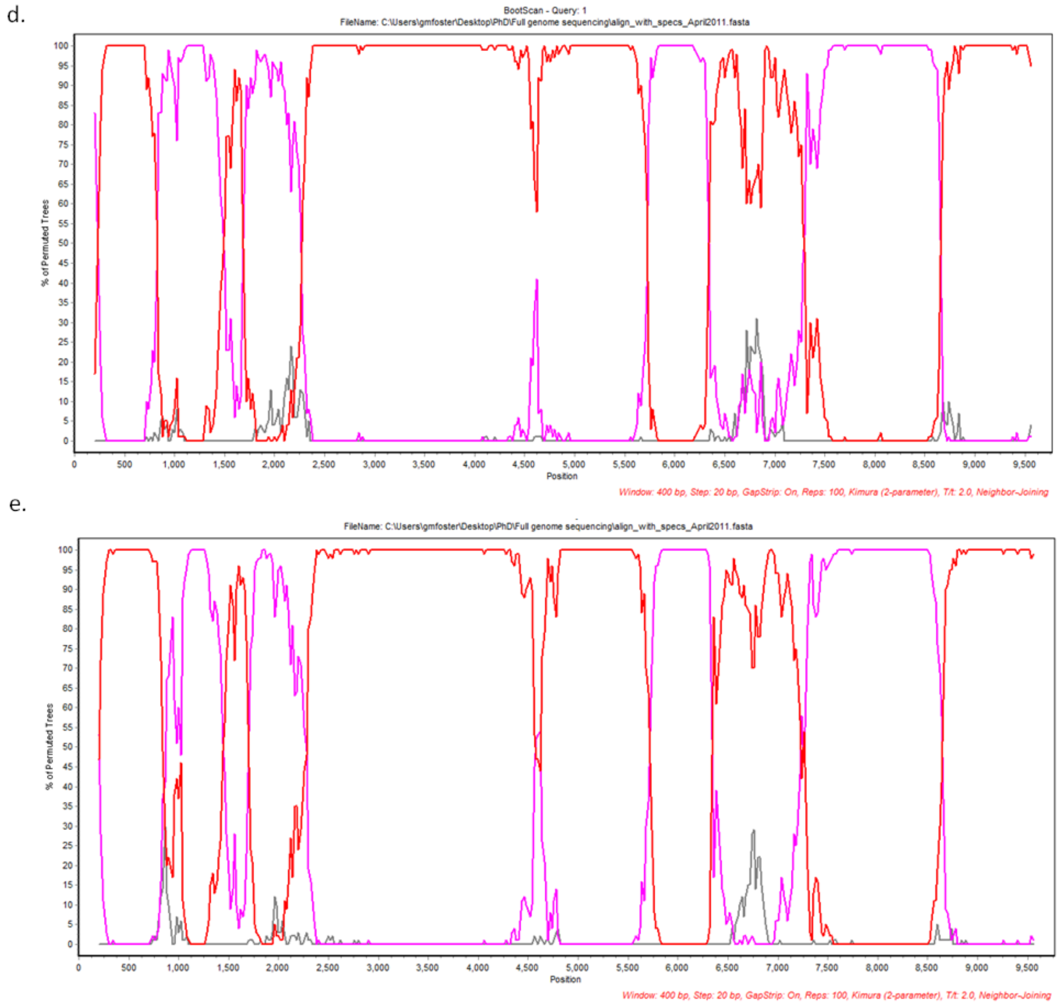
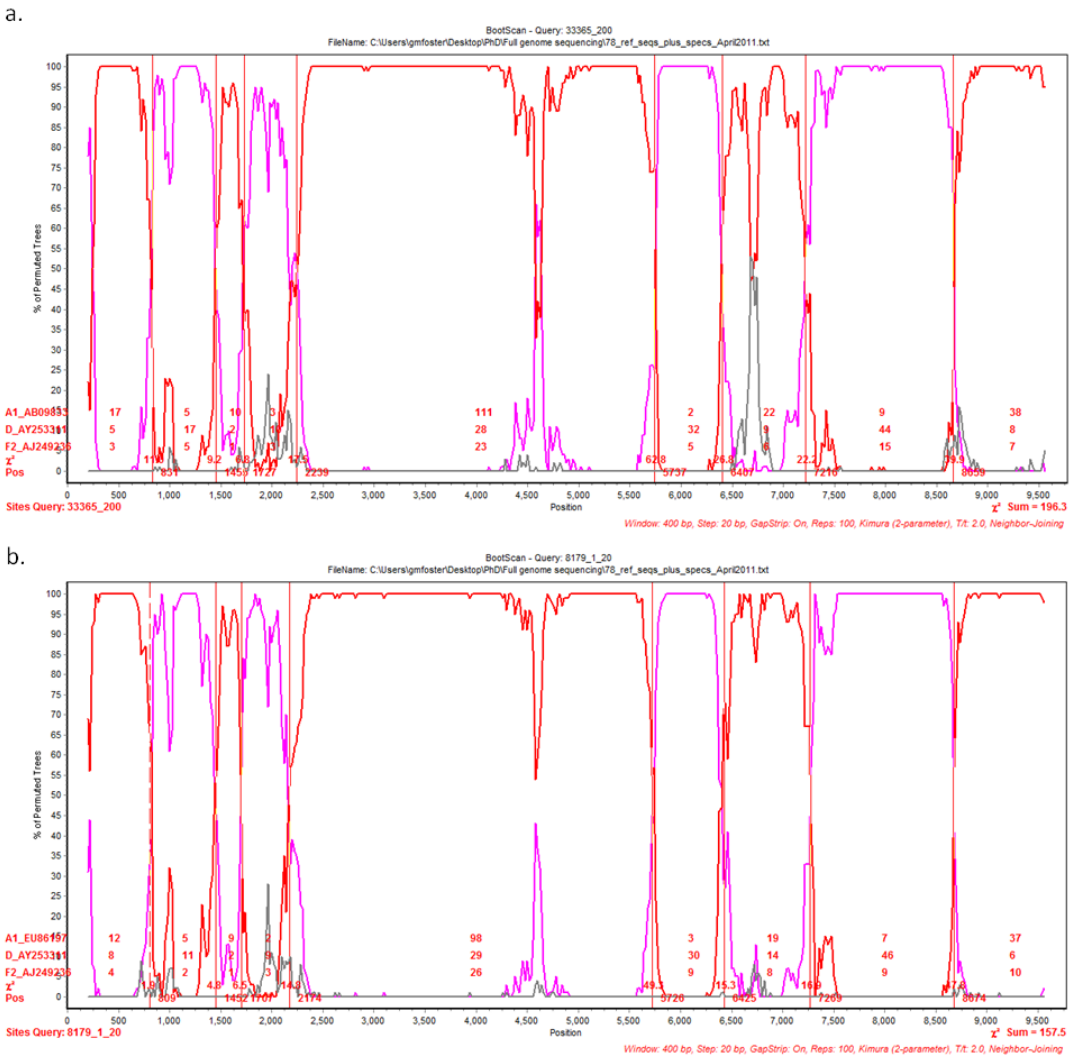


Figure 5_3. Bootscanning plots for the A1/D recombinant specimens (33365, 8179, 40534, 11762, 12792). Bootscanning plots from Simplot sliding window analysis using a window size of 400bp, a step size of 20bp and 100 bootscanning replicates. The y axis shows the percentage of permuted trees that the query sequence clustered with the closest subtype match from the reference alignment. The x axis shows the nucleotide position of the sequence (not HXB2 numbering). Subtype A is represented in red, subtype D in lavender, and subtype F (outgroup) in grey. All five specimens show identical bootscanning plots, with five subtype A1 regions and four subtype D regions. a) Specimen 33365; b) Specimen 8179; c) Specimen 40534; d) Specimen 11762; e) Specimen 12792

Figure 5_4 shows the same bootscan plots as Figure 5_3, but with informative sites analysis added. The informative sites plots show the breakpoint positions that maximised the x^2 score around the 50% crossover point between subtypes. All positions showing a subtype crossover, or a region of <70% identity, e.g. nucleotide positions 4575 - 4625 in specimen 33365 (Figure 5_4a), had putative breakpoint/s placed, and the statistical significance of the informative sites at that location was assessed using Fisher's exact test. The breakpoints remaining on the plots were those that showed statistical significance of <0.05. The areas of uncertainty introduced by Simplot were not replicated in jpHMM and RIP and were not statistically significant.



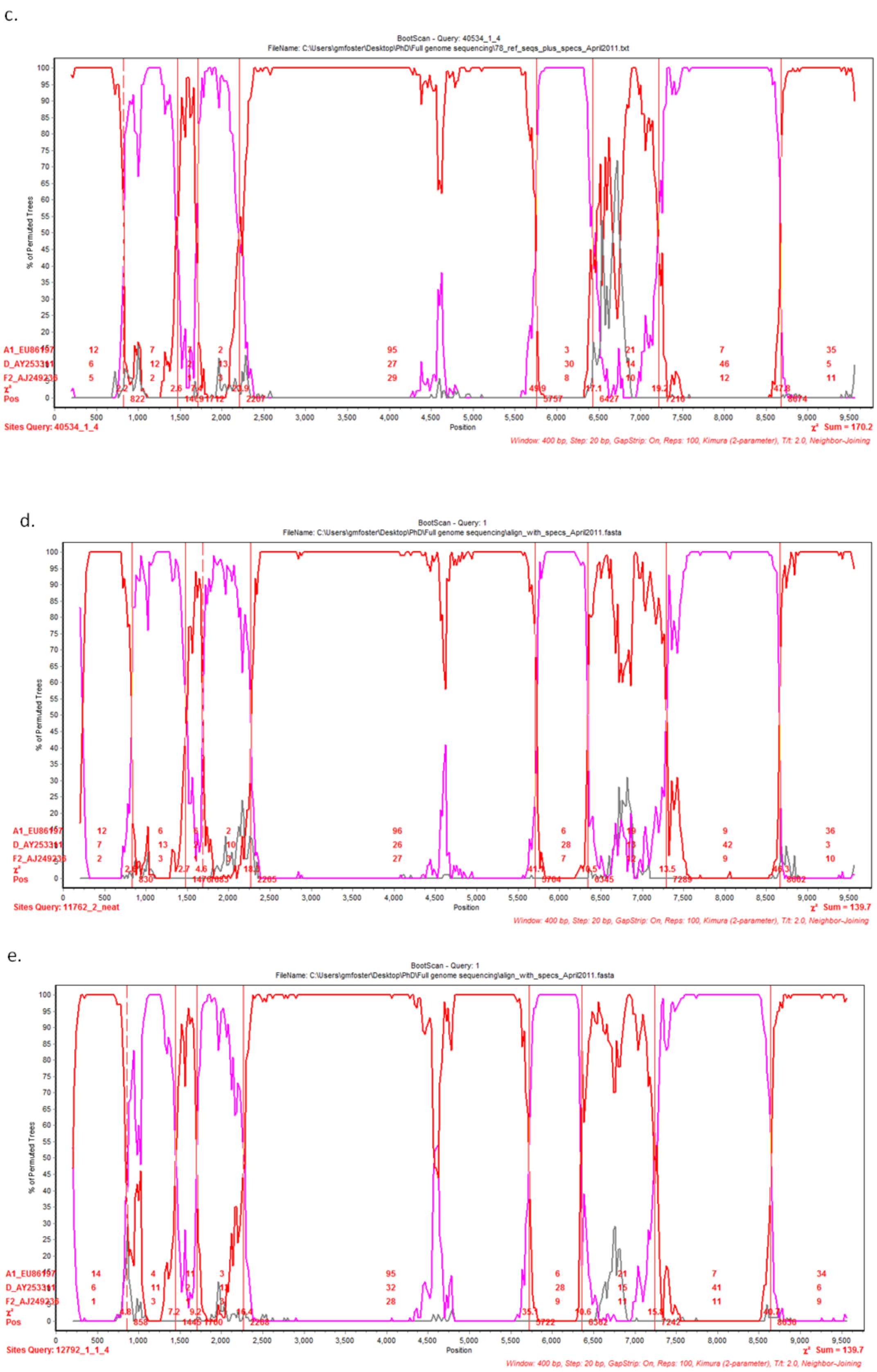


Figure 5_4. Bootscanning plots with informative sites analysis for the A1/D recombinant specimens 33365, 8179, 40534, 11762 and 12792. Bootscanning plots with informative sites analysis. The x and y axes are labelled as for Figure 5_3, as are the

colours representing each subtype. The vertical red lines represent putative breakpoint locations. Reference sequences used for informative sites analysis are shown in red at the left of the plot. Red numbers in the plot show the number of informative sites between the reference sequence, the query sequence and the outgroup sequence for that section of the alignment. The total x^2 sum across the whole genome is shown on the right. Breakpoints were positioned to maximise the x^2 at each location around the 50% crossover point between subtypes. The statistical significance of each breakpoint was checked using Fisher's exact test. These breakpoints were used to create alignment slices that were used for downstream phylogenetic analyses. a) Specimen 33365; b) Specimen 8179; c) Specimen 40534; d) Specimen 11762; e) Specimen 12792.

Table 5_3 summarises the breakpoint locations for specimens 33365, 8179, 40534, 11762 and 12792. Breakpoint locations were consistent across all three specimens, and were similar to those identified by jpHMM; minor differences in these can be attributed to differences between the jpHMM and sliding window algorithms.

Breakpoint	Specimen ID				
	33365	8179	40534	11762	12792
1	1273 (0.0007)	1272 (<0.0001)	1272 (<0.0001)	1275	1284 (0.0176)
2	1883 (0.0011)	1851 (<0.0001)	1875 (<0.0001)	1854	1874 (0.0032)
3	2100 (0.0048)	2100 (0.005)	2097 (0.0001)	2097 (0.0194)	2069 (0.0018)
4	2503 (<0.0001)	2503 (<0.0001)	2493 (<0.0001)	2523 (<0.0001)	2553 (0.0001)
5	6007 (<0.0001)	6004 (<0.0001)	6007 (<0.0001)	6009 (<0.0001)	6003 (<0.0001)
6	6611 (<0.0001)	6602 (<0.0001)	6621 (<0.0001)	6585 (0.0008)	6614 (0.0006)
7	7417 (<0.0001)	7440 (<0.0001)	7407 (<0.0001)	7400 (0.0001)	7363 (<0.0001)
8	8567 (<0.0001)	8573 (<0.0001)	8562 (<0.0001)	8596 (<0.0001)	8548 (<0.0001)

Table 5_3. Sliding window breakpoint locations and statistical significance for the five A1/D specimens. Breakpoint locations determined using bootscanning analysis for the five A1/D recombinant specimens. Breakpoint locations are displayed in HXB2 numbering. The statistical significance of each breakpoint was assessed using informative sites analysis and Fisher's exact test and is shown in parentheses under each breakpoint. Breakpoints were consistent among the five specimens.

Three A1/D specimens (33365, 8179 and 40534) were taken forward for maximum likelihood analysis of the putative subtype A1 and D regions of the genome. The identified breakpoints in the alignment were used to create ‘slices’ of the alignment that corresponded to the putative pure subtype regions of the recombinant genome. These slices were then used for subsequent downstream phylogenetic analysis. Each specimen was sliced and analysed separately, according to the breakpoints in Table 5_3. Therefore, nine slices of the alignment were made, corresponding to HXB2 coordinates 552 - 1272 (slice 1), 1273 - 1851 (slice 2), 1852 - 2100 (slice 3), 2101 - 2503 (slice 4), 2504 - 6004 (slice 5), 6005 - 6602 (slice 6), 6603 - 7440 (slice 7), 7441 - 8573 (slice 8), and 8574 - 9636 (slice 9), respectively (using specimen 8179 as reference coordinates).

Likelihood mapping to assess phylogenetic signal was performed on each slice of each specimen. Table 5_4 shows the likelihood mapping of each slice for specimens 3365, 8179 and 40534. There was a high percentage of unresolved quartets for slice 3 across all three specimens (10.3, 9.8, 10.6 for specimens 33365, 8179 and 40534, respectively); slice 4 also showed a uniformly high proportion of unresolved quartets (4.3, 4.4, and 4.8, respectively). Slice 1 also showed uniformly moderately high unresolved quartets of 2.1, 2.5, and 2.3, respectively. These high proportions of unresolved quartets indicated that any downstream phylogenetic analysis would need to be interpreted with caution, given the low phylogenetic signal for these regions. This lack of signal was attributable to the short length of the slices (248 and 402 nucleotides, respectively), and, in the case of slice 4, the highly conserved nature of the region of *pol* that this slice covered. No other slice showed a uniform percentage of unresolved quartets greater than 2.0; no individual figure was greater than 2.5%.

Entropy testing between subtypes B and D for the region covered by alignment slice 4 showed that the number of informative sites present between the two subtypes was only six. This indicated that downstream phylogenetic analyses would potentially be weak at distinguishing between subtypes B and D in this genomic region.

Slice	Study number								
	33365			8179			40534		
	HXB2 coordinates (length)	Unresolved quartets (%)	Partly resolved quartets (%)	HXB2 coordinates (length)	Unresolved quartets (%)	Partly resolved quartets (%)	HXB2 coordinates (length)	Unresolved quartets (%)	Partly resolved quartets (%)
1	552-1273 (721)	2.1	5.6	552-1272 (720)	2.5	5.7	552-1272 (720)	2.3	5.7
2	1274-1883 (609)	1.6	5.1	1273-1851 (578)	1.5	4.8	1273-1875 (602)	1.5	4.5
3	1884-2100 (216)	10.3	6.0	1852-2100 (248)	9.8	6.3	1876-2097 (221)	10.6	6.0
4	2101-2503 (402)	4.3	6.2	2101-2503 (402)	4.4	6.0	2098-2493 (395)	4.8	6.2
5	2504-6007 (3503)	0.1	1.3	2504-6004 (3500)	0.1	1.8	2494-6007 (3513)	0.1	1.6
6	6008-6611 (603)	2.0	4.4	6005-6602 (597)	2.3	4.8	6008-6621 (613)	1.9	4.8
7	6612-7417 (805)	2.1	4.4	6603-7440 (837)	1.9	5.0	6622-7407 (785)	2.0	4.2
8	7418-8567 (1149)	0.3	2.6	7441-8573 (1132)	0.3	2.6	7408-8562 (1154)	0.3	2.5
9	8568-9636 (1068)	0.6	3.8	8574-9636 (1062)	0.7	3.7	8563-9636 (1073)	0.6	3.6

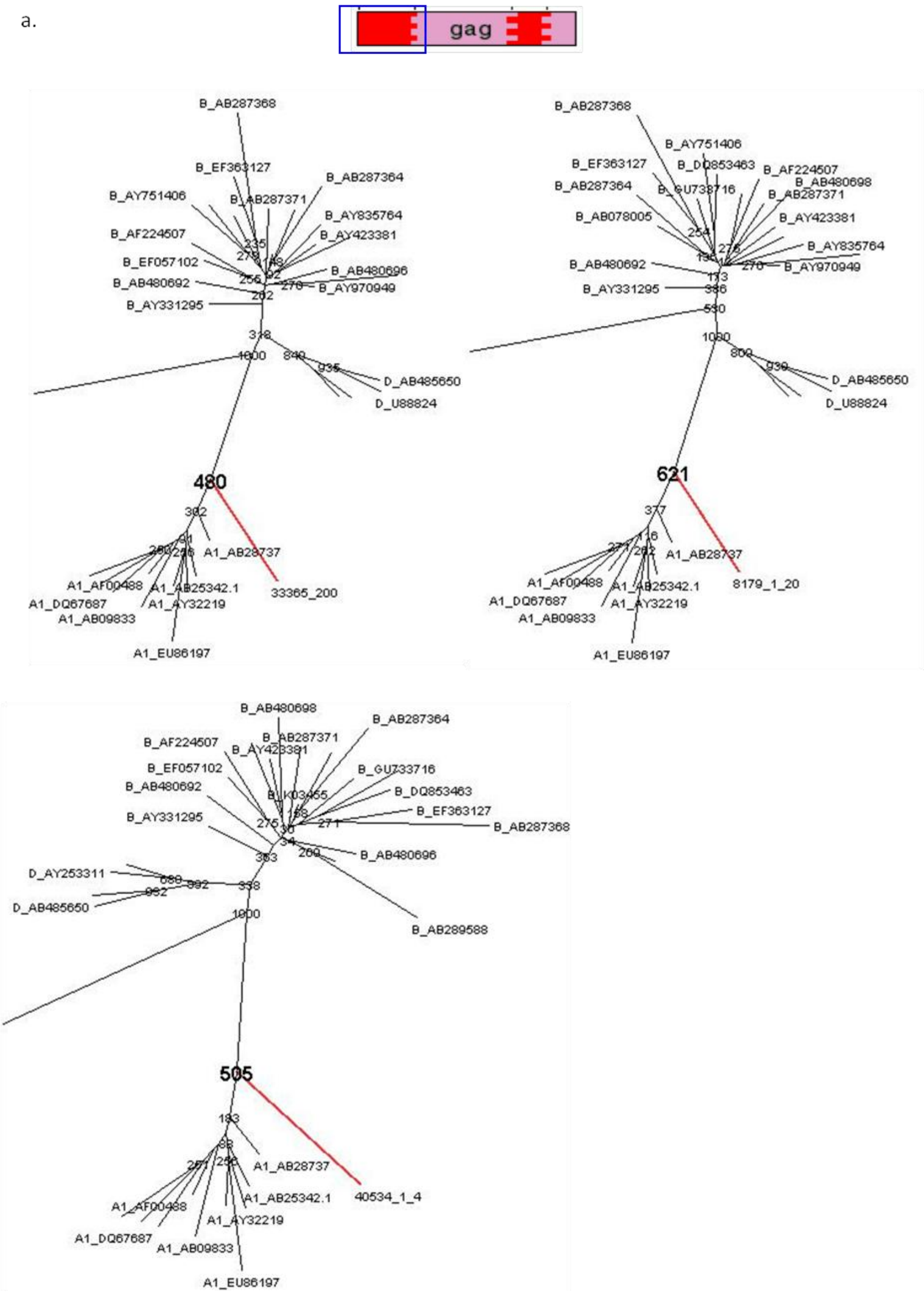
Table 5_4.Likelihood mapping of three A1/D specimens. Alignment slice numbers, HXB2 co-ordinates and nucleotide length, and likelihood mapping results for specimens 33365, 8179 and 40534. Likelihood mapping results are expressed as the number of unresolved and partly resolved quartets in each slice. A high percentage of un- and partly-resolved quartets indicates a lack of phylogenetic signal in the alignment. Slices 3 and 4 showed the highest percentages of un- and partly-resolved quartets, which was potentially due to them being the shortest alignments.

Following likelihood mapping, PhyML maximum likelihood analysis was performed. Figure 5_5 shows the PhyML maximum likelihood trees for specimens 33365, 8179 and 40534; Table 5_5 summarises the bootstrap support for each slice of each query specimen. Each fragment of each query specimen clustered with the pure subtype (A1 or D) indicated by the bootscanning analysis. As expected, the maximum likelihood trees for the slices with the highest consistent percentage of unresolved quartets returned the lowest bootstrapping support, i.e. slices 1 (Specimen 33365: 2.1% unresolved, 48% bootstrapping support; Specimen 8179: 2.5% unresolved, 62% bootstrapping support, Specimen 40534: 2.3% unresolved, 51% bootstrap support), 3 (Specimen 33365: 10.3% unresolved, 39% bootstrap support; Specimen 8179: 9.8% unresolved, 38% bootstrap support; Specimen 40534: 10.6% unresolved, 22% bootstrap support), and 4 (Specimen 33365: 4.3% unresolved, 58% bootstrap support; Specimen 8179: 4.4% unresolved, 57% bootstrapping support; Specimen 40534: 4.8% unresolved, 33% bootstrap support), respectively. To confirm the subtype classification for these fragments, the regions were first analysed using a higher range of subtypes from the reference alignment. When this did not substantially improve the bootstrap support (owing to the low phylogenetic signal), the posterior probabilities for these regions from the jpHMM analysis were examined. These results, in conjunction with the results from the entropy testing for slice 4, were sufficiently high to confirm the final subtype classification.

The final A1/D structure was predominantly subtype A1 in *pol* and the accessory genes, subtype D in *env*, and was fairly evenly split between subtype A1 and D in *gag*. There were three breakpoints in *gag,* one in *pol*, one in *tat/rev* and three in *env*, respectively. In *gag*, a breakpoint was located at either end of p24, suggesting that the entire coding region for the antigen was swapped in a recombination event. Similarly, the third breakpoint was located at the junction of the p7/p1 regions, again suggesting that entire coding regions were swapped in the formation of this recombinant. The distribution of subtypes in *gag* by protein were A1 (p17, p2, p7) and D (p24, p1, p6).

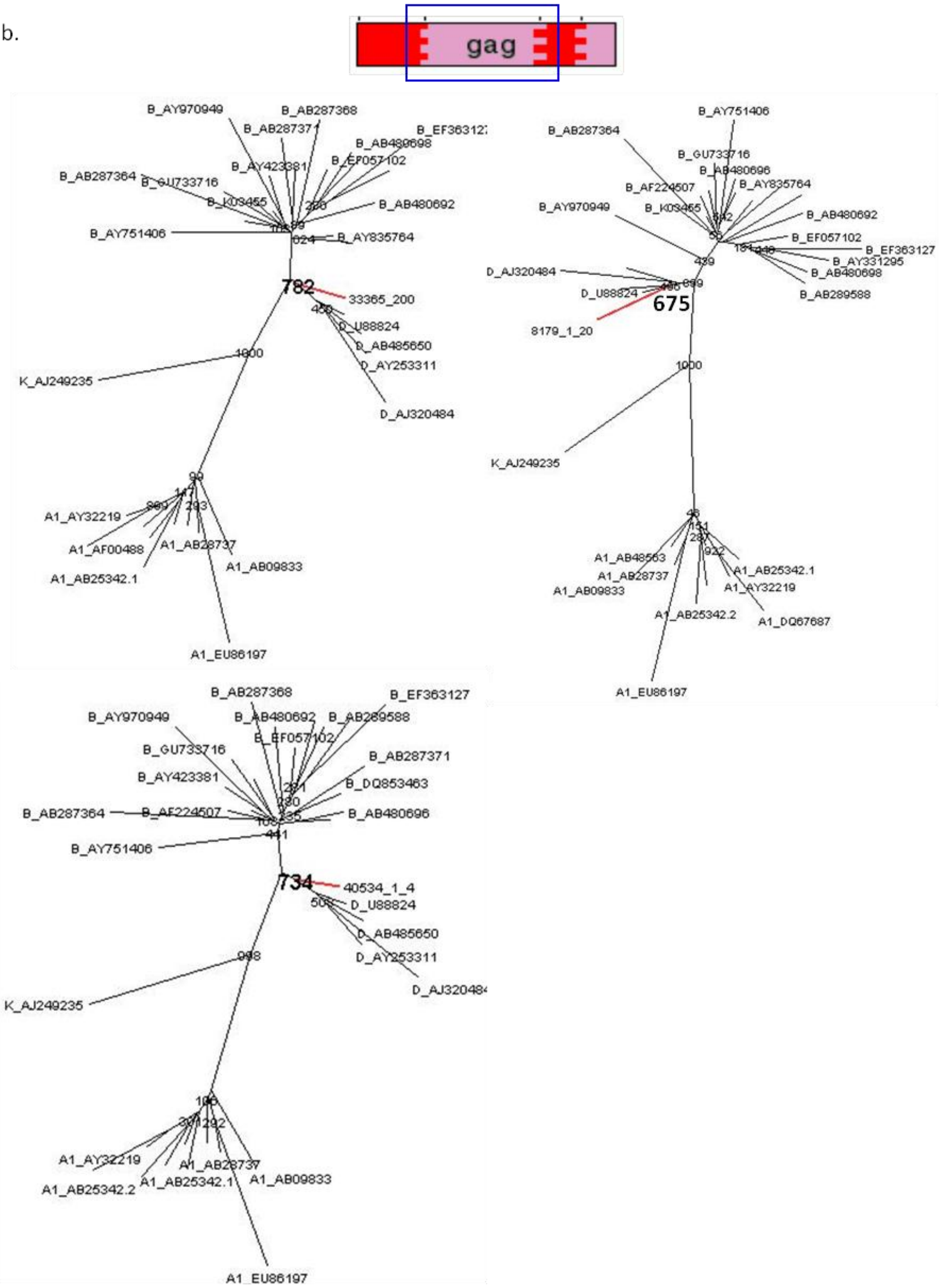
The single D/A1 breakpoint in *pol* was located approximately 250bp from the start of the protease. The remainder of the *pol* gene was subtype A1, as was *vif* and *vpr*. The breakpoint located at HXB2 6007 fell in the overlap of *tat* and *rev*, meaning that both of these genes were A1/D mosaics. *Vpu* was solely subtype D. Although the envelope gene was largely subtype D, three of the hypervariable regions, V1-V3 were subtype A1.

a.



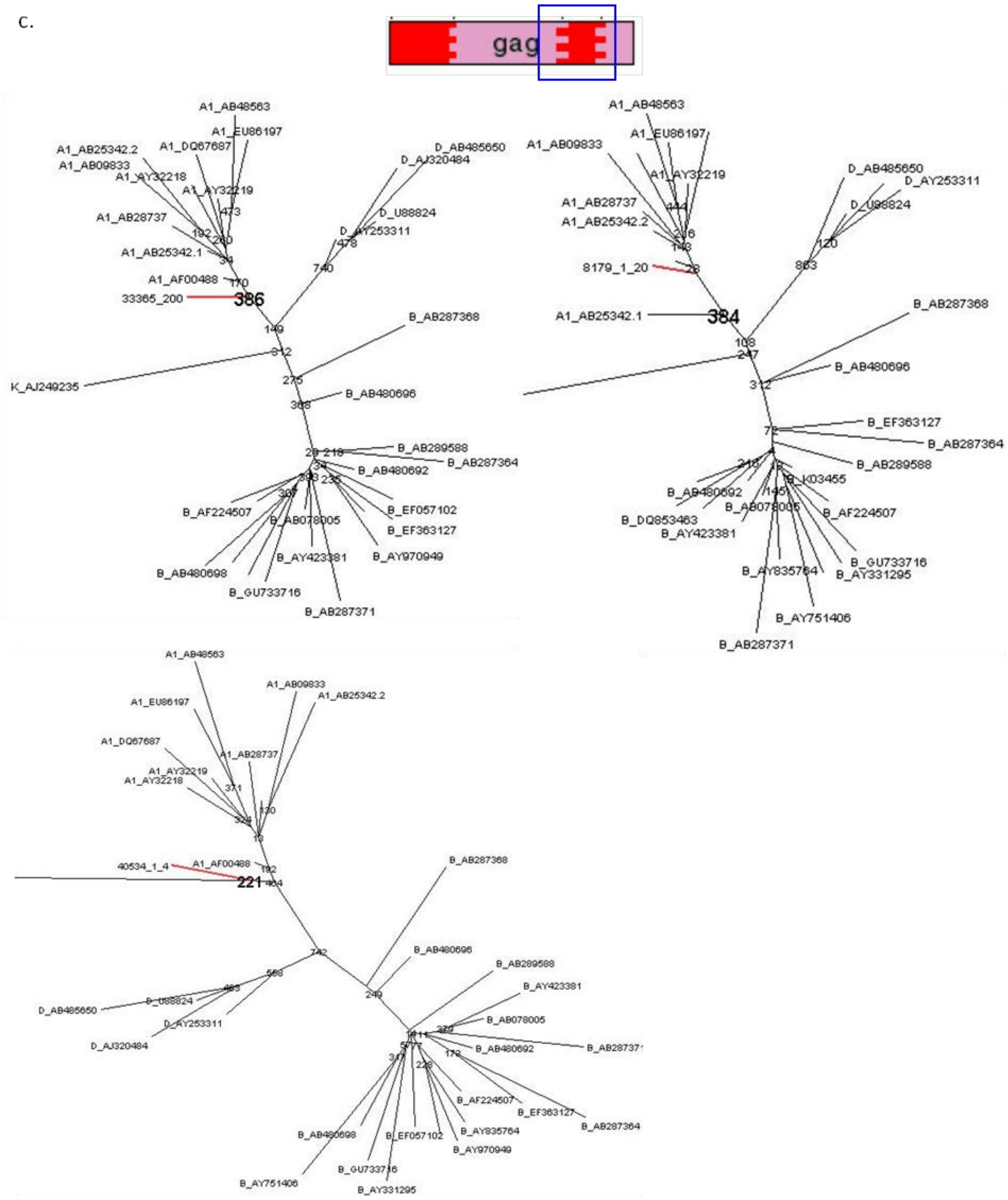
gag
B_AB287368
B_EF363127
B_AB287364
B_AY751406
B_AB287371
B_AY835764
B_AF224507
B_AY423381
B_EF057102
B_AB480696
B_AB480692
B_AY970949
B_AY331295
D_AB485650
D_U88824
480
A1_AB28737
33365_200
A1_AF00488
A1_AB25342.1
A1_DQ67687
A1_AY32219
A1_AB09833
A1_EU86197
B_DQ853463
B_AB480698
B_GU733716
B_AB078005
621
8179_1_20
B_K03455
D_AY253311
B_AB289588
505
40534_1_4

b.

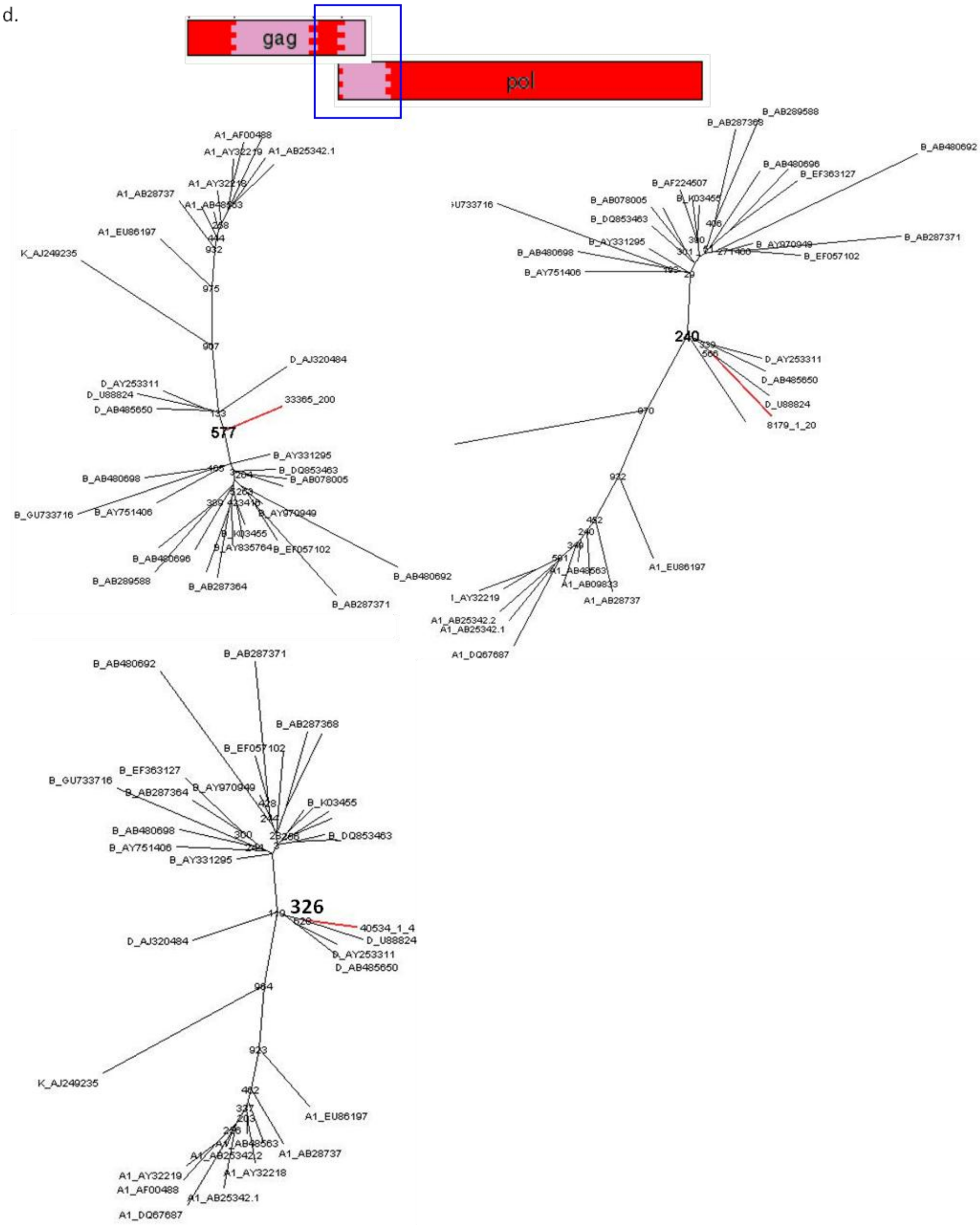


gag
B_AY970949
B_AB287368
B_AB287371
B_EF363127
B_AB480698
B_AY423381
B_EF057102
B_AB287364
B_GU733716
B_K03455
B_AB480692
B_AY751406
B_AY835764
782
33365_200
D_U88824
D_AB485650
D_AY253311
D_AJ320484
K_AJ249235
1000
A1_AY32219
A1_AF00488
A1_AB28737
A1_AB09833
A1_AB25342.1
A1_EU86197
B_AB480696
B_AF224507
B_AB289588
B_AY331295
675
8179_1_20
A1_AB48563
A1_AB25342.2
A1_DQ67687
B_DQ853463
734
40534_1_4
998

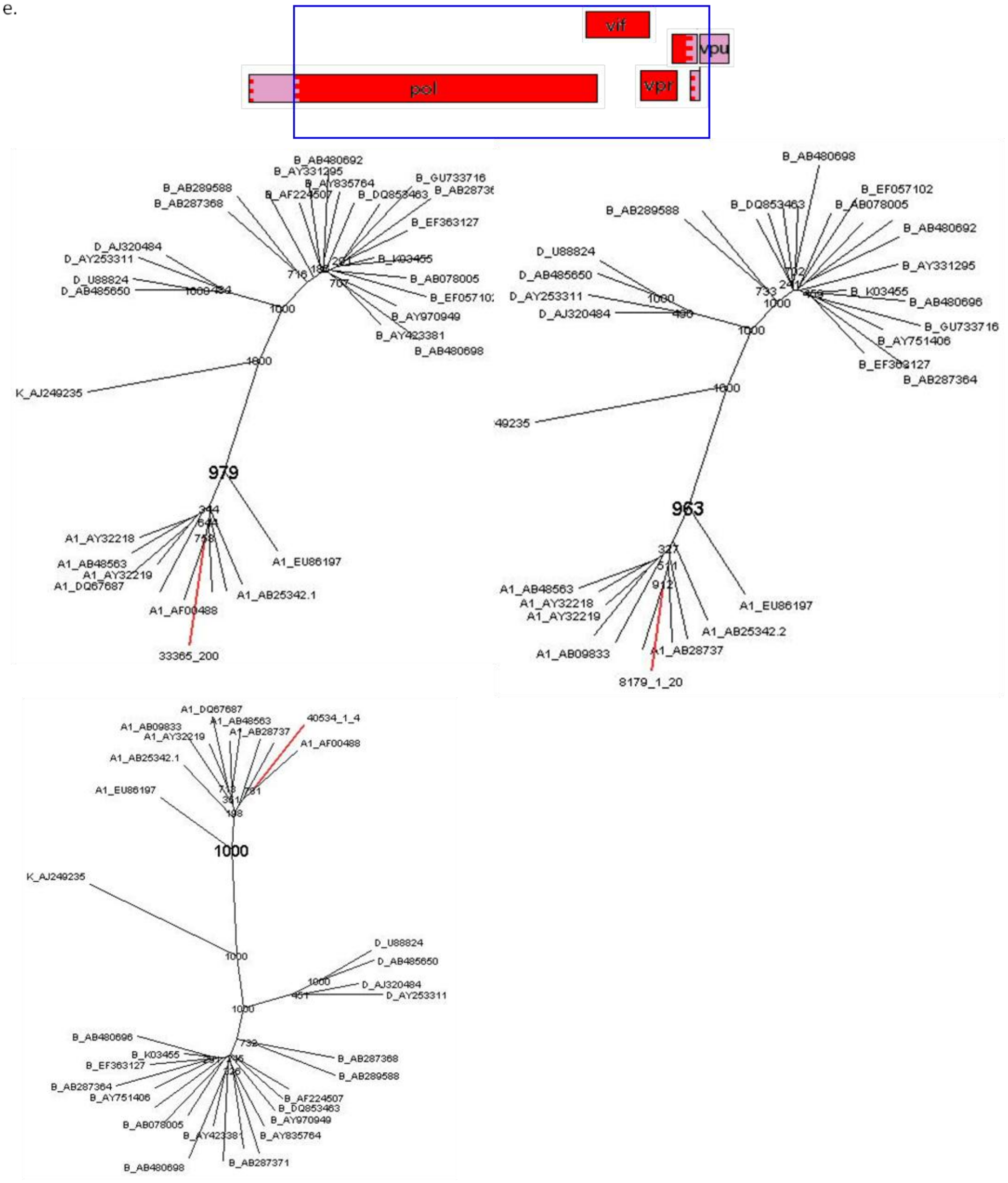
c.


gag
A1_AB48563
A1_EU86197
A1_DQ67687
A1_AB25342.2
A1_AB09833
A1_AY32218
A1_AY32219
A1_AB28737
A1_AB25342.1
A1_AF00488
33365_200
386
D_AB485650
D_AJ320484
D_U88824
D_AY253311
B_AB287368
K_AJ249235
B_AB480696
B_AB289588
B_AB287364
B_AB480692
B_AF224507
B_AB078005
B_EF057102
B_EF363127
B_AY423381
B_AY970949
B_AB480698
B_GU733716
B_AB287371
8179_1_20
384
B_K03455
B_DQ853463
B_AY835764
B_AY331295
B_AY751406
40534_1_4
221

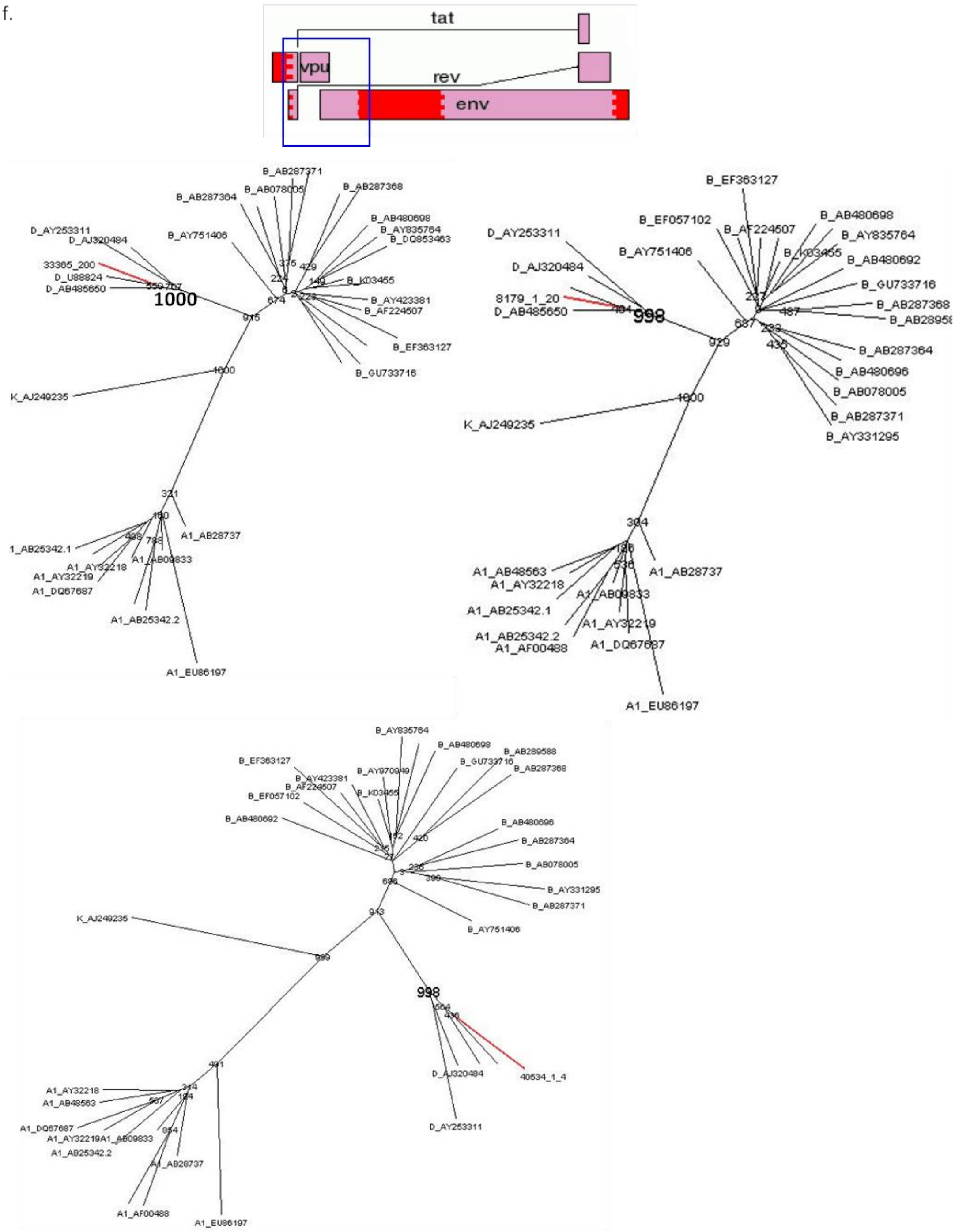
d.


gag
pol
A1_AF00488
A1_AY32219
A1_AB25342.1
A1_AY32218
A1_AB28737
A1_AB48563
A1_EU86197
K_AJ249235
975
907
D_AJ320484
D_AY253311
D_U88824
D_AB485650
133
33365_200
577
B_AY331295
B_AB480698
B_DQ853463
B_AB078005
B_GU733716
B_AY751406
B_AY970949
B_K03455
B_AY835764
B_EF057102
B_AB480696
B_AB480692
B_AB289588
B_AB287364
B_AB287371
B_AB287368
B_EF363127
B_AF224507
B_AB078005
B_DQ853463
B_AY331295
B_AB480698
B_AY751406
B_AY970949
B_AB287371
B_EF057102
240
D_AY253311
D_AB485650
D_U88824
8179_1_20
970
932
A1_EU86197
A1_AY32219
A1_AB25342.2
A1_AB25342.1
A1_DQ67687
A1_AB48563
A1_AB09833
A1_AB28737
B_AB480692
B_AB287371
B_AB287368
B_EF057102
B_EF363127
B_GU733716
B_AB287364
B_AY970949
B_K03455
B_AB480698
B_DQ853463
B_AY751406
B_AY331295
326
40534_1_4
D_U88824
D_AY253311
D_AB485650
D_AJ320484
964
K_AJ249235
923
A1_EU86197
A1_AB48563
A1_AB25342.2
A1_AB28737
A1_AY32219
A1_AY32218
A1_AF00488
A1_AB25342.1
A1_DQ67687

e.

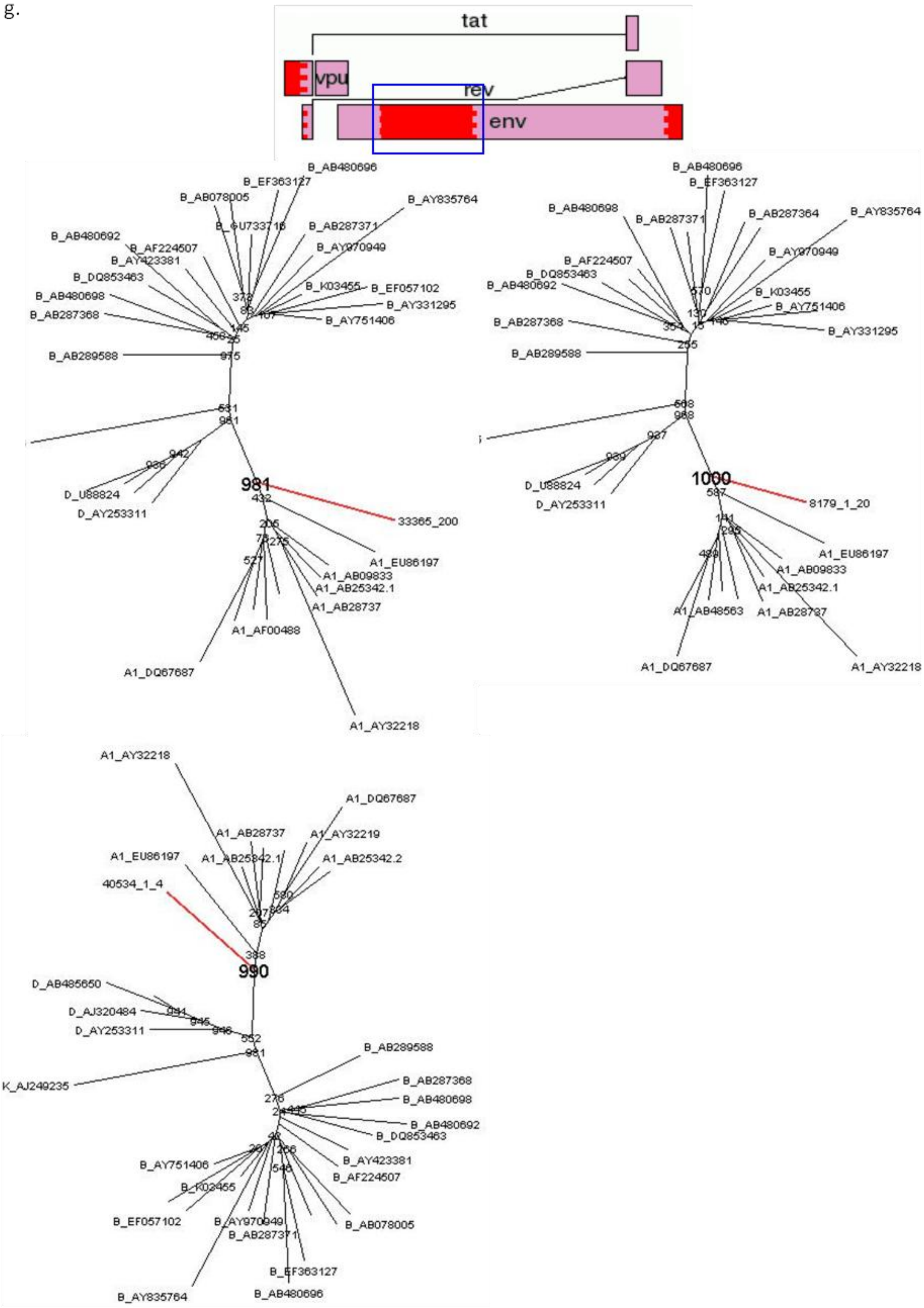

vif
vpu
pol
vpr
B_AB480692
B_AY331295
B_AY835764
B_GU733716
B_AB28736
B_AB289588
B_AB287368
B_AF224507
B_DQ853463
B_EF363127
D_AJ320484
D_AY253311
B_K03455
716
182
201
707
D_U88824
D_AB485650
1000
434
B_AB078005
B_EF05710
B_AY970949
B_AY423381
B_AB480698
K_AJ249235
979
344
644
758
A1_AY32218
A1_AB48563
A1_AY32219
A1_DQ67687
A1_AF00488
A1_EU86197
A1_AB25342.1
33365_200
B_AB480698
B_EF057102
B_DQ853463
B_AB078005
B_AB289588
B_AB480692
D_U88824
D_AB485650
D_AY253311
D_AJ320484
1000
490
733
702
241
459
B_AY331295
B_K03455
B_AB480696
B_GU733716
B_AY751406
B_EF363127
B_AB287364
49235
963
327
511
912
A1_AB48563
A1_AY32218
A1_AY32219
A1_AB09833
A1_AB28737
A1_AB25342.2
A1_EU86197
8179_1_20
A1_DQ67687
A1_AB48563
A1_AB09833
A1_AY32219
A1_AB28737
40534_1_4
A1_AF00488
A1_AB25342.1
A1_EU86197
713
351
781
198
1000
K_AJ249235
D_U88824
D_AB485650
D_AJ320484
D_AY253311
451
732
B_AB480696
B_K03455
B_EF363127
B_AB287364
B_AY751406
B_AB078005
B_AY423381
B_AB480698
B_AB287371
B_AY835764
B_AY970949
B_DQ853463
B_AF224507
B_AB289588
B_AB287368
245
226

f.



