THE UNIVERSITY
of EDINBURGH

Characterisation of the Ugandan HIV epidemic with full-length genome sequence data from 1986 to 2016

Heather E. Grant

Doctor of Philosophy

THE UNIVERSITY OF EDINBURGH

2022

Abstract

Presented here are two large Ugandan HIV genome datasets, one from the modern period (by the MRC/UVRI & LSHTM group and PANGEA), and another generated from stored 1986 serum samples using target-capture next generation sequencing. Uganda uniquely has two HIV subtypes at similar proportions, and although subtype A1 is older than subtype D, both subtypes are well established in all cohorts and risk groups, which has facilitated comparison of the two. Previously, HIV sequence data from East Africa has typically been short gene sequences, thus the full-length genome data here has presented an opportunity to compare the evolutionary histories of the two subtypes across the genome, and carry out an examination of inter-subtype recombination patterns. The majority of the modern HIV genomes are unique recombinant forms (URFs), representing a large number of independent superinfection events, which is consistent with the size and age of the epidemic. There are wide scale patterns of recombination along the genome, which are described. Specifically, the region of envelope from C2 of gp120 to the transmembrane region of gp41 is almost always found intact since disruption of these protein interactions is expected to be highly detrimental. Rediscovered serum samples from 1986 yielded 109 full-length HIV 'historical' genomes. The subtype distribution is shown to significantly change over time: subtype D fell from 67% in 1986 to 17% in the modern PANGEA sample. Furthermore, co-receptor tropism (CXCR4 or CCR5) was predicted with geno2pheno and a significant difference between the historical subtypes was observed: 63% of subtype D genomes are X4 tropic (known to be associated with faster progression to AIDS) whilst 0% of A1 sequences are X4 tropic. Therefore, co-receptor tropism may have reduced the effective reproductive number of subtype D by reducing the duration of potential onward exposure (due to faster time to death) compared with A1, and can explain a drop in subtype D prevalence over time. Finally, BEAST1 methods are applied to reconstruct the demographic histories of the two subtypes over time using *gag*, *pol*, and *env* gene data, and place the subtypes in their wider East African context. These findings characterise a highly diverse and complex epidemic in Uganda that has shifted from predominantly subtype D to predominantly subtype A1 between 1986 and 2016, whilst pervasive and ongoing recombination has generated a wide variety of URFs.

Lay Summary

Over time HIV has formed distinct lineages or ‘subtypes’. When an individual is infected with more than one HIV subtype, genetic material can be exchanged (at any point along the genome) creating hybrid viruses or ‘inter-subtype recombinants’. When a new recombinant virus is generated it may be called a ‘unique recombinant form’ (URF), unless it is transmitted and is seen at least three times in unlinked cases, then it may be designated a ‘circulating recombinant form’ (CRF). Using full-length genomes from the PANGEA project, the modern epidemic in Uganda is composed ’pure’ subtypes A1 and D and their recombinants. There is a conspicuous lack of clear CRF, which points to continual and ongoing inter-subtype recombination, rather than the expansion of any one URF. The location and frequency of detectable recombination events along the genome is described and a section of the envelope gene is found to be a ‘cold-spot’ for recombination. This envelope region is almost always found intact likely because the translated protein undergoes intricate folding, disruption of which would render the virus unable to infect new cells. Highly degraded HIV samples from 1986 were rediscovered in storage and subsequently sequenced with new and highly sensitive techniques to obtain 109 new full-length genomes. The subtype distribution is significantly different to the modern day PANGEA sample, containing a high proportion of subtype D (67%). HIV uses the cell receptor CD4 and a secondary co-receptor to gain entry into cells, most often a receptor called CCR5. However, some viruses can use the co-receptor CXCR4 which is linked to faster progression to AIDS in those patients. The subtype D historical sequences have remarkably high predicted CXCR4 usage (63%) whereas the subtype A1 historical sequences are all predicted to use CCR5, suggesting subtype D infections in the early epidemic progressed faster to AIDS. We then use phylodynamics to examine the two subtypes and show that many of the subtype D lineages present in 1986 are no longer found in the modern day. We propose that subtype D had a shorter time to AIDS, thus decreasing its fitness in the 1990s and 2000s when AIDS education lead to fewer infection opportunities for the virus. These findings characterise a highly diverse and complex epidemic in Uganda that has shifted from a higher proportion of subtype D to a higher proportion of subtype A1 between the 1980s and the modern day, whilst recombination has generated a wide variety of unique recombinant forms, without any clear expansion of any form in particular.