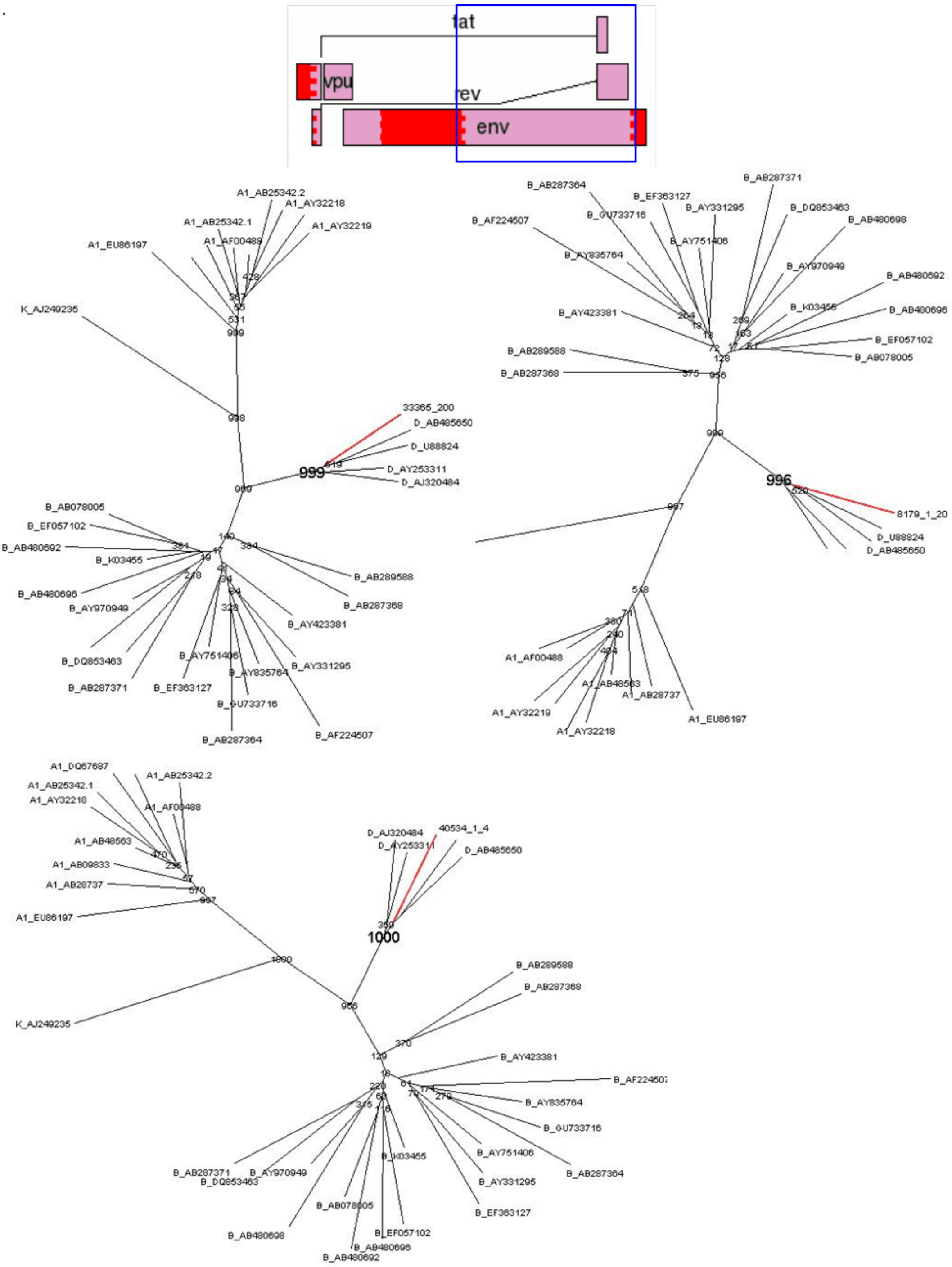
tat
vpu
rev
env
B_AB287371
B_AB078005
B_AB287368
B_AB287364
B_AB480698
B_AY835764
B_DQ853463
D_AY253311
D_AJ320484
B_AY751406
33365_200
D_U88824
D_AB485650
1000
B_K03455
B_AY423381
B_AF224507
B_EF363127
B_GU733716
K_AJ249235
A1_AB28737
1_AB25342.1
A1_AY32218
A1_AB09833
A1_AY32219
A1_DQ67687
A1_AB25342.2
A1_EU86197
B_EF363127
B_EF057102
B_AF224507
B_AB480698
B_AY835764
D_AY253311
B_AY751406
B_K03455
B_AB480692
D_AJ320484
B_GU733716
8179_1_20
B_AB287368
D_AB485650
998
B_AB28958
B_AB287364
B_AB480696
B_AB078005
K_AJ249235
B_AB287371
B_AY331295
A1_AB48563
A1_AY32218
A1_AB28737
A1_AB09833
A1_AB25342.1
A1_AY32219
A1_AB25342.2
A1_AF00488
A1_DQ67687
A1_EU86197
B_AY835764
B_AB480698
B_AB289588
B_EF363127
B_AY423381
B_AY970940
B_GU733716
B_AF224507
B_AB287368
B_EF057102
B_K03455
B_AB480692
B_AB480696
B_AB287364
B_AB078005
B_AY331295
B_AB287371
K_AJ249235
B_AY751406
998
D_AJ320484
40534_1_4
A1_AY32218
A1_AB48563
A1_DQ67687
A1_AY32219
A1_AB09833
A1_AB25342.2
A1_AB28737
D_AY253311
A1_AF00488
A1_EU86197

g.



tat
vpu
rev
env
B_AB480696
B_EF363127
B_AB078005
B_AY835764
B_GU733716
B_AB287371
B_AB480692
B_AF224507
B_AY423381
B_AY970949
B_DQ853463
B_K03455
B_EF057102
B_AB480698
B_AY331295
B_AB287368
B_AY751406
B_AB289588
D_U88824
D_AY253311
981
33365_200
A1_EU86197
A1_AB09833
A1_AB25342.1
A1_AB28737
A1_AF00488
A1_DQ67687
A1_AY32218
B_AB287364
1000
8179_1_20
A1_AB48563
A1_AY32219
A1_AB25342.2
40534_1_4
990
D_AB485650
D_AJ320484
K_AJ249235

h.


tat
vpu
rev
env
A1_AB25342.2
A1_AY32218
A1_AB25342.1
A1_AY32219
A1_AF00488
A1_EU86197
K_AJ249235
33365_200
D_AB485650
D_U88824
D_AY253311
D_AJ320484
999
B_AB078005
B_EF057102
B_AB480692
B_K03455
B_AB480696
B_AY970949
B_DQ853463
B_AB287371
B_EF363127
B_AY751406
B_AY835764
B_GU733716
B_AB287364
B_AF224507
B_AY331295
B_AY423381
B_AB287368
B_AB289588
B_AB480698
8179_1_20
996
A1_AB48563
A1_AB28737
1000
40534_1_4
A1_DQ67687
A1_AB09833

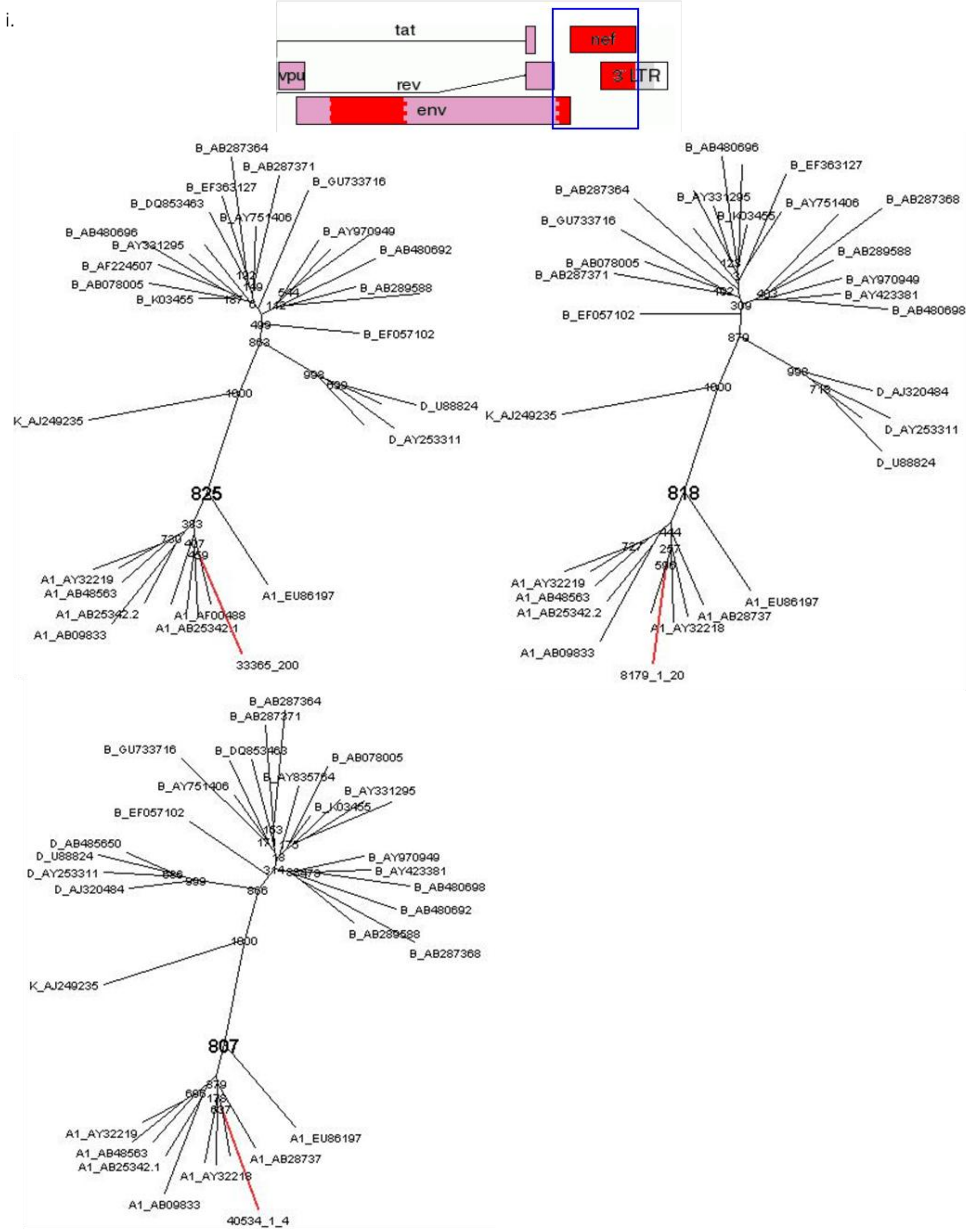


Figure 5_5. Maximum likelihood analysis of the A1/D recombinant (specimens 33365, 8179 and 40534).

Maximum likelihood trees of putative non-recombinant fragments from specimens 33365, 8179 and 49534 drawn using PhyML with PAUP-defined parameters. Subtypes used for analysis were A1, D, B and K (outgroup). Blue boxes indicate the genomic region analysed. Numbers indicate bootstrapping support from 1000 replicates (excepting slice 5; 100 replicates). 70% support was the cut-off for acceptable clustering. For each fragment, trees from specimen 33365 are shown in the upper left, specimen 8179 the upper right and 40534 lowermost. a) Fragment 1. This fragment clusters with subtype A1, with low to moderate bootstrap support across all three specimens (48%, 62% and 51%, respectively); b) Fragment 2. This fragment clusters strongly with subtype D across all three specimens

(78%, 90% and 73% support, respectively); c) Fragment 3. This fragment clusters with subtype A1 in specimens 33365 and 8179, although bootstrap support is low (39%, and 38%, respectively). However, the length of this fragment is short, and likelihood mapping for all three specimens showed a high percentage of unresolved quartets, which suggested a lack of phylogenetic signal for analysis. Specimen 40534 clusters with the K outgroup in this fragment. However, the bootstrap support is very low at 22%; d) Fragment 4. This highly conserved region of *pol* shows a low bootstrap support (58%, 57% and 33%, respectively), but clear clustering with subtype D; e) Fragment 5. This region, which encompasses the majority of the *pol* gene, shows strong support for subtype A1 across all three specimens (98%, 96%, 100%, respectively); f) Fragment 6. Strong support for subtype D is seen across the accessory genes and initial region of gp120 (100%, 77%, 100%, respectively); g) Fragment 7. This region, which codes for the V1-V3 regions, clusters strongly with subtype A1 (98%, 100%, 99%, respectively); h) Fragment 8. This subtype D region codes for the V4 and V5 loops and gp41 (100%, 97%, and 100%, respectively); i) Fragment 9. The end of gp41 and *nef* cluster strongly with subtype A1 (83%, 82% and 81% support, respectively).

Slice	Study number					
	33365		8179		40534	
	Subtype	Bootstrap support (%)	Subtype	Bootstrap support (%)	Subtype	Bootstrap support (%)
1	A1	48	A1	62	A1	51
2	D	78	D	90	D	73
3	A1	39	A1	38	A1	22
4	D	58	D	57	D	33
5	A1	98	A1	96	A1	100
6	D	100	D	77	D	100
7	A1	98	A1	100	A1	99
8	D	100	D	97	D	100
9	A1	83	A1	82	A1	81

Table 5_5. Bootstrap support from PhyML maximum likelihood analyses for each alignment slice of specimens 33365, 8179 and 40534. 1000 bootstrapping replicates were performed for each slice*. 70% support was the cut-off for subtype classification. Slices 2, 5, 6, 7, 8, and 9 showed sufficient support across all three specimens for subtype classification. Slices 1, 3 and 4 showed lower support values, but still clustered with the subtypes indicated by the preceding analyses. These slices were analysed using a greater range of subtypes before final subtype classification was assigned.

*Excepting slice 5 (100 replicates only)

5.2.3.2 Analysis of specimen 34567

The RIP and jpHMM analyses of specimen 34567 showed a unique, complex recombinant structure. The initial Simplot bootscanning analysis of this specimen showed an A1/B/D structure, with three substantial regions with no clear subtype (Figure 5_6a). Repeated scanning with the full range of pure HIV-1 subtypes did not improve the subtype designations (Figure 5_6b). However, all three analysis methods had suggested that specimen 34567 included some A1/D regions with identical breakpoints as the three London specimens.   Bootscanning analysis of specimen 34567 against the A1/D specimen 33365 showed three clear regions of A1/D identity and two subtype B regions (Figure 5_6c). With this in mind, two downstream analyses of this specimen were performed: informative sites, alignment slicing and subsequent downstream PhyML analyses, and alignment slicing and downstream PhyML analyses according to the breakpoints identified for the A1/D specimens.

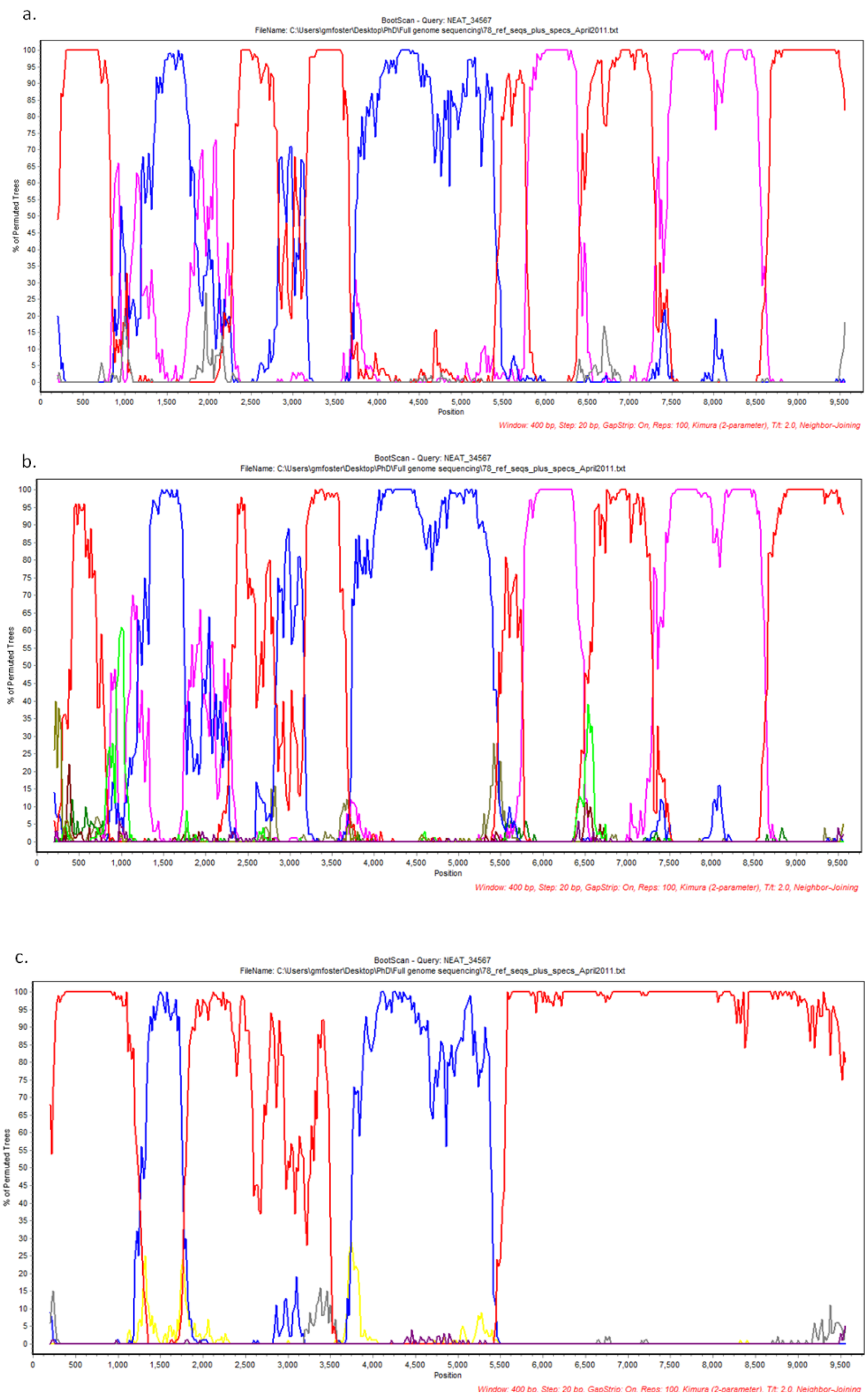


Figure 5_6. Bootscanning analysis of specimen 34567. Bootscanning plots for specimen 34567. The y axis shows the percentage of permuted trees that the query sequence

clustered with the closest subtype match from the reference alignment. The x axis shows the nucleotide position of the sequence (not HXB2 numbering). **a)** Bootscanning plot of specimen 34567 against subtypes A1, B, D and F (outgroup).Subtype A is represented in red, subtype B in blue, subtype D in lavender, and subtype F in grey. Although the majority of the genome shows clear subtype identity, there are three regions (800-1,300, 1,700-2250, 2750-3250) that have no clear subtype classification. **b)** Bootscanning plot of specimen 34567 against all HIV-1 pure subtypes. Subtype A is represented in red, subtype B in Blue, subtype C in light green, subtype D in lavender, subtype F in purple, subtype G in dark green, subtype H in burgundy, subtype K in light brown and subtype J in dark brown. Despite the larger range of reference subtypes, the unclassified regions from A) are still present. **c)** Bootscanning plot of specimen 34567 against specimen 33365, and reference subtype A1, B, D and F specimens. Specimen 33365 is represented in red, subtype A1 in grey, subtype B in blue, subtype D in yellow, and subtype F in purple. Three regions of identity with 33365, and two regions of subtype B, are present, indicating that this specimen is a recombinant of the A1/D recombinant.

Figure 5_7 shows the bootscanning plot with informative sites for specimen 34567. The informative sites analysis showed a structure similar to that shown by the jpHMM and RIP analyses, especially in those regions of the genome judged to be identical to the A1/D recombinant. However, the region between positions 770 and 2284 showed no clear subtype identity, and the x^2 was maximised when there were no breakpoints placed in this region. When extra slices of the alignment were created to explore this region more closely, no breakpoints placed within this region were statistically significant. Those breakpoints that were statistically significant are summarised in Table 5_6.

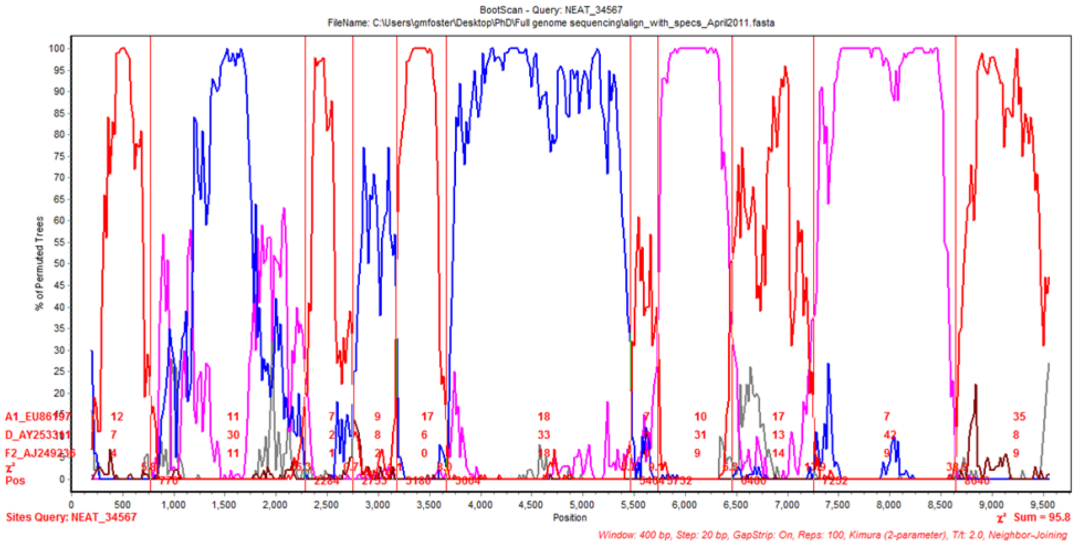


Figure 5_7. Bootscanning and informative sites analysis of specimen 34567. Bootscanning and informative sites analysis of specimen 34567 was performed using subtypes A1, A2, B, D and F2. Subtypes A1, A2, B and D were indicated from the jpHMM analysis of the specimen, and subtype F2 was included as an outgroup. The x axis shows the nucleotide position of the sequence (not HXB2 numbering. The y axis shows the percentage of permuted trees that the query specimen clustered with the subtype reference. Subtype A1 is shown in red, subtype A2 in grey, subtype B in blue, subtype D in lavender and subtype F2 in brown. Breakpoints are indicated by vertical red lines. The breakpoints shown on this plot were those that were subsequently shown to be statistically significant.

Breakpoint	jpHMM breakpoint	Sliding window breakpoint	Gene	Region	*p*
1	1162 (±8)	1196	*gag*	p24	<0.001
2	1479 (±7)	1688	Extra slices created to explore region of uncertainty more closely – not supported by statistical significance		
3	1625 (±27)	2070			
4		2257			
5	2535 (±16)	2622	*pol*	RT	0.0143
6	3296 (±23)	3095	*pol*	RT	1.0*
7	3488 (±30)	3471	*pol*	RT	0.0670
8	3866 (±31)	3875	*pol*	RNase	<0.001
9	5785 (±25)	5746	*vpr*	-	<0.001
10	5983 (±27)	6014	*tat/rev*	-	0.0017
11	6551 (±8)	6664	*env*	gp120	0.0011
12	7359 (±128)	7314	*env*	gp120	0.0011
13	8674 (±21)	8655	*env*	gp41	<0.0001

Table 5_6 jpHMM and sliding window breakpoints for specimen 34567. Breakpoints for specimen 34567 identified by jpHMM and sliding window analyses. Sliding window nucleotide positions have been converted into HXB2 numbering for ease of comparison. Statistical significance was assessed using informative sites analysis and Fisher's exact test. Breakpoints 2 - 4 (HXB2 1688 - 2257) correspond to the uncertain B/D region in the sliding window analysis.

*This corresponds to the unclassified region in the jpHMM analysis

The results from the informative sites analysis were used to create a tentative schematic of specimen 34567's structure (Figure 5_8). This structure recognised that areas of the *gag* and *pol* genes could not be reliably distinguished between subtypes B and D. The alignment was sliced according to the breakpoints in Figure 5_7, and PhyML analysis was performed to confirm the subtype of each region (Figure 5_9). Some of the trees were poorly supported, especially those shown in Figures 5_9c, 5_9d and 5_9g. This was most likely due at least partly to the short length of the fragments under analysis. Paraphyletic clustering of Specimen 34567 was seen in Figures 5_9b, 5_9c and 5_9d (HXB2 1196-3471), which corresponded to the uncertain B/D region seen in the jpHMM analysis. Despite not resolving the uncertain B/D region sufficiently, the maximum likelihood analysis did confirm the tentative structure seen in Figure 5_8, pending further analysis of the uncertain B/D region.

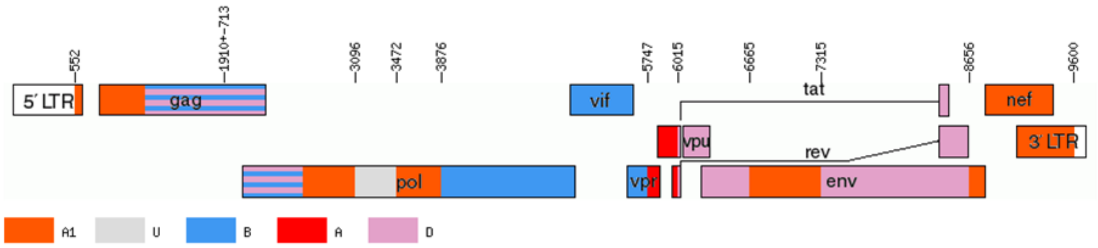
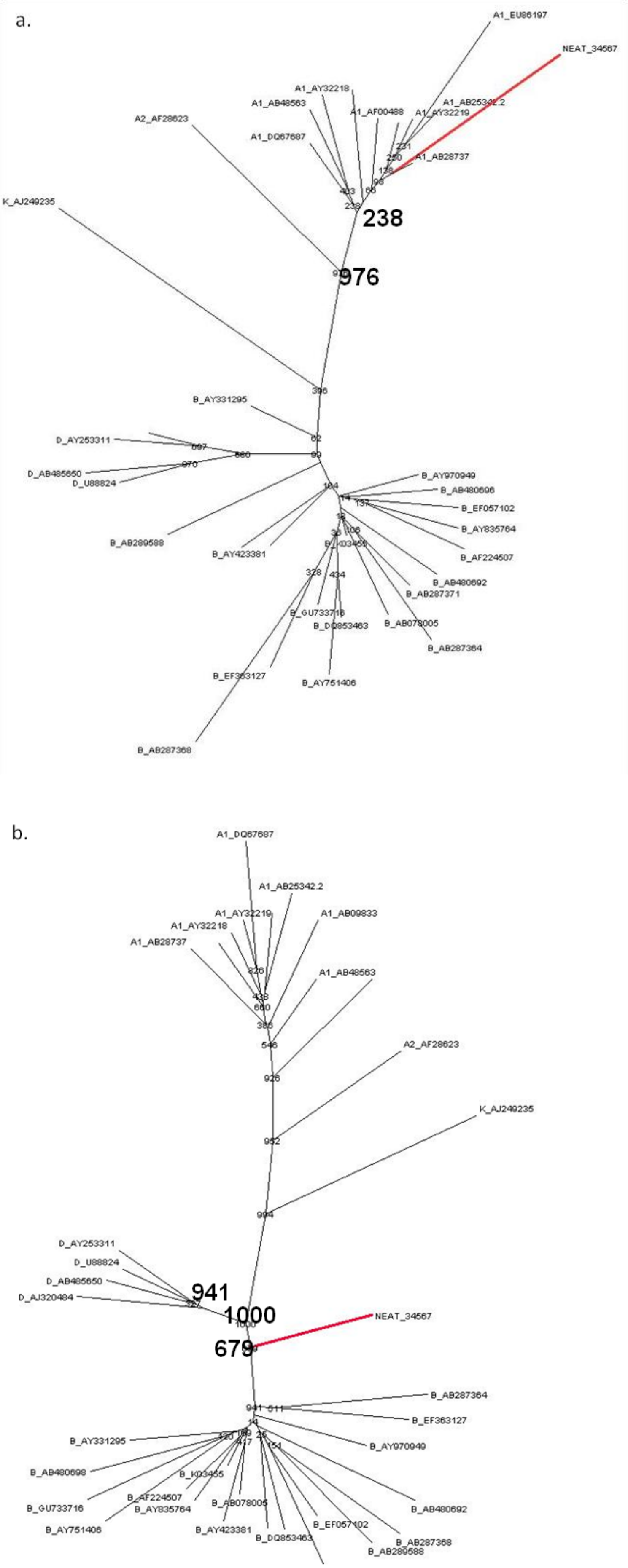
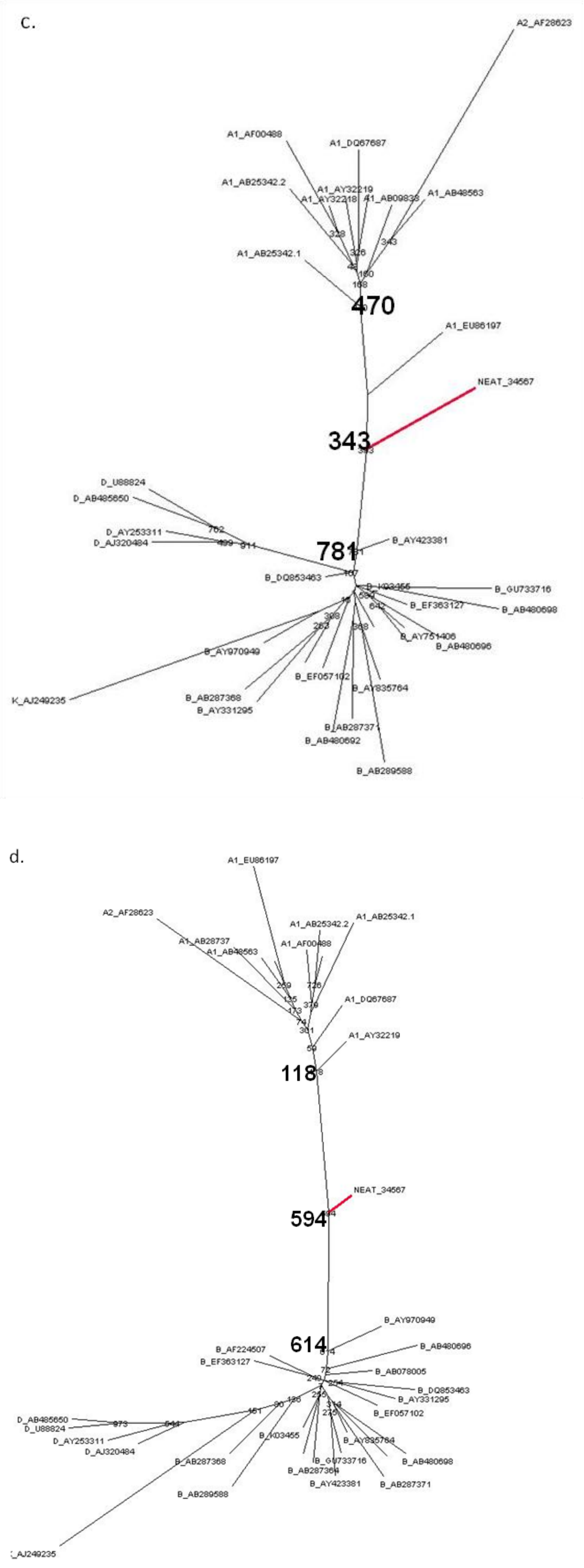
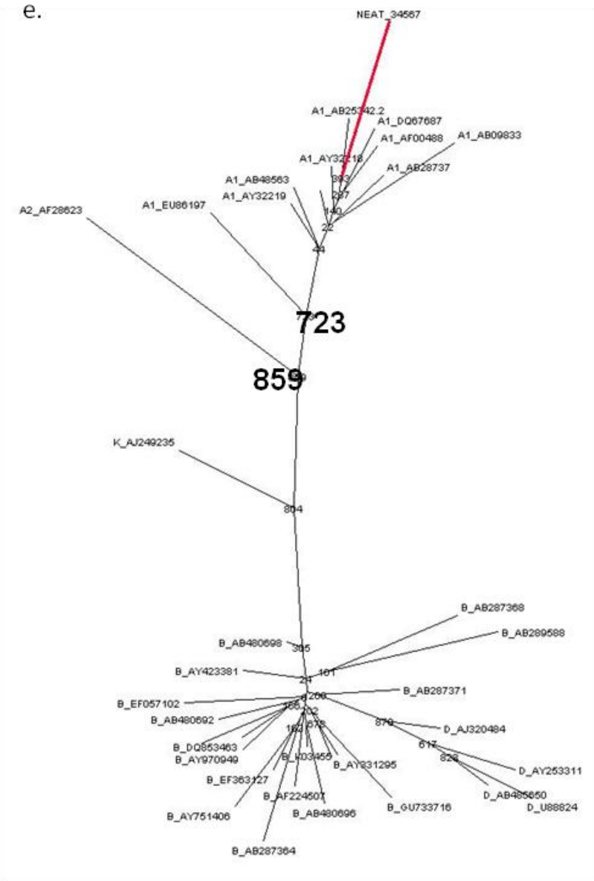
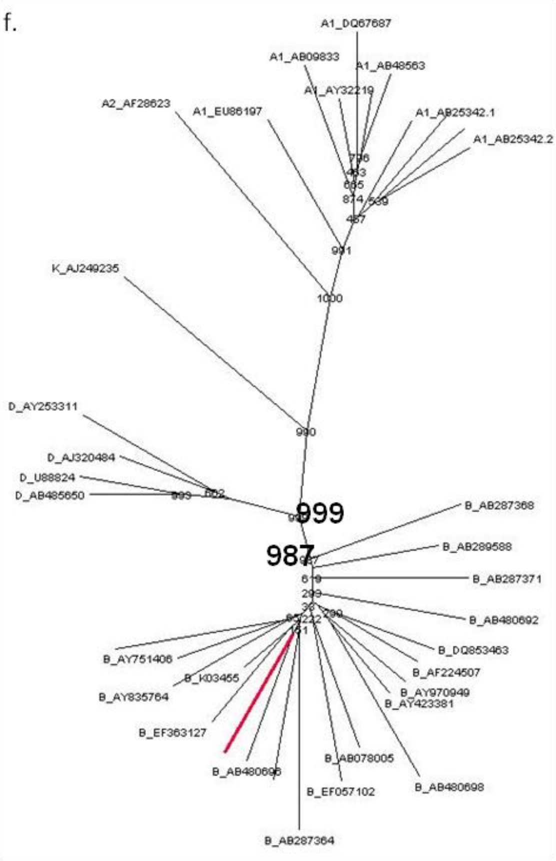


Figure 5_8. RDT schematic of specimen 34567. The results from the informative sites analysis were used to create a tentative structure for specimen 34567. This structure shows clear similarities to the A1/D recombinant specimens in the latter half of the genome, but recognised that the *gag* and *pol* genes cannot be clearly distinguished between subtypes B and D.

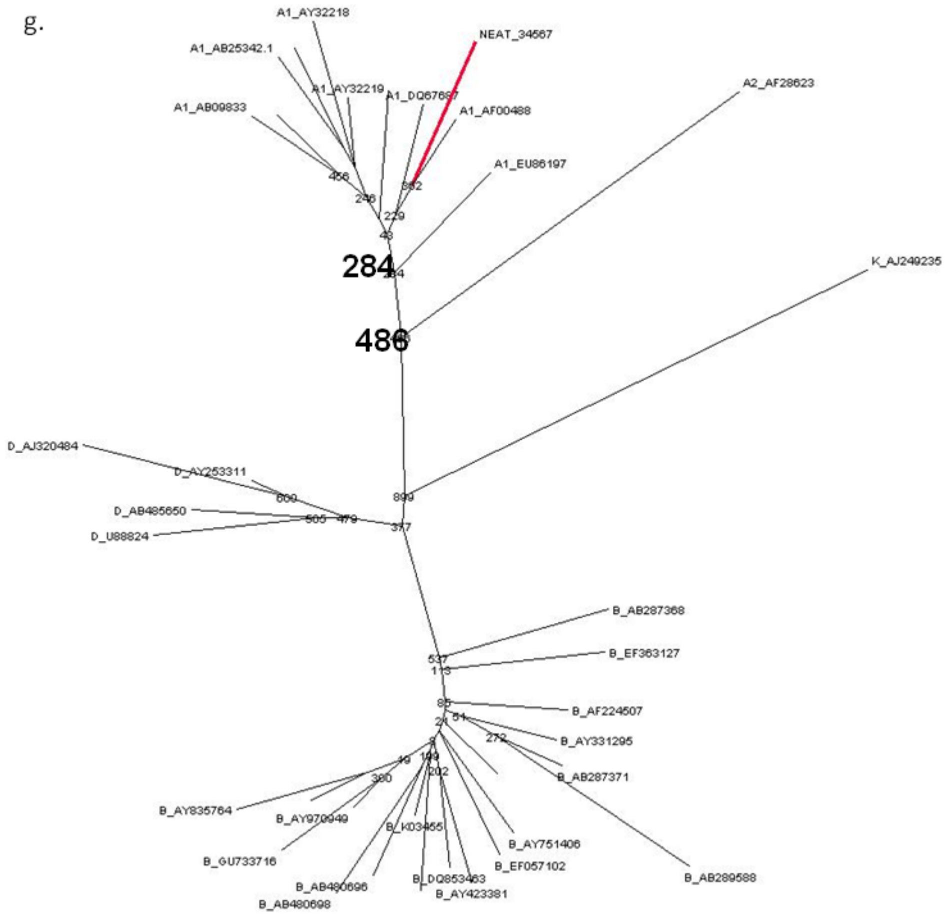

a.
A1_EU86197
NEAT_34567
A1_AY32218
A1_AB48563
A1_AF00488
A1_AB25342.2
A1_AY32219
A2_AF28623
A1_DQ67687
231
250
A1_AB28737
138
98
403
66
238
K_AJ249235
238
976
396
B_AY331295
62
D_AY253311
697
860
99
D_AB485650
970
D_U88824
B_AY970949
104
B_AB480696
14
137
B_EF057102
18
B_AY835764
106
B_AB289588
33
B_AY423381
B_K03455
B_AF224507
328
434
B_AB480692
B_AB287371
B_GU733716
B_DQ853463
B_AB078005
B_AB287364
B_EF363127
B_AY751406
B_AB287368
b.
A1_DQ67687
A1_AB25342.2
A1_AY32219
A1_AB09833
A1_AY32218
A1_AB28737
826
A1_AB48563
438
660
385
546
A2_AF28623
926
K_AJ249235
952
994
D_AY253311
D_U88824
D_AB485650
941
D_AJ320484
327
1000
1000
NEAT_34567
679
B_AB287364
941
511
B_EF363127
14
B_AY331295
420
189
25
B_AY970949
417
151
B_AB480698
B_K03455
B_AF224507
B_GU733716
B_AY835764
B_AB078005
B_AB480692
B_EF057102
B_AY751406
B_AY423381
B_DQ853463
B_AB287368
B_AB289588


c.
A2_AF28623
A1_AF00488
A1_DQ67687
A1_AB25342.2
A1_AY32219
A1_AY32218
A1_AB09833
A1_AB48563
328
343
326
A1_AB25342.1
42
160
168
470
A1_EU86197
NEAT_34567
343
D_U88824
D_AB485650
762
D_AY253311
D_AJ320484
499
911
781
B_AY423381
107
B_DQ853463
B_K03455
B_GU733716
B_EF363127
B_AB480698
642
398
263
368
B_AY751406
B_AB480696
B_AY970949
B_EF057102
B_AY835764
K_AJ249235
B_AB287368
B_AY331295
B_AB287371
B_AB480692
B_AB289588
d.
A1_EU86197
A2_AF28623
A1_AB25342.1
A1_AB25342.2
A1_AB28737
A1_AB48563
A1_AF00488
269
726
125
379
173
A1_DQ67687
74
301
59
A1_AY32219
118
NEAT_34567
594
B_AY970949
614
B_AB480696
B_AF224507
B_EF363127
72
B_AB078005
240
254
B_DQ853463
B_AY331295
255
136
314
B_EF057102
D_AB485650
973
544
451
90
275
D_U88824
D_AY253311
D_AJ320484
B_K03455
B_AY835764
B_AB287368
B_GU733716
B_AB480698
B_AB287364
B_AB289588
B_AY423381
B_AB287371
_AJ249235

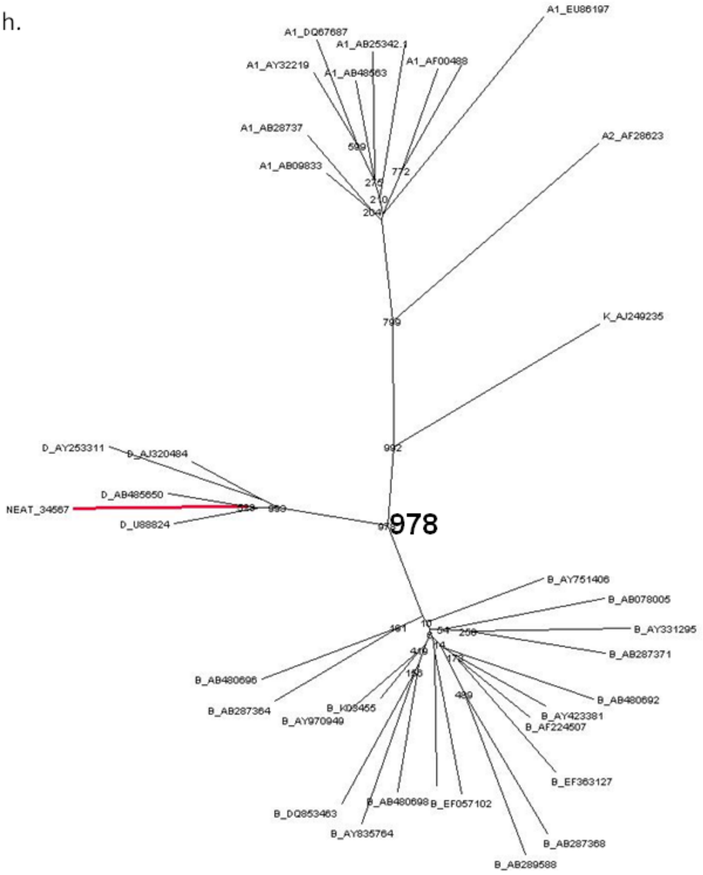

e.
NEAT_34567
A1_AB25342.2
A1_DQ67687
A1_AF00488
A1_AB09833
A1_AY32218
A1_AB28737
A1_AB48563
A1_AY32219
A1_EU86197
A2_AF28623
393
297
140
22
44
723
859
K_AJ249235
804
B_AB287368
B_AB289588
B_AB480698
305
B_AY423381
161
24
B_AB287371
B_EF057102
200
155
B_AB480692
202
162
678
879
D_AJ320484
B_DQ853463
B_AY970949
617
828
B_K03455
B_AY331295
D_AY253311
B_EF363127
D_AB485650
B_AF224507
D_U88824
B_AY751406
B_AB480696
B_GU733716
B_AB287364


f.
A1_DQ67687
A1_AB09833
A1_AB48563
A1_AY32219
A2_AF28623
A1_EU86197
A1_AB25342.1
A1_AB25342.2
776
463
685
874
539
467
991
K_AJ249235
1000
D_AY253311
990
D_AJ320484
D_U88824
602
D_AB485650
999
B_AB287368
999
987
B_AB289588
619
B_AB287371
209
36
290
B_AB480692
222
151
B_AY751406
B_DQ853463
B_K03455
B_AF224507
B_AY970949
B_AY835764
B_AY423381
B_EF363127
B_AB480696
B_AB078005
B_EF057102
B_AB480698
B_AB287364

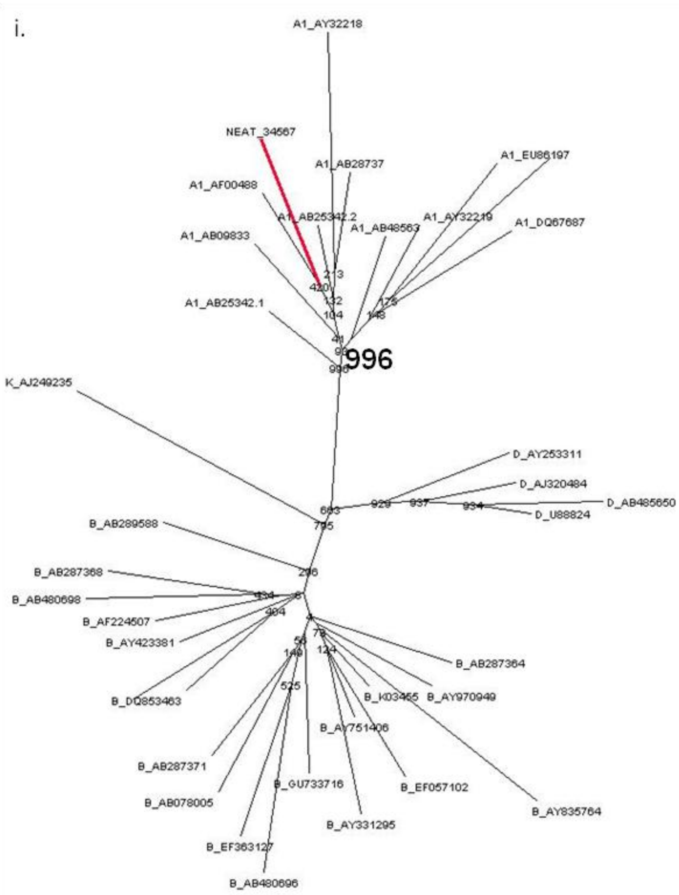
g.


A1_AY32218
NEAT_34567
A1_AB25342.1
A1_AY32219
A1_DQ67687
A2_AF28623
A1_AB09833
A1_AF00488
A1_EU86197
456
246
352
229
43
284
284
K_AJ249235
486
D_AJ320484
D_AY253311
600
D_AB485650
505
479
899
D_U88824
377
B_AB287368
537
113
B_EF363127
85
B_AF224507
21
51
272
B_AY331295
9
49
199
202
B_AB287371
380
B_AY835764
B_AY970949
B_K03455
B_AY751406
B_GU733716
B_EF057102
B_AB480696
B_DQ853463
B_AB289588
B_AY423381
B_AB480698

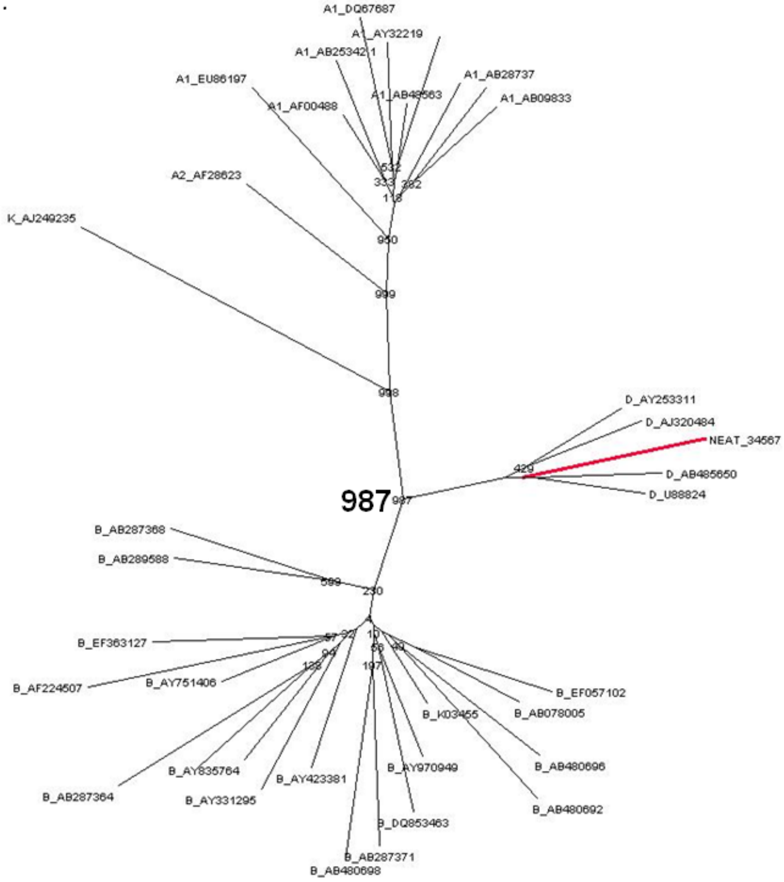
h.


A1_EU86197
A1_DQ67687
A1_AB25342.1
A1_AF00488
A1_AY32219
A1_AB48563
A2_AF28623
A1_AB28737
599
A1_AB09833
772
276
210
204
799
K_AJ249235
992
D_AY253311
D_AJ320484
D_AB485650
NEAT_34567
253
999
D_U88824
978
978
B_AY751406
B_AB078005
161
10
54
250
B_AY331295
416
14
B_AB287371
173
B_AB480696
153
B_AB480692
B_AB287364
B_K03455
489
B_AY423381
B_AY970949
B_AF224507
B_EF363127
B_DQ853463
B_AB480698
B_EF057102
B_AB287368
B_AY835764
B_AB289588

i.


A1_AY32218
NEAT_34567
A1_AB28737
A1_EU86197
A1_AF00488
A1_AB25342.2
A1_AY32219
A1_DQ67687
A1_AB48563
A1_AB09833
213
420
132
175
104
148
A1_AB25342.1
41
93
996
996
K_AJ249235
D_AY253311
D_AJ320484
929
937
934
D_AB485650
D_U88824
863
795
B_AB289588
296
B_AB287368
B_AB480698
334
81
404
B_AF224507
B_AY423381
4
79
58
140
124
B_AB287364
B_K03455
B_AY970949
525
B_DQ853463
B_AY751406
B_AB287371
B_GU733716
B_EF057102
B_AB078005
B_AY331295
B_AY835764
B_EF363127
B_AB480696

j.


A1_DQ67687
A1_AY32219
A1_AB25342.1
A1_AB28737
A1_EU86197
A1_AB48563
A1_AB09833
A1_AF00488
532
333
392
118
A2_AF28623
950
K_AJ249235
999
998
D_AY253311
D_AJ320484
NEAT_34567
429
D_AB485650
D_U88824
987
987
B_AB287368
B_AB289588
589
230
B_EF363127
57
92
10
56
49
94
128
197
B_AF224507
B_AY751406
B_EF057102
B_K03455
B_AB078005
B_AY835764
B_AY423381
B_AY970949
B_AB480696
B_AB287364
B_AY331295
B_AB480692
B_DQ853463
B_AB287371
B_AB480698

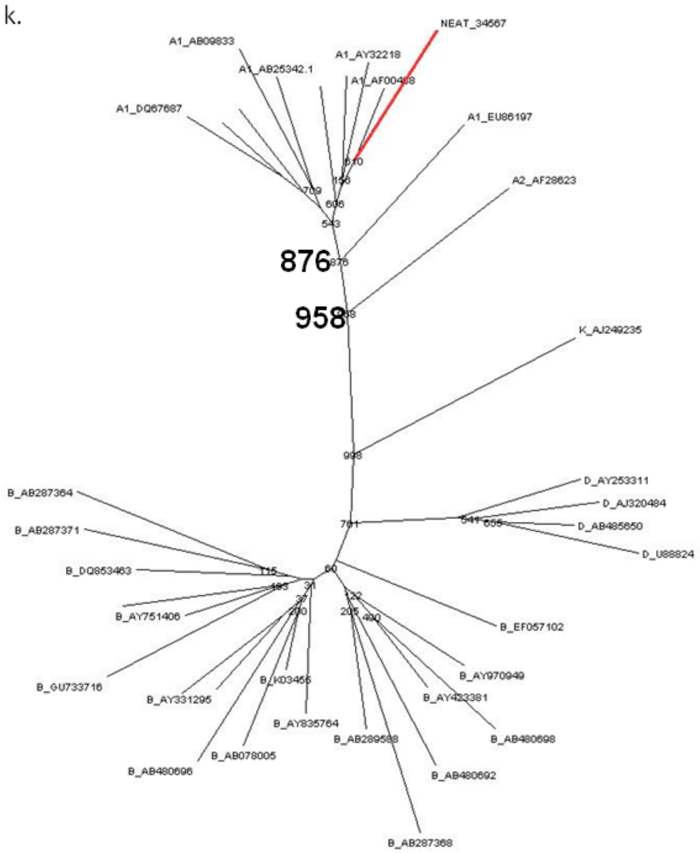
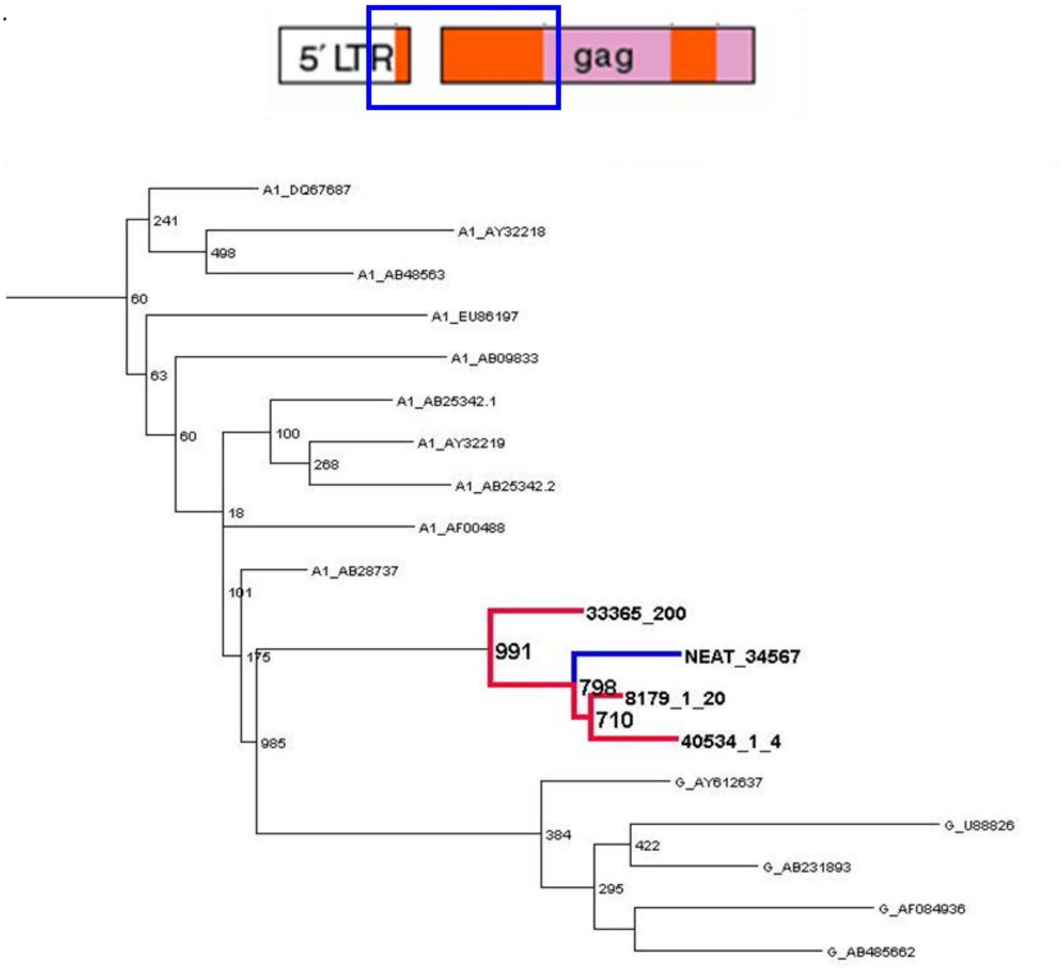


Figure 5_9. Maximum likelihood analysis of specimen 34567. Maximum likelihood analysis of slices of specimen 34567 was performed to classify pure subtype regions using the breakpoints identified in the information sites analysis (Figure 5_7). **a)** Slice 1. This regions showed strong support for subtype A (>90%), and weak support for sub-sutype A1 (23.8%); **b)** Slice 2. This region corresponds to the uncertain B/D region from the bootscanning analysis (HXB2 1688 - 2257). Specimen 34567 clusters paraphyletically with subtypes B and D, regardless of the number of sub-regions into which it is split. It was therefore classified as B/D uncertain; **c)** Slice 3 was eventually classified as subtype A1, despite the low bootstrapping support. The region showed statistically significant breakpoints differentiating it from the adjacent uncertain regions **d)** Slice 4 did not show significant clustering with any subtype, and was classified as unknown; **e)** Slice 5 showed 72% bootstrapping support for subtype A1, which was sufficient for subtype classification; **f)** Slice 6, which corresponds to the end region of *pol, vpu* and the beginning region of *vpr*, clustered with 99% support for subtype B; **g)** Slice 7 shows strong support for subtype A1; **h)-k)** Slices 8-11 show clustering and bootstrap support consistent with the corresponding regions seen in the five A1/D recombinant specimens.

In order to confirm the similarities between specimen 34567 and the five A1/D recombinant specimens that were seen in the bootstrap analysis (Figure 5_6c), slices of the alignment containing specimens 33365, 8179, 40534 and 34567 were made according to the A1/D recombinant breakpoints and submitted for PhyML analysis. The trees from this analysis showed that the complex 34567 clustered with the A1/D recombinant in 7/9 genomic regions (Figure 5_10). In the remaining 2/9 regions (slices 3 and 5), specimen 34567 clustered with subtype B (Figures 5_10c and e, respectively), although both trees showed low bootstrap support of approximately 44%. In the case of slice 3, the entire tree was poorly resolved, which was at least in part due to the short fragment length and low phylogenetic signal of the slice, and was similar to the trees of this slice seen in Figure 5_5c. In the case of slice 5, however, the poor bootstrap support seen for the clustering with subtype B was most likely due to the presence of a subtype-unresolved region within specimen 34567 in this slice (HXB2 3097 - 3473). This region was seen consistently throughout the analysis of this specimen and was present in the jpHMM analysis (Figure 5_2d) as a region of uncertain subtype A1/B and the sliding window analysis (Figure 5_8) as a subtype-unresolved region.

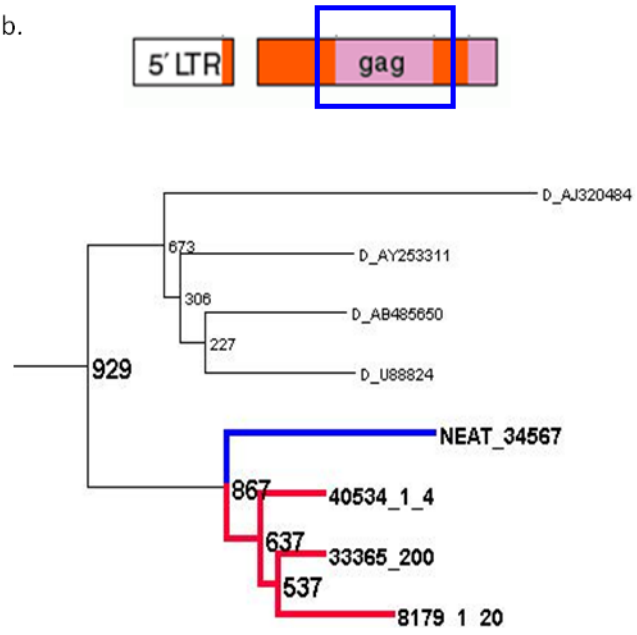
The results from the maximum likelihood analysis were sufficiently robust to confirm the structure for specimen 34567 that was suggested by the bootscan analysis (Figure 5_6c). This largely resolved the subtype B/D uncertain regions seen in earlier stages of the analysis and resulted in the structure seen in Figure 5_11, which shows the structure in relation to its similarity with the A1/D recombinant strain. This confirmed that specimen 34567 was either a recombinant arising from the strain shared by the five A1/D specimens, or that this structure was a precursor of the A1/D recombinant strain.

a.



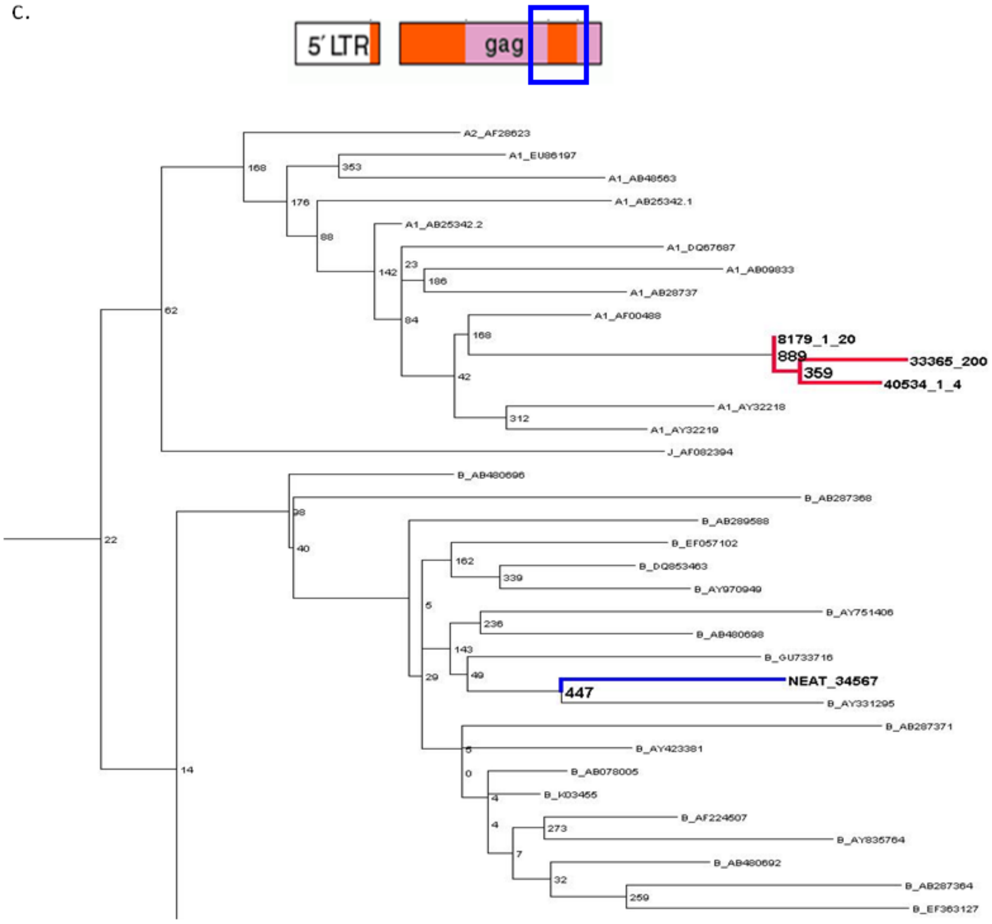
5′ LTR
gag
A1_DQ67687
241
A1_AY32218
498
A1_AB48563
60
A1_EU86197
A1_AB09833
63
A1_AB25342.1
60
100
A1_AY32219
268
A1_AB25342.2
18
A1_AF00488
A1_AB28737
101
33365_200
991
NEAT_34567
175
798
8179_1_20
710
40534_1_4
985
G_AY612637
G_U88826
384
422
G_AB231893
295
G_AF084936
G_AB485662

b.



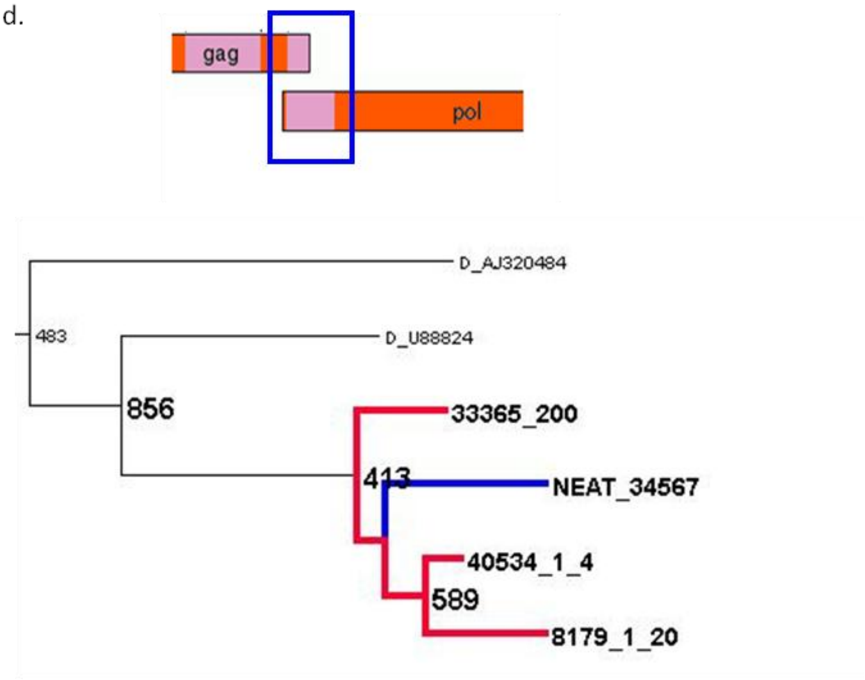
5′ LTR
gag
D_AJ320484
673
D_AY253311
306
D_AB485650
227
929
D_U88824
NEAT_34567
867
40534_1_4
637
33365_200
537
8179 1 20

c.



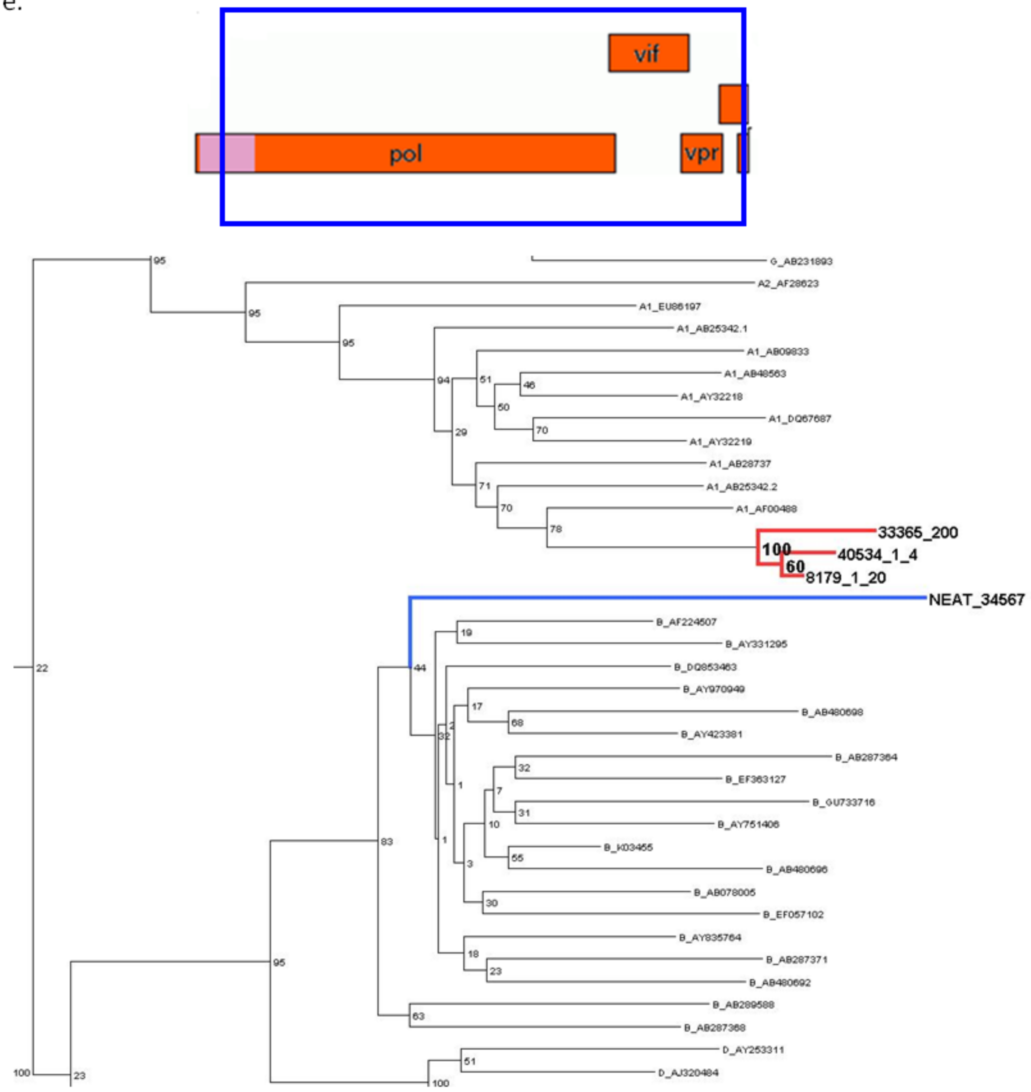
5′ LTR
gag
A2_AF28623
168
A1_EU86197
353
A1_AB48563
176
A1_AB25342.1
A1_AB25342.2
88
A1_DQ67687
142
23
A1_AB09833
186
A1_AB28737
84
A1_AF00488
168
8179_1_20
889
33365_200
359
40534_1_4
42
A1_AY32218
312
A1_AY32219
62
J_AF082394
B_AB480696
B_AB287368
98
B_AB289588
22
40
B_EF057102
162
B_DQ853463
339
B_AY970949
5
B_AY751406
236
B_AB480698
143
B_GU733716
29
49
NEAT_34567
447
B_AY331295
B_AB287371
5
B_AY423381
0
B_AB078005
14
4
B_K03455
B_AF224507
4
273
B_AY835764
7
B_AB480692
32
B_AB287364
259
B_EF363127

d.

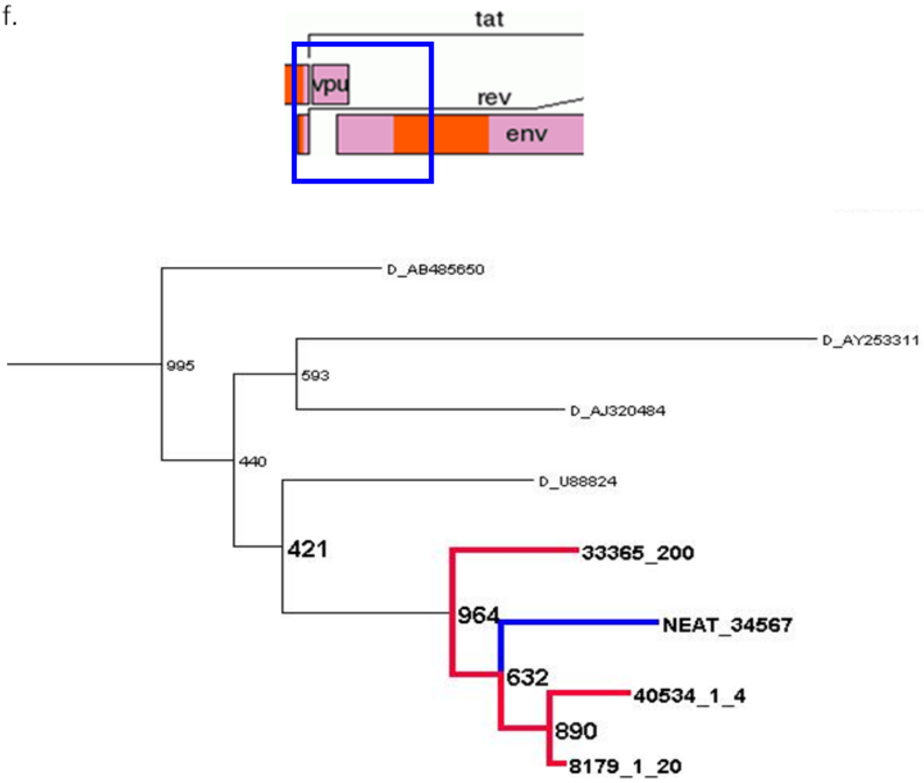


gag
pol
D_AJ320484
483
D_U88824
856
33365_200
413
NEAT_34567
40534_1_4
589
8179_1_20

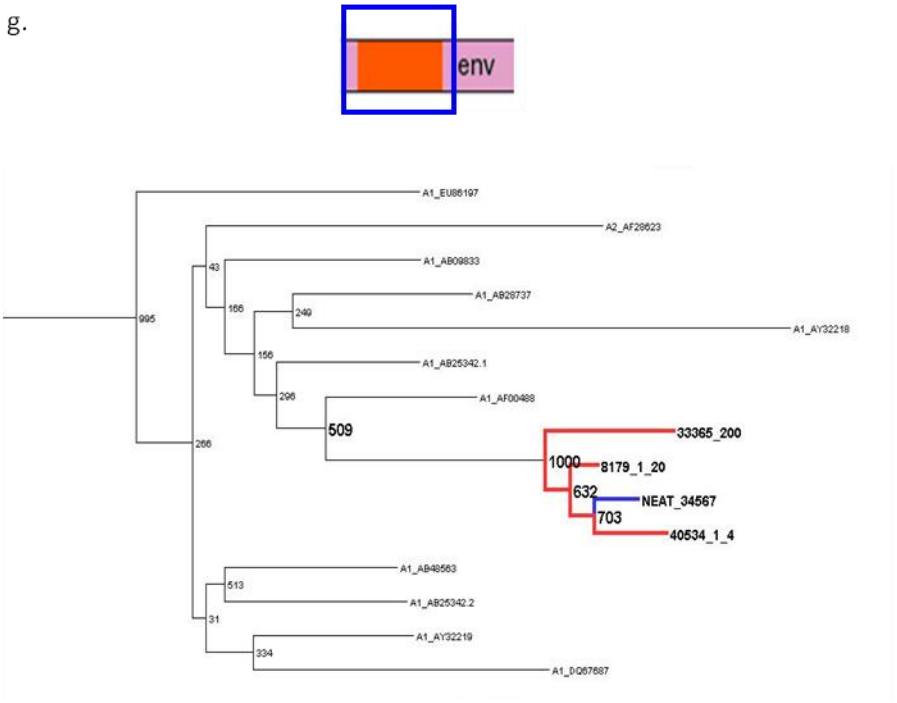
e.


vif
pol
vpr
95
G_AB231893
A2_AF28623
95
A1_EU86197
95
A1_AB25342.1
A1_AB09833
94
51
A1_AB48563
46
A1_AY32218
50
A1_DQ67687
70
A1_AY32219
29
A1_AB28737
71
A1_AB25342.2
70
A1_AF00488
78
33365_200
100
40534_1_4
60
8179_1_20
NEAT_34567
B_AF224507
19
B_AY331295
22
44
B_DQ853463
B_AY970949
17
B_AB480698
2
68
32
B_AY423381
B_AB287364
32
B_EF363127
1
7
B_GU733716
31
B_AY751406
10
83
1
B_K03455
55
B_AB480696
3
B_AB078005
30
B_EF057102
B_AY835764
18
B_AB287371
23
B_AB480692
95
B_AB289588
63
B_AB287368
D_AY253311
51
D_AJ320484
100
23
100

f.


tat
vpu
rev
env
D_AB485650
995
D_AY253311
593
D_AJ320484
440
D_U88824
421
33365_200
964
NEAT_34567
632
40534_1_4
890
8179_1_20

g.


env
A1_EU86197
A2_AF28623
43
A1_AB09833
166
A1_AB28737
249
995
A1_AY32218
156
A1_AB25342.1
296
A1_AF00488
509
33365_200
266
1000
8179_1_20
632
NEAT_34567
703
40534_1_4
A1_AB48563
513
A1_AB25342.2
31
A1_AY32219
334
A1_DQ67687

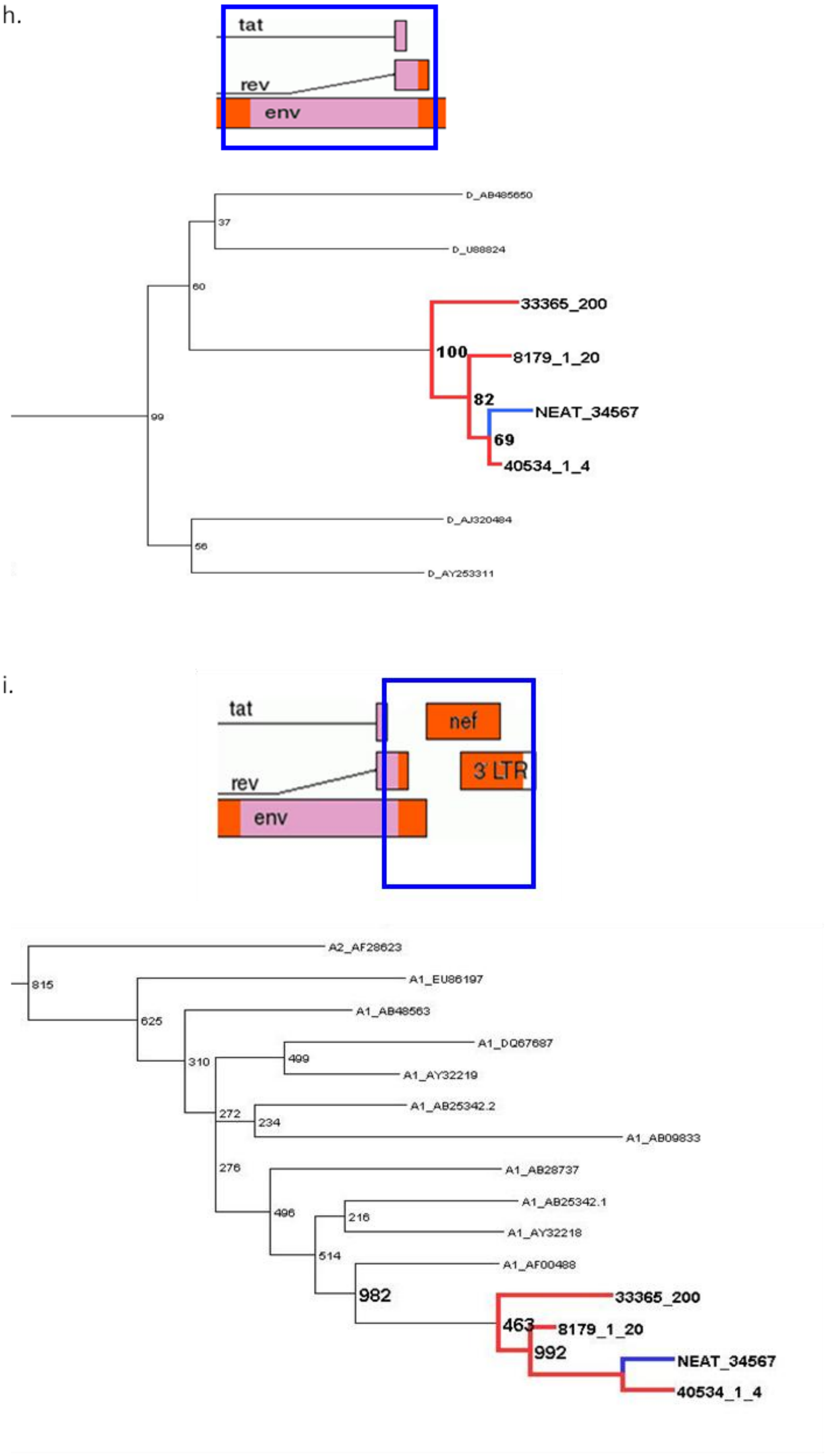


Figure 5_10. Maximum likelihood analysis of specimen 34567 with A1/D recombinant specimens 33365, 8179 and 40534. Following bootscanning analysis of the complex specimen 34567 against the A1/D recombinant specimen, specimen 34567 was cut into alignment slices according to the A1/D recombinant breakpoints, and maximum likelihood analysis performed to see if it clustered with the A1/D recombinant across the HIV-1

genome. 1000 bootstraps were performed, excepting slices 5 and 8, in which 100 bootstraps were performed to ease computational requirements. The A1/D specimens are highlighted in red, and 34567 is highlighted in blue. Partial trees are shown for easier visualisation. Relevant bootstrap values have been enlarged. Specimen 34567 clustered with the A1/D recombinant specimens in slices 1,2,4,6,7,8,and 9 (Figures 5_10a, b, d, f, g, h, and i, respectively). In slices 3 and 5 (Figures 5_10c and e), 34567 clustered with subtype B. Low bootstrap values (44%) were observed in slice 5 for the clustering with subtype B; this is most probably due to the presence of a subtype-unclassified region within this slice a) Slice 1; b) Slice 2; c) Slice 3; d) Slice 4; e) Slice 5; f) Slice 6; g) Slice 7; h) Slice 8**;** i) Slice 9.

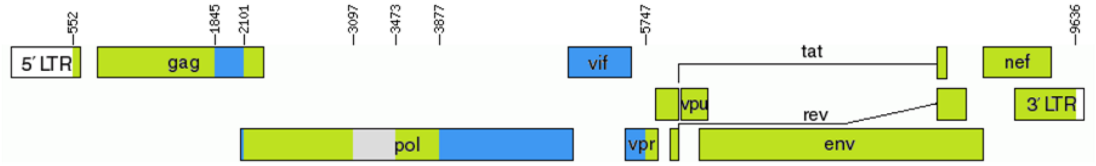


Figure 5_11. Confirmed structure of the complex recombinant 34567. The confirmed structure of the complex A1/B/D/U recombinant specimen 34567 following maximum likelihood analysis with the five A1/D specimens (Figure 5_10). The regions that correspond to the structure seen in the A1/D recombinants are shown in green, and subtype B regions are shown in blue. The region HXB2 3097 - 3473 was unable to be definitively resolved.

5.3 Confirmation of the new recombinant as a novel circulating recombinant from and registration as a new CRF.

The analysis performed to confirm the structure of 34567 also suggested that the A1/D recombinant was a new circulating recombinant form. To register a new CRF with the Los Alamos database, satisfaction of three criteria was required: a) that the recombinant was novel, b) that it clustered with itself across the whole genome and c) that it be identified in three epidemiologically unlinked patients. The three patients that were fully characterised (33365, 8179 and 40534) were from London centres, but were epidemiologically unlinked, thus criteria c was satisfied. Criteria b was satisfied during the clustering analysis of specimen 34567, as the A1/D recombinant specimens demonstrated clustering across the entire genome; an extra analysis was performed including specimens 11762 and 12792 using approximate maximum likelihood trees to see whether all five patients were infected with the same CRF. As all five specimens clustered together across the genome, we were satisfied that these five recombinant specimens were a shared strain. With two criteria satisfied, the final criteria, that of novelty, was assessed.

5.3.1 Global BLAST of sequence

Two checks of the A1/D recombinant were performed to ensure that the recombinant structure was novel. Firstly, the full-length sequence and *pol* gene sequence of specimen 33365 were submitted to BLAST to search for similar structures. This search returned no matches with the same recombinant structure as the A1/D structure for either the full-length sequence or the *pol* region.

5.3.2 Comparison with registered HIV-1 CRF list

The final check for novelty was to compare the recombinant structure with the published list of HIV-1 CRFs at the Los Alamos website. No matches were found. Accordingly, three A1/D sequences (specimens 33365, 8179 and 40534) were submitted to Los Alamos for registration as a new CRF. Following checks by Los Alamos, the recombinant was duly registered as CRF50_A1D, and can be found on the Los Alamos list of published CRFs (available at http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html). The official breakpoints for the recombinant were the breakpoints set using the jpHMM analysis, and the reference sequence for the strain was that of specimen 8179. All six full-length sequences were submitted to GenBank (accession numbers: JN417236-JN417241; reference sequence: JN417236).

Chapter 6: Emergence, distribution and likely origin of CRF50_A1D

Following the registration of CRF50_A1D as a new CRF, an investigation was performed into the origin of this recombinant. There were three goals for this investigation: 1) to identify how many cases of CRF50_A1D infection were present in the UK HIV DRD in order to assess the spreading potential of the novel strain; 2) to investigate the geographic origin of the component subtype A1 and D strains, in order to determine whether CRF50_A1D emerged from a recombination event in the UK or was imported as a formed recombinant; 3) to investigate the tMRCA of CRF50_A1D in the UK and the geographic and exposure group distribution across the country.

6.1 Identification of additional CRF50_A1D cases in the UK HIV DRD

6.1.1 BLAST, SCUEAL and jpHMM analysis of the UK HIV DRD

The original cluster analysis identified six potential cases of CRF 50_A1D, of which five were CRF50_A1D and the sixth a CRF50_A1D/B recombinant. Additional cases of CRF50_A1D present in the UK HIV DRB that were not captured during the original analysis were captured using local BLAST searches of the UK HIV DRB and genotyping using SCUEAL and jpHMM.

Summarised results for SCUEAL genotyping are in Table 6_1. In each of the three closest sequence match lists, the majority of sequences indentified as close genetic matches were pure subtype A1 sequences (305/500 (61.0%) for specimen 33365, 332/500 (66.4%) for specimen 8179 and 338/500 (67.6%) for specimen 40534, respectively). This was an expected result owing to the large proportion of subtype A1 in the *pol* gene of CRF50_A1D (~1kb in the 1302bp that was captured by the UK HIV DRB). In the remaining one third of sequences that were not subtype A1, A1/D recombinant results were the most numerous, with 90/500 (18.0%) A1/D recombinant results for specimen 33365, and 82/500 (16.4%) for specimens 8179 and 40534, respectively. The next largest fraction of results was A1/B recombinant sequences; overall 6.6% of tests produced an A1/B recombinant result.

	Reference sequence							
	33365		8179		40534		Overall	
Subtype	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion
A,A1 recombinant	2	0.40	2	0.40	3	0.60	7	0.47
A-ancestral	1	0.20	0	0	0	0	1	0.07
A-ancestral,A1 recombinant	3	0.60	4	0.80	6	1.20	13	0.87
A1	305	61.00	332	66.40	338	67.60	975	65.00
A1,A2 recombinant	3	0.60	2	0.40	2	0.40	7	0.47
A1,A4 recombinant	2	0.40	1	0.20	0	0.00	3	0.20
A1,AE recombinant	5	1.00	7	1.40	6	1.20	18	1.20
A1,B recombinant	35	7.00	38	7.60	26	5.20	99	6.60
A1,C recombinant	2	0.40	2	0.40	2	0.40	6	0.40
A1,D recombinant	90	18.00	82	16.40	82	16.40	254	16.93
A1,G recombinant	3	0.60	0	0	1	0.20	4	0.27
A1,J recombinant	1	0.20	0	0	1	0.20	2	0.13
A1,U recombinant	5	1.00	1	0.20	2	0.40	8	0.53
A3	1	0.20	1	0.20	1	0.20	3	0.20
AE	2	0.40	3	0.60	4	0.80	9	0.60
B	18	3.60	4	0.80	1	0.20	23	1.53
CRF02-like	1	0.20	0	0	0	0.00	1	0.07
CRF15	0	0	0	0	1	0.20	1	0.07
CRF22	1	0.20	1	0.20			2	0.13
Complex	19	3.80	20	4.00	24	4.80	63	4.20
D	1	0.20	0	0	0	0	1	0.07
Total	500		500		500		1500	100

Table 6_1. Summary of SCUEAL subtyping results. Three CRF50_A1D sequences (33365, 8179, 40534) were used as the reference sequence in three local BLAST searches of the UK HIV DRB. The 500 closest matches to each sequence were genotyped using SCUEAL. This method engendered a high degree of overlap between the three sequence lists, but allowed the entire range of genetic information contained in the three sequences to be used to capture additional cases of CRF50_A1D. The highest proportion of genotyping results were pure subtype A1, followed by A1/D recombinants and A1/B recombinants.

All A1/D and A1/B sequences from each list were also genotyped using jpHMM (Table 6_2). All sequences identified by SCUEAL as A1/D or A1/B were also identified as A1/D or A1/B by jpHMM. The jpHMM results show that the majority of sequences in the sequence lists had breakpoints located in a 20 nucleotide region between HXB2 2485 and 2505 (specimen 33365 71/90 (78.9%); specimen 8179 63/82 (76.8%); specimen 40534 70/91 (76.9%)). This 20 nucleotide region contained the jpHMM-identified breakpoint for CRF50_A1D of HXB2 2489, confirming that the closest sequence matches were likely to be additional cases of CRF50_A1D.

	Reference sequence					
	33365		**8179**		**40534**	
Total sequences analysed	90		82		91	
Subtype	**Count**	**Proportion**	**Count**	**Proportion**	**Count**	**Proportion**
A1/B	50	55.6	43	52.4	17	18.7
A1/D	40	44.4	39	47.6	74	81.3
Breakpoint	**Count**	**Proportion**	**Count**	**Proportion**	**Count**	**Proportion**
2385	0	0	1	1.2	0	0
2485-2490	22	24.4	19	23.2	56	61.5
2491-2495	3	3.3	4	4.9	8	8.8
2496-2500	4	4.4	4	4.9	1	1.1
2501-2505	42	46.7	36	43.9	5	5.5
2506-2510	2	2.2	1	1.2	7	7.7
2511-2515	2	2.2	3	3.7	2	2.2
2516-2520	0	0	0	0	1	1.1
2521-2525	0	0	0	0	0	0
2526-2530	1	1.1	1	1.2	3	3.3
2530-3115	14	15.5	13	15.9	8	8.8

Table 6_2. jpHMM subtyping results for A1/D and A1/B recombinant sequences. All A1/D and A1/B sequences identified by SCUEAL genotyping were further subtyped using jpHMM. Results are summarised first by subtype result (A1/D or A1/B) and then by breakpoint location. The majority of sequences had breakpoints located between HXB2 2485 and 2505.

All sequences with a breakpoint location which either matched the SCUEAL (HXB2 2465) or jpHMM (HXB2 2489) CRF50 breakpoint, or contained one of these breakpoint positions in the 95% confidence interval produced with breakpoint predictions (e.g. a breakpoint of HXB2 2505 with a 95% confidence interval of HXB2 2485-2525, which contains HXB2 2489) were classified as CRF50_A1D cases. Following removal of duplicate sequences from the same patient, 72 sequences remained (Appendix 6_1).

6.1.2 Addition of demographic information to the 72 CRF50_A1D sequences.

Where available, demographic information collected by UK CHIC was matched to the 72 CRF50_A1D patients. Of 72 patients, 53/72 (74%) had sex available, 51/72 (71%) had risk group available, and 49/72 (68.0%) had ethnicity information available (Table 6_3). All 53 patients with sex available were men. In terms of exposure route, 47/51 (92.1%) were MSM, 2/51 (3.9%) were heterosexual and 2/51 (3.9%) were IDU. The ethnicity results showed that 47/49 (95.9%) were white and 2/49 (2.0%) were Black-African and Indian/Pakistani, respectively. The overwhelming majority of cases were found in white MSM (46/51, 90%); however, the instances in heterosexual men and IVDU were significant. It should be noted, however, that the data capturing algorithms in UK CHIC only allow one risk group to be entered per individual, and that a risk group of IVDU is considered to be of higher priority than a risk group of MSM. Therefore, it is possible that the identified IVDU instances were also MSM. Finally, it should also be noted that risk groups are self-identified by the infected individuals, and that social and cultural pressures can influence self-reporting. Therefore, it is also possible that the sole heterosexual male also belonged to the MSM risk group.

Demographic characteristic	Classifier	Number of sequences
Sex	Male	53
	Female	0
	Unknown	19
Exposure group	MSM	47
	Heterosexual	2
	IDU	2
	Unknown	21
Ethnicity	White	47
	Black-African	1
	Indian/Pakistani	1
	Unknown	22

Table 6_3. Demographic breakdown of CRF50 patients. Where available, demographic data collected by UK CHIC was matched to the CRF50_A1D sequences. Approximately one third of patients had no demographic information available. Sequences with available demographics showed an exclusively male profile, with the overwhelming majority of sequences coming from white MSM. Two heterosexuals and two IDU were identified, and the majority of sequences came from white patients.

All sequences had associated geographic data. Most cases were located in NW England (39/72, 54.2%) and London/SE England (27/72, 37.5%). Three cases (4.2%) were located in SW England, one (1.4%) in NE England, and two (2.8%) in the Edinburgh region of Scotland.

6.2 Geographic origin of parental subtype A1 and D strains

The global BLAST search performed in Chapter 5 as part of the CRF50_A1D registration process identified no other recorded cases of the CRF50_A1D recombinant structure. In order to investigate the origin of this recombinant, the origin of the parental component subtype A1 and D strains was sought.

6.2.1 Construction of global alignments.

The alignments of subtype A and D sequences constructed for regions of *gag, pol* and *env* genes were analysed both in their entirety and following filtering by genetic distance. Although sequences were collected from any country with a proportion of subtype A or D sequences that exceeded 10%, in practice filtering the sequences using genetic distance resulted in alignments composed almost exclusively of sequences from East Africa and the UK (Tables 6_ 4, 6_5).

Gene					
gag		***pol***		***env***	
No. of sequences pre-screening	**No. of sequences post-screening**	**No. of sequences pre-screening**	**No. of sequences post-screening**	**No. of sequences pre-screening**	**No. of sequences post-screening**
1261	385	486	251	536	185
Countries pre-screening	**Countries in final alignment**	**Countries pre-screening**	**Countries in final alignment**	**Countries pre-screening**	**Countries in final alignment**
Belarus	Belarus	Belarus	Kenya	Burundi	Kenya
Cameroon	Latvia	Burundi	Rwanda	DRC	Rwanda
DRC	Kenya	Cameroon	Uganda	Kenya	Tanzania
Georgia	Rwanda	DRC	UK	Tanzania	Uganda
Kenya	Tanzania	Gabon		Rwanda	
Latvia	Uganda	Georgia		Uganda	
Russia		Kazakhstan			
Rwanda		Kenya			
Tanzania		Moldova			
Uganda		Russian Federation			
Ukraine		Rwanda			
		Tanzania			
		Uganda			
		Ukraine			
		UK			

Table 6_4. Composition of global subtype A alignments. Sequences were obtained from public repositories of HIV-1 sequences, and were collected from any country where the proportion of either subtype A or subtype D sequences was >10%. Filtering by genetic distance using CRF50 sequences as the reference resulted in alignments composed almost exclusively of sequences from East Africa and the UK.

Gene					
gag		***pol***		***env***	
No. of sequences pre-screening	**No. of sequences post-screening**	**No. of sequences pre-screening**	**No. of sequences post-screening**	**No. of sequences pre-screening**	**No. of sequences post-screening**
1091	138	360	216	375	215
Countries pre-screening	**Countries in final alignment**	**Countries pre-screening**	**Countries in final alignment**	**Countries pre-screening**	**Countries in final alignment**
Chad	Chad	Botswana	Kenya	Chad	Chad
DRC	DRC	Chad	Tanzania	DRC	DRC
Kenya	Kenya	DRC	Uganda	Kenya	Kenya
Rwanda	Tanzania	Ethiopia	UK	Tanzania	Tanzania
Tanzania	Uganda	India		Uganda	Uganda
Uganda	UK	Kenya			
UK		Sudan			
		Tanzania			
		Uganda			
		UK			

Table 6_5. Composition of global subtype D alignments. Sequences were obtained from public repositories of HIV-1 sequences using the conditions detailed in Table 6_4. In common with the subtype A alignments, filtering the subtype D alignments by genetic distance also resulted in alignments composed almost exclusively of sequences from East Africa and the UK.

6.2.2 FastTree analysis

6.2.2.1 Subtype A1.

Although the majority of sequences in the unfiltered *gag* alignment came from East Africa, a limited number of sequences from Central African countries (DRC, Cameroon) and Eastern Europe (Russia, Latvia, Belarus, Ukraine) were also available (Table 6_4). The FastTree analysis of the unfiltered alignment showed all CRF50_A1D specimens clustered together with an approximate maximum likelihood support value of 0.99 (Figure 6_1a). The sequences that the CRF50_A1D sequences clustered most closely with were from East Africa, with a support value of 0.97; support for the whole branch was 0.84. The closest sequence was from Rwanda in 1992 (accession number U86548), followed by a sequence from Kenya in 2000 (accession number AF457067, support value 0.74). All of the Eastern European sequences clustered closely with sequences from Central Africa, with a branch support value of 0.87.

The filtered alignment for *gag* contained almost exclusively sequences from East Africa, apart from two sequences from Eastern Europe, one from Belarus and one from Latvia (Table 6_4). The CRF50_A1D sequences also clustered together in this tree, with a support value of 0.99 (Figure 6_1b). The topology of the FastTree tree did not show clear grouping of sequences by country; however, the CRF50_A1D sequences clustered closest to the same sequence from Rwanda as in the unfiltered tree (support value 0.97). The next closest sequence was the same sequence from Kenya as in the unfiltered tree, however the support value for this branch was low (0.062). Overall, although the branch containing the CRF50_A1D sequences had high support (0.96) the internal branch support values for the topology were low, excepting the CRF50_A1D sequences.

In terms of international sequences, the unfiltered alignment for *pol* predominantly contained sequences from East Africa, although a limited number of sequences were also available from Central Africa, West Africa, Eastern Europe and Central Asia (Table 6_4). Pure subtype A1 sequences from the UK HIV DRB were available, and all of these were initially included. The FastTree analysis of this alignment showed that the CRF50_A1D sequences still formed a single cluster of sequences (support value = 1), even within the background of extensive sequences from the UK (Figure 6_1c). The CRF50_A1D sequences were clustered closest to East African sequences, particularly sequence AF457067, an A1 sequence from

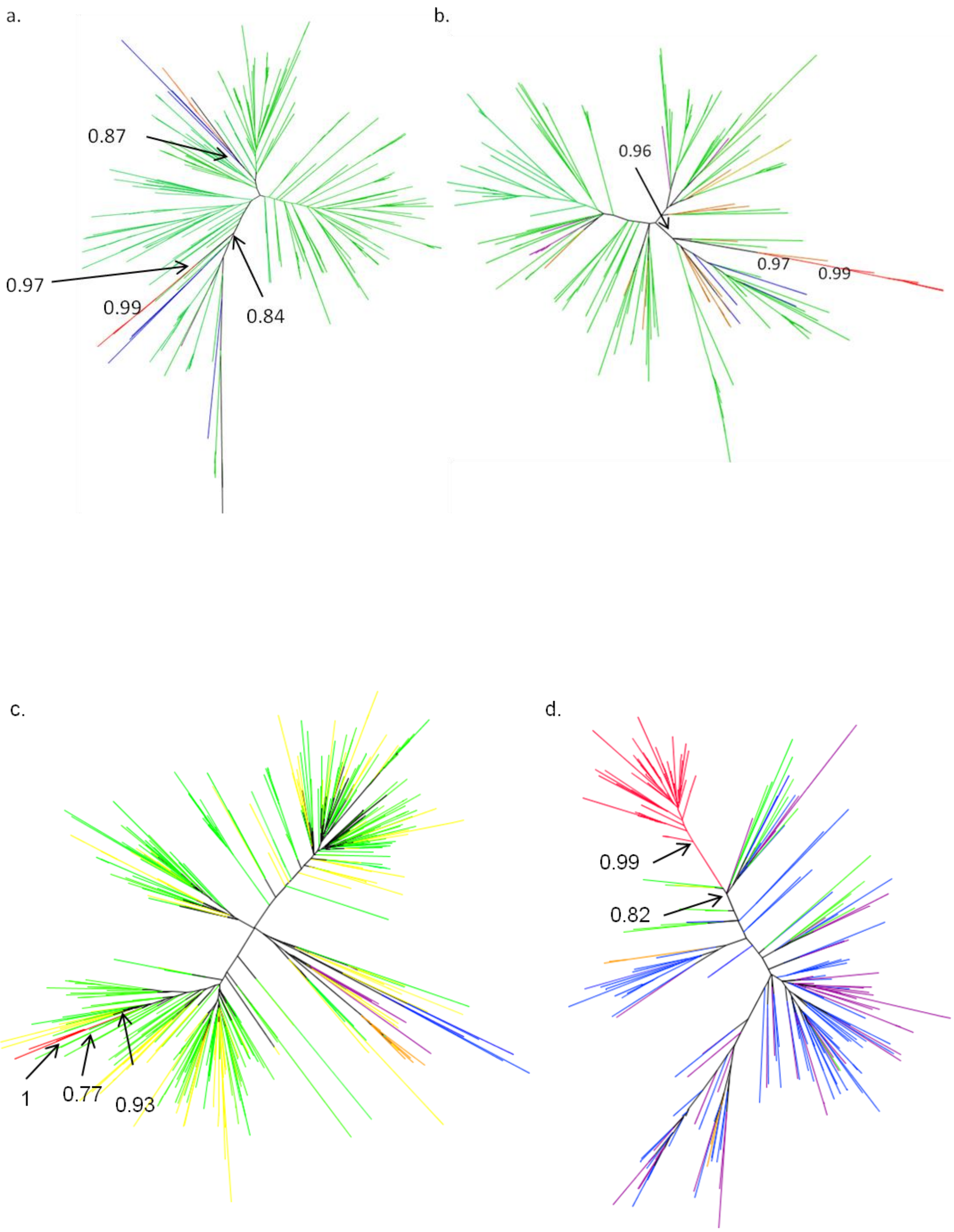
Kenya (support value = 0.77). The overall tree was well supported and showed high branch support values throughout.

The filtered alignment contained sequences from East Africa and the UK only (Table 6_4). In order to evaluate the relationship between the CRF50_A1D sequences, the East African sequences and the UK sequences, all 72 CRF50_A1D sequences were included in the analysis. The FastTree analysis showed that, even with all 72 sequences included, the CRF50_A1D sequences clustered in a monophyletic group (support value = 0.99; Figure 6_1d). The closest sequence to the CRF50_A1D sequences was sequence AF457067 (support value = 0.82) similar to the unfiltered analysis.

The subtype A1 region of *env* in CRF50_A1D was not a genetic region for which sequences have been collected from a wide variety of geographic regions. Even in the unfiltered alignment, almost all of the sequences were from East Africa, with the exception of three sequences from the DRC (Figure 6_1e). The CRF50_A1D sequences did not cluster with the DRC sequences. The CRF50_A1D sequences clustered together with a strong support value of 0.99. The closest sequences to the CRF50_A1D sequences were a cluster of 19 sequences from Kenya; 12 were from 1996 and collected as part of a mother-to-child transmission study, 2 were from 1997 and collected as part of a superinfection study (accession numbers EU164115 - EU 164116), and 5 were from 1997 and collected as part of a viral evolution study (accession numbers FJ641711 - FJ641715). The support value for this branch with CRF50_A1D was 0.88; support for the entire branch was 0.74 (Piantadosi et al., 2007, 2009).

The filtered alignment contained sequences from East Africa only (Table (6_4). The FastTree analysis showed the CRF50_A1D sequences clustered together with a support value of 1 (Figure 6_1f). The closest sequences were a cluster of sequences from Kenya made up of sequences from the superinfection and evolution studies mentioned in the paragraph above (support value = 0.87). The support value for the whole branch was 0.77. In this analysis, the branch containing the sequences from the mother-to-child transmission study were no longer located close to the CRF50_A1D sequences. In common with the trees for the subtype A1 *gag* region, the filtered tree did not display clear grouping of sequences into specific countries; however, almost all of the sequences from Tanzania grouped in a tight cluster. This was also seen in the filtered tree for *gag*, where nearly all of the sequences from Tanzania apart from two clustered together (Figure 6_1b).

Overall, the analysis of the subtype A global alignments showed that the CRF50_A1D sequences were most closely related to sequences from East Africa. This was the case even when an extensive background of sequences from the UK were included in the analysis.



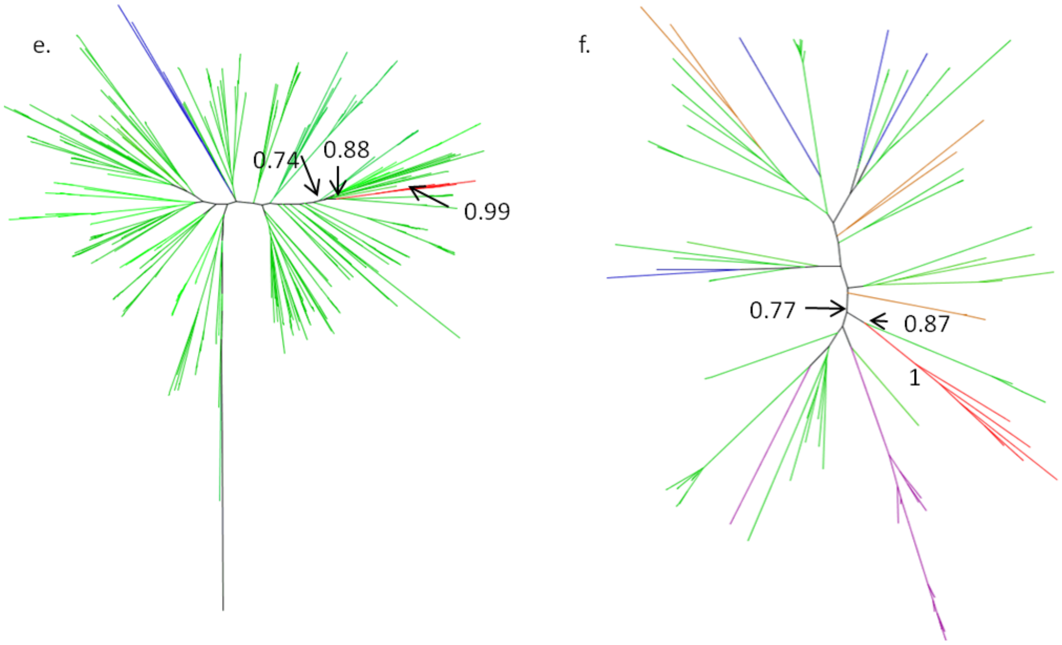


Figure 6_1. Approximate maximum likelihood trees for subtype A1 *gag, pol* and *env* regions of CRF50_A1D. Analysis was performed using unfiltered alignments containing all identified sequences and alignments filtered using genetic distance screening. In the unfiltered trees (a, c and e) sequences are coloured by geographic region, where East Africa = green, West Africa = purple, Central Africa = blue, Eastern Europe = orange, Central Asia = turquoise, UK = yellow. In the filtered trees for *gag* and *env* (b and f), sequences are coloured by country, where Kenya = green, Uganda = blue, Rwanda = orange, Tanzania = purple, Latvia = yellow. In the filtered tree for *pol* (d) sequences are coloured by country, where Kenya = green, Uganda = blue, Rwanda = orange, UK = purple. CRF50 sequences are shown in red all trees. Numbers represent approximate maximum likelihood support values. **a)** Unfiltered tree for *gag*. The CRF50_A1D sequences cluster closest to sequences from East Africa (support value for CRF50_A1D clustering = 0.99, support value for closest sequences = 0.97). The support for the branch containing the CRF50_A1D sequences and the closest East African sequences is 0.84. Sequences from Eastern Europe cluster closest to sequences from Central Africa (support value = 0.87); **b)** Filtered tree for *gag*. The CRF50_A1D sequences cluster together (support value = 0.99), and closest to a sequence from Rwanda (support value = 0.97). The support value for the branch containing CRF50_A1D and Rwanda sequences was 0.96; **c)** Unfiltered tree for *pol*. The CRF50_A1D sequences clustered closest to a sequence from Kenya (support value for CRF50_A1D sequences = 1, support value for branch with Kenyan sequence = 0.77, support value for entire branch = 0.93); **d)** Filtered tree for *pol*. This tree contains all 72 CRF50 sequences. The CRF50_A1D sequences form a monophyletic cluster with a support value of 0.99. The closest sequence is the same Kenyan sequence seen in the unfiltered tree; the support value for this branch is 0.82; **e)** Unfiltered tree for *env*. The CRF50_A1D sequences cluster together with a support value of 0.99. The closest sequences were a cluster of 19 sequences from Kenya; the support value for this was 0.88. The support value for the branch containing the CRF50_A1D and Kenyan sequences was 0.74; **f)** Filtered tree for *env*. The CRF50_A1D sequences cluster together with a support value of 1. The closest sequences were a cluster containing 12 of the same Kenyan sequences as seen in the unfiltered tree (support value = 0.87). The support value for the branch containing the CRF50_A1D sequences and the Kenyan sequences was 0.77.

6.2.2.2 Subtype D.

The unfiltered alignment for the *gag* subtype D region contained mainly sequences from East Africa, but also contained a good proportion of sequences from Central African countries (Table 6_5, Figure 6_2a). In the FastTree analysis, the pure CRF50_A1D sequences clustered together with a support value of 0.98; however the sequence from the URF, specimen 34567, clustered in a close, but separate, branch (Figure 6_2a). The closest sequence to the URF was from Uganda from 1995 (accession number K21221A6). This branch had weak support of 0.38. The support value for the branch containing the CRF50_A1D sequences, the Ugandan sequence and the URF was 0.26. The closest sequence to this branch was a sequence from the UK in 2000 (accession number FJ712794) and a sequence from Uganda in 1995 (accession number A12410A1). The support value for this branch was 0.87.

The filtered alignment contained sequences from both East and Central African countries (Table 6_5). The FastTree analysis also showed the pure CRF50_A1D sequences clustered together (Figure 6_2b, support value = 0.93). The closest sequence was a sequence from Uganda from 1994 (accession number A00336A1); the branch support was 0.71. In the filtered tree, the URF sequence clustered much further away from the CRF50_A1D sequences than in the unfiltered tree (Figure 6_2b). The URF was located in a branch that contained the sequence from the UK from 2000 with a support value of 0.79 (accession number FJ712794). The support value for the whole branch was 0.82.

The region of *gag* from the URF that was contained in this alignment was located in one of the subtype B regions of the URF genome. The clustering with the UK sequence in the filtered tree indicated that the subtype B regions of the URF came from a different geographic region than the subtype D regions of the URF, further indicating that the event that recombined the URF with a CRF50_A1D sequence took place in a different geographic region than the recombination event that resulted in CRF50_A1D.

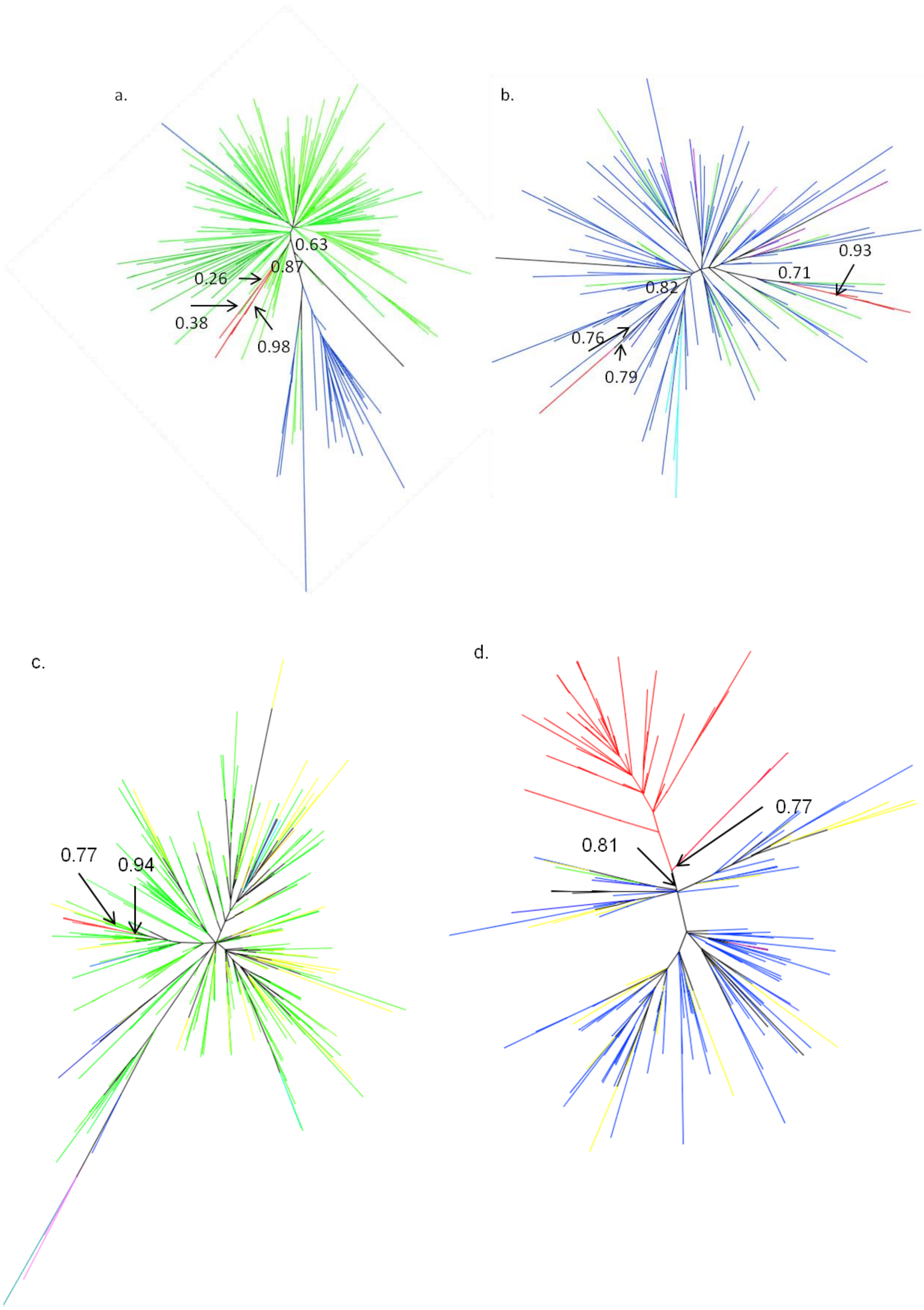
In common with the subtype A alignments, the unfiltered and filtered alignments for *pol* contained almost equal numbers of sequences from East Africa and the UK (Table 6_5). The FastTree analysis for the unfiltered alignment showed the CRF50_A1D sequences clustered together with a support value of 0.77. The closest sequences to the CRF50_A1D sequences were a cluster of five sequences from

Uganda. The support value for the branch containing the CRF50_A1D sequences and the Ugandan sequences was 0.94 (Figure 6_2c).

The filtered alignment for *pol* was comprised predominantly of sequences from Uganda (Figure 6_2d). This alignment contained all 72 CRF50_A1D sequences. The CRF50_A1D sequences clustered together with a support value of 0.77. The closest sequence was a sequence from Uganda dated 1997; the branch containing this sequence and the CRF50_A1D sequences had a support value of 0.81.

Both the unfiltered and filtered alignments for *env* contained sequences from East Africa, and Central Africa (Table 6_5). The FastTree analysis for the unfiltered alignment showed the CRF50_A1D sequences clustering together with a support value of 1 (Figure 6_2e). The closest sequence was a sequence from Kenya dated 1997; the support value for the branch containing this sequence and the CRF50_A1D sequences was 0.91. The filtered alignment contained predominantly sequences from Tanzania and Kenya (Figure 6_2f). The support value for the CRF50_A1D sequences was 1. The closest sequences were the same sequence from Kenya seen in the unfiltered tree and a sequence from Tanzania dated 1995. The support value for the branch containing these sequences and the CRF50_A1D sequences was 0.96.

Overall, each of the subtype A1 and D *gag, pol,* and *env* regions of CRF50_A1D showed strongly supported clustering with sequences from East Africa; unfortunately, we were not able to more precisely define the geographic region when using sequences from publicly available repositories. The CRF50_A1D sequences displayed monophyletic clustering in all of the 12 analyses. Monophyletic clustering in the *pol* trees, which contained all of the pure subtype A1 and D sequences from the UK HIV DRD, supports the assertion that this recombinant was imported into the UK as an existing recombinant strain rather than arising from an in-country recombination event. The exception to this is the URF sequence, specimen 34567. This sequence clustered monophyletically with CRF50_A1D in all trees apart from the subtype D *gag* trees, where it was located on a separate branch. That the closest sequence to this branch was a sequence from the UK supports the assertion that this CRF50_A1D/B recombinant arose from a UK recombination event with a subtype B strain.


a.
0.63
0.26
0.87
0.38
0.98
b.
0.93
0.71
0.82
0.76
0.79
c.
0.77
0.94
d.
0.77
0.81

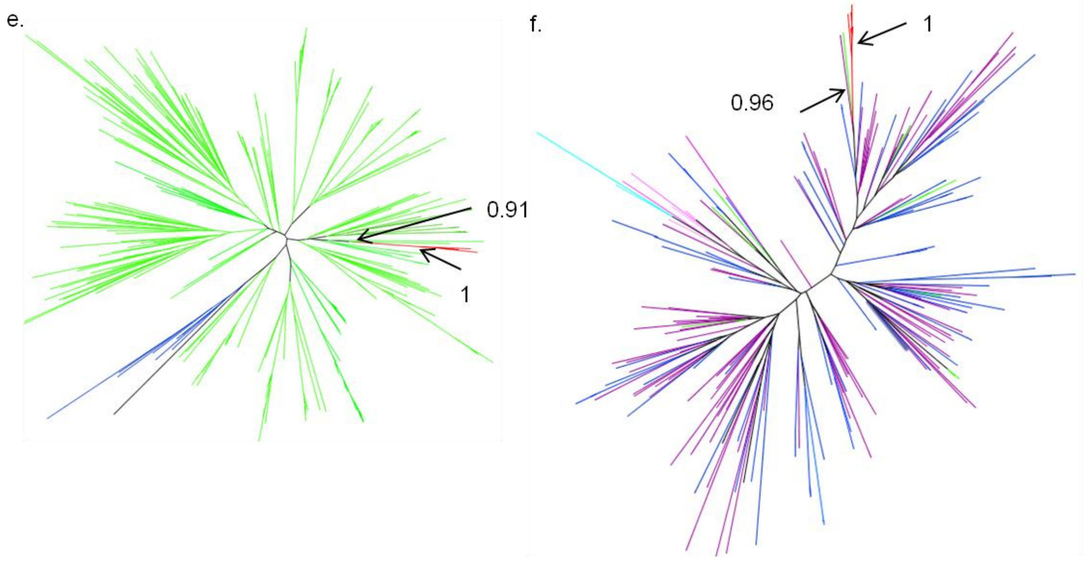


Figure 6_2. Approximate maximum likelihood trees for subtype D *gag, pol* and *env* regions of CRF50_A1D. Analysis was performed using unfiltered alignments and filtered alignments that had been filtered by genetic distance. In the unfiltered trees sequences are coloured by geographic region, where East Africa = green, West Africa = purple, Central Africa = blue, North Africa = turquoise, Southern Africa = Lavender, Eastern Europe = orange, India = teal, UK = yellow. In the filtered trees, sequences are coloured by country, where Kenya = green, Uganda = blue, Rwanda = Orange, Tanzania = purple, Chad = turquoise, DRC = Lavender, UK = yellow. Numbers represent approximate maximum likelihood support values. The trees for *pol* contain all 72 CRF50_A1D sequences. **a)** Unfiltered tree for *gag*. The pure CRF50_A1D sequences cluster closest to sequences from East Africa (support value = 0.98). Specimen 34567 (the URF) is located in a close, but separate branch; **b)** Filtered tree for *gag*. The CRF50_A1D sequences cluster together with a support value of 0.93. The closest sequence is a sequence from Uganda (support value = 0.71). The URF is located on a separate branch which contains a specimen from the UK (support value = 0.79); **c)** Unfiltered tree for *pol*. The CRF50_A1D sequences (support value = 0.77) cluster closest to sequences from Uganda (support value = 0.94); **d)** Filtered tree for *pol*. The CRF50_A1D sequences (support value (0.77) cluster closest to sequences from Uganda (support value = 0.81); **e)** Unfiltered tree for *env*. The CRF50_A1D sequences (support value = 1) cluster closest to a sequence from Kenya (support value = 0.91); **f)** Filtered tree for *env*. The CRF50_A1D sequences (support value = 1) cluster closest to a sequence from Kenya and a sequence from Tanzania (support value = 0.96).

6.3 Emergence and distribution of CRF50_A1D in the UK.

Having established that the parental subtype A1 and D strains of CRF50_A1D were related to strains from East Africa, and that CRF50 itself was likely to have been imported into Britain as a recombinant, an investigation was performed into the time of introduction of CRF50_A1D in Britain and its distribution throughout the UK. This was performed using the six full-length sequences (five CRF50_A1D plus one URF), the additional 67 CRF50_A1D *pol* sequences identified from the UK HIV DRB; from the full-length sequences, subtype A1 and D regions of *gag, pol* and *env* genes, and the concatenated A1/D *pol* gene region were investigated independently.

6.3.1 Model selection.

6.3.1.1 Alignments containing all 72 CRF50_A1D sequences (*pol* gene)

6.3.1.1.1 A/D concatenated alignment.

The FindModel results for the concatenated A1/D *pol* alignment were summarised in Table 6_6. The nucleotide substitution model with the lowest AIC score (and therefore the best fit for the data) was GTR plus gamma. Similar AIC scores were also seen for the HKY plus gamma and the Tamura-Nei plus gamma models, respectively.

Nucleotide substitution model	AIC score		
	A1	D	A1/D
Jukes-Cantor	87904.25	10628.46	125603.07
Jukes-Cantor plus gamma	5133.67	1855.31	7065.46
Felsenstein 1981	5320.99	4577.41	7340.82
Felsenstein 1981 plus gamma	5105.41	1731.40	7021.76
Kimura 2-parameter	5150.97	1768.25	7134.52
Kimura 2-parameter plus gamma	4933.75	1690.54	6816.78
Hasegawa-Kishino-Yano	5117.20	1766.59	7093.91
Hasegawa-Kishino-Yano plus gamma	4893.00	1686.11	6762.24
Tamura-Nei	5099.32	1768.02	7093.91
Tamura-Nei plus gamma	4869.08	1687.51	6745.54
GTR	5090.09	1763.93	7061.83
GTR plus gamma	4858.38	1685.00	6726.42

Table 6_6. FindModel results for 72 CRF50_A1D *pol* sequences. FindModel best fit results for subtype A1, subtype D, and A1/D concatenated sequences for all 72 CRF50_A1D sequences. The nucleotide substitution model with the lowest AIC score was considered the best fit. The lowest AIC score for each alignment was the GTR plus gamma substitution model (highlighted in grey).

Following the FindModel testing, the GTR plus gamma model and HKY plus gamma models were selected for further comparison using BEAST. GTR plus gamma and HKY plus gamma models were compared using both strict molecular clocks and uncorrelated lognormal relaxed molecular clocks, and Bayes factors calculated for each permutation. Table 6_7 showed the $\log_{10}$ Bayes factors for the model comparisons. The model with the best fit was the GTR plus gamma substitution model with a relaxed uncorrelated lognormal molecular clock; this had a $\log_{10}$ Bayes factor of 226.56 compared to a HKY plus gamma model with a strict molecular clock, 201.28 compared to the HKY plus gamma model with a relaxed uncorrelated lognormal molecular clock and 23.14 compared to the GTR plus gamma model with a strict molecular clock. The GTR plus gamma model with a relaxed uncorrelated lognormal molecular clock was selected as the best model to analyse the concatenated A1/D *pol* gene alignment.

Model	HKY strict clock	HKY relaxed clock	GTR relaxed clock	GTR strict clock
HKY strict clock	-	-25.28	-226.56	-203.42
HKY relaxed clock	25.28	-	-201.83	-178.15
GTR relaxed clock	226.56	201.28	-	23.14
GTR strict clock	203.42	178.15	-23.14	-

Table 6_7. $\log_{10}$ Bayes factors for A1/D segment. GTR and HKY nucleotide substitution models were compared using both strict and relaxed molecular clocks. The Bayes factor of each combination of models is shown; models listed vertically are compared to models listed horizontally. A score of >20 is considered to be significantly more in favour of the model in question. The GTR with a relaxed molecular clock had the highest Bayes factor in each comparison.

6.3.1.1.2 Subtype A1 and subtype D alignments

In common with the concatenated A1/D alignment, the FindModel results for the component subtype A1 and subtype D regions showed that the GTR plus gamma model had the best predicted fit (Table 6_6). This model was compared to the HKY plus gamma model using a relaxed uncorrelated molecular clock, and Bayes factors were calculated to compare the results (Table 6_8). For both A1 and D regions of the *pol* gene, the GTR plus gamma model with a relaxed uncorrelated lognormal molecular clock was the best model; Bayes scores were 141.01 and 39.50 for subtypes A1 and D, respectively when compared to the HKY plus gamma model.

Therefore, the GTR plus gamma model with a relaxed uncorrelated lognormal molecular clock was selected as the best model to use.

Model	**GTR relaxed**	**HKY relaxed**
Subtype A		
GTR relaxed	-	141.01
HKY relaxed	-141.01	-
Subtype D		
GTR relaxed	-	39.50
HKY relaxed	-39.50	-

Table 6_8 log$_{10}$ Bayes factors for subtype A1 and subtype D segments of CRF50_A1D *pol* gene. GTR and HKY nucleotide substitution models were compared using relaxed molecular clocks. In both cases, the GTR model had the highest Bayes factor.

6.3.1.2 Alignments containing CRF50_A1D full-length sequences only (gag, pol, env genes; full-length sequences)

Although the alignment that contained all 72 sequences was the most apposite alignment to use in terms of giving enough genetic information to estimate a tMRCA, it was restricted by only having a partial *pol* gene. Therefore, estimates for tMRCA were also calculated using the full-length sequences, and the *gag, pol* and *env* genes. Subtype A1 and D regions were analysed separately in order to confirm that the genes and subtypes had a shared evolutionary path. Alignments were analysed both with and without the URF (specimen 34567) in order to ascertain whether this sequence represented a precursor of CRF50_A1D or an onward recombination event.

6.3.1.2.1 Full-length sequences

The FindModel results for the alignment containing the full-length CRF50_A1D sequences indicated that the GTR plus gamma model was the best fit (Table 6_9). This model was compared with both strict and relaxed uncorrelated molecular clock models; the Bayes factor for the strict clock compared with the relaxed uncorrelated lognormal clock was 0.26. This result was not consistent with a significant difference in the molecular clock chosen, and so for reasons of consistency the relaxed molecular clock was selected for use.

6.3.1.2.2 Subtype A1 and D *gag, pol* and *env* alignments

Unlike the preceding alignments, the FindModel results for the subtype A1 and subtype D regions of the *gag, pol* and *env* genes were not uniformly in agreement (Table 6_9). The GTR plus gamma model was the predicted best fit in 5/12 (41.7%) cases, the HKY plus gamma model in 4/12 (33%) of cases, the Tamura Nei plus gamma model in 2/12 cases (16.7%) and the Tamura Nei model in 1/12 (8.3%) of cases. Accordingly, analysis began using the GTR plus gamma model. However, for those alignments in which GTR plus gamma was not the preferred substitution model, it proved difficult to achieve consistent ESS >200, even when combining multiple runs and using a strict molecular clock. Therefore, the decision was made to use the simpler HKY plus gamma model for all of the alignments, so that the results were comparable across each gene. The HKY plus gamma model was used with a relaxed uncorrelated lognormal molecular clock.

Model	Full-length sequences	Subtype A1 genes						Subtype D genes					
		gag		*pol*		*env*		*gag*		*pol*		*env*	
		Pure CRF50	**With 34567**	**Pure CRF50**	**With 34567**	**Pure CRF50**	**With 34567**	**Pure CRF50**	**With 34567**	**Pure CRF50**	**With 34567**	**Pure CRF50**	**With 34567**
Jukes-Cantor	33865.69	1802.76	2377.20	2575.86	1496.59	31137.71	4265.84	2110.36	2236.75	914.09	1068.85	2504.66	3183.13
Jukes-Cantor plus gamma	33694.28	1803.21	2369.06	2566.93	1488.97	3116.70	4196.05	2102.99	2224.20	915.43	1066.85	2505.50	3162.11
Felsenstein 1981	33194.72	1756.70	2322.92	2494.65	1451.58	3071.53	4196.19	2075.80	2204.21	896.86	1051.38	2469.13	3142.88
Felsenstein 1981 plus gamma	33020.53	1756.90	2313.03	2489.55	1443.81	3049.80	4120.66	2068.21	2190.92	898.03	1048.63	2469.94	3120.73
Kimura-2 parameter	33531.00	1788.48	2333.84	2543.99	1465.09	3092.69	4203.30	2089.82	2187.63	902.72	1043.04	2490.16	3141.60
Kimura-2 parameter plus gamma	33348.91	1788.60	2325.65	2534.75	1456.64	3070.83	4131.31	2082.04	2173.30	903.87	1040.50	2490.82	3120.34
HKY	32883.23	1744.60	2284.67	2466.23	1419.57	3020.86	4122.69	2056.55	2158.27	886.26	1026.89	2455.95	3104.69
HKY plus gamma	32699.73	1744.41	2273.90	2456.49	1408.02	2998.84	4048.25	2048.61	2143.86	887.44	1024.20	2456.46	3081.25
Tamura-Nei 93	32884.69	1745.63	2282.87	2466.29	1415.94	3018.36	4121.58	2058.47	2160.14	884.09	1025.30	2457.12	3105.97
Tamura-Nei 93 plus gamma	32701.47	1745.45	2272.59	2456.63	1406.18	2995.82	4047.63	2050.54	2145.62	885.53	1023.84	2457.74	3082.92
GTR	32886.29	1745.19	2276.97	2470.71	1419.81	3018.46	4115.25	2063.46	2164.54	884.63	1022.40	2455.22	3096.53
GTR plus gamma	32683.43	1745.10	2267.49	2461.06	1410.34	2996.03	4040.82	2055.46	2149.88	886.25	1020.86	2455.64	3071.82

Table 6_9. FindModel AIC scores for subtype A1 and D alignments for *gag, pol* and *env* genes. The reduced set of substitution models were considered for each alignment, and the model with the lowest AIC score was considered the best fit (highlighted). Please note that the results for the subtype A1 alignment that included specimen 34567 appeared inconsistent with the results for the alignment containing pure CRF50_A1D sequences only. This is not a discrepancy; in the region of *pol* that was analysed, specimen 34567 is not pure subtype A1, and therefore the alignment area was trimmed to ensure only subtype A1 genetic evolution was analysed.

6.3.2 Time scaled phylogenies

6.3.2.1 MCMC analysis of all 72 CRF50_A1D sequences

Time scaled phylogenies were performed for *pol* using the subtype A1 region of the genome, the subtype D region of the genome and the concatenated A1/D region (Figure 6_3). Summarised tMRCA results are shown in Table 6_10.

6.3.2.1.1 A1/D concatenated alignment.

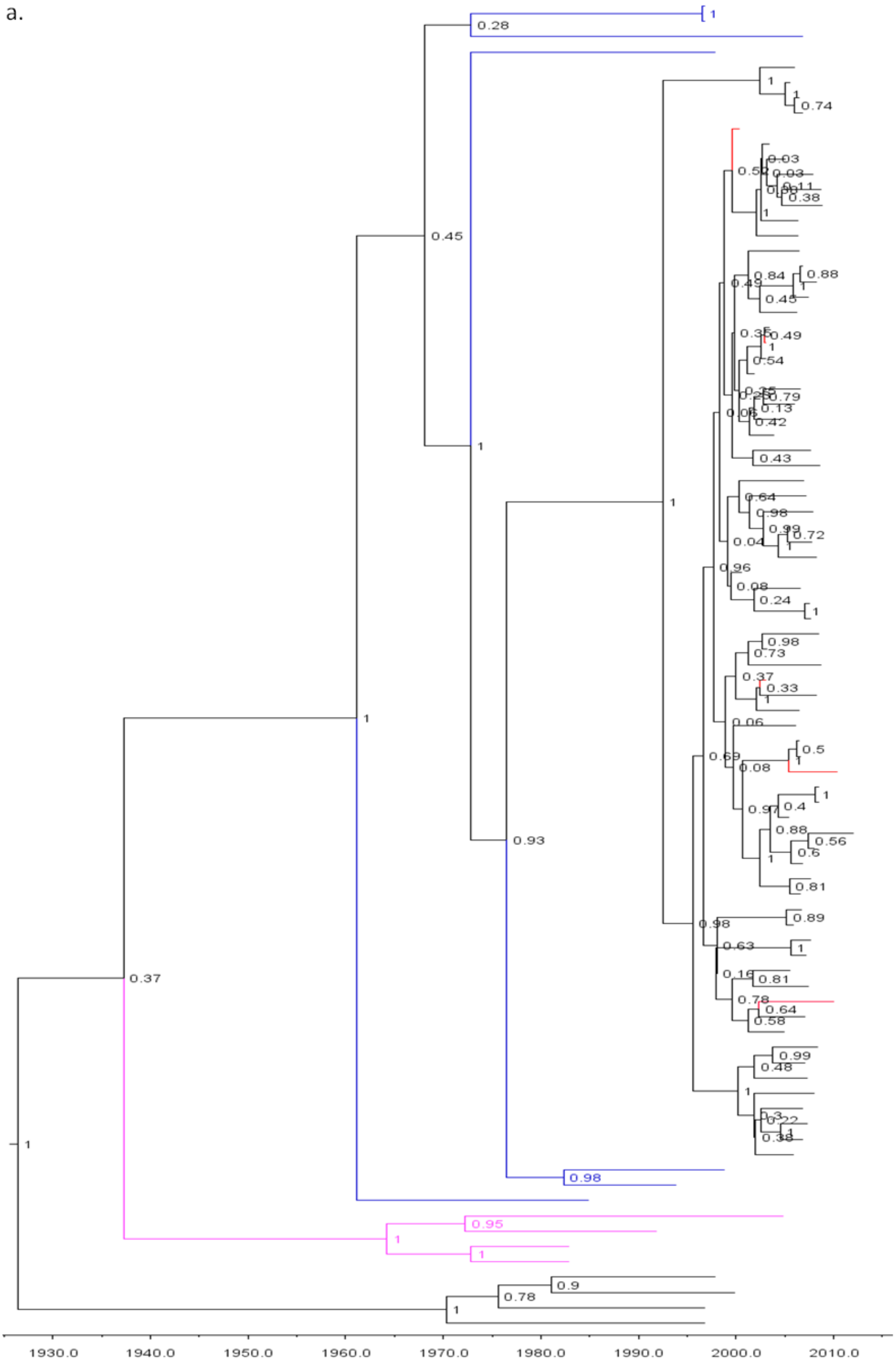
MCMC analysis of the A1/D concatenated alignment showed a tMRCA of 1992.46 (95% HPD 1987.47-1998.57) (Table 6_10, Figure 6_3a). The sequences from the full-length sequencing were distributed evenly across the CRF50_A1D cluster. The tree had consistent strong support for the topology throughout.

6.3.2.1.2 Subtype A1 and subtype D alignments

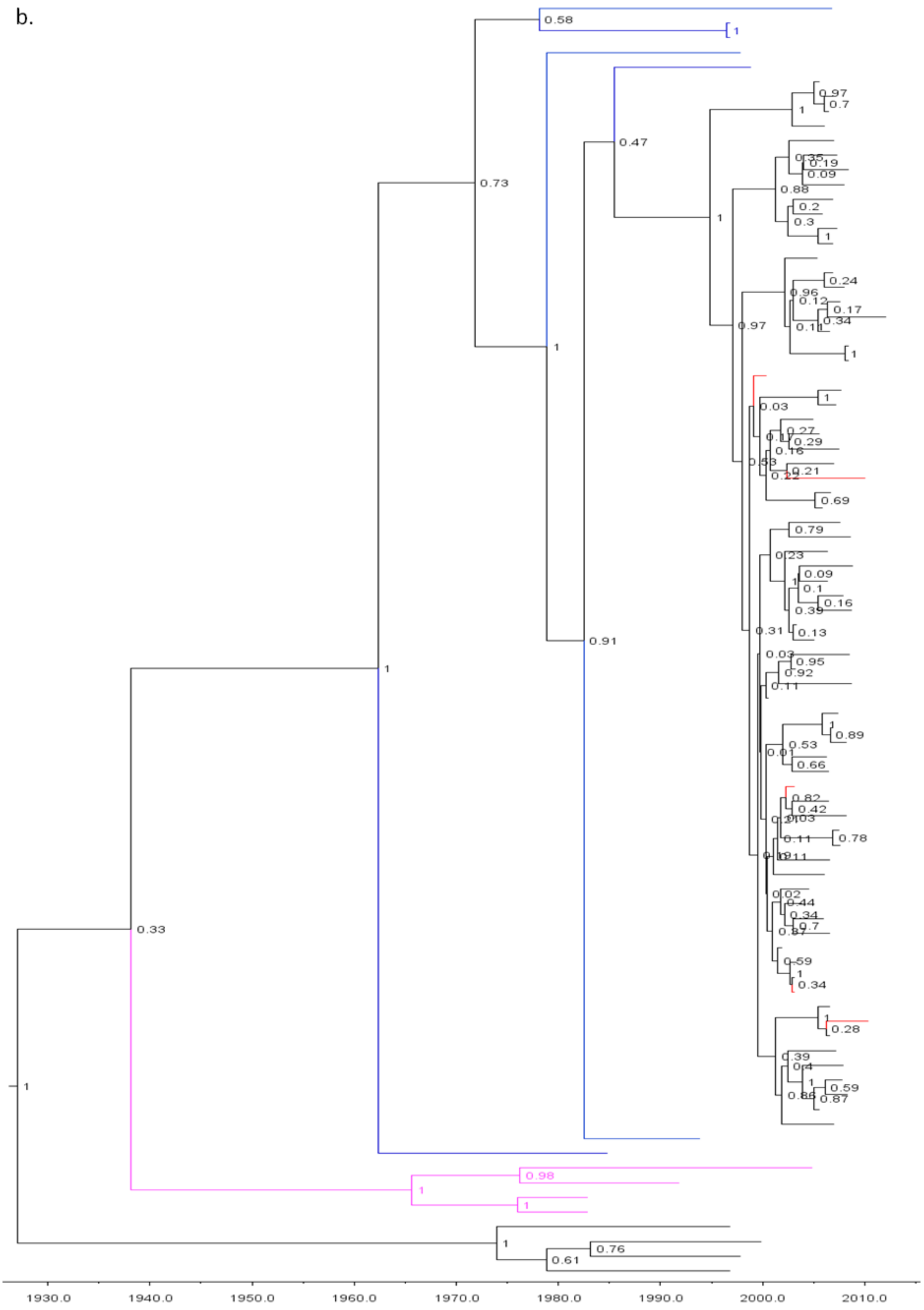
The tMRCA for subtype A1 was 1994.88 (95% HPD 1990.78-2000); the tMRCA for subtype D was 1994.36 (1989.80-2000.37) (Table 6_10). In both trees the sequences from the full length sequencing were distributed evenly across the CRF50_A1D cluster (Figure 6_3b, c). In the subtype D tree the CRF50_A1D cluster was split into two branches. The smaller of the two branches contained 25 sequences, but had very low topological support (posterior probability =0.02).

The A1 tree showed fairly consistent strong support across the tree. The D tree showed strong support for a uniform cluster of CRF50_A1D (posterior probability =0.89) but generally weak support for the topology within the CRF50_A1D cluster. This is not unexpected as the fragment of subtype D was very short and in a highly conserved area of the genome, resulting in a limited amount of variable genetic information.

a.


1
0.28
1
0.74
0.03
0.03
0.52
0.99
0.11
0.38
1
0.84
0.88
0.49
1
0.45
0.35
0.49
1
0.54
0.35
0.79
0.13
0.42
0.43
0.45
0.64
0.98
0.99
0.72
0.04
1
1
0.96
0.08
0.24
1
0.98
0.73
0.37
0.33
1
0.06
0.5
0.69
1
0.08
1
0.97
0.4
0.88
0.56
0.6
1
0.81
0.93
0.89
0.98
0.63
1
0.16
0.81
0.78
0.64
0.58
0.99
0.48
1
0.3
0.22
0.38
0.37
0.98
1
0.95
1
1
0.9
0.78
1
1930.0
1940.0
1950.0
1960.0
1970.0
1980.0
1990.0
2000.0
2010.0

b.


0.58
1
0.73
0.47
0.97
0.7
1
0.35
0.19
0.09
0.88
0.2
0.3
1
1
0.24
0.96
0.12
0.17
0.34
0.11
0.97
1
1
0.03
1
0.27
0.29
0.16
0.53
0.21
0.22
0.69
0.79
0.23
0.09
0.1
0.16
0.39
0.13
0.31
0.91
1
0.03
0.95
0.92
0.11
1
0.89
0.53
0.01
0.66
0.82
0.42
0.03
0.11
0.78
0.11
0.02
0.44
0.34
0.7
0.87
0.59
1
0.34
0.33
1
1
0.28
0.39
0.4
1
0.59
0.86
0.87
0.98
1
1
1
0.76
0.61
1930.0
1940.0
1950.0
1960.0
1970.0
1980.0
1990.0
2000.0
2010.0

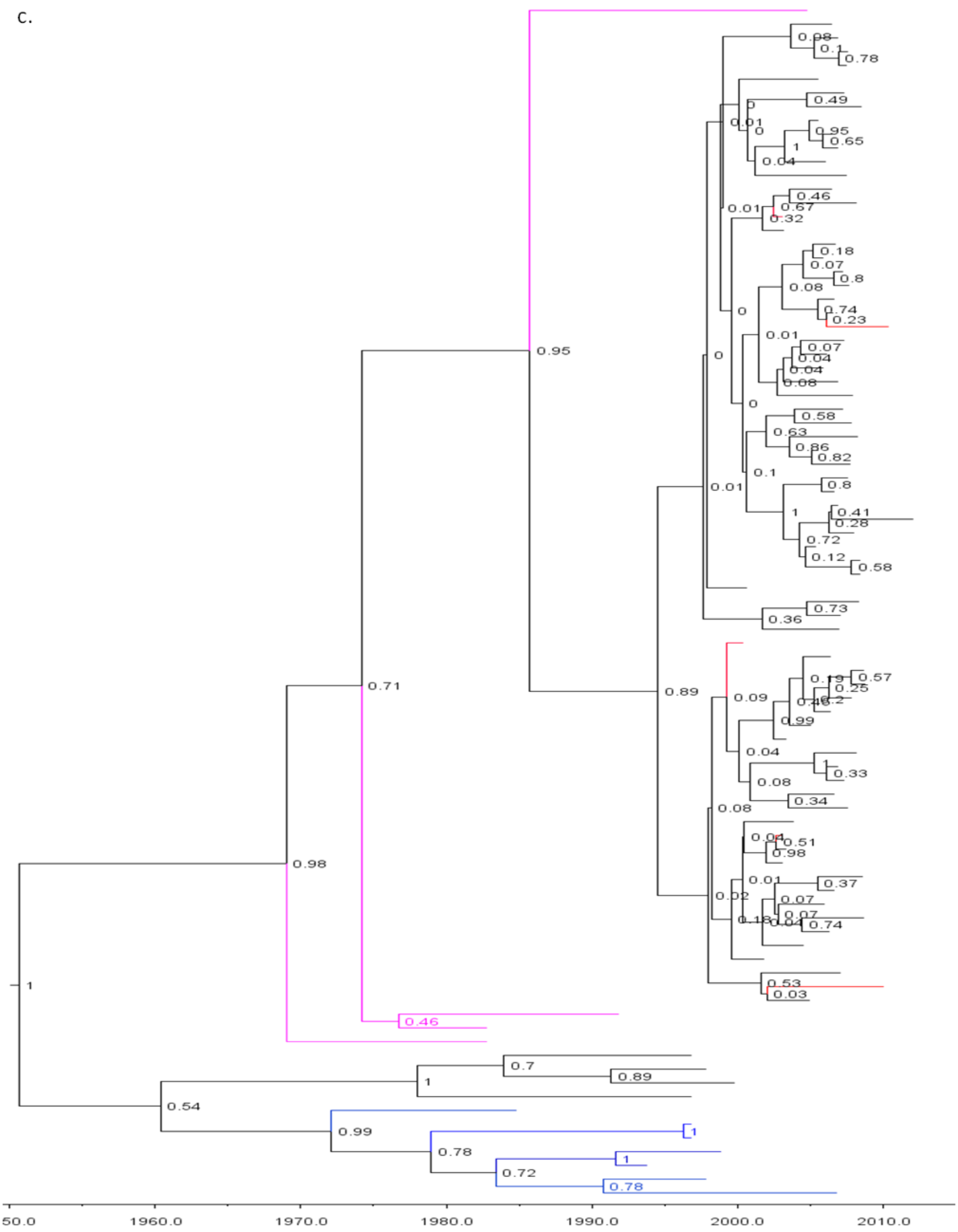


Figure 6_3. Time scaled phylogenies of all 72 CRF50_A1D sequences. Subtype A1 reference sequences are shown in blue. Subtype D sequences are shown in lavender. Sequences for which there was a full-length sequence available are shown in red. Posterior probabilities are displayed at nodes. The outgroup (subtype C) is shown in black. **a)** A1/D time scaled phylogeny. CRF50_A1D sequences have a tMRCA of approximately 1992; **b)** Subtype A1 region of *pol.* The topology of the tree is similar to that of the A1/D tree and shows CRF50_A1D with a tMRCA of approximately 1993; **c)** Subtype D region of *pol*. The topology of this tree is different to that of the A1/D and A1 trees, and the tree as a whole shows weaker support.

6.3.2.2 Demographic reconstruction

Demographic reconstruction of the number of effective CRF50_A1D infections, which estimates the number of infections contributing to new transmissions, grew exponentially in the period 2001-2002 (Figure 6_4). The reconstruction showed that aside from a small increase in the mid-1990s when CRF50_A1D first came to the UK, the number of effective infections remained stable until 2001-2002. Following the increase during this year, the number of effective infections returned to a more stable state. As of 2010, the number of effective infections was estimated at greater than 100, higher than the 72 instances identified in the UK HIV DRD. This indicated that a number of CRF50_A1D infections potentially remained uncaptured by current routine screening protocols.

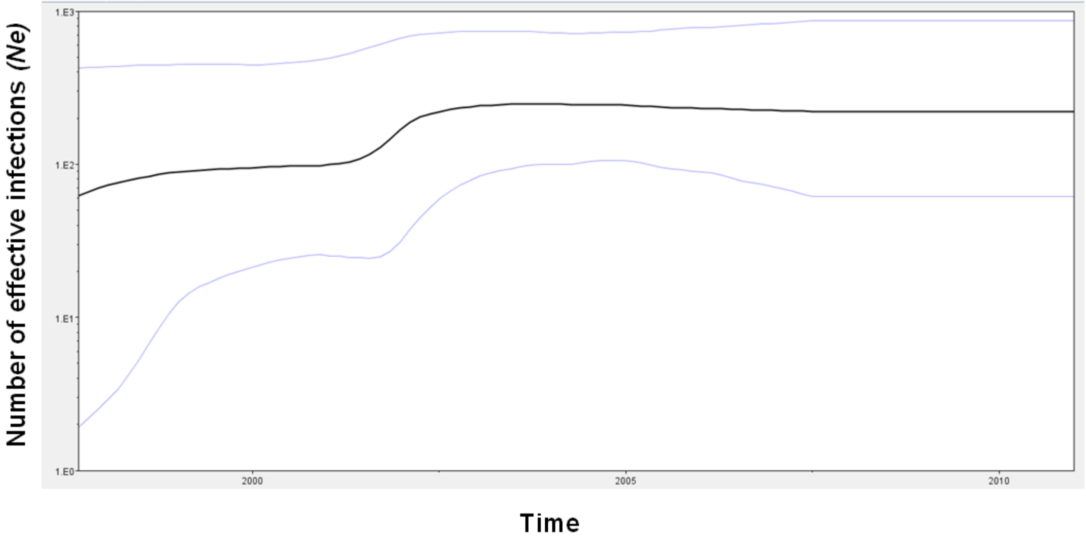


Figure 6_4. Bayesian Skyline reconstruction of the demographic history of CRF50_A1D in the UK. The black line represents the median; blue lines represent the 95% lower and upper HPD estimates of the effective population size. The plot shows an exponential increase in the number of effective infections between 2001 and 2002. As of 2010, the number of effective CRF50_A1D infections in the UK was estimated at greater than 100.

6.3.2.3 MCMC analysis of alignments containing CRF50_A1D full-length sequences and associated regions only (*gag, pol, env* genes; full-length sequences)

6.3.2.3.1 Full-length sequences

The tMRCA for the BEAST analysis of the full-length sequences was 1995.56 (1990.91-1999.60), consistent with the tMRCA of the trees using all 72 sequences

(Figure 6_4). There was strong support for the branches connecting 40534 and 8179 (posterior probability = 0.98) and for the branch connecting these sequences to 12792 (posterior probability = 0.88), but weak support for the branch connecting specimens 11762 and 33365 (posterior probability = 0.52). The branch lengths for specimens 12797 and 11762 were much longer than the branch lengths connecting the other sequences; this was as expected owing to the recent sampling date of these specimens compared to the specimens.

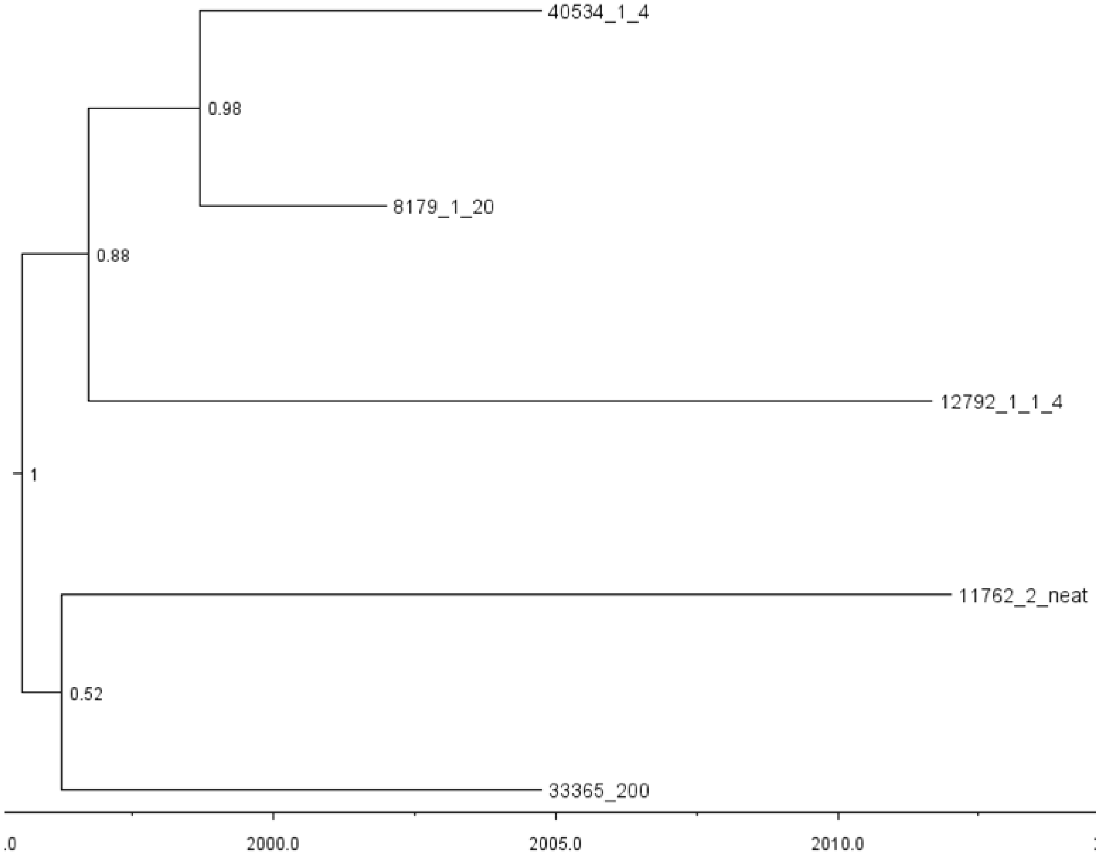


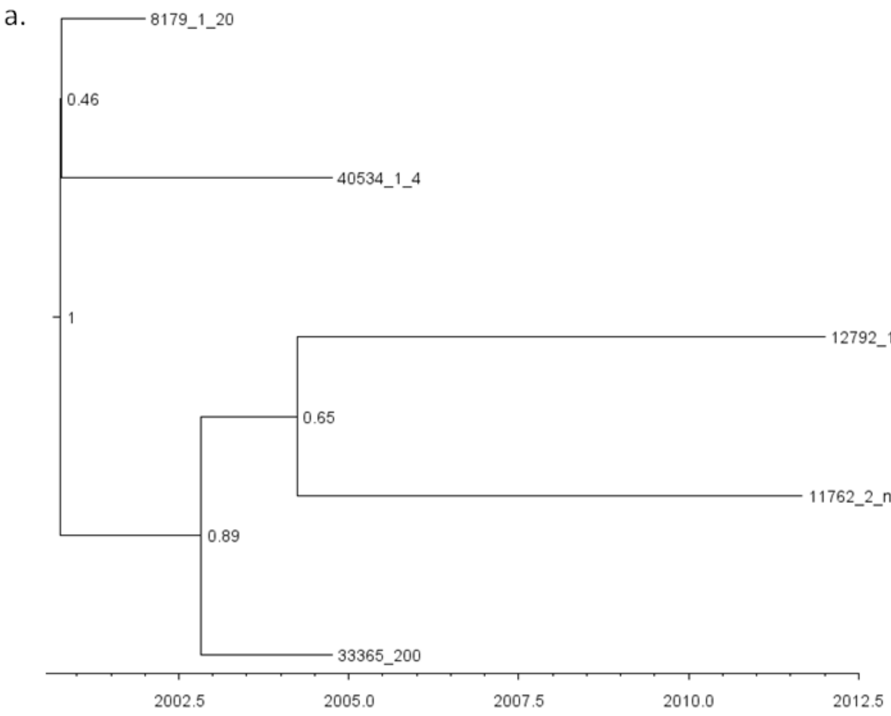
Figure 6_5. Time-scaled phylogeny using full-length CRF50_A1D sequences. The tMRCA using full-length sequences was 1995.56 (95% HPD 1990.91-1999.60), consistent with tMRCA from the A1/D *pol* analysis of 1992.46.

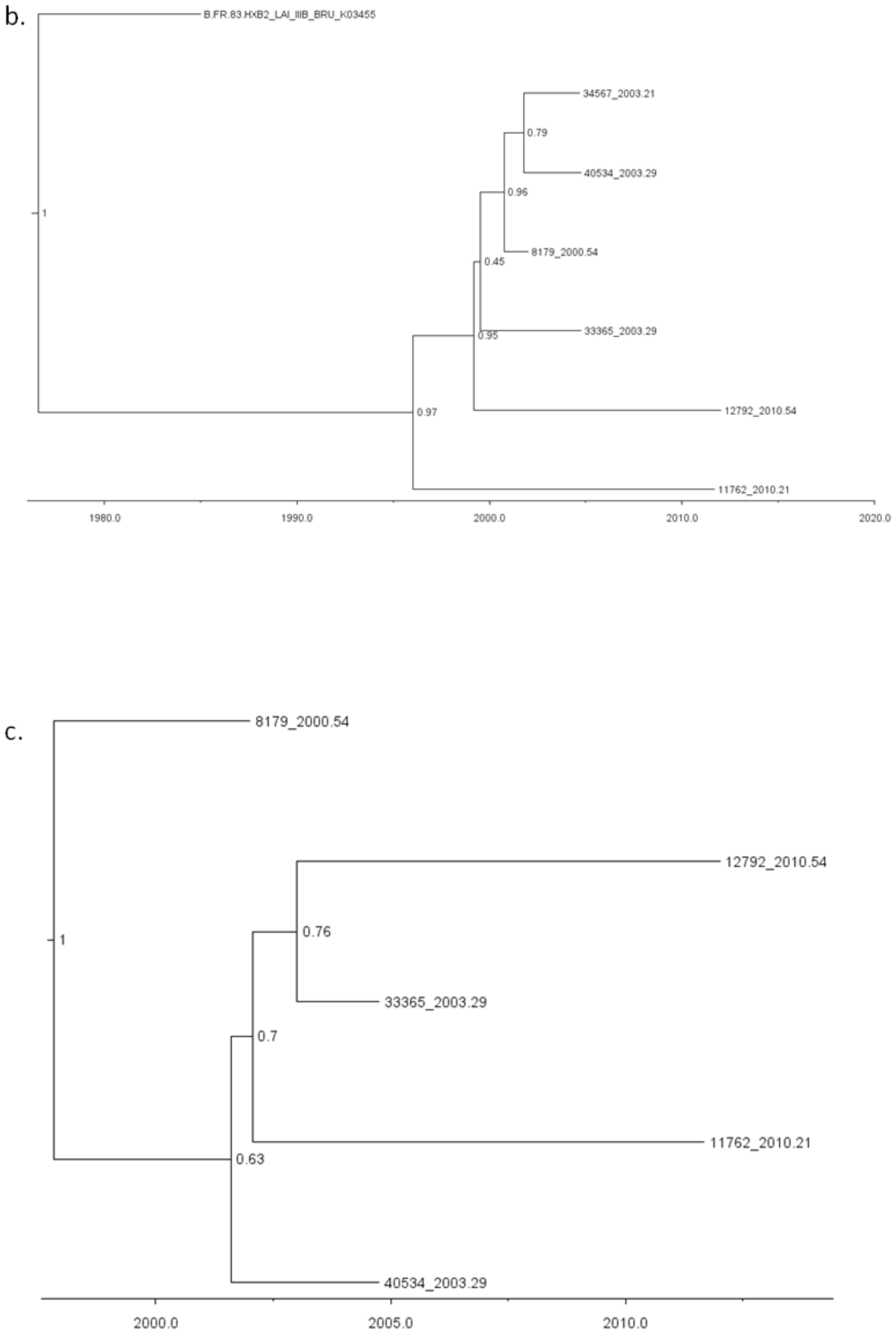
6.3.2.3.2 *gag , pol and env* alignments

The analysis of the *gag* subtype A1 alignment showed a tMRCA of 2000.76 (95% HPD 1998.84 – 2004.86) (Table 6_10). The tree showed generally weak support, with only one branch exceeding posterior probability support of 0.7 (Figure 6_6a). The corresponding tree containing specimen 34567 showed tMRCA of 1996.05 (95% HPD 1987.21-2010.49). This tree showed consistently strong support >0.7

except for the branch containing specimen 34567 (Figure 6_6b). The branch containing specimen 34567 showed that the sequence had a later divergence date than the pure CRF50_A1D sequences.

The subtype D trees for *gag* had similar topologies as the subtype A1 trees (Figure 6_6c, d). The tMRCA for the tree containing the pure CRF50_A1D sequences was 1997.83 (95% HPD 1991.69 – 209.60). Stronger support for the subtype D tree in general was shown compared to the subtype A1 tree. The tree containing specimen 34567 showed a later tMRCA than the pure CRF50_A1D specimens, but the branch had weak support.




b.
B.FR.83.HXB2_LAI_IIIB_BRU_K03455
34567_2003.21
0.79
40534_2003.29
0.96
1
8179_2000.54
0.45
33365_2003.29
0.95
12792_2010.54
0.97
11762_2010.21
1980.0
1990.0
2000.0
2010.0
2020.0
c.
8179_2000.54
12792_2010.54
1
0.76
33365_2003.29
0.7
11762_2010.21
0.63
40534_2003.29
2000.0
2005.0
2010.0

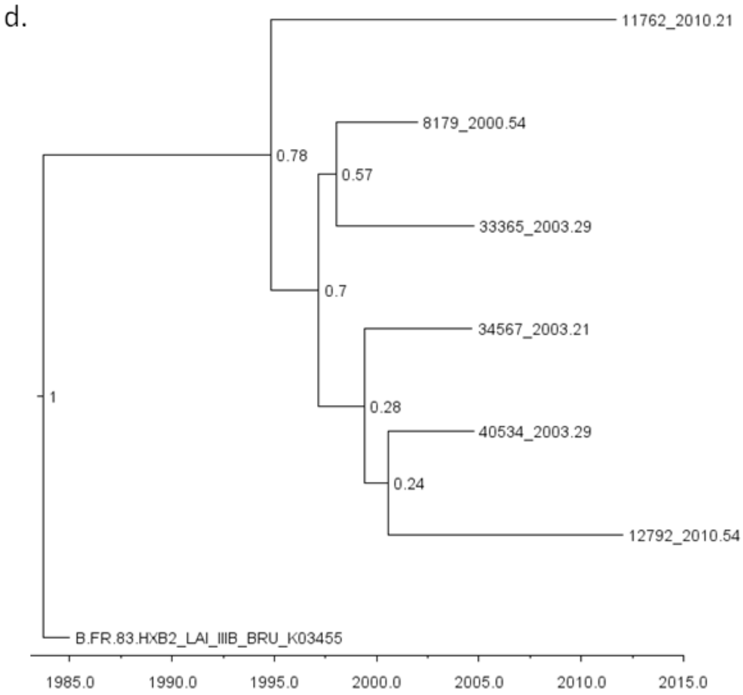
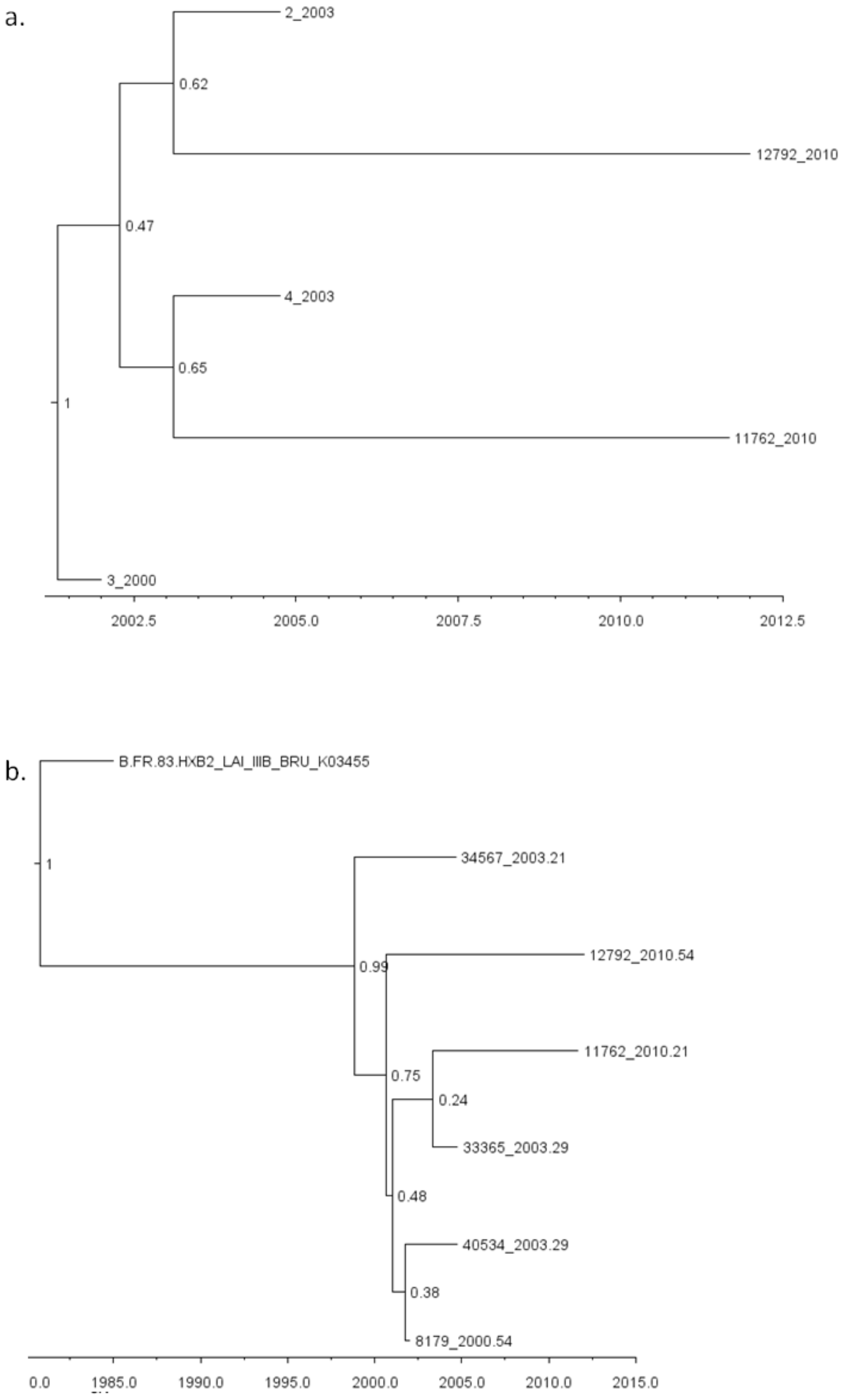


Figure 6_6. Time scaled phylogenies for CRF50_A1D *gag* regions. a) Subtype A1 region (pure CRF50_A1D sequences only) b) Subtype A1 region (including specimen 34567; c) Subtype D region (pure CRF50_A1D sequences only); d) Subtype D region (including specimen 34567).

Analysis of the subtype A1 *pol* alignment showed a tMRCA of 2001.33 (95%HPD 2000.22 – 2003.88) (Table 6_10). All branches in this tree had support values <0.7 (Figure 6_7a). The corresponding tree containing specimen 34567 showed stronger support but showed specimen 34567 with a divergence date prior to the pure CRF50_A1D sequences (Figure 6_7b).

The subtype D *pol* tree showed a tMRCA of 1998.83 (95% HPD 1994.46 – 2006.72) (Table6_10). There was good support throughout the tree (Figure 6_7c). The tree containing specimen 34567 showed a tMRCA of 2001.16 (95% HPD 1999.73 – 2004.64) (Table 6_9). The branch containing 34567 had weak support but showed a later divergence date than the pure CRF50_A1D sequences (Figure 6_7d).


a.
2_2003
0.62
12792_2010
0.47
4_2003
0.65
1
11762_2010
3_2000
2002.5
2005.0
2007.5
2010.0
2012.5
b.
B.FR.83.HXB2_LAI_IIIB_BRU_K03455
1
34567_2003.21
12792_2010.54
0.99
11762_2010.21
0.75
0.24
33365_2003.29
0.48
40534_2003.29
0.38
8179_2000.54
0.0
1985.0
1990.0
1995.0
2000.0
2005.0
2010.0
2015.0

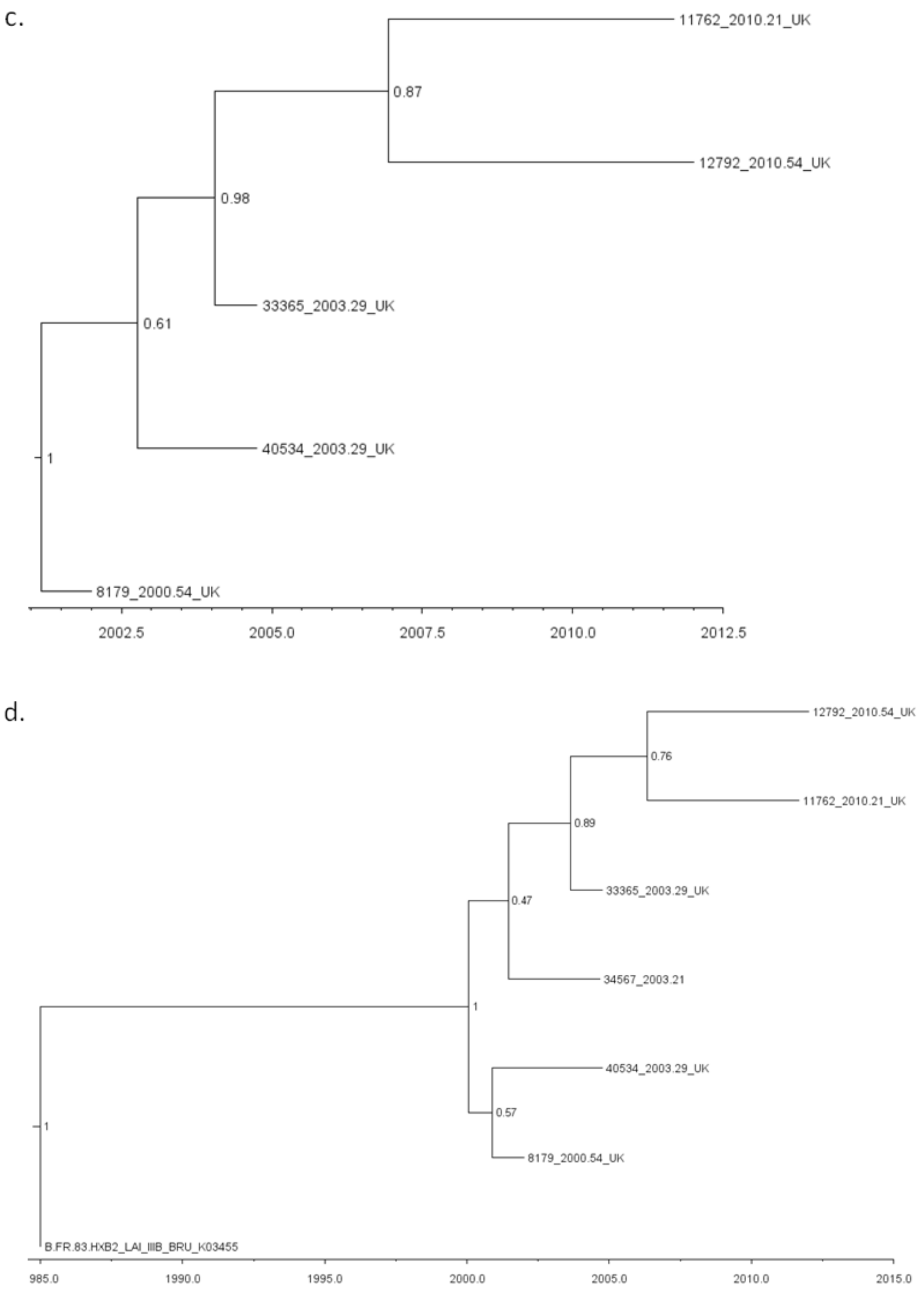
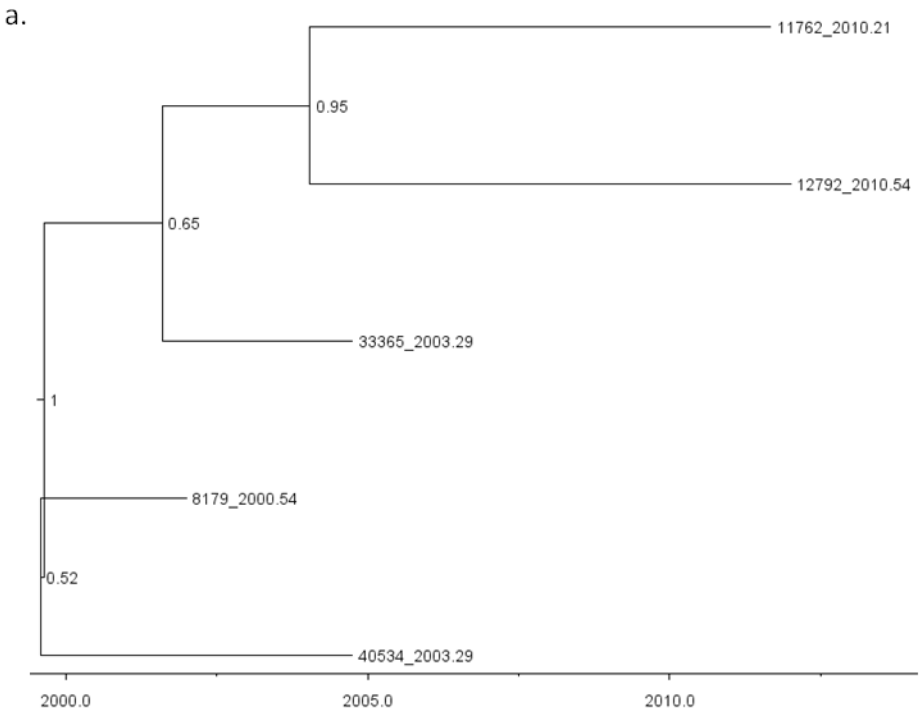


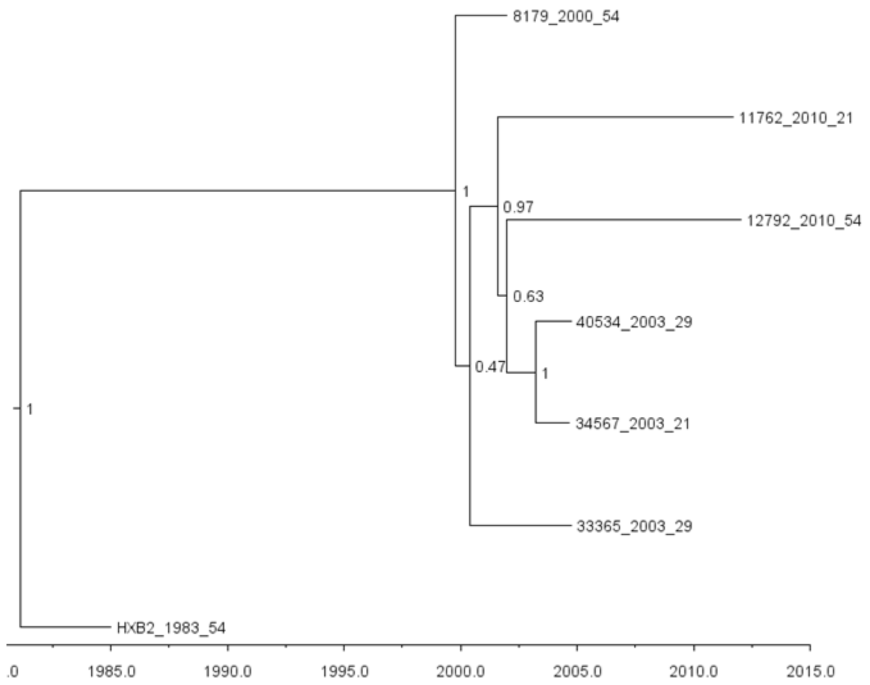
Figure 6_7. Time scaled phylogenies for CRF50_A1D *pol* regions. a) Subtype A1 region (pure CRF50_A1D sequences only) b) Subtype A1 region (including specimen 34567; c) Subtype D region(pure CRF50_A1D sequences only); d) Subtype D region (including specimen 34567).

The subtype A1 *env* tree had a tMRCA of 1999.64 (95% HPD 1996.47 – 2005.68). This tree was very weakly supported and contained negative branch lengths (Figure 6_8a). The tree containing specimen 34567 had consistent strong support especially for 34567 (posterior probability = 1) and showed a later divergence date of the URF compared to the pure CRF50_A1D sequences (Figure 6_8b).

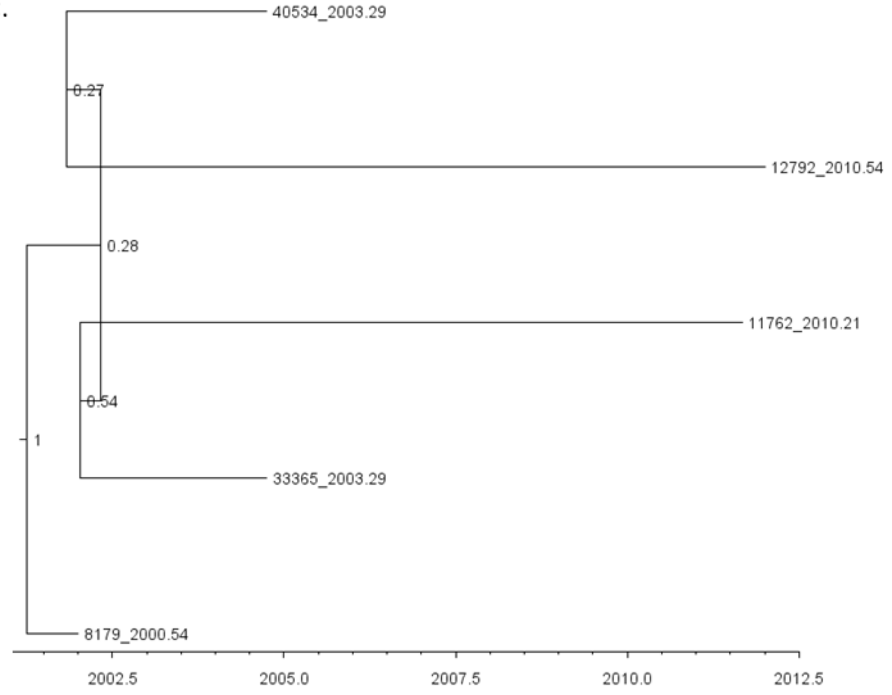
The subtype D *env* trees also showed negative branch lengths and very weak support (Figure 6_8c). In this case, the tree containing specimen 34567 also showed weak support throughout (Figure 6_8d).



b.


8179_2000_54
11762_2010_21
1
0.97
12792_2010_54
0.63
40534_2003_29
0.47
1
34567_2003_21
1
33365_2003_29
HXB2_1983_54
.0
1985.0
1990.0
1995.0
2000.0
2005.0
2010.0
2015.0

c.


40534_2003.29
0.27
12792_2010.54
0.28
11762_2010.21
1
0.54
33365_2003.29
8179_2000.54
2002.5
2005.0
2007.5
2010.0
2012.5

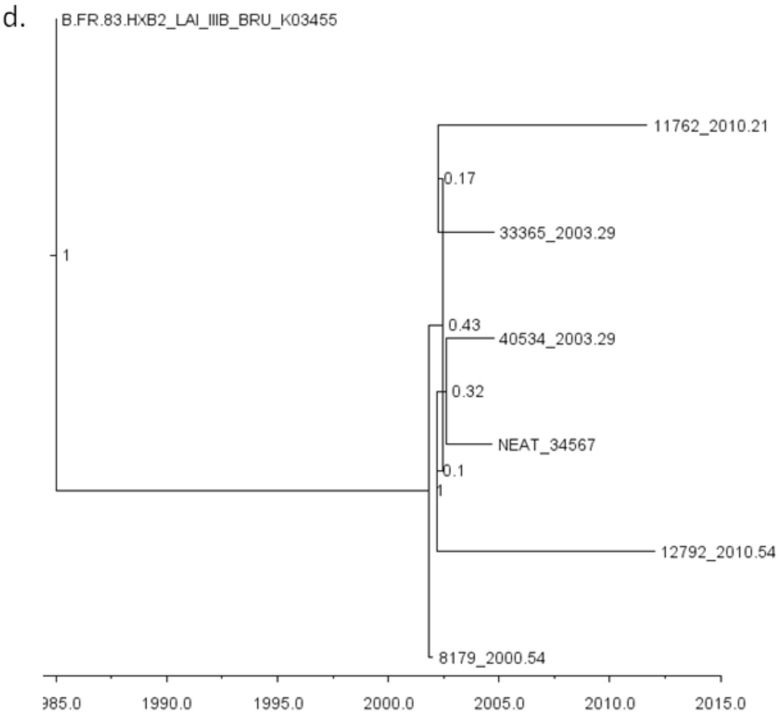


Figure 6_8. Time scaled phylogenies for CRF50_A1D *env* regions. a) Subtype A1 region (pure CRF50_A1D sequences only) b) Subtype A1 region (including specimen 34567); c) Subtype D region (pure CRF50_A1D sequences only); d) Subtype D region (including specimen 34567).

Overall, the median tMRCA for the individual gene analysis was 2000.76 for the pure CRF50_A1D sequences and 1999.31 for the alignments containing specimen 34567, respectively. The standard deviation of the tMRCA for the pure sequences was 1.32 and the standard deviation for alignments with 34567 was 2.58. The standard deviation over all individual gene dates was 2.15. The tMRCA dates from the individual gene analyses were consistent with those for the full-length sequences and for the alignments containing all 72 CRF50_A1D sequences. The divergence date of the URF specimen was consistently later than the divergence date of the pure CRF50_A1D sequences, confirming that this strain was the product of an onward recombination event between a CRF50_A1D strain and a subtype B strain.

Gene	A1		D	
	Pure CRF50	With 34567	Pure CRF50	With 34567
gag	2000.76 (1998.84 – 2004.86)	1996.05 (1987.21-2010.49)	1997.83 (1991.69 – 209.60)	1994.82 (1986.43 – 2005.27)
pol	2001.33 (2000.22 – 2003.88)	1998.83 (1994.46 – 2006.72)	2001.16 (1999.73 – 2004.64)	2000.07 (1997.73 – 2003.10)
env	1999.64 (1996.47 – 2005.68)	1999.79 (1995.59 – 2009.53)	2000.76 (1999.63 – 2003.31)	2001.57 (2000.94 – 2003.91)

Table 6_10. CRF50_A1D tMRCA dates. Time to most recent common ancestor for each subtype A1 and subtype D region of the CRF50_A1D genome that was analysed independently. Each separate region was analysed both with and without the URF specimen (34567) in order to ascertain whether the URF was precursor of CRF50_A1D or the product of an onward recombination event with a subtype B strain.

6.3.3 Phylogeographic analysis

The phylogeographic analysis was performed using all 72 UK sequences identified as CRF50_A1D infections i.e. five sequences that were obtained from full-length sequencing and the additional 67 sequences identified in the UK HIV DRD. Only *pol* sequences were used for this analysis. The phylogeographic analysis investigated the spread of CRF50_A1D throughout the UK following emergence in 1992 and was captured using video 6_1. Briefly, the video showed emergence of CRF50_A1D in Northwest England in 1992, followed by travel to London/Southeast England in approximately 1995; in 1998 independent spread in London/Southeast England and Northwest England was seen, followed by travel from Northwest England to Southwest England in approximately 2000. Also in 2000 was travel from Northwest England back into London/Southeast England and more accelerated spread within London/Southeast England followed by spread from London/Southeast England back into Northwest England. By 2003 there was spread into three areas of London/Southeast England, followed by transmissions from London into a new region of Northwest England, transmission from London and the Northwest to a new area of Southwest England followed by spread from Northwest England to Northeast England and Scotland.

The plotted inferences were coloured by rate; the rate of transmission from 1993 to 1998 remained stable, followed by slightly faster transmissions in 1999-2001, and then much faster transmissions from 2002 onward. The rate of spread was fastest within London and between London and Northwest England; the fastest rate seen was from Northwest England to Scotland in approximately 2007. The faster transmissions seen from 2002 onward correlate with the Bayesian Skyline reconstruction, which showed the number of effective infections increasing exponentially in the period 2001-2002.

The plotted inferences in the video showed CRF50_A1D emerging in Northwest England prior to travelling to London/Southeast England. Subsequent transmissions of the strain and further travel between London/Southeast England and Northwest England occurred before the video showed spread from Northwest England to Southwest England, Northeast England and Scotland. This indicated that the primary epicentre of CRF50_A1D was located in Northwest England, and that transmission of this strain was driven by repeated migration of the strain out of this region, rather than transmission to new regions creating new epicentres for subsequent transmissions to extra geographic regions.

Chapter 7: Influence of CRF50_A1D on disease pathogenesis

7.1 Obtainment of CD4 cell count data from CRF50_A1D patients captured by the UK CHIC database

Anonymised CHIC IDs for 16 CRF50_A1D patients that had UK CHIC identification numbers were transferred to the MRC-CTU for retrieval of CD4 cell count data stored in the UK CHIC database. CD4 cell count data was available for 15/16 patients; 2/15 (13.3%) patients had no pre-ART CD4 cell counts recorded; a further 1/15 (6.7%) had only one pre-ART count available (Table 7_1). The mean time range for pre-ART counts was 1.53 years; the range covered was 0 - 7.48 years. The median number of total CD4 cell counts available was 16, and the median number of post-ART CD4 cell counts was 10. The median number of pre-ART CD4 cell counts available was five and the median time range of pre-ART CD4 cell counts was 0.7 years.

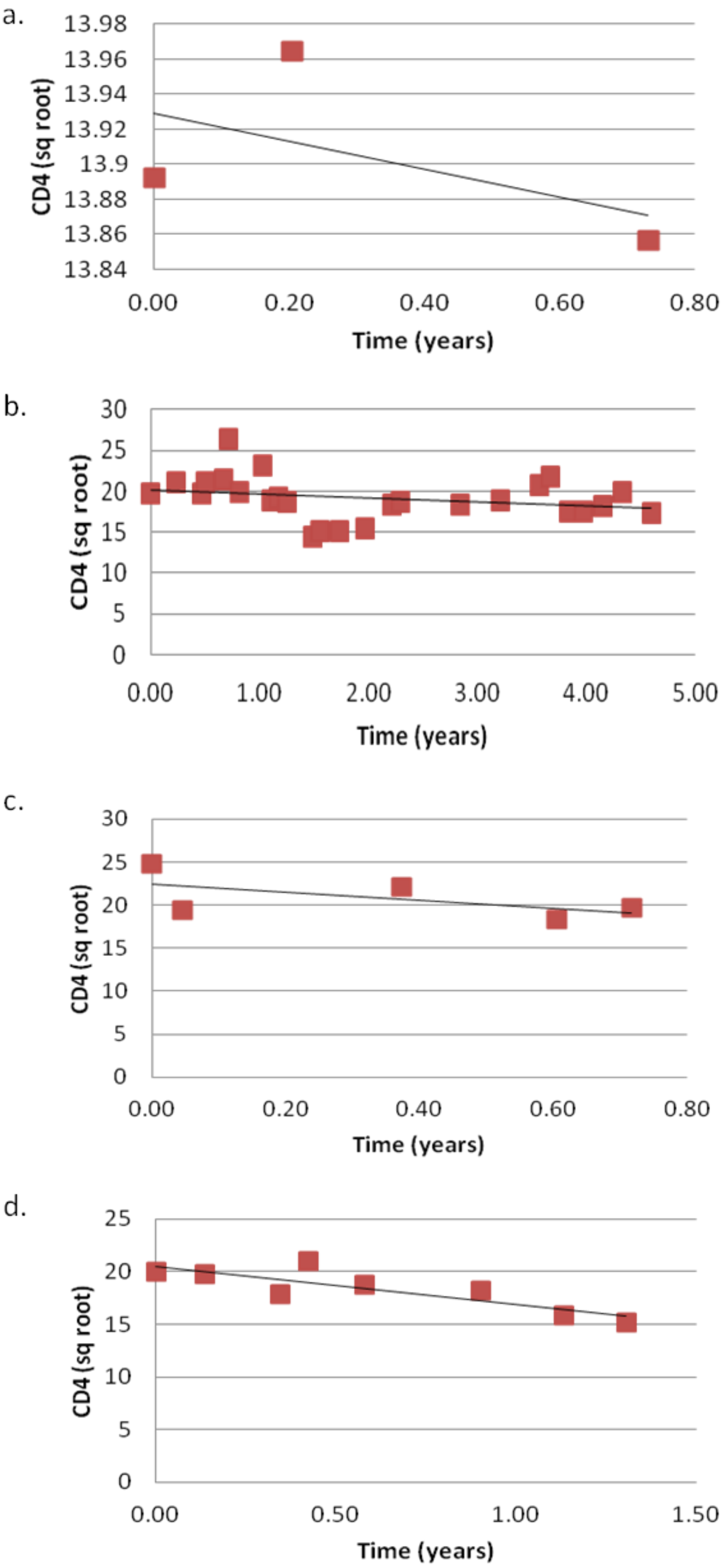
Patients with less than three pre-ART CD4 cell counts and/or less than two months between available pre-ART counts were excluded from further analysis; 6/15 (40%) of patients were excluded. The remaining nine patients were taken forward for further analysis.

Patient number	CD4 cell count measurements			
	Total	Pre-ART	Duration pre-ART (years)	Post-ART
37829	8	0	0	8
22720	12	3	0.7	9
27497	40	26	4.6	14
15732	34	5	0.7	29
5033	32	8	1.3	24
20545	16	5	0.7	11
27082	25	23	7.5	2
15846	17	7	2.5	10
27797	7	1	0	6
31	17	3	0.1	14
29820	16	0	0	16
20893	22	6	0.6	16
33119	2	2	0.7	0
40107	12	2	0	10
25529	5	5	3.5	0

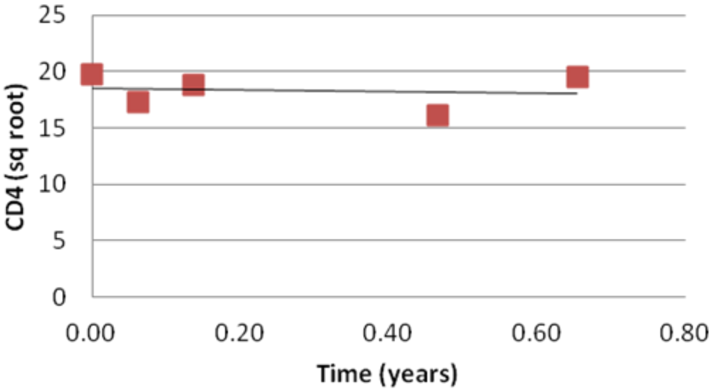
Table 7_1. CD4 cell counts for CRF50_A1D patients. Anonymised CD4 cell count data were obtained from the UK CHIC database. CD4 cell count data were available for 15 patients. The length of pre-ART follow-up ranged from 0 to 7.5.

7.2 Linear regression analysis of CD4 slopes

Figure 7_1 shows the linear regression slopes for the nine remaining patients. There was no uniform pattern apparent across the patients. The regression slope for each of the nine patients is shown in Table 7_2. The median slope was -0.52. When the data was restricted to those patients for whom the time range of CD4 cell counts was greater than one year, the median slope was also -0.52.

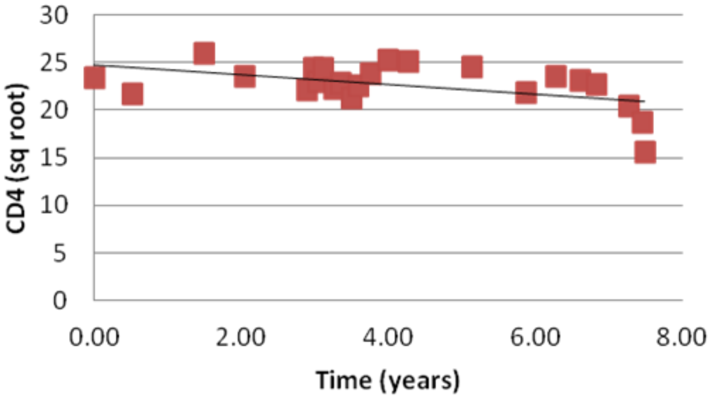

a.
13.98
13.96
13.94
13.92
13.9
13.88
13.86
13.84
CD4 (sq root)
0.00
0.20
0.40
0.60
0.80
Time (years)
b.
30
25
20
15
10
5
0
CD4 (sq root)
0.00
1.00
2.00
3.00
4.00
5.00
Time (years)
c.
30
25
20
15
10
5
0
CD4 (sq root)
0.00
0.20
0.40
0.60
0.80
Time (years)
d.
25
20
15
10
5
0
CD4 (sq root)
0.00
0.50
1.00
1.50
Time (years)

e.



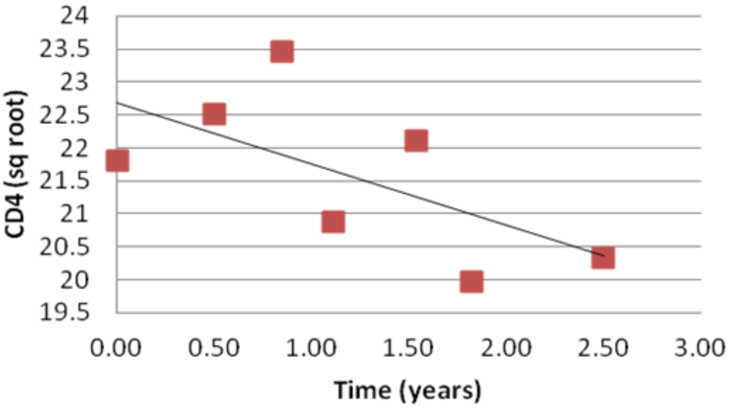
CD4 (sq root)
25
20
15
10
5
0
0.00
0.20
0.40
0.60
0.80
Time (years)

f.



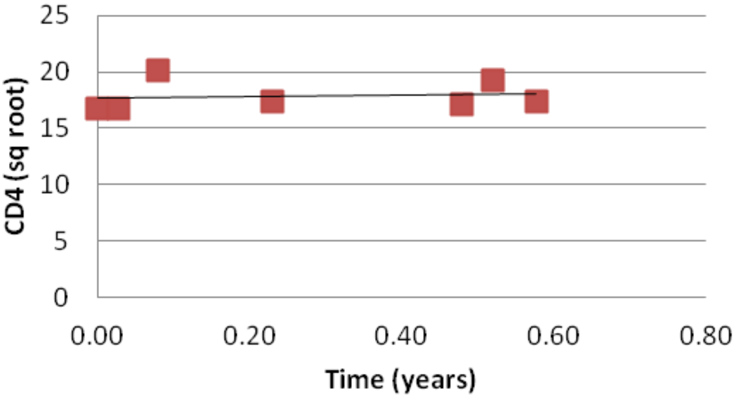
CD4 (sq root)
30
25
20
15
10
5
0
0.00
2.00
4.00
6.00
8.00
Time (years)

g.



CD4 (sq root)
24
23.5
23
22.5
22
21.5
21
20.5
20
19.5
0.00
0.50
1.00
1.50
2.00
2.50
3.00
Time (years)

h.



CD4 (sq root)
25
20
15
10
5
0
0.00
0.20
0.40
0.60
0.80
Time (years)

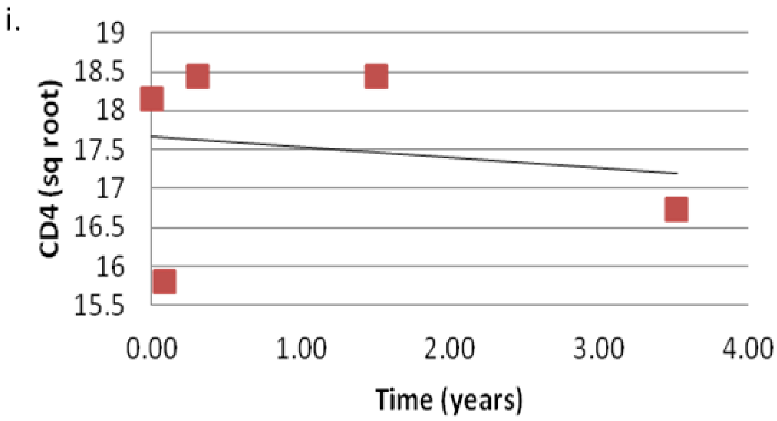


Figure 7_1. Pre-ART CD4 cell count slopes for 9 CRF50_A1D patients. Linear regression slopes a) Patient 22720. b) Patient 27497. c) Patient 15732. d) Patient 5033. e) Patient 20545. f) Patient 27082 g) Patient 15846 h) Patient 20893 i) Patient 25529

CHIC ID	CD4 cell count slope	r^2
22720	-0.08	0.30
27497	-0.49	0.07
15732	-4.76	0.35
5033	-3.66	0.72
20545	-0.64	0.01
27082	-0.51	0.24
15846	-0.93	0.39
20893	0.68	0.01
25529	-0.14	0.02

Table 7_2. Linear regression slopes for CRF50_A1D patients. Pre-ART CD4 cell slope results for each of the 9 patients eligible for analysis.

7.3 Comparison to subtype B slopes

Linear regression slopes for patients infected with subtype B strains of HIV-1 were obtained from Dr. Marina Klein at McGill University Health Centre, Montreal. The mean slope for subtype B patients screened to exclude early and late presenters was -1.22 (95% CI -1.26, -1.18); the slope for unselected patients was -1.33 (95% CI -1.37, -1.29). A CD4 cell decline slope was also available for subtype A1, which was -0.61 (95% CI -0.79, -0.43) (Klein et al., 2010, CROI 2011 abstract B-131). Slopes for subtype D were not available.

The mean CD4 cell decline slope for the nine CRF50_A1D patients was -1.17 (Range -0.08, -4.76; 95% CI -2.29, -0.05). When the slopes were restricted to those patients for whom the time range of CD4 cell counts was greater than one year, the mean CD4 decline was -1.15 (Range -0.49, -3.66; 95% CI not calculated due to low number of measurements available). When compared to the slopes for subtype B and subtype A1 infections, the CRF50_A1D patients appear to experience CD4 cell decline in a manner more similar to subtype B than subtype A1 infections.

Chapter 8: Discussion

Traditionally, HIV-1 infections in the UK displayed a pattern of parallel epidemics: a subtype B epidemic, which affected almost exclusively MSM and IVDU (predominantly acquired in-country), and a heterosexual epidemic, which showed a diverse range of non-B subtypes and was dominated by infections acquired whilst abroad (Geretti, 2006; HPA, 2012). However, these infection dynamics have changed in recent years, and the number of indigenously acquired infections currently exceeds the number of imported infections (HPA, 2012). This, coupled with new increases in rates of infection in both MSM and heterosexuals following years of decline, raised the question of whether these new infection dynamics would lead to the emergence of new population networks that had the potential to change the genetic landscape of HIV-1 infections in the UK.

This study was therefore designed to investigate the emergence of such networks, using subtype- unclassified strains of HIV-1 as a proxy for newly emerging population dynamics. We sought to answer the following questions: is the composition of the UK HIV-1 landscape changing on a genetic level? Are there novel HIV-1 CRFs circulating in Britain, and, if so, what are the origin and likely pathogenesis of these novel strains?

The study was performed using sequences collected from routine HIV care and captured by the UK HIV DRD and sequences generated using an in house protocol for full length, single genome analysis of HIV-1.

1. **Is the composition of the UK HIV-1 landscape changing on a genetic level?**

Constant intermixing of HIV-1 strains within and between regions and populations has resulted in an epidemic that is continually increasing in genetic complexity (Geretti, 2006; Vidal et al., 2009). This presents an unending challenge to existing phylogenetic classification systems, which is reflected in the number of sequences that are considered “untypeable” by these programs (Gifford et al., 2007). Accordingly, the method of genotyping chosen in this investigation was crucial, given variation between different methods and the unclassified nature of the sequences under investigation.

Genotyping complex recombinant sequences is necessarily a difficult task; different genotyping methods use different algorithms to identify the component genotypes of recombinant sequences. The UK HIV DRB was customarily genotyped using two methods: Rega and SCUEAL. Rega has high specificity but low sensitivity for recombination detection, and therefore was not considered a suitable screening method for this investigation (Abecasis et al., 2013). SCUEAL was designed to detect subtype recombination and evolutionary relationships, and was chosen as the primary screening method (Kosakovsky Pond et al., 2009). Owing to the lack of recombinant breakpoint locations in the data held by the UK HIV DRD for unclassified sequences, repeat genotyping of sequences of interest using SCUEAL was performed (referred to as SCUEAL 2012 in text). As breakpoints identified by SCUEAL were considered invitations for further analysis rather than confirmed breakpoints (Pond et al., 2006), the jpHMM program was also used to genotype unclassified strains and predict recombinant breakpoint locations.

The results from each iteration of SCUEAL (2010 and 2012) and jpHMM were very different. Greater agreement was seen between the SCUEAL 2012 results and the jpHMM results than the SCUEAL 2010 and SCUEAL 2012 results or the SCUEAL 2010 and jpHMM results. Differences between the SCUEAL 2010 and SCUEAL 2012 results were partially explained by a change in the background sequences comprising the reference alignment for the program, which resulted in the reference alignment more accurately reflecting genetic variation within subtypes (conversation with S. Frost, 2013; Gifford et al., 2006). The differences between the SCUEAL 2012 results and the jpHMM results could be investigated in more depth by running each program locally using the same background reference alignment; this approach will be considered for further investigations of this nature.

By stratifying the number of subtype unassigned sequences by year, and comparing them to the number of pure and recognised CRF strains, we found a statistically significant increase in the number of unassigned sequences between the years of 1997-2008 (0-1140, respectively, p=<0.001) and the total number of recombinant records (0-1,377, respectively, p=<0.001). During this period the proportion of unassigned sequences increased from 0% in 1996 to 12.6% in 2008 and the proportion of CRFs increased from 0% in 1996 to 2.6% in 2008, showing that the unassigned sequences increased at a faster rate than the recognised CRF sequences.

There were two caveats to these numbers. Firstly, SCUEAL rarely classified CRF02 sequences as such, meaning that the number of unclassified sequences might be artificially inflated. However, the proportion of CRF02 infections in the UK is comparatively small, and therefore the numbers of these infections was not considered to substantially affect the overall figures. Second, the overall numbers were calculated using the SCUEAL simplified subtype field. This is a more conservative classification than the detailed subtype field, and as such, the numbers of recombinant strains identified were likely to be an under- rather than over-estimate.

Given evidence that the proportion of unassigned strains in the UK was increasing, the transmission networks within these strains was investigated. Subtype B-containing recombinants were selected for this due to their traditional constraint by exposure group. Using the detailed subtype field in SCUEAL 2010, 2,030 sequences were identified as potential B-recombinant strains. Although it became apparent upon genotyping using SCUEAL and jpHMM that only 444 - 699 of these sequences were true B-recombinants, the results from the two genotyping methods were sufficiently different such that the entire 2,030 putative B-recombinant sequences were used as input into an approximate maximum likelihood tree. As genetic interpatient variability is less when infection comes from a common source, the maximum likelihood tree was screened for clusters using genetic distance measurements; this identified 28 potential transmission clusters (Simmonds et al., 1990). As the presence of recombination can affect the accuracy of phylogeny estimation (Posada and Crandall, 2002), confirmation of transmission clusters was performed by analysing each putative pure region separately, in order to confirm clustering, before concatenating sequences and performing the time scaled phylogenies.

The BEAST cluster analysis found 15/28 (53.6%) confirmed clusters. Of the 15 clusters, the smallest clusters contained 3 members (4 clusters) and the largest contained 23 members (1 cluster). 5/15 (33.3%) were mixed male and female, of which one was exclusively heterosexual black-African with six members, one was heterosexual white with four members, one was IVDU with three members, one was mixed IVDU and heterosexuals with five members and one was mixed IVDU, heterosexual and MSM with 23 members (all white); 20% of clusters overall showed mixed exposure profiles.

There was one all male cluster with mixed heterosexual/MSM with five members, and one all male MSM/IVDU cluster with three members. Eight clusters were exclusively MSM, of these five were all white, one was white/black-African/black-

Caribbean with six members and two were mixed white black-African with 3 and 5 members respectively.

In terms of geographic location, seven clusters were located exclusively in London/Southeast England, two in East Anglia, one in London/Southeast England and Northwest England, one in Northwest England, one in Scotland/Northwest England, and three in Southwest England/Wales. Overall, 75% of clusters showing mixed ethnicities were located in London/Southeast England and the only cluster involving exclusively black Africans was located in London/Southeast England.

Time-scaled phylogenies were performed on clusters showing potential extra-UK geographic involvement, including the largest cluster (Cluster 1) which contained 23 members. Geographic matching identified a number of sequences from Portugal and Spain displaying identical recombination profiles; the more recent tMRCA of the UK cluster relative to these sequences indicates possible import of this strain during the late 1990s, consistent with accepted geographic distribution of G/B recombinants, which have traditionally been found in Spain and Cuba (Thomson et al., 2012). Interestingly, subtype G in Portugal has been associated with indigenous transmission in heterosexual and IVDU populations (Abecasis et al., 2013); this pattern was continued in 1/3 G/B recombinant clusters identified in this study (Cluster 2, four heterosexual members), but MSM involvement was present in 2/3 clusters (Cluster 1, Cluster 5, four MSM members). The high number of G/B recombinant clusters found, and the links to Portugal, support data from Gifford et al., who also identified UK-Portugal transmission networks (2007).

Cluster 15, which also contained mixed exposure profiles, showed an identical recombinant profile to sequences previously identified in Spain, Paraguay and Brazil. Evolutionary analyses showed no linkage to the Brazilian sequences, but co-circulation of this novel strain in three distinct geographic regions (UK, Spain and Paraguay). Co-circulation in multiple countries was also found in Cluster 2, a B/G recombinant circulating exclusively in heterosexuals in the UK. Co-circulation of this strain was identified in the UK, Portugal and Luxembourg.

Evidence of three further novel circulating strains formed from onward recombination of CRF50 strains with subtype B strains was found. The tMRCAs of two of these strains indicated that onward recombination with subtype B infections took place relatively quickly following the emergence of CRF50_A1D in the UK (1997.60 and 1999.55 vs. 1992.46). Evolutionary analysis that included Canadian strains showing nearly identical recombination profiles showed a closer relationship with the pure CRF50 sequences, indicating that CRF50 may have given rise to at three further CRFs and been exported from the UK to North America during the period 1992-2007.

The final cluster exhibiting extra-UK linkages was Cluster 19, which showed a B/C recombinant circulating in White and Black-African MSM in the UK. The evolutionary analyses showed a tMRCA during the early 2000s, followed by probable export from the UK to Italy in approximately 2008. Non-B subtypes in Italy have been estimated to comprise 2.4-19.4% of all HIV-1 infections, but to date this has been dominated by B/F genetic forms from introduction events associated with South America (Ciccozzi et al., 2012; Lai et al., 2012). This is the first report of novel strains from the UK being imported into Italy.

Crucially, of the eight clusters with time scaled phylogenies, 5/8 (62.5%) show tMRCAs in 2004-2005, a period in which there was a statistically significant year-on-year increase in the number of unassigned sequences in the UK HIV DRD. Cluster 1, which had a tMRCA of 1999.72, showed short branch lengths around 2004-2006, indicating that the greatest number of transmissions took place during this time. This confirms that periods of increasing genetic diversity are the periods in which novel circulating strains are generated.

Overall, 66.7% of the UK clusters contained sequences from MSM. Given that this analysis focussed on B-containing recombinants, the high proportion of MSM sequences in not surprising; what is surprising is the 33.3% of clusters that were composed of B-containing recombinants that did not contain sequences from MSM. This shows parallel diversification of the UK epidemic whereby the non-MSM epidemic has been incorporating B-strains at the same time as the MSM epidemic has been incorporating strains more commonly associated with heterosexual infections.

Significantly higher proportions of males, IVDUs and people of white ethnicity were found to be associated with clusters of novel recombinants when compared to the proportion of individuals in the UK HIV DRD infected with pure subtypes and recognised CRFs. These figures indicate that emerging transmission of HIV-1 in the UK is being driven by new transmission patterns and is an area that would benefit from further study.

Finally, 33.3% of clusters were found to be imported into the UK or have evidence of co-circulation in other countries; however, 60% showed no geographic linkages outside of the UK, confirming that as the number of indigenous transmissions within the UK increases, so too does the number of new CRFs generated in-country.

2. **Are there novel HIV-1 CRFs circulating in the UK?**

The main limitation of the cluster study was that in any study using only one genomic region the presence or absence of recombination outside of the sequenced region remains unknown (Abecasis et al., 2013; Sanabani et al., 2009). Confirmation of a new circulating recombinant form of HIV-1 required characterisation of substantial regions of the genome (Robertson et al., 2000). Following the analysis that identified a potential novel A1/D CRF in the UK, full-length HIV-1 sequencing was performed on six specimens that were retrieved from routine storage from three centres. Although identification of novel CRFs can be performed using viral DNA from PBMC samples, plasma sampling was selected in this case in order to characterise the circulating, rather than the archived, virus.

To develop the protocol used for the full length sequencing, a protocol from Nadai et al. was selected for adaptation. Ultimately, adaptation of this protocol was not successful, and a protocol was developed using primers from CHAVI-MSBNC. The final protocol was optimised to work with subtypes A, B, C and D and specimens with low viral loads or reduced sample volume, unlike the original protocol which was optimised for use with subtypes B and C only. A number of modifications were trialled and adopted, including an increase in the number of cycles during the first round of PCR from 35 to 40; an optimised annealing temperature of 60ºC, and a primer concentration of 0.25µm.

Due to the age of the specimens, the likelihood of poor quality viral RNA was substantial. Automated nucleic acid extractions were trialled; however, these resulted in PCR template fragments of incorrect lengths. It is likely that the automated extraction process sheared the RNA, resulting in mis-priming and shorter fragments than desired. An RT enzyme with engineered reduced RNase H activity was selected for use to counteract an increase in *in vitro* errors due to RT strand switching during minus strand DNA synthesis, which is especially likely if the original RNA is of poor quality (Jansen and Ledley, 1990; Marton et al., 1991; Meyerhans et al., 1990; Shriner et al., 2004).

In order to detect the existence of *Taq* polymerase-incorporated errors in the amplified specimens, two specimens were sequenced at more than one cDNA dilution beyond that required by Poisson's distribution. One specimen was sequenced at the theoretical dilution of 0.5 copies/µl (1 copy/reaction) given circumstances of no viral RNA degradation during 7 years of storage, and perfect extraction and RT results. The specimen amplified at a rate of 6% positive reactions, which is below the 30% required for Poisson's distribution, but still sufficient to achieve full genome sequencing. Only one specimen amplified successfully at this dilution, but a combination of low volume and low viral load made this testing impractical for the other recombinant specimens. No differences were seen between the sequences obtained at any dilution below 30% positive, demonstrating the success of the optimised protocol.

The protocol was also optimised to work with samples of low viral load and volume. Of the study samples received, the range of viral loads was 9,148-500,000 copies/ml and the range of samples volumes was 270-1500ul. Conversely, the sample with the highest viral load was also the sample with the most volume available. Although the protocol called for 20,000 copies of viral RNA to be extracted, this was only possible for 3/6 specimens, owing to the low sample volumes available. The number of viral RNA copies extracted per specimen ranged from 548 – 20,000.

PCR amplification of the specimens showed relatively consistent amplification across all six specimens. One specimen amplified when using a cDNA dilution such that only one copy of cDNA was present in each well; a further specimen amplified at 1.37 copies per well. Each of the remaining four specimens were able to be sequenced at dilutions lower than that producing the Poisson distribution result, demonstrating the robustness of the optimised protocol with these specimens. The highest input into a reaction was for specimen 11762, which had a theoretical input of 200 copies/reaction well. However, this is based on the viral load at the time of sample, not a measurement of the sample once received; as this sample did not amplify at all at any dilution, and only amplified with 19/90 (21.1%) positive wells when tested undiluted, we have concluded that the specimen was degraded during storage and/or transport. Overall, given that sample ages ranged from 10 years old (specimen 8179), to seven years old (specimens 33365, 40534, 34567) to two years old (specimen 12792) to one year old (specimen 11762), the procedure was remarkably successful at amplification. That the most recent specimen was used at the lowest dilution of cDNA adds weight to the argument that this specimen in particular was somehow degraded.

Following sequencing, the recombination analysis of the six full-length genomes showed identical putative A1/D structures and one genome that was an A1/D/B/U recombinant. The five identical structures showed a genome containing eight breakpoints, three in the *gag* gene, one in the *pol* gene, one in the accessory genes and a further three in the *env* gene.

There were regions of the putative pure subtype components that initially defied easy subtype classification. This was due to three main factors. The first was the low number of genomes and the high similarity between them. This resulted in low phylogenetic signal in some of the fragments, which was shown in the likelihood mapping results of three of the specimens (33365, 40543 and 8179). This was particularly evident in fragments three (HXB2 co-ordinates 1884-2100, fragment length 216 nucleotides, *gag* p2, p7) and four (HXB2 co-ordinates 2101 -2503, fragment length 402 nucleotides, *gag* p1, p6, *pol* protease amino acids 1-84). Whilst the short fragment length of 216 nucleotides was the most probable explanation for the low phylogenetic signal for fragment three, the low signal for fragment 4 was explained by something different, namely the remarkable similarity of subtypes B and D in this region. The entropy analysis of this region showed that the number of informative sites between the two subtypes was only six, thus explaining the poor likelihood mapping results.

Maximum likelihood analysis was used to genotype the putative pure subtype regions in the six genomes. Following this analysis, the genome was confirmed as an A1/D recombinant where the genome starts as subtype A1 through the 5' LTR and *gag* p17, and part of p24 and then has breakpoints at HXB2 1272 (within p24) , 1851 (end of p24, p2, p7), 2100 (*gag* p1, p6, *pol* protease amino acids 1-81), 2489 (*pol* RT, RNase, Integrase, *vif, vpr,* beginning of *tat/rev*), 6004 (*tat/rev, vpu, env* CDS, gp120), 6602 (*env* gp120, V1- V4 loops), 7440 (*env* gp120, V5 loop, gp41, *tat/rev* exon), 8573 (*tat/rev* exon, gp41, *nef,* 3'LTR).

Limited selective advantage is provided by the majority of HIV-1 recombination breakpoint locations, with the exception of breakpoints falling within regions associated with viral fitness such as the *env* gene (Archer et al., 2008; Palmer et al., 2005; Quiñones-Mateu and Arts, 2002; Tee et al., 2009; Troyer et al., 2009). The tendency of breakpoints in *env* to cluster around the 5' and 3' ends of the gene, has indicated a tendency for either the entire gene, or at least the gp120 coding region, to be preserved intact (Archer et al., 2008; Simon-Loriere et al., 2009). The *env* breakpoints in the identified A1/D recombinant structure fell within the recombination hot-spots for *env* identified by Simon-Loriere et al., in 2009, in regions 1 and 3, respectively. As the recombination hot zone 3 has been associated with a high probability of loss of functionality (Simon-Loriere et al., 2009), further research on this particular recombinant structure could be performed, in order to ascertain whether a breakpoint location within this narrow zone constrains onwards viral evolution and affects the phenotypic properties of the viral strain. Such studies could consider the position of the recombination breakpoint with regards to the boundaries of protein folds, as this can influence functionality; another aspect to consider is the degree of similarity between the parental subtype A1 and D strains around each breakpoint location, as the degree of local sequence similarity has been found to be crucial in determining the frequency and pattern of genetic recombination (Archer et al., 2008; Hu and Temin, 1990; Simon-Loriere et al., 2009, 2010; Zhang et al., 2010).

The final specimen sequenced was a URF specimen with a CRF50_A1D/B structure, where the subtype B regions were *gag* p2, p7 and *pol* RNase, Integrase, *vif, vpr,* and the beginning of the *tat/rev* intron. Given that data have indicated that recombination between genetically distinct HIV strains may not be an immediate or common outcome to dual infection *in vivo* and suggest critical roles for viral and host factors such as viral fitness, viral diversity and host immune responses (Powell et al., 2010), and that the URF specimen preserved the *env* breakpoints seen in CRF50_A1D, it is possible that breakpoints at these positions possess fitness determinants that favoured the preservation of the CRF50_A1D structure over a structure incorporating subtype B in *env*.

There were limitations involved in performing full-length sequencing in the manner undertaken in this study. Amplification of HIV-1 in a single ~9kb fragment meant that shared nucleotide polymorphisms and mixed bases can be detected more frequently than when methods involving shorter fragments are used (Salazar-Gonzalez et al., 2009). To control for this, additional safeguards were added to the sequencing protocol to ensure that the PCR template fragments were not subject to PCR-mediated recombination. These were: to sequence each specimen below the point of Poisson’s distribution and to pool the DNA from the positive PCR wells before sequencing.

Pooling product from positive PCR wells was performed because the 50µl product from a single positive well was not sufficient to sequence the entire genome, even after dilution. Pooling of positive wells prior to sequencing therefore allowed the detection of potential chimeric products across the length of the entire genome. Any partial sequence that looked as if it came from a mixed source i.e. showed evidence of ambiguous bases that could not be explained by poor sequencing signal were repeated. In addition, proof-reading enzymes and high fidelity PCR kits were used throughout to reduce the chance of PCR-mediated recombination.

Instead of performing full-length sequencing using a single fragment, there was the option of developing a protocol that involved amplification of multiple overlapping fragments. However, we did not choose to take this approach. Although the sensitivity of the PCR would have been increased, other drawbacks made this an unattractive option. The drawbacks included the unknown structure of the recombinant, and therefore the risk that breakpoints fell within or close to the end of the overlapping fragments, clouding the downstream analyses.

Following the phylogenetic confirmation of the five A1/D structures, further instances of this strain were sought by performing global BLAST searches in both NCBI and the Los Alamos National HIV Database. When no matches for this structure were found, the structure was registered as CRF50_A1D.

Additional interrogation of the UK HIV DRD using three of the CRF50_A1D sequences as reference sequences found an additional 67 instances of this recombinant in the UK, leading to a total of 72 confirmed cases. Investigation of demographic information (where available) showed 46 cases in white MSM and transmission to two IVDU and two heterosexuals. Ethnicity results showed transmission to one black-African MSM and to one white heterosexual male; the IVDU's were white; there was one Indian-Pakistani MSM. The range of country of origins showed cases from patients originating from UK, Ireland, Norway, the Philippines, and the US.

All 72 sequences had associated aggregated centre data to indicate geographic origin; this showed that most cases were located in Northwest England (39/72, 54.2%) and London/Southeast England (27/72, 37.5%). There were three cases in Southwest England, one in Northeast England and two in the Edinburgh region of Scotland.

3. What are the origin and likely pathogenesis of these novel strains?

The investigation into the geographic origin of the component subtype A1 and D parental strains showed that the strains were most closely related to strains from

East Africa. The analysis was unable to determine a specific country of origin with the East African region; further studies that might address this could include reconstruction of the ancestral subtype A1 and D sequences in order to more precisely determine the geographic origin of the component CRF50_A1D subtypes. Epidemic origins can be difficult to pinpoint if the source is geographically or temporally remote (Pybus and Rambaut, 2009); more extensive sampling of these strains, in the right genetic region, would also assist in arriving at a more precise determination. Speculation regarding likely candidate countries would involve Kenya, as the predominant circulating subtype is A1, but a high recombination rate is evident and the country borders others where other subtypes are prominent (Dowling et al., 2002; Hué et al., 2012; Land et al., 2008).

The analysis of the *pol* gene global trees showed that the CRF50_A1D sequences clustered closet to the East African strains even in the presence of subtype A1 and D sequences from the UK, demonstrating that CRF50_A1D was most likely imported into the UK as a fully-formed recombinant, rather than arising from recombination events in the UK itself.

Phylogeographic analysis of the emergence and spread of CRF50_A1D showed emergence in Northwest England, followed by subsequent spread to London and Southeast England. Northwest England remained the epicentre of the strain's spread, as spread to Southeast England, Northeast England and Scotland all occurred from this region. The analysis showed that the spread of CRF50_A1D was fastest within London/Southeast England. It is important to note that the likelihood of every person in the UK infected with CRF50 being included in the analysis is extremely low, meaning that the time between transmissions has almost certainly been overestimated (Lewis et al., 2008). Similar patterns of spread of CRFs have been reported in China, particularly the exchange of CRF01_AE circulating in MSM in the geographically close regions of Liaoning and Beijing (An et al., 2012). The phylogeography was modelled using discrete transitions; using continuous diffusions could allow a more nuanced description of the geographic emergence of CRF50_A1D (Lemey et al., 2010).

Time scaled phylogenies were performed in order to determine the time to most recent common ancestor of CRF50_A1D in the UK. Multiple methods were used in order to ensure that the date arrived at was robust. The five full-length sequences were split into component subtype A1 and D regions in the *gag, pol,* and *env* genes, with subsequent analysis of each component; the full-length sequences were analysed as a whole, and the *pol* region for all 72 UK CRF50_A1D cases was analysed as both component subtype A1 and D regions and as a concatenated A1/D region.

Results for all 72 CRF50_A1D cases showed a tMRCA of 1992.46 (95%HPD 1987.47-1998.57), which was consistent with the tMRCAs for the component A1 and D regions of 1994.88 and 1994.36, respectively. Low topological support was seen in parts of the subtype D tree, but this was due to the region being highly conserved, as previously noted, and also the short length of the fragment under analysis.

In general, the results for the individual gene subtype A1 and D regions showed low posterior probabilities. However, the results for the tMRCA across the regions were consistent overall, with a standard deviation across all genes of 2.15. The tMRCA when using the full-length sequences was 1995.56 (95% HPD 1990.91-1999.60), which was also consistent with the other dates. Conclusions drawn from small and local samples will underestimate dynamic complexity (Pybus and Rambaut, 2009). This explains why later dates of divergence are seen when using the genetic regions of the full-length sequences only. Therefore, we are confident that the predicted tMRCA for CRF50_A1D is robust.

Gifford et al. predicted that a novel subtype A recombinant was circulating in MSM in the UK, with an origin date of the late 1980s/early 1990s (2007). Given the results of this analysis, we are confident that CRF50_A1D is the predicted novel recombinant. The period of origin was the period during which the HIV epidemic in Africa was growing particularly quickly, and is, crucially, prior to the introduction of HAART in the UK (Gifford at al., 2007). This period also saw a doubling of prevalence of HIV-1 in MSM in London, along with multiple transmissions of clusters (Lewis et al., 2008).

The final sample characterised using full-length sequencing proved to be a B/CRF50_A1D URF. This recombinant structure was substantially more complex than the A1D recombinant. However, the complexity was evident outside the region of *pol* that is routinely sequenced for the evaluation of drug resistance. This further indicates that the proportion of recombinant specimens circulating in the UK may be much higher than previously suspected, and also highlights the inherent difficulties in resolving recombinant HIV-1 sequences to parental strains.

The time-scaled analyses including the URF showed that the URF was the product of onward recombination between a CRF50_A1D infection and a subtype B strain. The index CRF50_A1D case for this recombination was not identified during this study, indicating that additional instances of this strain exist in the UK. This was supported by the Bayesian skyline analysis, which predicted a number of effective infections in the low hundreds. This, in addition to the evidence for CRF50_A1D creating further novel CRFs and the possible export to Canada of CRF50_A1D that was discovered during the cluster analysis, demonstrates that newly recombinant strains can rapidly become established and significantly change the genetic landscape of HIV-1 infections.

It has been suggested that the ever-growing number of CRFs will be become blurred and unmanageable to the HIV research community, and that a nomenclature system that recognises recombination 'families' would more accurately reflect the dynamic reality of HIV-1 evolution (Zhang et al., 2010). The data uncovered during the course of this study supports this suggestion. This study was conducted using the current CRF nomenclature, which requires all sequences to have near-identical breakpoints. However, there were approximately 200 A1/D recombinant sequences in the UK HIV DRD that were very similar to CRF50_A1D, but did not meet the strict nomenclature criteria. As A1/D recombinants have been described in Europe and the UK for nearly two decades (Leitner et al., 1995), it seems probable that ongoing evolution has created a family of similar A1/D recombinants and that full elucidation of the various lineages present would be better understood following an investigation concentrating on the entirety of A1/D recombinants rather than a single group that meets the current CRF nomenclature criteria.

Thorough analysis of the process by which new CRFs become established in communities cannot be considered complete without an investigation into the phenotypic properties of the recombinant virus, as mentioned above. The process of recombination may disrupt linked compensatory point mutations that have evolved to preserve viral fitness (Drummond et al., 2002; Simon-Loriere et al., 2010; Voigt et al., 2002), meaning that recombinants may be more, or less, fit than the sum of fitness of the parental strains. The final aspect of this study was a preliminary investigation into the likely pathogenesis of CRF50_A1D infections. Pre-ART CD4 cell counts were obtained for 9 patients and linear regression of the CD4 decline slopes showed that the mean CD4 cell decline was -1.17 sq root cells/year (95% CI -2.29 - -0.05). When compared to the decline slopes for subtype B (-1.22, 95% CI -

1.26 - -1.18) and subtype A (-0.61, 95% -0.79, -0.43) infections, it appeared that CRF50_A1D patients experience disease progression in a manner more akin to a subtype B infection than a subtype A infection.

The scarcity of the data available to make this calculation was reflected in the wide 95% CI for the mean CD4 cell decline, which meant that the data must be interpreted with caution. Additionally, no slopes were available for subtype D for comparison. Therefore, although it appears that CRF50_A1D patients experience disease progression in a manner more similar to subtype B than subtype A infections, we were unable to make any comparisons as to whether this is faster or slower than progression with a subtype D infection, although the current data tends toward agreement with that from earlier studies (Baeten et al., 2007; Huang et al., 2007; Kaleebu et al., 2002; Kiwanuka et al., 2010). Of interest was the specific structure of CRF50_A1D in the *env* gene, which was subtype A1 in the hypervariable loop region. This region has specific fitness determinants, and so data indicating CRF50_A1D infections behave more similarly to subtype A1 than subtype D would be expected (Fouchier et al., 1992).

The unusual structure of CRF50_A1D in the *env* gene further indicated that this recombinant structure may possess interesting phenotypic or functional properties associated with specific breakpoint locations. This could be investigated using growth competition assays and other assays designed to measure specific determinants of viral fitness. This work was beyond the scope of the current investigation, but should be considered for the future.

Another aspect for further investigation is whether the structure of CRF50_A1D (particularly in *env*) possesses properties that elicit unique host immune responses, such as particular adaptive mutations driven by early cytotoxic T-cell responses (Salazar-Gonzalez et al., 2009). Again, this work was outside the scope of the current study, but should be considered for the future.

The UK HIV DRD collects nearly all sequences identified in the course of routine HIV care in the UK. The coverage of the database is estimated at approximately 80% of the known HIV infected population in the UK. However, a natural limitation of this type of work is that only the sequences that are available can be used, i.e. there are inevitable gaps in the data from those who are infected and know it but have not been captured by routine care (or have been but have not had sequences submitted to the database), but more importantly, those who are infected but not yet diagnosed, which is still estimated as a third of the UK HIV population (HPA, 2012).

Therefore, work such as mapping the networks of populations or of estimating how many people are infected with a particular strain of HIV are necessarily incomplete. As universal genotyping at diagnosis increases (particularly when closer to the time of infection), the genotypic coverage contained in the database will improve, and some of the caveats that still remain in Bayesian analysis will be resolved (Brenner et al., 2013).

Finally, it is important to note that in performing this type of analysis, corrections for the presence of drug resistance mutations in sequences is necessary, as ART leads to suppression of quasispecies evolution (Redd et al., 2012). Although ART history was not available for all patients in the UK HIV DRD, we controlled for the presence of ART-associated drug resistance mutations by using only the earliest sequence available for each patient. All downstream analyses such as full-length sequencing and the CD4 decline slopes were performed using confirmed pre-ART data and samples.

8.1 Final conclusions

This study reported a comprehensive investigation into emerging population dynamics in HIV-1 infections in the UK, using the molecular epidemiological analysis of subtype-unclassified HIV-1 strains captured by the UK HIV DRD. Investigations of this nature have the potential to have significant impacts on public health interventions to reduce the incidence of HIV-1 transmission. Mathematical modelling of HIV-1 transmission has shown that the resurgence in HIV infections among MSM in the UK and Europe is the result of continuing risky sexual behaviour, and that long term reductions in incidence will require increased and sustained uptake of effective interventions (van Sighem et al., 2012). However, interventions can be effective at either the community or the individual level, and a systematic review of infections in MSM in Western Europe, the US and Australia concluded that the mean incidence rate of 2.5% per year remained stable during 1998-2008, indicating that interventions that are effective at the individual level are not having an effect at the population level (DeGruttola et al., 2010; Stall et al., 2009). Phylogenetic analyses can elucidate the longitudinal phylodynamic structure of a local epidemic and help to identify potential targets for interventions, for example targeted Treatment as Prevention programmes (DeGruttola et al., 2010; Hughes et al., 2009; Lewis et al., 2008; von Wyl et al., 2011). Integrating these analyses with epidemiological, clinical and demographic data will be crucial in delineating the role of linkage to care, behaviour, socioeconomic factors and migration on transmission dynamics (Brenner et al., 2013).

The potential for phylogenetic analyses to improve the impact of public health interventions is particularly relevant in the UK, where the transmission network among UK MSM is characterised by preferential association such that a randomly distributed intervention would not be expected to stop the epidemic (Leigh Brown et al., 2011). Many countries, including the UK, currently perform routine surveillance of HIV-1 prevalence and drug resistance, however phylogenetic analyses have an added value above these existing methods, in that phylogeographic modelling can identify the source of continued transmission to new geographic areas. For example, the phylogeographic analysis of CRF50 showed that, although the greatest number of transmissions were occurring in London and Southeast England, the source of geographic spread to London and Southeast England, Southwest England, Northeast England, and Scotland was Northwest England in each instance, even after transmission to London had occurred. This implies that an effective public health intervention targeting Northwest England could have prevented the spread of CRF50_A1D to other areas of the UK. It is this information that is not captured by existing surveillance programmes, but which would be invaluable in designing effective public health interventions.

The spread of newly recombinant strains has implications beyond that of providing insights into population dynamics. Analysis of the SPREAD programme showed that MSM infected with a subtype B virus were significantly more likely to be infected with drug-resistant HIV-1 (Vercauteren et al., 2009). Increased transmission of newly recombinant strains, especially strains that recombine readily with subtype B, such as CRF50_A1D, could therefore potentially increase the rate of transmitted drug resistance in HIV-1 populations. There is the potential for newly recombinant strains to exhibit differing response profiles to newer HIV-1 drugs in development, or to accumulate drug resistance mutations at a faster rate than other subtypes (Schader et al., 2012; Soares et al., 2007); however, the evidence for this is weak as HIV-1 antiretroviral drugs are designed to work across the entire spectrum of HIV-1 genotypes, and clinical evidence suggests no genotype-based difference in the accumulation of drug resistance mutations.

In addition to the opportunities for further investigations that have been detailed above, there are two obvious further investigations regarding the transmission and progression of CRF50_A1D (or other newly recombinant) infections that could be performed. Firstly, next-generation sequencing to characterise the viral quasispecies in combination with replicative capacity assays could be used to incorporate a population model into an assessment of the fitness of this viral strain (Lauring and Andino, 2010). This approach has the advantage of combining directly observed phenotypic behaviour with factors such as immune escape, transmissibility and cellular tropism (Lauring and Andino, 2010). Quantitative assessment of HIV-1 evolution can be determined by measuring HIV sequence diversity and comparing it to a reference point e.g. the founder virus (Gall et al., 2013). Therefore, longitudinal deep sequencing would also provide a method of tracking the dynamic evolution of the HIV-1 quasispecies and, ultimately, viral progression (Tsibris et al., 2009).

The second investigation is an assessment of whether the burden of transmission of these novel recombinant strains is due to recent or established infections. Acute infections are particularly important in determining the evolutionary course of the HIV-1 epidemic, as the viruses present in acute infection are stored and "preferentially" transmitted (Lythgoe and Fraser, 2012). Until now, the temporal spread of clustered transmissions has been evaluated exclusively with respect to chronology and geography to improve our understanding of transmission dynamics in the UK (Brenner et al., 2013). The next stage to evaluate the data from this study would be to look at the stage of infection in of each sequence, in order to ascertain whether transmissions were occurring primarily between recent or established infections. This could be done using ambiguous nucleotides within sequences as a marker for viral diversity and therefore the age of infection (Kouyos et al., 2009).

This study proposed using newly recombinant strains of HIV-1 as a proxy for the changing landscape of HIV-1 infection in the UK. We identified CRF50_A1D, a subtype A1/D recombinant that emerged in the UK in approximately 1992 from East African origins. The study of the extent of recombination in the UK found one year, 2004-2005, in which a statistically significant increase in subtype-unclassified infections occurred. Further study of transmission clusters within these unassigned sequences found that 5/8 (62.5%) of clusters had a tMRCA in this time period. Additionally, the cluster study identified probable co-circulation of a number of unassigned strains in countries outside of the UK, and export of an unassigned strain from the UK to Italy. This, in combination with the finding that majority of unassigned strains investigated contained mixed ethnicities and exposure routes within Britain, suggests that indigenous HIV-1 transmission in the UK is of increasing importance, and that previous compartmentalised transmission patterns will continue to blur. Continuing detailed characterisation of local epidemics will enable the design of public health interventions suited to responding effectively to these changed dynamics.

Bibliography

Abecasis, A.B., Lemey, P., Vidal, N., Oliveira, T. de, Peeters, M., Camacho, R., Shapiro, B., Rambaut, A., and Vandamme, A.-M. (2007). Recombination Confounds the Early Evolutionary History of Human Immunodeficiency Virus Type 1: Subtype G Is a Circulating Recombinant Form. J. Virol. *81*, 8543–8551.

Abecasis, A.B., Wensing, A.M., Paraskevis, D., Vercauteren, J., Theys, K., Van de Vijver, D.A., Albert, J., Asjö, B., Balotta, C., Beshkov, D., et al. (2013). HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology *10*, 7.

Afonso, J.M., Morgado, M.G., and Bello, G. (2012). Evidence of multiple introductions of HIV-1 subtype C in Angola. Infect. Genet. Evol. *12*, 1458–1465.

Allan, J., Coligan, J., Barin, F., McLane, M., Sodroski, J., Rosen, C., Haseltine, W., Lee, T., and Essex, M. (1985). Major glycoprotein antigens that induce antibodies in AIDS patients are encoded by HTLV-III. Science *228*, 1091–1094.

Allen, T.M., O'Connor, D.H., Jing, P., Dzuris, J.L., Mothe, B.R., Vogel, T.U., Dunphy, E., Liebl, M.E., Emerson, C., Wilson, N., Kunstman, K.J., Wang, X., Allison, D.B., Hughes, A.L., Desrosiers, R.C., Altman, J.D., Wolinsky, S.M., Sette, A., and Watkins, D.I. (2000). *Tat* specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia. Nature, *407*, 386-390.

An, M., Han, X., Xu, J., Chu, Z., Jia, M., Wu, H., Lu, L., Takebe, Y., and Shang, H. (2012). Reconstituting the epidemic history of HIV strain CRF01_AE among men who have sex with men (MSM) in Liaoning, northeastern China: implications for the expanding epidemic among MSM in China. J. Virol. *86*, 12402–12406.

Archer, J., Pinney, J.W., Fan, J., Simon-Loriere, E., Arts, E.J., Negroni, M., and Robertson, D.L. (2008). Identifying the Important HIV-1 Recombination Breakpoints. PLoS Comput Biol *4*, e1000178.

Arnold, C., Barlow, K.L., Parry, J.V., and Clewley, J.P. (1995). At Least Five HIV-1 Sequence Subtypes (A, B, C, D, A/E) Occur in England. AIDS Res. Hum. Retroviruses *11*, 427–429.

Artenstein, A.W., VanCott, T.C., Mascola, J.R., Carr, J.K., Hegerich, P.A., Gaywee, J., Sanders-Buell, E., Robb, M.L., Dayhoff, D.E., Thitivichianlert, S., et al. (1995). Dual Infection with Human Immunodeficiency Virus Type 1 of Distinct Envelope Subtypes in Humans. J. Infect. Dis. *171*, 805–810.

Badri, M., Lawn, S.D., and Wood, R. (2008). Utility of CD4 cell counts for early prediction of virological failure during antiretroviral therapy in a resource-limited setting. BMC Infect. Dis. *8*, 89.

Baeten, J.M., Chohan, B., Lavreys, L., Chohan, V., McClelland, R.S., Certain, L., Mandaliya, K., Jaoko, W., and Julie, O. (2007). HIV-1 Subtype D Infection Is Associated with Faster Disease Progression than Subtype A in Spite of Similar Plasma HIV-1 Loads. J. Infect. Dis. *195*, 1177–1180.

Balotta, C., Facchi, G., Violin, M., Van Dooren, S., Cozzi-Lepri, A., Forbici, F., Bertoli, A., Riva, C., Senese, D., Caramello, P., et al. (2001). Increasing Prevalence of Non-Clade B HIV-1 Strains in Heterosexual Men and Women, as Monitored by Analysis of Reverse Transcriptase and Protease Sequences. J. Acquir. Immune Defic. Syndr. 2001 *27*, 499–505.

Baltimore, D. (1971). Expression of animal virus genomes. Bacteriol. Rev. *35*, 235–241.

Bebenek, K., Abbotts, J., Roberts, J.D., Wilson, S.H and Kunkel, T.A. (1989). Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase. The Journal of Biological Chemistry, *264,* 16948-16956.

Beemon, K., Duesberg, P., and Vogt, P. (1974). Evidence for Crossing-Over Between Avian Tumor Viruses Based on Analysis of Viral RNAs. Proc. Natl. Acad. Sci. *71*, 4254–4258.

Berry, I.M., Ribeiro, R., Kothari, M., Athreya, G., Daniels, M., Lee, H.Y., Bruno, W., and Leitner, T. (2007). Unequal Evolutionary Rates in the Human Immunodeficiency Virus Type 1 (HIV-1) Pandemic: the Evolutionary Rate of HIV-1 Slows Down When the Epidemic Rate Increases. J. Virol. *81*, 10625–10635.

Brenner, B., Wainberg, M.A., and Roger, M. (2013). Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions. AIDS *27*, 1045–1057.

Brenner, B.G., Roger, M., Routy, J.-P., Moisi, D., Ntemgwa, M., Matte, C., Baril, J.-G., Thomas, R., Rouleau, D., Bruneau, J., et al. (2007). High Rates of Forward Transmission Events after Acute/Early HIV-1 Infection. J. Infect. Dis. *195*, 951–959.

Buonaguro, L., Tornesello, M.L., and Buonaguro, F.M. (2007). Human Immunodeficiency Virus Type 1 Subtype Distribution in the Worldwide Epidemic: Pathogenetic and Therapeutic Implications. J. Virol. *81*, 10209–10219.

Buzón, M.J., Massanella, M., Llibre, J.M., Esteve, A., Dahl, V., Puertas, M.C., Gatell, J.M., Domingo, P., Paredes, R., Sharkey, M., et al. (2010). HIV-1 replication and immune dynamics are affected by raltegravir intensification of HAART-suppressed subjects. Nat. Med. *16*, 460–465.

Ciccozzi, M., Santoro, M.M., Giovanetti, M., Andrissi, L., Bertoli, A., and Ciotti, M. (2012). HIV-1 non-B subtypes in Italy: a growing trend. New Microbiol. *35*, 377–386.

Coffin, J.M. (1979). Structure, Replication, and Recombination of Retrovirus Genomes: Some Unifying Hypotheses. J. Gen. Virol. *42*, 1–26.

DeGruttola, V., Smith, D.M., Little, S.J., and Miller, V. (2010). Developing and Evaluating Comprehensive HIV Infection Control Strategies: Issues and Challenges. Clin. Infect. Dis. *50*, S102–S107.

Dittmar, M.T., Simmons, G., Donaldson, Y., Simmonds, P., Clapham, P.R., Schulz, T.F., and Weiss, R.A. (1997). Biological characterization of human immunodeficiency virus type 1 clones derived from different organs of an AIDS patient by long-range PCR. J. Virol. *71*, 5140–5147.

Dowling, W.E., Kim, B., Mason, C.J., Wasunna, K.M., Alam, U., Elson, L., Birx, D.L., Robb, M.L., McCutchan, F.E., and Carr, J.K. (2002). Forty-one near full-length HIV-

1 sequences from Kenya reveal an epidemic of subtype A and A-containing recombinants. AIDS Lond. Engl. *16*, 1809–1820.

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

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 Biol *4*, e88.

Easterbrook, P.J., Smith, M., Mullen, J., O'Shea, S., Chrystie, I., de Ruiter, A., Tatt, I.D., Geretti, A., and Zuckerman, M. (2010). Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy. J. Int. AIDS Soc. *13*, 4.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32,* 5, 1792-1797.

Eigen, M., and Schuster, P. (1979). The Hypercycle: a principle of natural self-organisation. Springer-Verlag; Berlin, ISBN 0-387-09293-5.

Esté, J.A., and Telenti, A. (2007). HIV entry inhibitors. The Lancet *370*, 81–88.

Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., et al. (2007). A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. Science *317*, 944–947.

Fischer, W., Apetrei, C., Santiago, M.L., Li, Y., Gautam, R., Pandrea, I., Shaw, G.M., Hahn, B.H., Letvin, N.L., Nabel, G.J., et al. (2012). Distinct Evolutionary Pressures Underlie Diversity in Simian Immunodeficiency Virus and Human Immunodeficiency Virus Lineages. J. Virol. *86*, 13217–13231.

Fouchier, R.A., Groenink, M., Kootstra, N.A., Tersmette, M., Huisman, H.G., Miedema, F., and Schuitemaker, H. (1992). Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. *66*, 3183–3187.

Frange, P., Meyer, L., Jung, M., Goujard, C., Zucman, D., Abel, S., Hochedez, P., Gousset, M., Gascuel, O., Rouzioux, C. and Chaix, M. (2013). Sexually-transmitted/founder HIV-1 cannot be directly predicted from plasma or PBMC-derived viral quasispecies in the transmitting partner. PLoS One *8*, e69144.

Frey, G., Chen, J., Rits-Volloch, S., Freeman, M.M., Zolla-Pazner, S., and Chen, B. (2010). Distinct conformational states of HIV-1 gp41 are recognized by neutralizing and non-neutralizing antibodies. Nat. Struct. Mol. Biol. *17*, 1486–1491.

Gall, A., Kaye, S., Hué, S., Bonsall, D., Rance, R., Baillie, G.J., Fidler, S.J., Weber, J.N., McClure, M.O., and Kellam, P. (2013). Restriction of V3 region sequence divergence in the HIV-1 envelope gene during antiretroviral treatment in a cohort of recent seroconverters. Retrovirology *10*, 8.

Galli, A., Kearney, M., Nikolaitchik, O.A., Yu, S., Chin, M.P.S., Maldarelli, F., Coffin, J.M., Pathak, V.K., and Hu, W.-S. (2010). Patterns of Human Immunodeficiency Virus Type 1 Recombination Ex Vivo Provide Evidence for Coadaptation of Distant Sites, Resulting in Purifying Selection for Intersubtype Recombinants during Replication. J. Virol. *84*, 7651–7661.

Geretti, A.M. (2006). HIV-1 subtypes: epidemiology and significance for HIV management : Current Opinion in Infectious Diseases. Curr. Opin. Infect. Dis. *19*.

Geretti, A.M., Harrison, L., Green, H., Sabin, C., Hill, T., Fearnhill, E., Pillay, D., and Dunn, D. (2009). Effect of HIV-1 Subtype on Virologic and Immunologic Response to Starting Highly Active Antiretroviral Therapy. Clin. Infect. Dis. *48*, 1296–1305.

Gifford, R., de Oliveira, T., Rambaut, A., Myers, R.E., Gale, C.V., Dunn, D., Shafer, R., Vandamme, A.-M., Kellam, P., and Pillay, D. (2006). Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity: AIDS *20*, 1521–1529.

Gifford, R.J., Oliveira, T. de, Rambaut, A., Pybus, O.G., Dunn, D., Vandamme, A.-M., Kellam, P., and Pillay, D. (2007). Phylogenetic Surveillance of Viral Genetic Diversity and the Evolving Molecular Epidemiology of Human Immunodeficiency Virus Type 1. J. Virol. *81*, 13050–13056.

Gilbert, M.T.P., Rambaut, A., Wlasiuk, G., Spira, T.J., Pitchenik, A.E., and Worobey, M. (2007). The emergence of HIV/AIDS in the Americas and beyond. Proc. Natl. Acad. Sci. *104*, 18566–18570.

Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O. (2005). PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. *33*, W557–W559.

Gupta, R.K., and Towers, G.J. (2009). A Tail of Tetherin: How Pandemic HIV-1 Conquered the World. Cell Host Microbe *6*, 393–395.

Gupta, R.K., Hué, S., Schaller, T., Verschoor, E., Pillay, D., and Towers, G.J. (2009). Mutation of a Single Residue Renders Human Tetherin Resistant to HIV-1 Vpu-Mediated Depletion. PLoS Pathog *5*, e1000443.

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. No 41 95–98.

Harris, M.E., Serwadda, D., Sewankambo, N., Kim, B., Kigozi, G., Kiwanuka, N., Phillips, J.B., Wabwire, F., Meehen, M., Lutalo, T., et al. (2002). Among 46 Near Full Length HIV Type 1 Genome Sequences from Rakai District, Uganda, Subtype D and AD Recombinants Predominate. AIDS Res. Hum. Retroviruses *18*, 1281–1290.

Harrison, S.C. (2008). Viral membrane fusion. Nat. Struct. Mol. Biol. *15*, 690–698.

Hill, C.P., Worthylake, D., Bancroft, D.P., Christensen, A.M., and Sundquist, W.I. (1996). Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: implications for membrane association and assembly. Proc. Natl. Acad. Sci. *93*, 3099–3104.

Hinz, A., Miguet, N., Natrajan, G., Usami, Y., Yamanaka, H., Renesto, P., Hartlieb, B., McCarthy, A.A., Simorre, J.-P., Göttlinger, H., et al. (2010). Structural Basis of

HIV-1 Tethering to Membranes by the BST-2/Tetherin Ectodomain. Cell Host Microbe *7*, 314–323.

Ho, S.Y.W., and Phillips, M.J. (2009). Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times. Syst. Biol. *58*, 367–380.

Ho, D.D., Neumann, A.U., Perelson, A.S., Chen, W., Leonard, J.M., and Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature *373*, 123–126.

Holguín, A., Lospitao, E., López, M., de Arellano, E.R., Pena, M.J., del Romero, J., Martín, C., and Soriano, V. (2008). Genetic characterization of complex inter-recombinant HIV-1 strains circulating in Spain and reliability of distinct rapid subtyping tools. J. Med. Virol. *80*, 383–391.

Holland, J.J., de la Torre, J.C., Clarke, D.C. and Duarte, E. (1991). Quantitation of relative fitness and great adaptability of clonal populations of RNA viruses. J Virol, *65,* 2960-2967.

Holmes, E.C. (2010). The RNA virus quasispecies: fact or fiction? J. Mol. Biol. *400*, 271-273.

HPA (2012). HIV in the United Kingdom: 2012 Report (Health Protection Agency).

Hu, W.S., and Temin, H.M. (1990). Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. Proc. Natl. Acad. Sci. *87*, 1556–1560.

Huang, W., Eshleman, S.H., Toma, J., Fransen, S., Stawiski, E., Paxinos, E.E., Whitcomb, J.M., Young, A.M., Donnell, D., Mmiro, F., et al. (2007). Coreceptor Tropism in Human Immunodeficiency Virus Type 1 Subtype D: High Prevalence of CXCR4 Tropism and Heterogeneous Composition of Viral Populations. J. Virol. *81*, 7885–7893.

Huber, H.E., McCoy, J.M., Seehra, J.S., and Richardson, C.C. (1989). Human immunodeficiency virus 1 reverse transcriptase. Template binding, processivity, strand displacement synthesis, and template switching. J. Biol. Chem. *264*, 4669–4678.

Hué, S., Clewley, J.P., Cane, P.A., and Pillay, D. (2004). HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS Lond. Engl. *18*, 719–728.

Hué, S., Pillay, D., Clewley, J.P., and Pybus, O.G. (2005). Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc. Natl. Acad. Sci. U. S. A. *102*, 4425–4429.

Hué, S., Hassan, A.S., Nabwera, H., Sanders, E.J., Pillay, D., Berkley, J.A., and Cane, P.A. (2012). HIV type 1 in a rural coastal town in Kenya shows multiple introductions with many subtypes and much recombination. AIDS Res. Hum. Retroviruses *28*, 220–224.

Hughes, G.J., Fearnhill, E., Dunn, D., Lycett, S.J., Rambaut, A., Leigh Brown, A.J., and on behalf of the UK HIV Drug Resistance Collaboration (2009). Molecular

Phylodynamics of the Heterosexual HIV Epidemic in the United Kingdom. PLoS Pathog *5*, e1000590.

Huthoff, H., Autore, F., Gallois-Montbrun, S., Fraternali, F., and Malim, M.H. (2009). RNA-Dependent Oligomerization of APOBEC3G Is Required for Restriction of HIV-1. PLoS Pathog *5*, e1000330.

Jansen, R., and Ledley, F.D. (1990). Disruption of phase during PCR amplification and cloning of heterozygous target sequences. Nucleic Acids Res. *18*, 5153–5156.

Jern, P., Russell, R.A., Pathak, V.K., and Coffin, J.M. (2009). Likely Role of APOBEC3G-Mediated G-to-A Mutations in HIV-1 Evolution and Drug Resistance. PLoS Pathog *5*, e1000367.

Jetzt, A.E., Yu, H., Klarmann, G.J., Ron, Y., Preston, B.D., and Dougherty, J.P. (2000). High Rate of Recombination throughout the Human Immunodeficiency Virus Type 1 Genome. J. Virol. *74*, 1234–1240.

Jost, S. and Altfeld, M., (2012). Evasion from NK cell-mediated immune responses by HIV-1. Microbes Infect. *14,* 11, 904-915.

Junghans, R.P., Boone, L.R., and Skalka, A.M. (1982). Retroviral DNA H structures: Displacement-assimilation model of recombination. Cell *30*, 53–62.

Kaleebu, P., French, N., Mahe, C., Yirrell, D., Watera, C., Lyagoba, F., Nakiyingi, J., Rutebemberwa, A., Morgan, D., Weber, J., et al. (2002). Effect of Human Immunodeficiency Virus (HIV) Type 1 Envelope Subtypes A and D on Disease Progression in a Large Cohort of HIV-1—Positive Persons in Uganda. J. Infect. Dis. *185*, 1244–1250.

Karlsson, A.C., Iversen, A.K.N., Chapman, J.M., de Oliveria, T., Spotts, G., McMichael, A.J., Davenport, M.P., Hecht, F.M. and Nixon, D.F. (2007). Sequential broadening of CTL responses in early HIV-1 infection is associated with viral escape. PLoS ONE, *2*, 2, e225.

Keele, B.F., and Derdeyn, C.A. (2009). Genetic and antigenic features of the transmitted virus: Curr. Opin. HIV AIDS *4*, 352–357.

Keele, B.F., Heuverswyn, F.V., Li, Y., Bailes, E., Takehisa, J., Santiago, M.L., Bibollet-Ruche, F., Chen, Y., Wain, L.V., Liegeois, F., et al. (2006). Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1. Science *313*, 523–526.

Kent S.J., Cameron P.U., Reece J.C., Thompson P.R. and Purcell D.F. (2001). Attenuated and wild-type HIV-1 infections and long terminal repeat-mediated gene expression from plasmids delivered by gene gun to human skin ex vivo and macaques in vivo. Virology, *287,* 71-78.

Kiwanuka, N., Laeyendecker, O., Quinn, T.C., J.Wawer, M., Shepherd, J., Robb, M., Kigozi, G., Kagaayi, J., Serwadda, D., Makumbi, F.E., et al. (2009). HIV-1 subtypes and differences in heterosexual HIV transmission among HIV-discordant couples in Rakai, Uganda. AIDS Lond. Engl. *23*, 2479–2484.

Kiwanuka, N., Robb, M., Laeyendecker, O., Kigozi, G., Wabwire-Mangen, F., Makumbi, F.E., Nalugoda, F., Kagaayi, J., Eller, M., Eller, L.A., et al. (2010). HIV-1

Viral Subtype Differences in the Rate of CD4+ T-Cell Decline Among HIV Seroincident Antiretroviral Naive Persons in Rakai District, Uganda. J. Acquir. Immune Defic. Syndr. 1999 *54*, 180–184.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. Science *288*, 1789–1796.

Kosakovsky Pond, S.L., Posada, D., Stawiski, E., Chappey, C., Poon, A.F.Y., Hughes, G., Fearnhill, E., Gravenor, M.B., Leigh Brown, A.J., and Frost, S.D.W. (2009). An Evolutionary Model-Based Algorithm for Accurate Phylogenetic Breakpoint Mapping and Subtype Prediction in HIV-1. PLoS Comput Biol *5*, e1000581.

Koup R.A., Safrit J.T., Cao Y., Andrews C.A., McLeod G., Borkowsky, W., Farthing, C. and Ho, D.D. (1994) Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. J Virol *68*, 4650–4655.

Kouyos, R.D., Fouchet, D., and Bonhoeffer, S. (2009). Recombination and drug resistance in HIV: Population dynamics and stochasticity. Epidemics *1*, 58–69.

Van Laethem, K., Schrooten, Y., Lemey, P., Wijngaerden, E.V., Wit, S.D., Ranst, M.V., and Vandamme, A.-M. (2005). A genotypic resistance assay for the detection of drug resistance in the human immunodeficiency virus type 1 envelope gene. J. Virol. Methods *123*, 25–34.

Van Laethem, K., Schrooten, Y., Dedecker, S., Van Heeswijck, L., Deforche, K., Van Wijngaerden, E., Van Ranst, M., and Vandamme, A.-M. (2006). A genotypic assay for the amplification and sequencing of gag and protease from diverse human immunodeficiency virus type 1 group M subtypes. J. Virol. Methods *132*, 181–186.

Lai, A., Simonetti, F.R., Zehender, G., De Luca, A., Micheli, V., Meraviglia, P., Corsi, P., Bagnarelli, P., Almi, P., Zoncada, A., et al. (2012). HIV-1 subtype F1 epidemiological networks among Italian heterosexual males are associated with introduction events from South America. PloS One *7*, e42223.

Land, A.M., Ball, T.B., Luo, M., Rutherford, J., Sarna, C., Wachihi, C., Kimani, J., and Plummer, F.A. (2008). Full-length HIV type 1 proviral sequencing of 10 highly exposed women from Nairobi, Kenya reveals a high proportion of intersubtype recombinants. AIDS Res. Hum. Retroviruses *24*, 865–872.

Langford, S.E., Ananworanich, J., and Cooper, D.A. (2007). Predictors of disease progression in HIV infection: a review. AIDS Res. Ther. *4*, 11.

Lauring, A.S., and Andino, R. (2010). Quasispecies Theory and the Behavior of RNA Viruses. PLoS Pathog *6*, e1001005.

Leigh Brown, A.J., Lycett, S.J., Weinert, L., Hughes, G.J., Fearnhill, E., Dunn, D.T., and UK HIV Drug Resistance Collaboration (2011). Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J. Infect. Dis. *204*, 1463–1469.

Leitner, T., Escanilla, D., Marquina, S., Wahlberg, J., Broström, C., Hansson, H.B., Uhlén, M., and Albert, J. (1995). Biological and Molecular Characterization of

Subtype D, G, and A/D Recombinant HIV-1 Transmissions in Sweden. Virology *209*, 136–146.

Lemey, P., Pybus, O.G., Rambaut, A., Drummond, A.J., Robertson, D.L., Roques, P., Worobey, M., and Vandamme, A.-M. (2004). The Molecular Population Genetics of HIV-1 Group O. Genetics *167*, 1059–1068.

Lemey, P., Rambaut, A., Drummond, A.J., and Suchard, M.A. (2009). Bayesian Phylogeography Finds Its Roots. PLoS Comput Biol *5*, e1000520.

Lemey, P., Rambaut, A., Welch, J.J., and Suchard, M.A. (2010). Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time. Mol. Biol. Evol. *27*, 1877–1885.

Lewis, F., Hughes, G.J., Rambaut, A., Pozniak, A., and Leigh Brown, A.J. (2008). Episodic Sexual Transmission of HIV Revealed by Molecular Phylodynamics. PLoS Med *5*, e50.

Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., and Ray, S.C. (1999). Full-Length Human Immunodeficiency Virus Type 1 Genomes from Subtype C-Infected Seroconverters in India, with Evidence of Intersubtype Recombination. J. Virol. *73*, 152–160.

Lorenzo-Redondo, R., Bordería, A.V., and Lopez-Galindez, C. (2011). Dynamics of In Vitro Fitness Recovery of HIV-1. J. Virol. *85*, 1861–1870.

Lythgoe, K.A., and Fraser, C. (2012). New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. Proc. R. Soc. B Biol. Sci. *279*, 3367–3375.

Marlink, R., Kanki, P., Thior, I., Travers, K., Eisen, G., Siby, T., Traore, I., Hsieh, C.C., Dia, M.C., and Gueye, E.H. (1994). Reduced rate of disease development after HIV-2 infection as compared to HIV-1. Science *265*, 1587–1590.

Marton, A., Delbecchi, L., and Bourgaux, P. (1991). DNA nicking favors PCR recombination. Nucleic Acids Res. *19*, 2423–2426.

Mbisa, J.L., Hué, S., Buckton, A.J., Myers, R.E., Duiculescu, D., Ene, L., Oprea, C., Tardei, G., Rugina, S., Mardarescu, M., et al. (2012). Phylodynamic and Phylogeographic Patterns of the HIV Type 1 Subtype F1 Parenteral Epidemic in Romania. AIDS Res. Hum. Retroviruses 120503121254004.

McNatt, M.W., Zang, T., Hatziioannou, T., Bartlett, M., Fofana, I.B., Johnson, W.E., Neil, S.J.D., and Bieniasz, P.D. (2009). Species-Specific Activity of HIV-1 Vpu and Positive Selection of Tetherin Transmembrane Domain Variants. PLoS Pathog *5*, e1000300.

Mellors, J.W., Munoz, A., Giorgi, J.V., Margolick, J.B., Tassoni, C.J., Gupta, P., Kingsley, L.A., Todd, J.A., Saah, A.J., Detels, R., Phair, J.P. and Rinaldo, C.R. (1997). Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. Ann Intern Med. *126,* 12, 946-954.

Meyerhans, A., Vartanian, J.-P., and Wain-Hobson, S. (1990). DNA recombination during PCR. Nucleic Acids Res. *18*, 1687–1691.

Mothe, B., Ibarrondo, J., Llano, A., and Brander, C. (2009). Virological, immune and host genetics markers in the control of HIV infection. Dis. Markers *27*, 105–120.

Nadai, Y., Eyzaguirre, L.M., Constantine, N.T., Sill, A.M., Cleghorn, F., Blattner, W.A., and Carr, J.K. (2008). Protocol for Nearly Full-Length Sequencing of HIV-1 RNA from Plasma. PLoS ONE *3*, e1420.

Ntale, R.S., Chopera, D.R., Ngandu, N.K., Rosa, D.A. de, Zembe, L., Gamieldien, H., Mlotshwa, M., Werner, L., Woodman, Z., Mlisana, K., et al. (2012). Temporal Association of HLA-B*81:01- and HLA-B*39:10-Mediated HIV-1 p24 Sequence Evolution with Disease Progression. J. Virol. *86*, 12013–12024.

Nuin, P.A.S., Wang, Z. and Tillier, E.R.M. (2006). The accuracy of several multiple sequence alignments for proteins. BMC Bioinformatics *7*, 471.

Oliveira, T. de, Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., Rensburg, E.J. van, Wensing, A.M.J., Vijver, D.A. van de, et al. (2005). An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics *21*, 3797–3800.

Ormsby, C.E., Sengupta, D., Tandon, R., Deeks, S.G., Martin, J.N., Jones, R.B., Ostrowski, M.A., Garrison, K.E., Vázquez-Pérez, J.A., Reyes-Terán, G., et al. (2012). Human endogenous retrovirus expression is inversely associated with chronic immune activation in HIV-1 infection. PloS One *7*, e41021.

Palmer, S., Kearney, M., Maldarelli, F., Halvas, E.K., Bixby, C.J., Bazmi, H., Rock, D., Falloon, J., Davey, R.T., Dewar, R.L., et al. (2005). Multiple, Linked Human Immunodeficiency Virus Type 1 Drug Resistance Mutations in Treatment-Experienced Patients Are Missed by Standard Genotype Analysis. J. Clin. Microbiol. *43*, 406–413.

Paraskevis, D., Magiorkinis, E., Magiorkinis, G., Kiosses, V.G., Lemey, P., Vandamme, A.-M., Rambaut, A., and Hatzakis, A. (2004). Phylogenetic Reconstruction of a Known HIV-1 CRF04_cpx Transmission Network Using Maximum Likelihood and Bayesian Methods. J. Mol. Evol. *59*, 709–717.

Peeters, M. (2001). The genetic variability of HIV-1 and its implications. Transfus. Clin. Biol. J. Société Fr. Transfus. Sang. *8*, 222–225.

Perez-Caballero, D., Hatziioannou, T., Yang, A., Cowan, S., and Bieniasz, P.D. (2005). Human Tripartite Motif 5α Domains Responsible for Retrovirus Restriction Activity and Specificity. J. Virol. *79*, 8969–8978.

Phillips, A.N., Lampe, F.C., Smith, C.J., Geretti, A.-M., Rodger, A., Lodwick, R.K., Cambiano, V., Tsintas, R., and Johnson, M.A. (2010). Ongoing changes in HIV RNA levels during untreated HIV infection: implications for CD4 cell count depletion: AIDS *24*, 1561–1567.

Piantadosi, A., Chohan, B., Chohan, V., McClelland, R.S., and Overbaugh, J. (2007). Chronic HIV-1 Infection Frequently Fails to Protect against Superinfection. PLoS Pathog *3*, e177.

Piantadosi, A., Chohan, B., Panteleeff, D., Baeten, J.M., Mandaliya, K., Ndinya-Achola, J.O., and Overbaugh, J. (2009). HIV-1 evolution in gag and env is highly correlated but exhibits different relationships with viral load and the immune response. AIDS Lond. Engl. *23*, 579–587.

Plantier, J.-C., Leoz, M., Dickerson, J.E., De Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., Robertson, D.L., and Simon, F. (2009). A new human immunodeficiency virus derived from gorillas. Nat. Med. *15*, 871–872.

Pond, S.L.K., Posada, D., Gravenor, M.B., Woelk, C.H., and Frost, S.D.W. (2006). Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. Mol. Biol. Evol. *23*, 1891–1901.

Poon, A.F.Y., Frost, S.D.W., and Pond, S.L.K. (2009). Detecting Signatures of Selection from DNA Sequences Using Datamonkey. In Bioinformatics for DNA Sequence Analysis, D. Posada, ed. (Humana Press), pp. 163–183.

Posada, D., and Crandall, K.A. (2002). The Effect of Recombination on the Accuracy of Phylogeny Estimation. J. Mol. Evol. *54*, 396–402.

Powell, R.L.R., Lezeau, L., Kinge, T., and Nyambi, P.N. (2010). Longitudinal Quasispecies Analysis of Viral Variants in HIV Type 1 Dually Infected Individuals Highlights the Importance of Sequence Identity in Viral Recombination. AIDS Res. Hum. Retroviruses *26*, 253–264.

Preston, B.D., Poiesz, B.J. and Loeb, L.A. (1988). Fidelity of HIV-1 Reverse Transcriptase. Science, *242,*1168-1171.

Preston, B.D., and Dougherty, J.P. (1996). Mechanisms of retroviral mutation. Trends Microbiol. *4*, 16–21.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2009). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. *26*, 7, 1641-1650.

Pybus, O.G., and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. Nat. Rev. Genet. *10*, 540–550.

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.

Quiñones-Mateu, M.E., and Arts, E.J. (2002). Fitness of drug resistant HIV-1: methodology and clinical implications. Drug Resist. Updat. *5*, 224–233.

Rambaut, A., Posada, D., Crandall, K.A., and Holmes, E.C. (2004). The causes and consequences of HIV evolution. Nat. Rev. Genet. *5*, 52–61.

Ranki, A., Lagerstedt, A., Ovod, V., Aavik, E. and Krohn, K.J.E. (1994). Expression kinetics and subcellular localisation of HIV-1 regulatory proteins Nef, Tat and Rev in acutely and chronically infected lymphoid cell lines. Arch Virol. *139,* 365-378.

Redd, A.D., Collinson-Streng, A.N., Chatziandreou, N., Mullis, C.E., Laeyendecker, O., Martens, C., Ricklefs, S., Kiwanuka, N., Nyein, P.H., Lutalo, T., et al. (2012). Previously transmitted HIV-1 strains are preferentially selected during subsequent sexual transmissions. J. Infect. Dis. *206*, 1433–1442.

Robertson, D.L., Hahn, B.H., and Sharp, P.M. (1995). Recombination in AIDS viruses. J. Mol. Evol. *40*, 249–259.

Robertson, D.L., Anderson, J.P., Bradac, J.A., Carr, J.K., Foley, B., Funkhouser, R.K., Gao, F., Hahn, B.H., Kalish, M.L., Kuiken, C., et al. (2000). HIV-1 nomenclature proposal. Science *288*, 55–56.

Rouzioux, C., Hubert, J., Burgard, M., Deveau, C., Goujard, C., Bary, M., Sereni, D., Viard, J., Delfraissy, J. and Meyer, J. (2005). Early levels of HIV-1 DNA in peripheral blood mononuclear cells are predictive of disease progression independently of HIV-1 RNA levels and CD4+ T cell counts. J Infect Dis. *192,* 1, 46-55.

Rusine, J., Jurriaans, S., van de Wijgert, J., Cornelissen, M., Kateera, B., Boer, K., Karita, E., Mukabayire, O., de Jong, M., and Ondoa, P. (2012). Molecular and Phylogeographic Analysis of Human Immuno-deficiency Virus Type 1 Strains Infecting Treatment-naive Patients from Kigali, Rwanda. PLoS ONE *7*.

Sagar, M., Laeyendecker, O., Lee, S., Gamiel, J., Wawer, M.J., Gray, R.H., Serwadda, D., Sewankambo, N.K., Shepherd, J.C., Toma, J., et al. (2009). Selection of HIV variants with signature genotypic characteristics during heterosexual transmission. J. Infect. Dis. *199*, 580–589.

Salazar-Gonzalez, J.F., Bailes, E., Pham, K.T., Salazar, M.G., Guffey, M.B., Keele, B.F., Derdeyn, C.A., Farmer, P., Hunter, E., Allen, S., et al. (2008). Deciphering Human Immunodeficiency Virus Type 1 Transmission and Early Envelope Diversification by Single-Genome Amplification and Sequencing. J. Virol. *82*, 3952–3970.

Salazar-Gonzalez, J.F., Salazar, M.G., Keele, B.F., Learn, G.H., Giorgi, E.E., Li, H., Decker, J.M., Wang, S., Baalwa, J., Kraus, M.H., et al. (2009). Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. J. Exp. Med. *206*, 1273–1289.

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 Res. Hum. Retroviruses *11*, 1423–1425.

Sanabani, S.S., Pastena, É.R. de S., Neto, W.K., Barreto, C.C., Ferrari, K.T., Kalmar, E.M., Ferreira, S., and Sabino, C.E. (2009). Near full-length genome analysis of low prevalent human immunodeficiency virus type 1 subclade F1 in São Paulo, Brazil. Virol. J. *6*, 1–11.

Sanders-Buell, E., Saad, M.D., Abed, A.M., Bose, M., Todd, C.S., Strathdee, S.A., Botros, B.A., Safi, N., Earhart, K.C., Scott, P.T., et al. (2007). A Nascent HIV Type 1 Epidemic among Injecting Drug Users in Kabul, Afghanistan Is Dominated by Complex AD Recombinant Strain, CRF35_AD. AIDS Res. Hum. Retroviruses *23*, 834–839.

Sauter, D., Schindler, M., Specht, A., Landford, W.N., Münch, J., Kim, K.-A., Votteler, J., Schubert, U., Bibollet-Ruche, F., Keele, B.F., et al. (2009). Tetherin-Driven Adaptation of Vpu and Nef Function and the Evolution of Pandemic and Nonpandemic HIV-1 Strains. Cell Host Microbe *6*, 409–421.

Schader, S.M., Colby-Germinario, S.P., Quashie, P.K., Oliveira, M., Ibanescu, R.-I., Moisi, D., Mesplède, T., and Wainberg, M.A. (2012). HIV gp120 H375 Is Unique to

HIV-1 Subtype CRF01_AE and Confers Strong Resistance to the Entry Inhibitor BMS-599793, a Candidate Microbicide Drug. Antimicrob. Agents Chemother. *56*, 4257–4267.

Schierup, M.H., and Hein, J. (2000). Consequences of Recombination on Traditional Phylogenetic Analysis. Genetics *156*, 879–891.

Schindler, M., Münch, J., Kutsch, O., Li, H., Santiago, M.L., Bibollet-Ruche, F., Müller-Trutwin, M.C., Novembre, F.J., Peeters, M., Courgnaud, V., et al. (2006). Nef-Mediated Suppression of T Cell Activation Was Lost in a Lentiviral Lineage that Gave Rise to HIV-1. Cell *125*, 1055–1067.

Schmidt, H.A., Strimmer, K., Vingron, M., and Haeseler, A. von (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502–504.

Schultz, A.-K., Zhang, M., Leitner, T., Kuiken, C., Korber, B., Morgenstern, B., and Stanke, M. (2006). A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. BMC Bioinformatics *7*, 265.

Schultz, A.-K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., and Stanke, M. (2009). jpHMM: Improving the reliability of recombination prediction in HIV-1. Nucleic Acids Res. *37*, W647–W651.

Sharp, P.M., and Hahn, B.H. (2011). Origin of HIV and the AIDS pandemic. Cold Spring Harb Perspect Med. *1,* 1, a006841.

Sharp, P.M., Shaw, G.M., and Hahn, B.H. (2005). Simian Immunodeficiency Virus Infection of Chimpanzees. J. Virol. *79*, 3891–3902.

Shriner, D., Rodrigo, A.G., Nickle, D.C., and Mullins, J.I. (2004). Pervasive Genomic Recombination of HIV-1 in Vivo. Genetics *167*, 1573–1583.

Siepel, A.C., Halpern, A.L., Macken, C., and Korber, B.T. (1995). A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. AIDS Res Hum Ret. *11,* 11, 1413-1416.

Van Sighem, A., Vidondo, B., Glass, T.R., Bucher, H.C., Vernazza, P., Gebhardt, M., de Wolf, F., Derendinger, S., Jeannin, A., Bezemer, D., et al. (2012). Resurgence of HIV Infection among Men Who Have Sex with Men in Switzerland: Mathematical Modelling Study. PLoS ONE *7*, e44819.

Simmonds, P., Balfe, P., Peutherer, J.F., Ludlam, C.A., Bishop, J.O., and Brown, A.J. (1990). Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. J. Virol. *64*, 864–872.

Simon-Loriere, E., Galetto, R., Hamoudi, M., Archer, J., Lefeuvre, P., Martin, D.P., Robertson, D.L., and Negroni, M. (2009). Molecular Mechanisms of Recombination Restriction in the Envelope Gene of the Human Immunodeficiency Virus. PLoS Pathog *5*, e1000418.

Simon-Loriere, E., Martin, D.P., Weeks, K.M., and Negroni, M. (2010). RNA Structures Facilitate Recombination-Mediated Gene Swapping in HIV-1. J. Virol. *84*, 12675–12682.

Skalka, A.M., Boone, L., Junghans, R., and Luk, D. (1982). Genetic recombination in avian retroviruses. J. Cell. Biochem. *19*, 293–304.

Smith, J.A. and Daniel, R. (2006). Following the path of the virus: the exploitation of host DNA repair mechanisms by retroviruses. ACS Chem Biol. *1,* 4, 217-226.

Soares, E.A.J.M., Santos, A.F.A., Sousa, T.M., Sprinz, E., Martinez, A.M.B., Silveira, J., Tanuri, A., and Soares, M.A. (2007). Differential Drug Resistance Acquisition in HIV-1 of Subtypes B and C. PLoS ONE *2*, e730.

Stall, R., Duran, L., Wisniewski, S.R., Friedman, M.S., Marshal, M.P., McFarland, W., Guadamuz, T.E., and Mills, T.C. (2009). Running in Place: Implications of HIV Incidence Estimates among Urban Men Who Have Sex with Men in the United States and Other Industrialized Countries. AIDS Behav. *13*, 615–629.

Stremlau, M., Perron, M., Welikala, S., and Sodroski, J. (2005). Species-Specific Variation in the B30.2(SPRY) Domain of TRIM5α Determines the Potency of Human Immunodeficiency Virus Restriction. J. Virol. *79*, 3139–3145.

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. *10*, 512–526.

Tang, J.W., and Pillay, D. (2004). Transmission of HIV-1 drug resistance. J. Clin. Virol. *30*, 1–10.

Tee, K.K., Pybus, O.G., Parker, J., Ng, K.P., Kamarulzaman, A., and Takebe, Y. (2009). Estimating the date of origin of an HIV-1 circulating recombinant form. Virology *387*, 229–234.

Thomson, M.M., Fernández-García, A., Delgado, E., Vega, Y., Díez-Fuertes, F., Sánchez-Martínez, M., Pinilla, M., Castro, M.Á., Mariño, A., Ordóñez, P., et al. (2012). Rapid expansion of a HIV-1 subtype F cluster of recent origin among men who have sex with men in Galicia, Spain. J. Acquir. Immune Defic. Syndr. 1999 *59*, e49–51.

Troyer, R.M., McNevin, J., Liu, Y., Zhang, S.C., Krizan, R.W., Abraha, A., Tebit, D.M., Zhao, H., Avila, S., Lobritz, M.A., et al. (2009). Variable Fitness Impact of HIV-1 Escape Mutations to Cytotoxic T Lymphocyte (CTL) Response. PLoS Pathog *5*, e1000365.

Truszkowski, J., and Brown, D.G. (2011). More accurate recombination prediction in HIV-1 using a robust decoding algorithm for HMMs. BMC Bioinformatics *12*, 168.

Tsibris, A.M.N., Korber, B., Arnaout, R., Russ, C., Lo, C.-C., Leitner, T., Gaschen, B., Theiler, J., Paredes, R., Su, Z., et al. (2009). Quantitative Deep Sequencing Reveals Dynamic HIV-1 Escape and Large Population Shifts during CCR5 Antagonist Therapy In Vivo. PLoS ONE *4*, e5683.

Vercauteren, J., Wensing, A.M.J., van de Vijver, D.A.M.C., Albert, J., Balotta, C., Hamouda, O., Kücherer, C., Struck, D., Schmit, J.-C., Asjö, B., et al. (2009). Transmission of drug-resistant HIV-1 is stabilizing in Europe. J. Infect. Dis. *200*, 1503–1508.

Vessière, A., Leoz, M., Brodard, V., Strady, C., Lemée, V., Depatureaux, A., Simon, F., and Plantier, J.-C. (2010). First evidence of a HIV-1 M/O recombinant form circulating outside Cameroon: AIDS *24*, 1079–1082.

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. J. Virol. *74*, 10498–10507.

Vidal, N., Bazepeo, S.E., Mulanga, C., Delaporte, E., and Peeters, M. (2009). Genetic Characterization of Eight Full-Length HIV Type 1 Genomes from the Democratic Republic of Congo (DRC) Reveal a New Subsubtype, A5, in the A Radiation That Predominates in the Recombinant Structure of CRF26_A5U. AIDS Res. Hum. Retroviruses *25*, 823–832.

Vidal, N., Diop, H., Montavon, C., Butel, C., Bosch, S., Ngole, E.M., Touré-Kane, C., Mboup, S., Delaporte, E., and Peeters, M. (2013). A novel multiregion hybridization assay reveals high frequency of dual inter-subtype infections among HIV-positive individuals in Cameroon, West Central Africa. Infect. Genet. Evol. *14*, 73–82.

Voigt, E., Wickesberg, A., Wasmuth, J.-C., Gute, P., Locher, L., Salzberger, B., Wöhrmann, A., Adam, A., Weitner, L., and Rockstroh, J. (2002). First-line ritonavir/indinavir 100/800 mg twice daily plus nucleoside reverse transcriptase inhibitors in a German multicentre study: 48-week results. HIV Med. *3*, 277–282.

Wagner, G.A., Pacold, M.E., Pond, S.L.K., Caballero, G., Chaillon, A., Rudolph, A.E., Morris, S.R., Little, S.J., Richman, D.D., and Smith, D.M. (2013). Incidence and Prevalence of Intrasubtype HIV-1 Dual Infection in At-Risk Men in the United States. J. Infect. Dis. jit633.

Ward, M.J., Lycett, S.J., Kalish, M.L., Rambaut, A., and Brown, A.J.L. (2013). Estimating the Rate of Intersubtype Recombination in Early HIV-1 Group M Strains. J. Virol. *87*, 1967–1973.

Wei, X., Decker, J.M., Wang, S., Hui, H., Kappes J.C., Wu, X., Salazar-Gonzalez, J.F., Salazar, M.G., Kilby, J.M., Saag, M.S., Komarova, N.L., Nowak, M.A., Hahn, B.H., Kwong, P.D. and Shaw, G.M. (2003). Antibody neutralisation and escape by HIV-1. Nature, *422,* 6929, 307-312.

Wieland, U., Seelhoff, A., Hofmann, A., K√°hn, J.E., Eggers, H.J., Mugyenyi, P., and Schwander, S. (1997). Diversity of the vif gene of human immunodeficiency virus type 1 in Uganda. J. Gen. Virol. *78*, 393–400.

Worobey, M., and Holmes, E.C. (2001). Homologous Recombination in GB Virus C/Hepatitis G Virus. Mol. Biol. Evol. *18*, 254–261.

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., et al. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature *455*, 661–664.

Wright, J.K., Brumme, Z.L., Carlson, J.M., Heckerman, D., Kadie, C.M., Brumme, C.J., Wang, B., Losina, E., Miura, T., Chonco, F., et al. (2010). Gag-Protease-

Mediated Replication Capacity in HIV-1 Subtype C Chronic Infection: Associations with HLA Type and Clinical Parameters. J. Virol. *84*, 10820–10831.

Von Wyl, V., Gianella, S., Fischer, M., Niederoest, B., Kuster, H., Battegay, M., Bernasconi, E., Cavassini, M., Rauch, A., Hirschel, B., et al. (2011). Early Antiretroviral Therapy During Primary HIV-1 Infection Results in a Transient Reduction of the Viral Setpoint upon Treatment Interruption. PLoS ONE *6*, e27463.

Zhang, J., and Temin, H.M. (1994). Retrovirus recombination depends on the length of sequence identity and is not error prone. J. Virol. *68*, 2409–2414.

Zhang, M., Foley, B., Schultz, A.-K., Macke, J.P., Bulla, I., Stanke, M., Morgenstern, B., Korber, B., and Leitner, T. (2010). The role of recombination in the emergence of a complex and dynamic HIV epidemic. Retrovirology *7*, 25.

Zhuang, J., Jetzt, A.E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B.D., and Dougherty, J.P. (2002). Human Immunodeficiency Virus Type 1 Recombination: Rate, Fidelity, and Putative Hot Spots. J. Virol. *76*, 11273–11282.

Chapter 9: Appendices

Appendix 2-1: UK HIV DRD proposal for the analysis of unassigned sequences

UK HIV
Drug Resistance Database

Analysis Proposal

1. Title of proposal

Analysis of the subtype unassigned sequences within the UK HIV Drug Resistance Database.

2. Study team

Please list all collaborators (lead investigator named first).

Geraldine Foster (UCL/University of Liverpool)

John Ambrose (UCL/University of Liverpool)

Anna Maria Geretti (UCL/University of Liverpool)

3. Background

The UK HIV-1 epidemic is becoming increasingly complex in terms of the subtypes involved[1,2,3]. This is largely due to increased immigration, especially from sub-Saharan Africa[4], but there is evidence that onward transmission of non-B subtypes is now occurring in the UK[5] and that they may be spreading in the MSM population[3,6].

The sequences in the UK HIV DRB have traditionally been subtyped using the Rega subtyping software tool. This is a conservative tool that returns many complex/unassigned sequences. More recently, the subtyping tool SCUEAL[7] has been used to subtype database sequences. Although it is common for different subtyping methodologies to exhibit discrepant classifications for a proportion of sequences[8,9], the results from using SCUEAL have been shown to be more accurate than Rega when characterising complex recombinant isolates[10]. However, even this method returns a significant proportion of unassigned sequences, which functions as an indicator of the increasing complexity of the epidemic, and may conceal unidentified circulating recombinant forms (CRFs).

CRF50_A1D, which is circulating in the Men who have Sex with Men (MSM) community, has been recently identified[11] from within the unclassified sequences in the UK HIV DRB, and has been shown to be contained within a sub-cluster of subtype A1 specimens analysed as part of the phylogeography of subtype A1 in the UK[12]. The identification of this CRF appears to confirm the hypothesis of Gifford *et al.* in 2007 that a novel subtype A recombinant is spreading within the UK via the MSM community[13].

Ongoing monitoring of the UK HIV epidemic is an important part of the battle against the disease. Knowledge of which subtypes are in circulation may inform vaccine design in the future[14], whilst a greater understanding of transmission dynamics and possible differences between risk groups will feed into decision making on prevention and intervention strategies within different populations.

References:

1. Arnold *et al.* (1995) AIDS Research and Human Retroviruses 11:427-429
2. Parry *et al.* (2001) Journal of Acquired Immune Deficiency Syndromes 26: 381-388
3. Gifford *et al.* (2007) Journal of Virology 81:13050-13056
4. Health Protection Agency (2009), HIV in the United Kingdom: 2009 report
5. Aggarwal *et al.* (2006). Journal of Acquired Immune Deficiency Syndromes; 41:201-209

6. Ambrose *et al.* 17th Conference on Retroviruses and Opportunistic Infections, San Francisco, February 2010, Abstract 98
7. Pond *et al.* (2009) PLoS Computational Biology 5 (11): e1000581
8. Gifford *et al.* (2006) AIDS 20: 1521-1529
9. Ntemgwa *et al.* (2008) AIDS Research and Human Retroviruses; 24 (7): 995-1002
10. Foster et al. 14th Annual Resistance Meeting: 'New Directions, New Challenges', London, September 2010, Abstract
11. Foster *et al.* 18th Conference on Retroviruses and Opportunistic Infections, Boston, February 2011, Abstract 456
12. Hue *et al.* Conference on Retroviruses and Opportunistic Infections, Boston, February 2011, Abstract 457
13. Gifford *et al.* (2007). Journal of Virology; 81: 13050-13056
14. Thomson *et al.* (2002). Lancet Infectious Diseases, 2:461-471

4. Proposed analysis

The SCUEAL analysis has indentified approximately 1400 unassigned and complex sequences in the UK HIV DRB.

In this proposal we wish to analyse the unassigned sequences in the database, in order to determine whether the proportion of unassigned sequences is increasing relative to 2007. Further questions will address whether the unassigned sequences represent the generation of new CRFs or the increasing complexity of the epidemic into complex URFs. The study population will be composed of people with complex/unassigned sequences in the UK HIV DRB.

The 4 main questions we hope to address are:

- Of the ~1400 patients with complex/unassigned sequences, what proportion of these can be said to represent new CRFs?

- When relating people with complex/unassigned infections to the whole UK HIV Drug Resistance Database population, can we identify further clusters or a trend towards the mixing of subtypes into complex URFs?
- Is there any evidence that these complex subtypes are transmitted preferentially to ‘pure’ infections?
- Is the contribution of recombination to the UK epidemic greater now than in 2007?

Our analysis will be based upon methods detailed in Gifford *et al.* 2007, and will focus on two aspects: 1) characterising the complex/unassigned sequences; 2) mapping these sequences to previously identified ‘pure’ subtype clusters, in order to assess the hitherto undetected contribution of recombination to the epidemic as a whole.

1. Sequences from patients with complex/unassigned subtypes by SCUEAL from both the 2007 and 2010 downloads will be requested from the UK HIV Drug Resistance Database (UKHDRD). In the case of patients with multiple sequences in the database, all sequences will be requested.
2. Where patients have multiple sequences in the database, the sequence closest to the time of infection will be used in the phylogenetic analysis.
3. Sequences will be analysed for recombination subtypes and breakpoints using jpHMM and Simplot, using a reference alignment comprised of sequences from the Los Alamos HIV Database that has been screened for recombination using GARD.
4. Phylogenetic analysis by maximum likelihood and Bayesian analysis will be used to identify any sets of complex sequences that cluster together.
5. Sets of complex sequences that cluster together will be split into subtype-specific fragments together with randomly selected subset of control sequences and all cohort and new sequences (ensuring duplicates are removed) will be analysed using maximum likelihood or Bayesian analyses.
6. Sequences that cluster together with apparently identical breakpoint locations will be requested and characterised using full-length, single genome analysis. *It is our intention to request no more than 9 specimens in total, in order to comply with our current ethics approval that limits our total number of characterised specimens to 15. Specimens will only be requested if there appears to be a previously unidentified CRF amongst the unidentified sequences.*

7. The number of complex sequences will be compared to the number of complex sequences from the 2007 download.
8. Next, a Bayesian Monte Carlo Markov Chain approach will be used to analyse reduced sets of sequences (those identified by the previous analyses as potentially linked in transmission clusters, together with selected control sequences).
9. BEAST (Drummond and Rambaut, BMC Evolutionary Biology 2007;7:214) will be used to construct time-scaled phylogenies of clusters, using sample dates to help date the tree.
10. Major sources of complex infections within the United Kingdom will be mapped using grouped centre data.

5. Intended start date and completion date

12 months from approval

6. Variables required

Variables required from UK HIV Drug Resistance Database:

Variable Name	Variable Label
id	Internal patient ID
labsampid	Test identifier
sampleno	Local sample identifier
centreid	Clinic requesting test (MRC coding)
dbsample	Date of blood sample
dob	Date of birth
lastcd4	
lastvl	
lastdate	
lastaids	

group	Treatment status when sample taken
dupsamp	Number of tests per sample
idcheck	Flag to indicate if sequence indicates wrongly matched patient
subtype	
method	Subtyping Method - (R)ega or (S)tar

Further variables required (some of which may need to come from UKCHIC):

1. *pol* sequences for all patients included in the database with date of sample/sequence (multiple sequences from the same patient to be included but identifiable as such)
2. Date of HIV diagnosis
3. Risk group
4. Ethnic group
5. Date of arrival in UK
6. Suspected country of infection

7. Feasibility assessment

Database scanning of the 2007 UK HIV DRB download identified 5 patients that appeared to contain a novel recombinant. 4 patients seen at clinics in London were characterised using full-length single genome analysis; 3 were found to be infected with a novel CRF, CRF50_A1D, and 1 was infected with a recombinant of CRF50_A1D and subtype B[1]. Geographic mapping identified a further 74 patients who appeared to be infected with the same recombinant. 32 of these patients were present in a putative subtype A cluster analysed by S. Hue.

Although the size of the cohort used in this study is relatively small, the results of the analysis allow us to tentatively suggest a number of conclusions. Firstly, unidentified CRFs can be located in the subtype unassigned sequences in the database. Secondly, sequences with a short recombinant fragment can be misclassified as ‘pure’, indicating that the contribution of recombination to epidemic diversity is likely to have been underestimated. Thirdly, non-B subtypes appear to be spreading within the MSM population.

References:

1. Foster et al. 18th Conference on Retroviruses and Opportunistic Infections, Boston, February 2011, Abstract 457

8. Resource required

We require support from the MRC-CTU and possibly UK CHIC for handling of the dataset.

Appendix 2_2: UK CHIC proposal for the analysis of CD4 decline in CRF50_A1D patients

Title: Investigating the pathogenesis of CRF50_A1D.

Submission Date: 04.05.2012

Study Team: Geraldine Foster, Anna Maria Geretti

Background:

We recently identified CRF50_A1D, a novel HIV recombinant circulating predominately in the UK MSM population[1]. Our research suggests that this recombinant currently infects 72 people in the UK, who are located almost exclusively in north west and south east England[2]. 17 of these patients have data held in UK CHIC.

Research from the Rakai cohort in Uganda has indicated that recombinant A/D infections progress faster than those with ‘pure’ A or D infections[3,4,5]. We would like to investigate the pathogenesis of CRF50 in the UK population, and propose to begin this analysis by analyzing the CD4 decline of the 17 patients with data in UK CHIC. This will be a preliminary analysis to allow us to decide whether to collect data from those patients who are not in CHIC.

Aims:

- To assess the rate of CD4 decline in CRF50 patients by mapping CD4 slopes onto the CD4 decline slopes presented at CROI 2011[6]

Inclusion/Exclusion criteria: Inclusion: Infected with CRF50_A1D

Variables required:

- CD4 counts; Viral loads; ART history

Possible limitations/issues that should be considered when interpreting the findings:

This analysis will only involve 17/72 (23.6%) CRF50 patients, and, as such, should be considered preliminary.

Proposed time scale for analysis: 3 months from proposal approval

Resources required: No extra resources will be required

References

1. *Foster et al. http://www.retroconference.org/2011/Abstracts/40621.htm* 2011
2. *Foster et al. Oral presentation 19th International HIV Dynamics and Evolution, 2012*
3. *Baeten et al. J. Infect Dis 2007*
4. *Kaleebu et al. J.Infect Dis 2002*
5. *Kiwanuka et al. J AIDS 2010*
6. *Klein et al. http://www.retroconference.org/2011/PDFs/463.pdf 2011*

Appendix 2-3: Sequencing primers used for near full-length HIV-1 single genome sequencing

Number	Primer Name	Sequence 5'-3'	HXB2 co-ordinates	Source
Forward Primers				
1	2.U5.B4F	AGTAGTGTGTGCCCGTCTGTTGTGTGACTC	552-581	CHAVI-MBSC, 2009, unpublished
2	msf12b(+)	AAATCTCTAGCAGTGGCGCCCGAACAG	623-649	Nadai et al., 2008
3	KVL066	TCTCTAGCAGTGGCGCCCGAACAG	626-649	Van Laetham et al., 2006
4	f2nst(+)	GCGGAGGCTAGAAGGAGAGAGATGG	769-793	Nadai et al., 2008
5	DD	GTATGGGCAAGCAGGGAGCTAGAA	892-915	Nadai et al., 2008
6	HH	ATGAGGAAGCTGCAGAATGGG	1406-1426	Nadai et al., 2008
7	II	ATAATCCACCTATCCCAGTAGGAGAAAT	1544-1571	Nadai et al., 2008
8	pro5F(+)	AGAAATTGCAGGGCCCCTAGGAA	1966-2018	Nadai et al., 2008
9	POLCLO1-	GAGAGACAGGCTAATTTTTTAGGGAA	2071-2096	Nadai et al., 2008
10	pro3F(+)	AGANCAGAGCCAACAGCCCCACCA	2143-2166	Nadai et al., 2008
11	POLoutF1(+)	CCTCAAATCACTCTTTGGCARCGAC	2253-2277	Nadai et al., 2008
12	BJPOL1	ACAGGAGCAGATGATACAGTA	2328-2348	Nadai et al., 2008
13	POLinF1	AGGACCTACRCCTGTCAACATAATTGG	2483-2509	Nadai et al., 2008
14	AZT3	CCAGGAATGGATGGACCAA	2589-2607	Nadai et al., 2008
15	SP4S	GGGCCTGAAAATCCATACAATACT	2700-2723	Nadai et al., 2008
16	AZT9	TGGATGTGGGTGATGCATA	2875-2893	Nadai et al., 2008

Number	Primer Name	Sequence 5’-3’	HXB2 co-ordinates	Source
17	SP5S	GGATTAGATATCAGTACAATGTGC	2971-2994	Nadai et al., 2008
18	AZT6	CAATACATGGATGATTTGTATGTAGG	3093-3118	Nadai et al., 2008
19	POLP	GGATGGGATATGAACTCCATCC	3235-3256	Nadai et al., 2008
20	DGPOLF7	GGAATATATTATGACCCATCAAAAGAC	3495-3521	Nadai et al., 2008
21	POLU	ACTTTCTATGTAGATGGGGCAGC	3864-3886	Nadai et al., 2008
22	POLI	GAGCAGTTAATAAAAAAGGAA	4116-4136	Nadai et al., 2008
23	POLJ	GAAGCCATGCATGGACAAGTAGA	4371-4393	Nadai et al., 2008
24	POLK	ACGGTTAAGGCCGCCTGTTGGTGG	4602-4625	Nadai et al., 2008
25	POLSEQ2	CGGGTTTATTACAGGGACAGC	4899-4919	Nadai et al., 2008
26	VIF1	GGGTTTATTACAGGGACAGCAGAG	4900-4923	CHAVI-MBSC, 2009, unpublished
27	ACC1	TTCAGAAGTATACATCCCACTAGG	5196-5219	Nadai et al., 2008
28	KVL008	GGTCAKGGRGTCTCCATAGAATGGA	5284-5308	Van Laetham et al., 2005
29	VIFB	ATATAGCACACAAGTAGACCCT	5319-5340	Nadai et al., 2008
30	AV317	TCAAGCAGGACATAAYAAGGTAGG	5445-5468	Van Laetham et al., 2005
31	VIFC	GAYAAAGCCACCTTTGCCTAGTGTT	5514-5538	Nadai et al., 2008
32	ENVoutF1 (+)	AGARGAYAGATGGAACAAGCCCCAG	5550-5574	Nadai et al., 2008
33	ACC5	TGAAACTTAYGGGGATACTTGG	5699-5720	Nadai et al., 2008
34	ENVinF1	TGGAAGCATCCRGGAAGTCAGCCT	5861-5884	Nadai et al., 2008
35	ENVA	GGCTTAGGCATCTCCTATGGCAGGAAGAA	5954-5982	CHAVI-MBSC, 2009, unpublished
36	ED3	TTAGGCATCTCCTATGGCAGGAAGAAGCGG	5957-5986	Nadai et al., 2008

Number	Primer Name	Sequence 5'-3'	HXB2 co-ordinates	Source
37	GP1205-	AGAGCAGAAGACAGTGGCAATGA	6206-6228	Nadai et al., 2008
38	Z1F	TGGGTCACAGTCTATTATGGGGTACCT	6327-6353	Nadai et al., 2008
39	ENVSEQ22	GTGTACCCACAGACCCCAGCCCACAAG	6445-6471	Nadai et al., 2008
40	ZFF	GGGATCAAAGCCTAAAGCCATGTGTAA	6559-6585	Nadai et al., 2008
41	793SEQ1	AACACCTCAGTCATTACACAGGCC	6813-6836	Nadai et al., 2008
42	E16	CCAATTCCCATACATTATTGTG	6858-6879	Nadai et al., 2008
43	AV318	TGCTGYTRAATGGCAGTCTAGCAGA	7000-7024	Van Laetham et al., 2005
44	E15	GTAGAAATTAATTGTACAAGACCC	7098-7121	Nadai et al., 2008
45	OFM54	TTTAATTGTGGAGGGGAATTTTTCT	7350-7374	Nadai et al., 2008
46	E13	ACAAATTATAAACATGTGGCAGG	7487-7509	Nadai et al., 2008
47	JL109	GTGAATTATATAAATATAAAGTAG	7668-7689	Nadai et al., 2008
48	TUG	GTCTGGTATAGTGCAACAGCA	7859-7879	Nadai et al., 2008
49	ZLF	GGGATAACATGACCTGGATGCAGTGGG	8092-8118	Nadai et al., 2008
50	JL104	GGAGGCTTGATAGGTTTAAGAATA	8292-8315	Nadai et al., 2008
51	JL106	TTCAGCTACCACCGCTTGAGAGACT	8520-8544	Nadai et al., 2008
52	NEF7	TAAGATGGGTGGCAAGTGGTCCAAAA	8793-8818	Nadai et al., 2008
53	NEF6	AGCAGCAGATGGGGTGGGAGCAG	8871-8893	Nadai et al., 2008
54	LTR2	TTTGGATGGTGCTACAAGCTA	9211-9231	Designed using Primer3
Reverse primers				
55	JL19	CTTCTATTACTTTTACCCATGC	1249-1270	Nadai et al., 2008

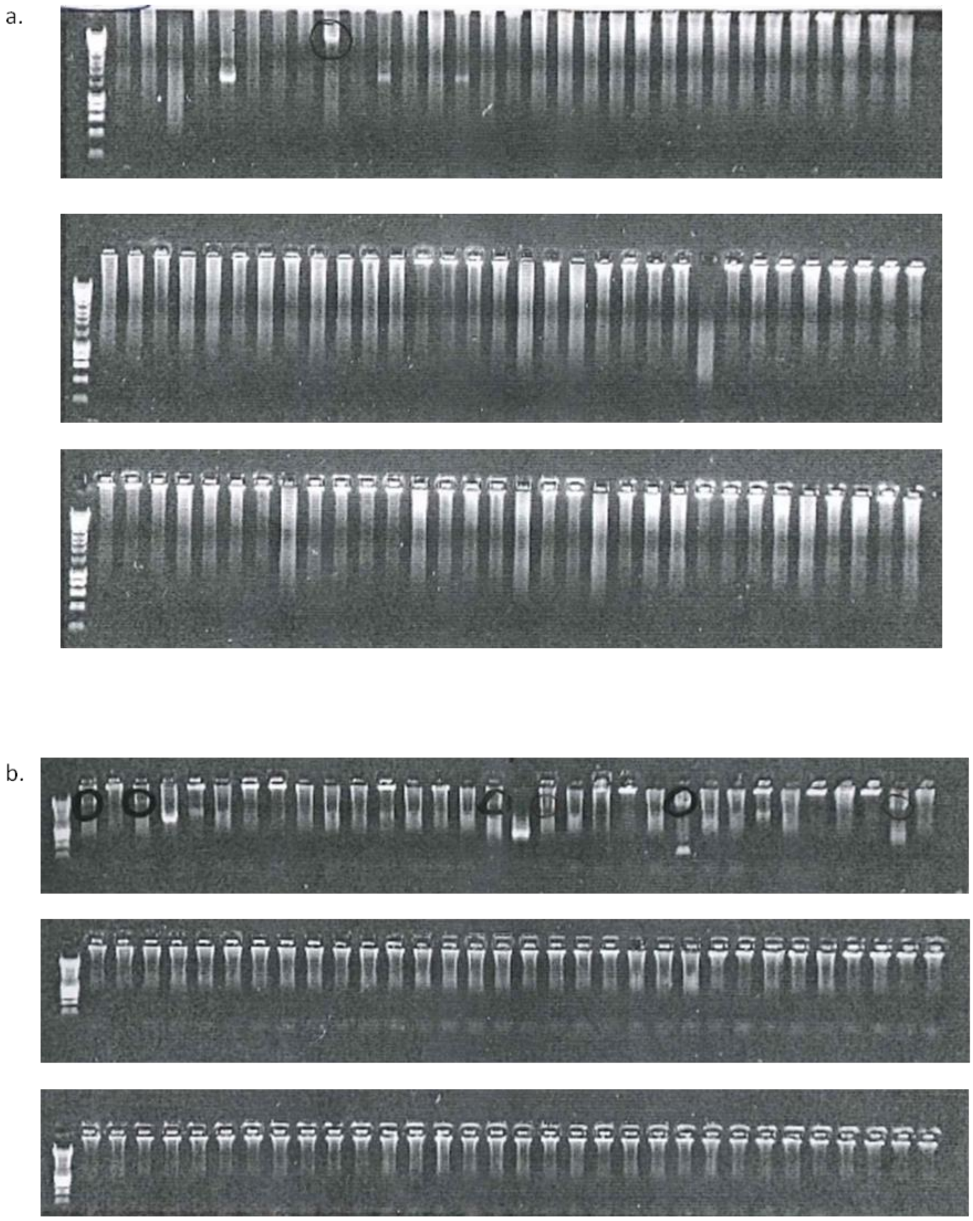
Number	Primer Name	Sequence 5'-3'	HXB2 co-ordinates	Source
56	JL17	CATTCTGCAGCTTCCTCATTGAT	1402-1424	Nadai et al., 2008
57	SP2AS	GGTGGGGCTGTTGGCTCTG	2147-2165	Nadai et al., 2008
58	SP3AS	CCTCCAATTCCCCCTATCATTTTTGG	2382-2407	Nadai et al., 2008
59	KVL067	GGCCATTGTTTAACYTTTGGDCCATCC	2597-2623	Van Laetham et al., 2006
60	SP4AS	AGTATTGTATGGATTTTCAGGCCC	2700-2723	Nadai et al., 2008
61	KVL065	TCCTAATTGAACYTCCCARAARTCYTGAGTTC	2797-2828	Van Laetham et al.
62	POLC-	CTAGGTATGGTAAATGCAGTATA	2928-2950	Nadai et al., 2008
63	AZT10	CCTACATACAAATCATCCATGTATTG	3093-3118	Nadai et al., 2008
64	AZT5	TCAGATCCTACATACAAATCATCCATGTATTG	3093-3124	Nadai et al., 2008
65	AZT4	TATAGGCTGTACTGTCCATTT	3261-3281	Nadai et al., 2008
66	POLEE-	TGTATGTCATTGACAGTCCAGCTG	3299-3322	Nadai et al., 2008
67	proRT	TTTCCCCACTAACTTCTGTATGTCATTGACA	3308-3338	Nadai et al., 2008
68	RT3473R (-)	GAATCTCTCTGTTTTCTGCCAGTTC	3453-3477	Nadai et al., 2008
69	POLSEQ3	GATATGWCCACTGGTCTTGCCC	3546-3567	Nadai et al., 2008
70	DGPOL3R	GTATTGACAAACTCCCAGTCAGGAAT	3780-3805	Nadai et al., 2008
71	POLI-	TTTGTGTGCTGGTACCCATGCCAG	4146-4169	Nadai et al., 2008
72	POLT-	GCAGTCTACTTGTCCATGCATGGC	4374-4397	Nadai et al., 2008
73	SP1AS	GGATGAATACTGCCATTTGTACTGC	4752-4776	Nadai et al., 2008
74	DGPOL2R	CACTATTGTCTTGTATTACTAC	4974-4995	Nadai et al., 2008
75	ACC2	AGGGTCTACTTGTGTGYTATAT	5319-5340	Nadai et al., 2008

Number	Primer Name	Sequence 5’-3’	HXB2 co-ordinates	Source
76	ACC6	GCTTGTTCCATCTRTCYTCTGTYAG	5545-5569	Nadai et al., 2008
77	ACC4	CCAAGTATCCCCRTAAGTTTCA	5699-5720	Nadai et al., 2008
78	ACC8R	TCTCCGCTTCTTCCTGCCATAG	5968-5989	Nadai et al., 2008
79	VIF-VPUinR1	CTCTCATTGCCACTGTCTTCTGCTC	6207-6231	Nadai et al., 2008
80	ES33	CATTGCCACTGTCTTCTGCTC	6207-6227	Nadai et al., 2008
81	VIF-VPUoutR1 (-)	GGTACCCCATAATAGACTGTRACCCACAA	6324-6352	Nadai et al., 2008
82	JL99	TTTAGCATCTGATGCACAAAATAG	6378-6401	Nadai et al., 2008
83	AENVSEQ4	CAAGCTTGTGTAATGGCTGAGG	6817-6838	Nadai et al., 2008
84	TUE3	TCCTTCTGCTAGACTGCCATTTA	7006-7028	Nadai et al., 2008
85	JL98	AGAAAAATTCCCCTCCACAATTAA	7351-7374	Nadai et al., 2008
86	JL102	GATGGGAGGGGCATACAT	7509-7524	Nadai et al., 2008
87	EDS8	CACTTCTCCAATTGTCCCTCA	7648-7668	Nadai et al., 2008
88	AV323	CTGCTCCYAAGAACCCAA	7783-7800	Van Laetham et al., 2005
89	TUH	GCCCCAGACTGTGAGTTGCAACAGATG	7914-7940	Nadai et al., 2008
90	FM116	CAGAGATTTATTACTCCAACTA	8060-8081	Nadai et al., 2008
91	ENVSEQ6	CCTGCCTAACTCTATTCAC	8337-8355	Nadai et al., 2008
92	E8	CTCTCTCTCCACCTTCTTCTTC	8424-8445	Nadai et al., 2008
93	JL71	TTTTGACCACTTGCCACCCAT	8797-8817	Nadai et al., 2008
94	AV319	GCTSCCTTRTAAGTCATTGGTCT	9025-9047	Van Laetham et al., 2005
95	JL89	TCCAGTCCCCCCTTTTCTTTTAAAAA	9064-9089	Nadai et al., 2008

Number	Primer Name	Sequence 5’-3’	HXB2 co-ordinates	Source
96	KVL009	GCCAATCAGGGAAGWAGCCTTGTGT	9145-9169	Van Laetham et al., 2005
97	nefyn05 (-)	GTGTGTAGTTCTGCCAATCAGGGAA	9157-9181	Nadai et al., 2008
98	UNINEF 7’ (-)	GCACTCAAGGCAAGCTTTATTGAGGCTT	9605-9632	Nadai et al., 2008
99	OFM19	GCACTCAAGGCAAGCTTTATTGAGGCTTA	9604-9632	CHAVI-MBSC, 2009, unpublished
100	2.R3.B6R	TGAAGCACTCAAGGCAAGCTTTATTGAGGC	9607-9636	CHAVI-MBSC, 2009, unpublished

Appendix 5_1: Amplification of specimen 11762

Although sufficient volume of specimen 11762 to extract 20,000 copies was received, the specimen did not amplify in a manner suggesting near-ideal extraction/RT success.



a) Initial amplification. 20,000 RNA copies were extracted and reverse transcribed. The specimen was amplified using 16 replicates of neat cDNA, 32 replicates of cDNA diluted 1:20 and 46 replicates of a 1:100 cDNA dilution. Two ntc wells were run. Lane 1 of each row contains Hyperladder I. Top row: lanes 2 - 17 contain the neat cDNA, lanes 18-33 contain 1:20 dilution. Middle row: lanes 2 - 17 contain the 1:20 dilution, and lanes 18 - 33 contain the 1:100 dilution. Bottom row: Lanes 2 - 31 contain the 1:100 dilution and lanes 32 - 33 contain the ntc replicates. Only one reaction using neat cDNA is positive; no other positive reactions are present.

b) Subsequent amplification. The remaining plasma was used to extract 10,443 RNA copies. Following reverse transcription, the specimen was amplified using 16 replicates of neat cDNA, 32 replicates of 1:4 diluted cDNA, and 32 replicates of 1:20 diluted cDNA. 16 replicates of the original extraction amplified in a) was amplified in parallel. Lane 1 of each row contains Hyperladder I. Top row: lanes 2 - 17 contain the original neat cDNA; lanes 18 - 33 contain the new neat cDNA. Middle row: lanes 2 - 33 contain the 1:4 diluted cDNA. Bottom row: lanes 2 - 33 contain the 1:100 dilution. 3/16 positive reactions are seen using the original cDNA extraction, 3/16 positive reactions are present using the new neat cDNA, and no positive reactions using the 1:4 or 1:20 dilution are present. Subsequent Sanger sequencing was performed using the positive reactions from the original extraction.

Appendix 6_1: CRF50 sequences, genotyping results and breakpoint location

ID	Sample date	jpHMM result	jpHMM breakpoint	SCUEAL result	SCUEAL breakpoint
33365	2003.29	A1/D	2489	A1/D	2465
8179	200.54	A1/D	2489	A1/D	2465
40534	2003.29	A1/D	2489	A1/D	2465
11762	2010.21	A1/D	2487	A1/D	2465
12792	2010.54	A1/D	2498	A1/D	2465
129767	2000.79	A1/B	2496	A1/D	2465
118019	2002.04	A1/D	2504	A1/D	2465
92050	2003.29	A1/B	2505	A1/D	2465
141509	2003.38	A1/B	2489	A1/D	2465
133564	2003.54	A1/D	2504	A1/D	2465
108052	2003.54	A1/B	2490	A1/D	2489
110018	2004.04	A1/D	2489	A1/D	2465
134946	2004.71	A1/D	2504	A1/D	2465
97391	2005.13	A1/B	2515	A1/D	2465
102936	2005.21	A1/D	2504	A1/D	2465
93054	2005.54	A1/B	2489	A1/D	2473
142422	2005.71	A1/B	2505	A1/D	2484
100178	2005.71	A1/B	2505	A1/D	2484
102174	2005.71	A1/D	2489	A1/D	2479
138730	2006.04	A1/D	2504	A1/D	2465
125935	2006.04	A1/B	2505	A1/D	2465
141321	2006.13	A1/D	2504	A1/D	2465
103799	2006.21	A1/B	2505	A1/D	2474
107137	2006.29	A1/B	2505	A1/D	2465
91340	2006.46	A1/B	2489	A1/D	2465
109054	2006.54	A1/B	2489	A1/D	2465
134450	2006.54	A1/B	2513	A1/D	2484
141058	2006.63	A1/B	2490	A1/D	2484
143060	2006.63	A1/B	2505	A1/D	2465
107838	2006.63	A1/D	2485	A1/D	2484
145478	2006.71	A1/B	2491	A1/D	2465
99870	2006.79	A1/B	2505	A1/D	2465
106138	2006.79	A1/B	2489	A1/D	2465
107730	2006.79	A1/B	2491	A1/D	2465
127533	2006.88	A1/B	2490	A1/D	2465
118453	2007.04	A1/B	2489	A1/D	2465
144994	2007.04	A1/B	2505	A1/D	2484
140410	2007.04	A1/D	2504	A1/D	2465
121724	2007.04	A1/B	2505	A1/D	2465
145903	2007.04	A1/D	2489	A1/D	2465
132904	2007.13	A1/B	2490	A1/D	2465
110110	2007.21	A1/B	2505	A1/D	2484
108721	2007.21	A1/D	2504	A1/D	2465
103364	2007.38	A1/B	2505	A1/D	2465
92971	2007.38	A1/D	2504	A1/D	2465
138786	2007.46	A1/B	2489	A1/D	2465
109993	2007.46	A1/B	2497	A1/D	2465
91853	2007.54	A1/B	2489	A1/D	2465

ID	Sample date	jpHMM result	jpHMM breakpoint	SCUEAL result	SCUEAL breakpoint
135349	2007.63	A1/D	2504	A1/D	2465
129432	2007.71	A1/B	2496	A1/D	2465
102695	2007.79	A1/B	2489	A1/D	2465
114465	2007.79	A1/B	2489	A1/D	2465
96930	2007.79	A1/B	2505	A1/D	2465
141595	2007.88	A1/D	2504	A1/D	2465
135766	2007.96	A1/D	2504	A1/D	2465
103741	2008.04	A1/D	2504	A1/D	2465
111029	2008.13	A1/D	2504	A1/D	2465
102569	2008.21	A1/D	2489	A1/D	2473
125047	2008.38	A1/B	2491	A1/D	2465
122007	2008.46	A1/D	2534	A1/D	2465
120030	2008.54	A1/B	2505	A1/D	2465
134766	2008.63	A1/D	2489	A1/D	2465
110602	2008.63	A1/D	2489	A1/D	2473
123511	2008.71	A1/B	2490	A1/D	2465
119404	2008.79	A1/D	2489	A1/D	2465
114012	2008.88	A1/B	2494	A1/D	2465
115813	2008.88	A1/B	2505	A1/D	2465
121059	2008.96	A1/B	2490	A1/D	2484
103294	2008.04	A1/B	2505	A1/D	2465
137893	2008.38	A1/B	2489	A1/D	2465
103571	2006.79	A1/B	2505	A1/D	2465
114916	2012.21	A1/D	2489	A1/D	2465

Appendix 6_2 PLoS ONE publication

Novel HIV-1 Recombinants Spreading across Multiple Risk Groups in the United Kingdom: The Identification and Phylogeography of Circulating Recombinant Form (CRF) 50_A1D

Geraldine M. Foster[1], John C. Ambrose[1], Stéphane Hué[2], Valerie C. Delpech[3], Esther Fearnhill[4], Ana B. Abecasis[5], Andrew J. Leigh Brown[6], Anna Maria Geretti[1]*, on behalf of the UK HIV Drug Resistance Database

1 University of Liverpool, Liverpool, United Kingdom, **2** University College London, London, United Kingdom, **3** Public Health England, London, United Kingdom, **4** MRC Clinical Trials Unit, London, United Kingdom, **5** Universidade Nova de Lisboa, Lisbon, Portugal, **6** University of Edinburgh, Edinburgh, United Kingdom

Abstract

Background: An increase in non-B HIV-1 infections among men who have sex with men (MSM) in the United Kingdom (UK) has created opportunities for novel recombinants to arise and become established. We used molecular mapping to characterize the importance of such recombinants to the UK HIV epidemic, in order to gain insights into transmission dynamics that can inform control strategies.

Methods and Results: A total of 55,556 *pol* (reverse transcriptase and protease) sequences in the UK HIV Drug Resistance Database were analyzed using Subtype Classification Using Evolutionary Algorithms (SCUEAL). Overall 72 patients shared the same A1/D recombination breakpoint in *pol*, comprising predominantly MSM but also heterosexuals and injecting drug users (IDUs). In six MSM, full-length single genome amplification of plasma HIV-1 RNA was performed in order to characterize the A1/D recombinant. Subtypes and recombination breakpoints were identified using sliding window and jumping profile hidden markov model approaches. Global maximum likelihood trees of *gag*, *pol* and *env* genes were drawn using FastTree version 2.1. Five of the six strains showed the same novel A1/D recombinant (8 breakpoints), which has been classified as CRF50_A1D. The sixth strain showed a complex CRF50_A1D/B/U structure. Divergence dates and phylogeographic inferences were determined using Bayesian Evolutionary Analysis using Sampling Trees (BEAST). This estimated that CRF50_A1D emerged in the UK around 1992 in MSM, with subsequent transmissions to heterosexuals and IDUs. Analysis of CRF50_A1D/B/U demonstrated that around the year 2000 CRF50_A1D underwent recombination with a subtype B strain.

Conclusions: We report the identification of CRF50_A1D, a novel circulating recombinant that emerged in UK MSM around 1992, with subsequent onward transmission to heterosexuals and IDUs, and more recent recombination with subtype B. These findings highlight the changing dynamics of HIV transmission in the UK and the converging of the two previously distinct MSM and heterosexual epidemics.

Citation: Foster GM, Ambrose JC, Hué S, Delpech VC, Fearnhill E, et al. (2014) Novel HIV-1 Recombinants Spreading across Multiple Risk Groups in the United Kingdom: The Identification and Phylogeography of Circulating Recombinant Form (CRF) 50_A1D. PLoS ONE 9(1): e83337. doi:10.1371/journal.pone.0083337

Editor: Chiyu Zhang, Institut Pasteur of Shanghai, Chinese Academy of Sciences, China

Received April 3, 2013; **Accepted** November 4, 2013; **Published** January 15, 2014

Funding: Primary funding of the UK HIV Drug Resistance Database is provided by the Medical Research Council; the views expressed in the publication are those of the authors and not necessarily those of the Medical Research Council. The Database was originally established through funding from the Department of Health; additional support has been provided by Janssen-Cilag Ltd, Bristol-Myers Squibb, Gilead, Pfizer, and Roche. The authors declare that no commercial source had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: geretti@liverpool.ac.uk

Introduction

The dynamics of the HIV epidemic are changing in the United Kingdom (UK). By the end of 2010, approximately 91,500 people were estimated to be living with HIV, including 40,100 men who have sex with men (MSM), 47,000 heterosexual men and women and 2,300 injecting drug users (IDUs) [1]. In previous years, heterosexual infections, which were mostly imported, had overtaken infections in MSM, 81% of which are indigenously acquired. However, the trend has now reversed, reflecting a decline in the number of infections acquired abroad, and a corresponding increase in the number of infections acquired in the UK [1]. These changes have considerable potential to modify established epidemic patterns.

Mapping the molecular epidemiology of HIV infection can provide valuable insights into transmission networks and thereby inform prevention and containment strategies.

As seen in other Western countries, including Italy and the United States [2,3], in the UK the HIV epidemic among MSM was traditionally composed of nearly uniformly subtype B infections, contrasting with the variety of non-B subtypes found in heterosexual infections, which are predominantly imported from sub-Saharan Africa [4,5]. Non-B infections have been increasing in recent years in MSM [2,6,7]. Under favorable conditions such as those found in populations at risk of multiple HIV exposures, novel recombinant strains can emerge and become established, supplanting previous patterns of infection. The emergence of recombinant HIV-1 strains among UK MSM was proposed by Gifford *et al.* in 2007, based upon the detection of a potentially novel subtype A recombinant in phylogenetic analyses of *pol* gene sequences [6].

The aim of this study was to seek firmer evidence that novel recombinant forms of HIV are emerging in the UK MSM population. Through the screening of a large national database containing reverse transcriptase and protease sequences from patients undergoing drug resistance testing in routine care, and subsequent near full-length, single genome sequencing (SGS) of clinical isolates, we identified a novel A1/D circulating recombinant form, now registered as CRF50_A1D, which first emerged in the UK around 1992. We show that over time, CRF50_A1D spread geographically and entered heterosexual and IDU transmission networks, followed by recombination with a subtype B strain and emergence of the unique recombinant form (URF) CRF50/B/U. These findings indicate that the two previously distinct HIV epidemics in MSM and heterosexuals have started to converge in the UK, creating opportunities for greater HIV genetic diversification.

Methods

Study population

The UK HIV Drug Resistance Database (HIV-DRD) (http://www.hivrdb.org.uk/) and Public Health England (previously the Health Protection Agency) provided access to *pol* sequences and demographic and clinical data. The HIV-DRD is a national repository of protease and reverse transcriptase sequences obtained by Sanger sequencing in patients undergoing drug resistance testing in routine care. At the time of the analysis, there were 55,556 sequences in the database from both antiretroviral treatment (ART)-naïve and ART-experienced patients. Based upon data from Gifford et al. indicating that a potentially novel subtype A (sub-subtype A1) recombinant was circulating among MSM in the UK [6], sequences from patients infected with sub-subtype A1 were selected from the database for further analysis. The REGA Subtyping tool and bootscanning analysis using Simplot v3.5.1 were used to subtype recombinant sequences [8]. Subsequent subtyping was performed using Subtype Classification Using Evolutionary Algorithms (SCUEAL) [9]. Stored plasma samples from six selected patients were retrieved for further sequence analysis.

Ethics statement

The Ethics Committee of the Royal Free Hospital in London approved the anonymized use of stored plasma samples collected during routine care. Personnel from the HIV-DRD selected patients with different identifiers that attended centers with more than 1000 patients in follow-up and were not known to be related, and communicated the HIV-DRD identifier to the center of care to allow sample retrieval from storage. Samples were anonymized prior to shipment to the laboratory for sequencing.

Near full-length single genome sequencing

Near full-length SGS was performed using a protocol adapted from the Centre for HIV/AIDS Vaccine Immunology (CHAVI-MBSC 2009, unpublished) and optimised for plasma specimens with low HIV-1 RNA levels. Briefly, 140 μl of plasma was adjusted through either dilution or centrifugation to contain 20,000 HIV-1 RNA copies. RNA was extracted using the QiAmp Viral RNA Mini kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions, with a final elution volume of 65 μl. All extracted RNA was immediately transcribed with Superscript III First Strand Synthesis Supermix (Life Technologies, Paisley, UK) using the following protocol per reaction: 0.25 μM of reverse primer 1.R3.B3R 5′-ACTACTTGAAGCACTCAAGGCAAGCTTTATTG (CHAVI-MBSC 2009, unpublished), 1.87 μl nuclease free water, 2.5 μl annealing buffer and 15 μl (5,000 copies) RNA template were denatured at 65°C for 5 minutes. Reactions were placed on ice for at least 1 minute before adding 25 μl of reaction mixture and 5 μl of enzyme mixture. Reverse transcription was performed at 50°C 90 minutes, 55°C 90 minutes, 85°C 5 minutes. An extra 2 μl of Superscript III RT Enzyme was added prior to increasing the temperature to 55°C.

A two-step protocol was used to ensure single genome amplification. The first step was a limiting dilution to assess that cDNA was amplifying at rates suggesting near-ideal extraction and reverse transcription conditions, i.e. that the undiluted cDNA contained 100 copies/μl. Following this, cDNA was amplified using a 1:200 cDNA dilution (a theoretical input of 1 copy/reaction). In circumstances where the extracted number of HIV-1 RNA copies was below 20,000 copies, or the limiting dilution plate suggested that the sample was amplifying suboptimally, dilutions were adjusted accordingly.

Both first round and nested PCR reactions used the Platinum PCR Supermix High Fidelity Kit (Life Technologies, Paisley, UK). First round reactions comprised 45 μl PCR supermix, 0.25 μM each forward primer 1.U5.B1F 5′CCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGT and reverse primer 1.R3.B3R, 0.5 μl nuclease free water and 2 μl cDNA. Cycling conditions were 94°C 2 minutes, followed by 40 cycles of 94°C 15 s, 60°C 30 s, 68°C 9.5 m, and a final extension at 68°C 20 minutes. Nested PCR reactions were identical to the first round reactions, excepting the use of forward primer 2.U5.B4F 5′-AGTAGTGTGTGCCCGTCTGTTGTGTGACTC, reverse primer 2.R3.B6R 5′-TGAAGCACTCAAGGCAAGCTTTATTGAGGC, and 45 cycles of PCR. The resulting 9 kb product spanned HXB2 nucleotides 552–9636.

Positive nested PCR reactions were identified using 1% agarose gel electrophoresis. Filtered PCR products were directly sequenced using fluorescently labeled dideoxy chain terminators (BigDye Terminator v3.1 Cycle Sequencing Assay, Life Technologies) and an automated ABI 3730×l sequencer. Sequencing primers were either sourced from the in-house protocols of the Molecular Biology and Sequencing Core at the Centre for HIV/AIDS Vaccine Immunology or protocols available in published literature [10–12], or designed using Primer3 version 4.0 (http://frodo.wi.mit.edu/primer3) (Supplementary information 1). Sequencing reactions were repeated until near full bi-directional coverage was obtained, and sequences were assembled using SeqScape version 2.6 software (Life Technologies). Fragment sequences from individual sequencing primers were examined for mixed bases; where evidence of amplification of >1 target molecule was found, amplification and sequencing was repeated.

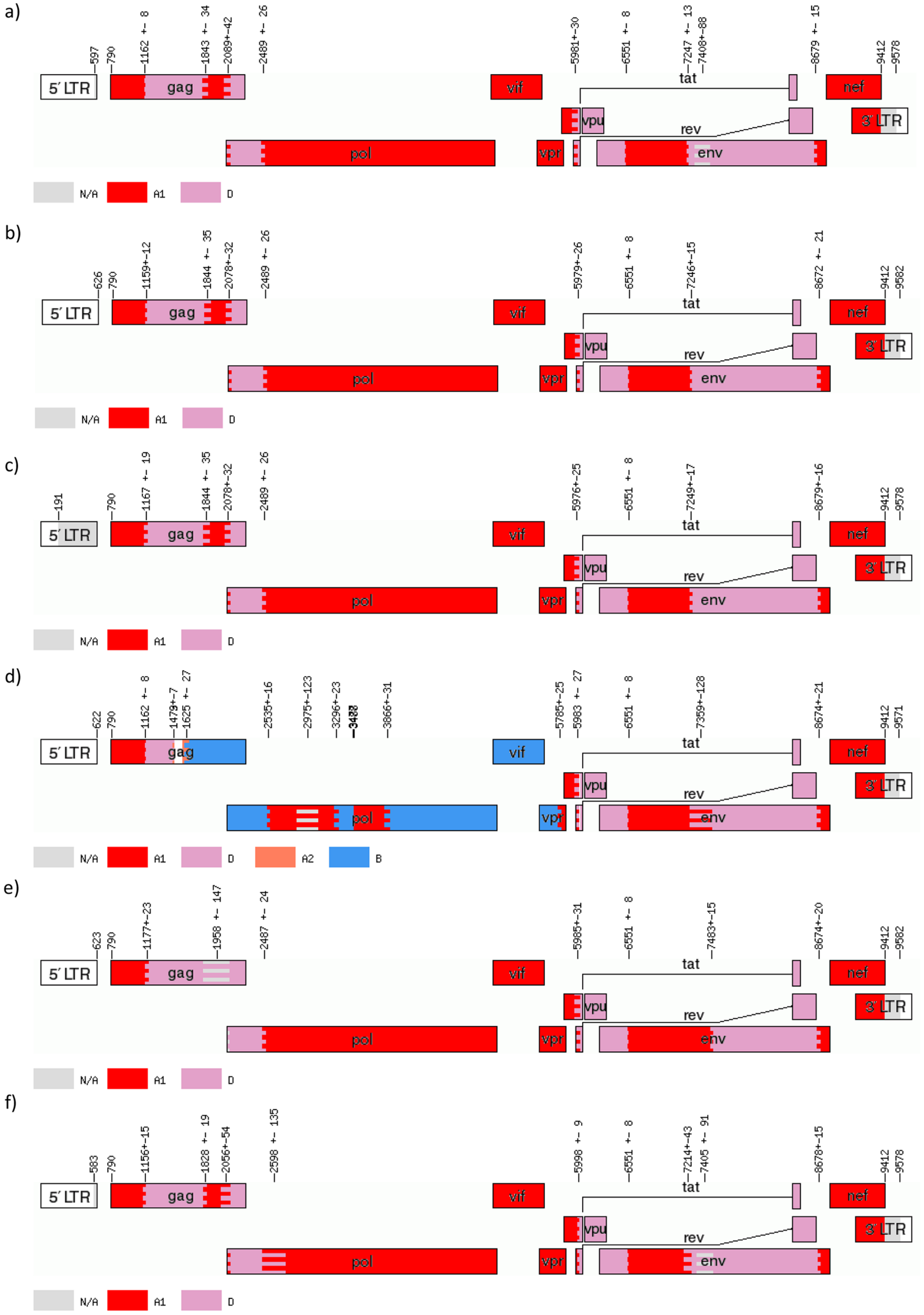

a)
597
790
1162 +- 8
1843 +- 34
2089+-42
2489 +- 26
5981+-30
6551 +- 8
7247 +- 13
7408+-88
8679 +- 15
9412
9578
5′ LTR
gag
pol
vif
vpr
tat
vpu
rev
env
nef
3′ LTR
N/A
A1
D
b)
626
790
1159+-12
1844 +- 35
2078+-32
2489 +- 26
5979+-26
6551 +- 8
7246+-15
8672 +- 21
9412
9582
5′ LTR
gag
pol
vif
vpr
tat
vpu
rev
env
nef
3′ LTR
N/A
A1
D
c)
191
790
1167 +- 19
1844 +- 35
2078+-32
2489 +- 26
5976+-25
6551 +- 8
7249+-17
8679+-16
9412
9578
5′ LTR
gag
pol
vif
vpr
tat
vpu
rev
env
nef
3′ LTR
N/A
A1
D
d)
622
790
1162 +- 8
1479+-7
1625 +- 27
2535+-16
2975+-123
3296+-23
3438
3866+-31
5785+-25
5983 +- 27
6551 +- 8
7359+-128
8674+-21
9412
9571
5′ LTR
gag
pol
vif
vpr
tat
vpu
rev
env
nef
3′ LTR
N/A
A1
D
A2
B
e)
623
790
1177+-23
1958 +- 147
2487 +- 24
5985+-31
6551 +- 8
7483+-15
8674+-20
9412
9582
5′ LTR
gag
pol
vif
vpr
tat
vpu
rev
env
nef
3′ LTR
N/A
A1
D
f)
583
790
1156+-15
1828 +- 19
2056+-54
2598 +- 135
5998 +- 9
6551 +- 8
7214+-43
7405 +- 91
8678+-15
9412
9578
5′ LTR
gag
pol
vif
vpr
tat
vpu
rev
env
nef
3′ LTR
N/A
A1
D

Figure 1. jpHMM analysis of six recombinant HIV-1 sequences. Putative recombinant HIV-1 sequences were submitted to the online implementation of jpHMM at the GOBICS server. The program used its own stored reference alignment and statistical algorithm to determine subtype classifications, breakpoint locations and 95% confidence intervals. Breakpoint locations and confidence intervals are marked on each plot and are equivalent to HXB2 numbering. In each plot, subtype A1 is represented in red, subtype A2 in coral (plot d only), subtype D in lavender, and subtype B in blue (plot d only). Areas of subtype uncertainty are grey. Five specimens (a, b, c, e, and f) showed largely identical A1/D structures, whereas one specimen (d) showed a complex A1/A2/D/B/U structure. a) Specimen 33365; b) Specimen 8179; c) Specimen 40534; d) Specimen 34567; e) Specimen 11762; f) Specimen 12792.
doi:10.1371/journal.pone.0083337.g001

Phylogenetic and recombination analyses

Recombination analyses and subtype assignation was performed using the Recombinant Identification Program (RIP) (http://www.hiv.lanl.gov/content/sequence/RIP/RIP.html), jpHMM (http://jphmm.gobics.de) and Simplot [13]. For RIP and Simplot analyses, a window size of 400 bp and a step size of 20 were used. Sequences were gap-stripped and genetic distances were calculated using Kimura 2-p parameters. Simplot analyses were performed using a full-length reference alignment of 78 pure subtype sequences. Bootscanning of the query sequence was performed using subtypes A1, B, D, and F2 with informative sites analysis. Recombination breakpoints were set using the highest statistically significant X^2 value around the 50% crossover point between subtypes. The statistical significance of the identified breakpoints was assessed using Fisher's exact test. Following breakpoint assignment, slices of the alignment corresponding to putative pure subtype regions between each breakpoint were created and saved for downstream analyses. Likelihood mapping of each slice was used to assess phylogenetic signal prior to maximum likelihood analysis and was performed using TreePuzzle [14].

Likelihood parameters for each putatively pure subtype region of the HIV genome were estimated using PAUP version 4.0 (Sinauer Associates, Massachusetts, USA). Maximum likelihood analysis was performed using the PhyML implementation housed at the ATGC server (http://www.atgc-montpellier.fr/phyml/). 1000 bootstrapping replicates were performed, with the exception of alignment slices 5 and 8, which were restricted to 100 replicates to limit computational requirements.

Phylogenetic trees were visualized using Dendroscope version 2.3 (available from http://ab.inf.uni-tuebingen.de/data/software/dendroscope3/download/welcome.html) and FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). Schematics of finalized recombinant structures were drawn using the Recombinant HIV-1 Drawing Tool (RDT), available from the Los Alamos website (http://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html).

Beyond full-length sequencing, further instances of CRF50_A1D infections were identified using BLAST to compare three representative CRF50_A1D sequences (33365, 8179, 40534) to the sequences contained in the HIV-DRD. The top 500 hits for each sequence were analyzed for recombination profiles and breakpoints using jpHMM and SCUEAL. Sequences with identical subtype classifications and with a jpHMM breakpoint that fell within the SCUEAL 95% confidence interval were considered CRF50_A1D matches for further investigation.

The likely global origin of the parental subtype A1 and D strains of CRF50 was investigated using global subtype alignments containing subtype A1 or D sequences from every country in the Los Alamos National HIV Database with a greater than 10% representation of either subtype. These sequences were selected by geographical region only and no further data was sought. One alignment for each subtype was generated for partial *gag*, *pol*, and *env* genes. The *pol* gene trees were supplemented with pure subtype A and D sequences from the HIV-DRD. Approximate maximum likelihood analysis was performed using FastTree 2.1 using a GTR+CAT model (http://meta.microbesonline.org/fasttree/).

The emergence and distribution of identified CRF50_A1D sequences in the UK was analyzed using time-scaled analyses implemented in BEAST. The 72 putative CRF50_A1D sequences were aligned with 8 reference A1 and D sequences from East Africa, the 4 closest sequence matches in the NCBI database, and 4 subtype C sequences as an outgroup. A total of 3× MCMC runs of 1×10^8 states were performed and combined for each analysis. The GTR+Γ nucleotide model was used with a relaxed, log-normal molecular clock and a Bayesian skyline coalescent with a constant population distribution and 10 skyline groups. For discrete phylogeographic analyses phylogeographic operators as detailed in (http://beast.bio.ed.ac.uk/Discrete_Phylogeographic_Analysis) were used with a resampled time-scaled tree as input. In order to preserve patient anonymity, the locations of individual clinics were not used as inputs into the phylogeographic analysis. Instead, the geographic location for patients was determined using aggregated center data which groups clinics together in approximate locations; the central latitude/longitude point of each aggregate was used as patient location. Following BEAST analysis, phylogeographic trees were converted to .kml format and visualized in Google Earth.

The A1/D recombinant structure was registered with the Los Alamos National Database as CRF50_A1D. All six full-length sequences were submitted to Genbank (accession numbers: JN417236-JN417241); the reference sequence for CRF50_A1D is JN417236.

Results

Identification and amplification of putative novel recombinant sequences

Following screening of 55,556 HIV-1 *pol* gene sequences in the HIV-DRD, sequences from eight subjects were identified that appeared to share a novel recombinant structure. Stored plasma samples from six of these eight subjects were retrieved from three centers in the UK. The samples had been collected between 2000 and 2011 and stored at −80°C under routine conditions. The HIV-1 RNA load measured at the time of sample collection ranged from 9,148 to 500,000 copies/ml and the available sample volumes ranged from 270 to 1500 µl. The optimal 20,000 HIV-1 RNA copies for sequencing were recovered from three of the six samples; all six specimens, however, were successfully amplified at lower than the 30% Poisson distribution set-point for single genome amplification following limiting dilution.

Recombination analyses

RIP analysis of six specimens showed a putatively identical A1/D structure with five of the six clinical isolates analyzed (33365, 8179, 40534, 11762, 12792); the sixth isolate (34567) showed a complex A1/B/D structure (data not shown). jpHMM analysis similarly identified five isolates with largely identical A1/D structures (33365, 8179, 40543, 11762, 12792) and one isolate with a complex A1/A2/D/B/U structure (34567) (Figure 1). The

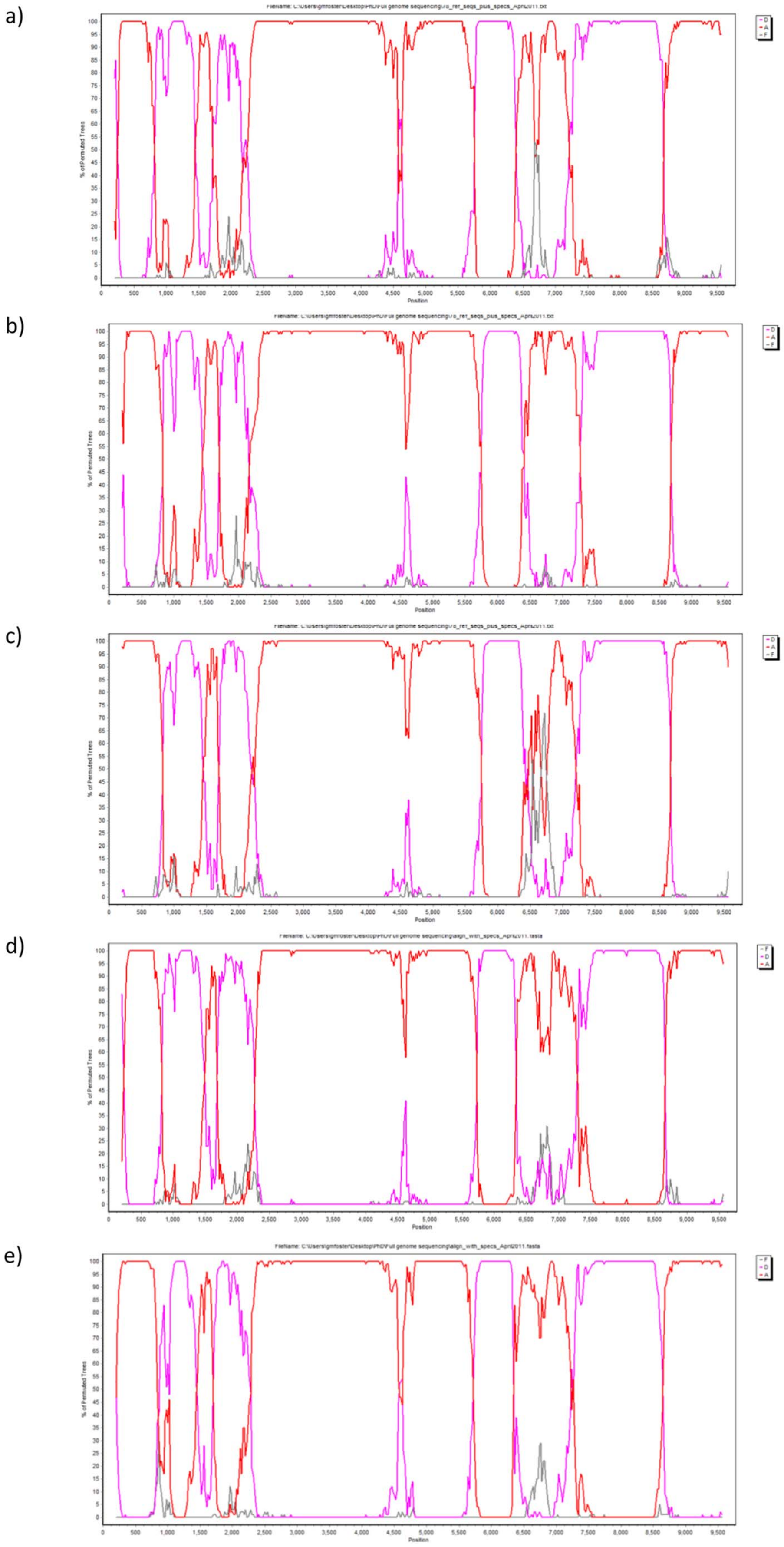

a)
FileName: C:\Users\gmfoster\Desktop\PhD\Full genome sequencing\78_ref_seqs_plus_specs_April2011.txt
b)
FileName: C:\Users\gmfoster\Desktop\PhD\Full genome sequencing\78_ref_seqs_plus_specs_April2011.txt
c)
FileName: C:\Users\gmfoster\Desktop\PhD\Full genome sequencing\78_ref_seqs_plus_specs_April2011.txt
d)
FileName: C:\Users\gmfoster\Desktop\PhD\Full genome sequencing\align_with_specs_April2011.fasta
e)
FileName: C:\Users\gmfoster\Desktop\PhD\Full genome sequencing\align_with_specs_April2011.fasta
% of Permuted Trees
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Position
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500
6,000
6,500
7,000
7,500
8,000
8,500
9,000
9,500
D
A
F

Figure 2. Bootscanning plots for five A1/D recombinants. Bootscanning plots from Simplot sliding window analysis using a window size of 400 bp, a step size of 20 bp and 100 bootscanning replicates. The y axis shows the percentage of permuted trees that the query sequence clustered with the closest subtype match from the reference alignment. The x axis shows the nucleotide position of the sequence (not HXB2 numbering). Subtype A is represented in red, subtype D in lavender, and subtype F (outgroup) in grey. All five specimens (33365, 8179, 40534, 11762, 12792) showed identical bootscanning plots, with five subtype A1 regions and four subtype D regions. A) Specimen 33365; b) Specimen 8179; c) Specimen 40534; d) Specimen 11762; e) Specimen 12792.
doi:10.1371/journal.pone.0083337.g002

breakpoint locations for the five A1/D specimens are summarized in Table 1. Generally, the jpHMM breakpoint locations and subtype classifications showed a good level of consistency among the five A1/D specimens, and with the structure suggested by the RIP screening. Two potential structural discrepancies were suggested by jpHMM. With specimen 11762, the p2–p7 regions of *gag* showed a lower degree of subtype A1 identity than observed with the other four A1/D specimens; however, overlapping confidence intervals indicated that the uncertainty was unlikely to reflect a true structural difference. With specimens 33365 and 12792, two regions of *env* were designated as subtype D/uncertain in the jpHMM plots; however bootscanning of these regions confirmed the subtype D classification (Figure 2).

Breakpoints identified using bootscanning and informative sites analyses were consistent with those identified using jpHMM. All five A1/D specimens (33365, 8179, 40534, 11762, 12792) showed identical bootscanning plots, with five subtype A1 regions and four subtype D regions.

The jpHMM analysis of specimen 34567 further clarified the recombinant structure of this complex isolate. Two clear regions with the same structure as the five A1/D specimens were identified, at the very beginning of *gag*, which had an identical A1/D breakpoint (1162±8), and from the breakpoint in *tat/rev* (5983±23) to the end of the genome. This suggested that this specimen resulted from a further recombination event between the A1/D recombinant and a subtype B strain.

Maximum likelihood analyses

Maximum likelihood trees of putative non-recombinant fragments drawn using PhyML with PAUP-defined parameters showed that each fragment of each specimen clustered with the pure subtype (A1 or D) indicated by the bootscanning analysis. Results obtained with A1/D specimens 33365, 8179, and 40534, and complex specimen 34567 are shown in Figure 3. The A1/D structure was predominantly subtype A1 in *pol* and the accessory genes; subtype D in *env*; and fairly evenly split between subtype A1 and D in *gag*. Three breakpoints were located in *gag*, one in *pol*, one in *tat/rev*, and three in *env*, respectively. In *gag*, a breakpoint was located at either end of p24, suggesting that the entire coding region for the antigen was swapped in the recombination event. Similarly, the third breakpoint was located at the junction of the p7/p1 regions, suggesting that entire coding regions were swapped in the recombination event. The distribution of subtypes in *gag* by protein was A1 (p17, p2, p7) and D (p24, p1, p6).

The single D/A1 breakpoint in *pol* was located approximately 250 bp from the start of the protease; the remainder of the *pol* gene was subtype A1, as were *vif* and *vpr*. The breakpoint located at HXB2 6007 fell in the overlap of *tat* and *rev*; both of these genes were A1/D mosaics. *Vpu* was solely subtype D. Although *env* was largely subtype D, three of the hypervariable regions (V1–V3) were subtype A1.

The maximum likelihood analysis confirmed that the A1/D isolates clustered monophyletically across the entire genome (Figure 3). The complex isolate 34567 clustered with the A1/D isolates in 7/9 genomic regions; in 2/9 regions this specimen clustered with subtype B reference sequences, confirming that this specimen was a recombinant of the A1/D structure and a subtype B infection (Figures 3 and 4).

Emergence and distribution of CRF50_A1D

Analysis of the UK HIV-DRD identified a further 67 sequences showing a recombination profile that matched that of CRF50_A1D. The global approximate maximum likelihood trees were built using sequences from East Africa (Kenya, Tanzania, Rwanda, Uganda, Burundi), Central Africa (DRC), Western Africa (Cameroon), Eastern Europe (Latvia, Belarus, Georgia, Russia) and the UK, due to the prevalence of subtypes A and D in these regions. In the approximate maximum likelihood analysis of global alignments of *gag, pol and env* gene subtype A and D sequences the CRF50 sequences clustered monophyletically with

Table 1. jpHMM-assigned breakpoint locations (with 95% confidence intervals) for five HIV-1 A1/D recombinant sequences[a].

Break point	Study number					Gene	Region
	33365	8179	40534	11762	12792		
1	1162 (1154–1170)	1159 (1147–1171)	1167 (1148–1186)	1177 (1154–1200)	1156 (1141–1171)	*gag*	p24
2	1843 (1809–1877)	1844 (1809–1879)	1844 (1809–1879)	1958* (1811–2105)	1828 (1809–1847)	*gag*	p24
3	2089 (2047–2131)	2078 (2046–2110)	2078 (2046–2110)	2078 (2046–2110)	2056 (2002–2110)	*gag*	p1
4	2489 (2463–2515)	2489 (2465–2515)	2489 (2465–2515)	2487 (2475–2499)	2598† (2463–2733)	*pol*	PR
5	5981 (5951–6011)	5979 (5953–6005)	5976 (5951–6001)	5985 (5973–5997)	5998 (5989–6007)	*tat/rev*	
6	6551 (6543–6559)	6551 (6543–6559)	6551 (6543–6559)	6551 (6539–6563)	6551 (6543–6559)	*env*	gp120
7	7247 (7234–7260)	7246 (7231–7261)	7249 (7232–7266)	7483 (7471–7495)	7214 (7171–7257)	*env*	gp120
8	8679 (8664–8694)	8672 (8651–8693)	8679 (8663–8695)	8674 (8662–8686)	8678 (8663–8693)	*env*	gp41

[a]Breakpoint locations as determined by jpHMM with HXB2 numbering. The breakpoint locations are generally consistent across the five specimens, indicating that the same A1/D recombinant structure is shared.
*This corresponds to a region of subtype D uncertainty;
†This corresponds to a region of subtype A uncertainty; refer to Figure 1. PR = Protease.
doi:10.1371/journal.pone.0083337.t001

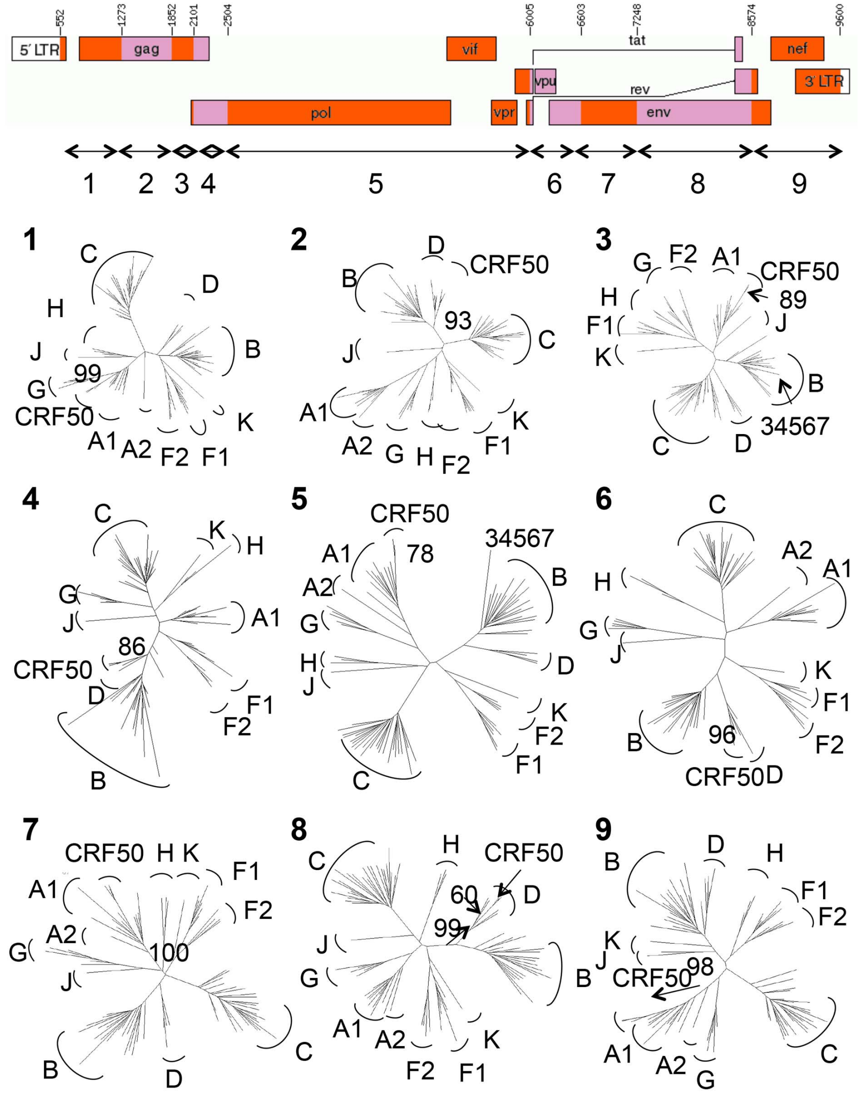


Figure 3. Recombinant map of CRF50_A1D and maximum likelihood phylogenetic trees of non-recombinant fragments. Maximum likelihood trees of putative non-recombinant fragments from specimens 33365, 8179, 40534 and 34567 drawn using PhyML with PAUP-defined parameters. HIV-1 subtypes used for analysis were A–D, F, G, H, J, K. Numbers indicate bootstrapping support from 1000 replicates (excepting slice 5; 100 replicates). 70% bootstrap support and monophyletic clustering were the criteria for subtype classification. The recombinant map was drawn using the RDT program at Los Alamos. Component subtype fragments are labeled 1–9 on the genome map, corresponding with numbered phylogenetic trees. The genomic regions in which the URF specimen 34567 did not cluster with the CRF50_A1D specimens are indicated in the appropriate trees.

doi:10.1371/journal.pone.0083337.g003

the East African sequences in both the subtype A and subtype D trees (data not shown). No clustering was observed with subtype A1 or D sequences from the UK. This suggested that CRF0_A1D probably originated in East Africa and was possibly introduced to the UK as a recombinant, rather than emerging from subtype A1 and D strains circulating in the UK.

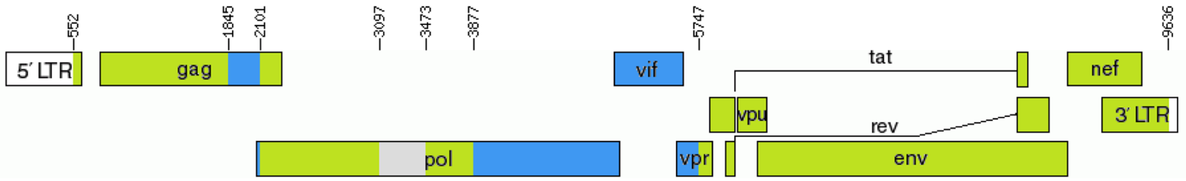


Figure 4. Confirmed structure of the complex recombinant. The confirmed structure of the complex A1/B/D/U recombinant specimen 34567 following maximum likelihood analysis with the CRF50_A1D specimens. CRF50_A1D regions are shown in green and subtype B regions are shown in blue.

doi:10.1371/journal.pone.0083337.g004

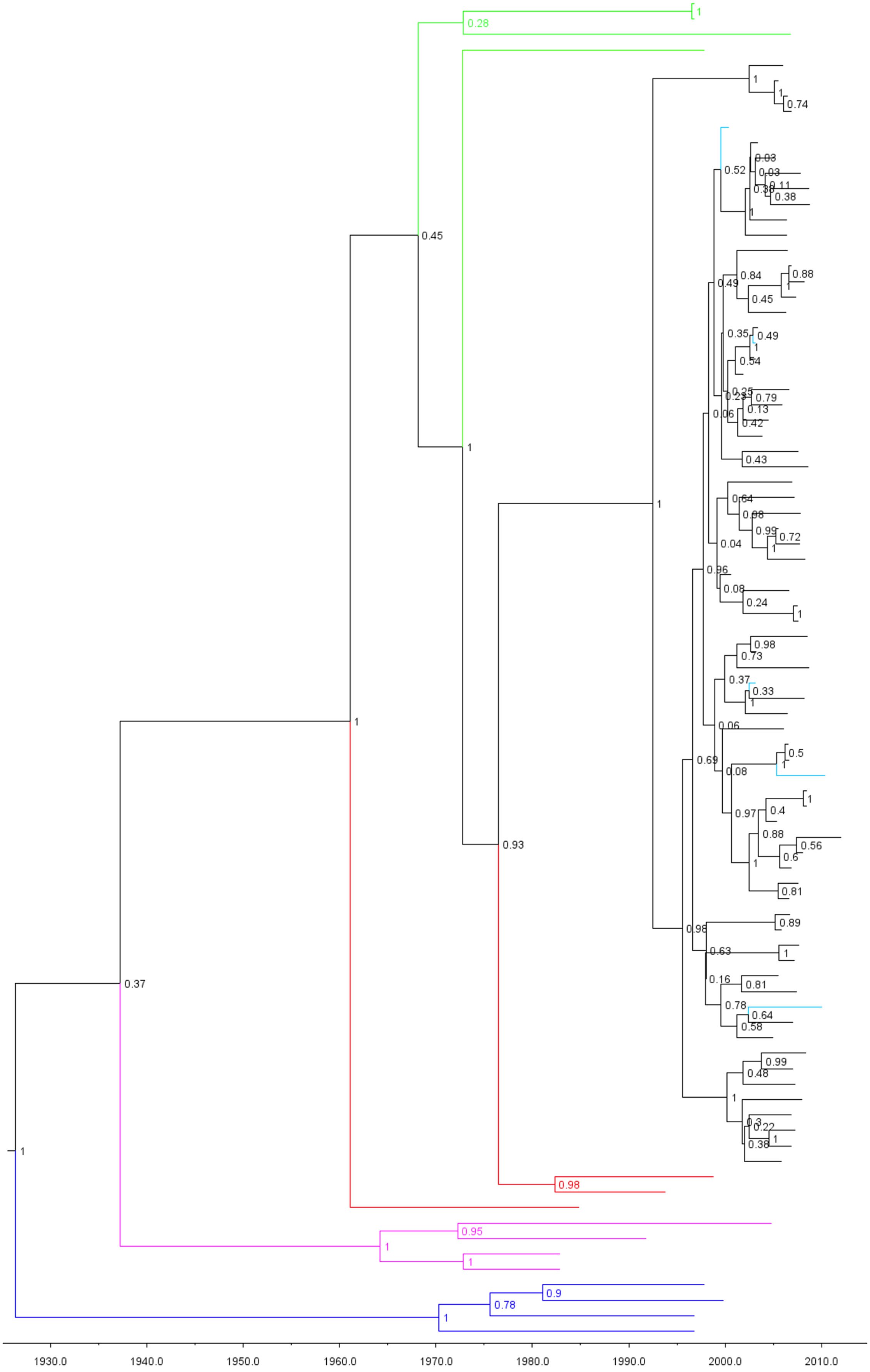

1
0.28
1
1
0.74
0.03
0.52
0.03
0.11
0.38
0.38
1
0.45
0.84
0.88
0.49
1
0.45
0.35
0.49
1
0.54
0.25
0.23
0.79
0.13
0.06
0.42
1
0.43
0.64
0.98
0.99
0.72
0.04
1
0.96
0.08
0.24
1
0.98
0.73
0.37
0.33
1
0.06
1
0.5
0.69
0.08
1
0.97
0.4
0.88
0.56
0.93
0.6
1
0.81
0.89
0.98
0.63
1
0.16
0.81
0.78
0.64
0.58
0.99
0.48
1
0.3
0.22
0.38
1
0.37
1
0.98
0.95
1
1
0.9
0.78
1
1930.0
1940.0
1950.0
1960.0
1970.0
1980.0
1990.0
2000.0
2010.0

Figure 5. Emergence of CRF50_A1D. Time-scaled analysis of CRF50_A1D *pol* gene sequences using BEAST. Green sequences indicate the closest NCBI BLAST matches to the CRF50 sequences (all subtype A1); red sequences indicate A1 reference sequences; purple sequences indicate subtype D reference sequences. Outgroup subtype C sequences are shown in blue. Node values indicate the posterior probability of each node. The five full-length genomes labeled in turquoise. The emergence of CRF50_A1D in Britain is dated to 1992.
doi:10.1371/journal.pone.0083337.g005

Time-scaled analysis dated the emergence of CRF50_A1D in the UK to 1992 [95% highest posterior density (HPD) 1966–2007; posterior probability 0.9933] (Figure 5). Phylogeographic analysis showed probable emergence in northwest England followed by spread to London and southeast England, with further, limited transmission events in southwest and northeast England and Scotland. Demographic information was available for 51/72 patients infected with CRF50_A1D. Of these, 45 (88.2%) were MSM, 3 (5.9%) were heterosexual males, and 2 (3.9%) were IDUs.

Analysis including the CRF50/B/U unique recombinant form (URF) sequence showed a median divergence date of 2000, indicating an onward recombination event between CRF50_A1D and a subtype B strain. The CRF50_A1D/B/U sequence came from an MSM.

Discussion

The HIV epidemic in MSM in the UK continues to diversify, creating opportunities for the emergence of novel recombinant forms. By scanning a large national sequence repository, we identified 72 patients who all appeared to be infected with the same novel A1/D recombinant. Near full-length SGS of plasma HIV-1 RNA was performed to characterize the structure of the recombinant. Five patients were found to carry the same A1/D recombinant, which was classified as CRF50_A1D. Based on the recombinant profile, we conclude that CRF50_A1D is the subtype A recombinant that Gifford *et al* hypothesized was circulating among MSM in 2007 [6]. It should be noted that some recombination breakpoints were not identical among the five CRF50_A1D isolates in the jpHMM plots. However, the confidence intervals of the identity estimations and the subsequent analyses indicated that the uncertainties were unlikely to reflect a true alternative recombinant structure. We also found evidence of further genetic evolution of CRF50_A1D through recombination with subtype B, which is the predominant HIV-1 subtype circulating among MSM in the UK. This complex URF was classified as CRF50_A1D/B/U. Crucially, the estimated emergence date of 1992 was both prior to the introduction of highly active antiretroviral therapy and during a period when HIV infections were spreading exponentially in African countries, creating ideal conditions for the creation of novel HIV recombinants which could move into the wider epidemic.

We found a relatively low number of patients infected with this strain within a database that at the time of screening contained 55,556 sequences from 43,002 patients. This relatively modest spread could reflect fitness properties of the CRF. We detected an unusual structure of the *env* gene in this recombinant, in which three out of the five hypervariable regions belonged to subtype A1, whereas the remaining two regions belonged to subtype D. Available data indicate that recombination events in *env* tend to include either the entire gene or at least the entirety of gp120, and this has been related to the functional impact of this protein on viral fitness [15].

We found that CRF50_A1D was related to A1 and D strains of East African origin. A1/D recombinants detected in East Africa have been associated with a fast disease progression, which may limit the number of infections in the community [16–19]. It should be noted however that in a sub-analysis, the CD4 cell count slopes before starting ART were similar in MSM infected with subtype B or CRF50_A1D (data not shown). These considerations indicate that CRF50_A1D has potentially interesting phenotypic properties, which would bear further investigation. Further studies are required to indicate whether there is an influence on clinical outcomes or treatment responses.

There are limitations to this study. Our phylogeographic approach dated the emergence of CRF50_A1D in the UK to mid-1992. This study had a limited number of sequences with which to draw this inference. The prevalence of CRF50_A1D was low in the dataset (72/43,002 or 0.17%) with no evidence for rate increase over time. While the UK-DRB comprehensively collects *pol* gene sequences from patients undergoing drug resistance in routine care in the UK, not all HIV centers contribute to the dataset. Furthermore, the database contains only protease and reverse transcriptase sequences and there are similarity between subtype B and subtype D in these genetic regions. Thus it may be proposed that the 72 CRF50_A1D infections identified represent an underestimate. This in turn may potentially bias the estimated date of emergence. Evolutionary analysis of the individual gene using the available full-length sequences and reconstruction of the ancestral subtype A1 and D strains may yield a more precise elucidation of the emergence date and help to determine whether single or multiple introductions occurred in the UK. Furthermore, given that the majority of 72 individuals infected with CRF50_A1D had only partial *pol* sequences available for analysis, it may be postulated that some of these cases may have shown a more complex viral genomic structure if full-length genome analysis had been performed.

The study of novel HIV variants such as CRF50_A1D and the URF CRF50/B/U provides a tool for studying transmission networks and interactions between populations and risk groups, thus producing valuable epidemiological insights [20,21]. Although in the early years of the HIV epidemic in Western Europe it was rare to find non-B infections in MSM [22], more recent data indicate that non-B infections are not only increasingly important, but are being transmitted indigenously among this population [23]. The use of molecular epidemiological techniques to map these variants can add to our understanding of data gathered using traditional epidemiological means and provides valuable insights into the dynamics of the HIV epidemic that can be used to guide control strategies.

Supporting Information

Table S1 Sequencing primers used for near full-length HIV-1 single genome sequencing.
(DOCX)

Acknowledgments

Steering Committee

Celia Aitken, Gartnavel General Hospital, Glasgow; David Asboe, Anton Pozniak, Chelsea & Westminster Hospital, London; Daniel Webster, Royal Free NHS Trust, London; Patricia Cane, Health Protection Agency, Porton Down; Hannah Castro, David Dunn, David Dolling, Esther Fearnhill, Kholoud Porter, MRC Clinical Trials Unit, London; David Chadwick, South Tees Hospitals NHS Trust, Middlesbrough; Duncan Churchill, Brighton and Sussex University Hospitals NHS Trust; Duncan Clark, St Bartholomew's and The London NHS Trust;

Simon Collins, HIV i-Base, London; Valerie Delpech, Health Protection Agency, Centre for Infections, London; Anna Maria Geretti, University of Liverpool; David Goldberg, Health Protection Scotland, Glasgow; Antony Hale, Leeds Teaching Hospitals NHS Trust; Stéphane Hué, University College London; Steve Kaye, Imperial College London; Paul Kellam, Wellcome Trust Sanger Institute & UCL Medical School; Linda Lazarus, Expert Advisory Group on AIDS Secretariat, Health Protection Agency, London; Andrew Leigh-Brown, University of Edinburgh; Nicola Mackie, Imperial NHS Trust; Chloe Orkin, St. Bartholomew's Hospital, London; Philip Rice, St George's Healthcare Trust, London; Deenan Pillay, Andrew Phillips, Caroline Sabin, University College London Medical School; Erasmus Smit, Health Protection Agency, Birmingham Heartlands Hospital; Kate Templeton, Royal Infirmary of Edinburgh; Peter Tilston, Manchester Royal Infirmary; William Tong, Guy's and St. Thomas' NHS Foundation Trust, London; Ian Williams, Mortimer Market Centre, London; Hongyi Zhang, Addenbrooke's Hospital, Cambridge; Mark Zuckerman, King's College Hospital, London.

Centres contributing data

Clinical Microbiology and Public Health Laboratory, Addenbrooke's Hospital, Cambridge (Jane Greatorex); HIV/GUM Research Laboratory, Chelsea and Westminster Hospital, London (Adrian Wildfire); Guy's and St. Thomas' NHS Foundation Trust, London (Siobhan O'Shea, Jane Mullen); HPA – Public Health Laboratory, Birmingham Heartlands Hospital, Birmingham (Erasmus Smit); HPA London (Tamyo Mbisa); Imperial College Health NHS Trust, London (Alison Cox); King's College Hospital, London (Richard Tandy); Medical Microbiology Laboratory, Leeds Teaching Hospitals NHS Trust (Tony Hale, Tracy Fawcett); Specialist Virology Centre, Liverpool (Mark Hopkins, Lynn Ashton); Department of Clinical Virology, Manchester Royal Infirmary, Manchester (Peter Tilston); Department of Virology, Royal Free Hospital, London (Daniel Webster, Ana Garcia-Diaz); Edinburgh Specialist Virology Centre, Royal Infirmary of Edinburgh (Jill Shepherd); Department of Infection & Tropical Medicine, Royal Victoria Infirmary, Newcastle (Matthias L Schmid, Brendan Payne); South Tees Hospitals NHS Trust, Middlesbrough (David Chadwick); St George's Hospital, London (Phillip Hay, Phillip Rice, Mary Paynter); Department of Virology, St Bartholomew's and The London NHS Trust (Duncan Clark, David Bibby); Molecular Diagnostic Unit, Imperial College, London (Steve Kaye); University College London Hospitals (Stuart Kirk); West of Scotland Specialist Virology Lab Gartnavel, Glasgow (Alasdair MacLean, Celia Aitken, Rory Gunson).

Coordinating Centre

Medical Research Council Clinical Trials Unit (MRC CTU), London (Kate Coughlin, David Dolling, David Dunn, Esther Fearnhill, Lorraine Fradette, Kholoud Porter).

Author Contributions

Conceived and designed the experiments: AMG GMF. Performed the experiments: GMF. Analyzed the data: AMG GMF JCA SH VD EF AA ALB. Contributed reagents/materials/analysis tools: VD ALB EF. Wrote the paper: GMF AMG.

References

1. Health Protection Agency (n.d.) Sexually transmitted infections in men who have sex with men in the United Kingdom: 2011 report. Available: http://www.hpa.org.uk/Publications/InfectiousDiseases/HIVAndSTIs/1111STIsinMSMintheUK2011report/. Accessed 6 March 2013.
2. Balotta C, Facchi G, Violin M, Van Dooren S, Cozzi-Lepri A, et al. (2001) Increasing prevalence of non-clade B HIV-1 strains in heterosexual men and women, as monitored by analysis of reverse transcriptase and protease sequences. J Acquir Immune Defic Syndr 27: 499–505.
3. Brodine SK, Garland FC, Mascola JR, Porter KR, Mascola JR, et al. (1995) Detection of diverse HIV-1 genetic subtypes in the USA. The Lancet 346: 1198–1199.
4. Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, et al. (2000) High Rate of recombination throughout the Human Immunodeficiency Virus Type 1 genome. J Virol 74: 1234–1240.
5. Kuiken C, Leitner T, Foley B, Hahn BH, Marx P, et al., editors (n.d.) HIV sequence compendium 2009 Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR 08-03719.
6. Gifford RJ, Oliveira T de, Rambaut A, Pybus OG, Dunn D, et al. (2007) Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of Human Immunodeficiency Virus type 1. J Virol 81: 13050–13056.
7. Buonaguro L, Tornesello ML, Buonaguro FM (2007) Human Immunodeficiency Virus Type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. J Virol 81: 10209–10219.
8. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, et al. (1999) Full-length Human Immunodeficiency Virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 73: 152–160.
9. Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23: 1891–1901.
10. Nadai Y, Eyzaguirre LM, Constantine NT, Sill AM, Cleghorn F, et al. (2008) Protocol for nearly full-length sequencing of HIV-1 RNA from plasma. PLoS ONE 3: e1420.
11. Van Laethem K, Schrooten Y, Lemey P, Wijngaerden EV, Wit SD, et al. (2005) A genotypic resistance assay for the detection of drug resistance in the human immunodeficiency virus type 1 envelope gene. J Virol Meth 123: 25–34.
12. Van Laethem K, Schrooten Y, Dedecker S, Van Heeswijck L, Deforche K, et al. (2006) A genotypic assay for the amplification and sequencing of gag and protease from diverse human immunodeficiency virus type 1 group M subtypes. J Virol Meth 132: 181–186.
13. Schultz A-K, Zhang M, Bulla I, Leitner T, Korber B, et al. (2009) jpHMM: Improving the reliability of recombination prediction in HIV-1. Nucleic Acids Res 37: W647–W651.
14. Schmidt HA, Strimmer K, Vingron M, Haeseler A von (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
15. Archer J, Pinney JW, Fan J, Simon-Loriere E, Arts EJ, et al. (2008) Identifying the important HIV-1 recombination breakpoints. PLoS Comput Biol 4: e1000178.
16. Kaleebu P, French N, Mahe C, Yirrell D, Watera C, et al. (2002) Effect of Human Immunodeficiency Virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1 positive persons in Uganda. J Infect Dis 185: 1244–1250.
17. Kaleebu P, Nankya IL, Yirrell DL, Shafer LA, Kyosiimire-Lugemwa J, et al. (2007) Relation between chemokine receptor use, disease stage, and HIV-1 subtypes A and D. J Acquir Immune Defic Syndr 45: 28–33.
18. Baeten JM, Chohan B, Lavreys L, Chohan V, McClelland RS, et al. (2007) HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. J Infect Dis 195: 1177–1180.
19. Kiwanuka N, Laeyendecker O, Quinn TC, Wawer JM, Shepherd J, et al. (2009) HIV-1 subtypes and differences in heterosexual HIV transmission among HIV-discordant couples in Rakai, Uganda. AIDS 23: 2479–2484.
20. Doherty IA, Padian NS, Marlow C, Aral SO (2005) Determinants and consequences of cexual networks as they affect the spread of sexually transmitted infections. J Infect Dis 191: S42–S54.
21. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ (2008) Episodic Sexual Transmission of HIV Revealed by Molecular Phylodynamics. PLoS Med 5: e50.
22. Boni J, Pyra H, Gebhardt M, Perrin L, Burgisser P, et al. (1999) High Frequency of non-B subtypes in newly diagnosed HIV-1 infections in Switzerland. J Acquir Immune Defic Syndr 1: 174–179.
23. Semaille C, Barin F, Cazein F, Pillonel J, Lot F, et al. (2007) Monitoring the dynamics of the HIV epidemic using assays for recent infection and serotyping among new HIV diagnoses: experience after 2 years in France. J Infect Dis 196: 377–383.