Acknowledgements

Thanks are due first and foremost to my primary supervisor Prof Andy Leigh Brown for his steadfast guidance, eternal patience, and kind support, particularly during covid. Similarly I am also very grateful to my secondary supervisors, Dr Sam Lycett, Prof Rowland Kao, Prof David Robertson, and Dr Emma Hodcroft for their insight, suggestions, and support. I also had the honour of working with great scientists at the Uganda Virus Research Institute in Entebbe, particularly Prof Pontiano Kaleebu, Dr Deo Ssemwanga, Dr Nicholas Bbosa, and Prof Matt Cotton, whose knowledge and input vastly improved this work. I am extremely grateful to Prof Judy Breuer at the Pathogens Genome Unit at University College London for her enthusiasm in facilitating the sequencing of historical samples, to Dr Rachel Williams for explaining the target-capture sequencing method, and Dr Sunando Roy sharing his bioinformatics pipeline. It was an honour to talk to Dr Wilson Carswell on the phone about the patients he treated and the samples collected in 1986, and it is pleasing to know his efforts are proving still fruitful to the scientific community decades later. I gratefully acknowledge the Precision Medicine (MRC) Doctoral Training Program for funding my PhD research, and providing key training and essential practical support. I was very fortunate to receive an award from the Canada-UK Globalink (NERC) Doctoral Training Exchange Program which allowing me the opportunity to work for three months with Prof Art Poon, Dr Abayomi Olabode, and the rest of the Poon Lab at Western University. Even though I couldn't attend in person, the exchange was extremely enjoyable, and provoked many of the ideas expressed this thesis. I would also like to express my gratitude to Dr Jarrod Hadfield who very generously assisted with the mixed effects linear model for recombination breakpoints. It was a special privilege to experience the Ashworth environment, seminars, symposiums, trips to Leslie's, and coffee breaks (before the covid pandemic generally made everything a lot more difficult). It was wonderful to get to know so many fellow students and 'grown ups' alike in the department (and special thanks go to the 'grown ups' who paid for the coffee). To my dear friend and fellow PhD student Dr Lucy Peters, thank you for being the best flatmate and making the lockdowns bearable. It was a pleasure to get to know my lab mates Emma Pujol Hodge and James Baxter who kept me company in our covid 'virtual office'. Last but not least, thank you to my wonderful friends and family who have supported me wholeheartedly throughout, and my parents who looked after me so well whilst I wrote up in the spare room.

Publications as chapters

Chapter 2 describes full-length HIV genome sequences from the MRC/UVRI & LSHTM Uganda Research Unit between 2009-2016. It has been published in *Virus Evolution* on behalf of the PANGEA consortium who facilitated sequencing and assembly.

Grant *et al.*, (2020) Pervasive and non-random recombination in near full-length HIV genomes from Uganda. *Virus Evolution*, 6(1), 1–12. https://doi.org/10.1093/ve/veaa004

Chapter 3 describes new HIV genomes from 1986 in collaboration with the UVRI and UCL. Judith Brueur at the Pathogen Genomes Unit at UCL facilitated next generation sequencing of historical samples, Helena Tutil, Rachel Williams, and Bridget Ferns carried out extraction and sequencing work, and Sunando Roy wrote the initial bioinformatics pipeline. This chapter will be submitted to BioRXV and subsequently *Retrovirology*.

Chapter 4 is a continuation of the long term collaboration with the UVRI on the history of the generalized epidemic in Uganda as can be inferred from sequence data. This work forms a phylodynamic characterisation with use of datasets from chapters 2 and 3, and will be submitted to to BioRXV and subsequently *Virus Evolution*.

Chapter 5 came about from discussions with David Robertson, Art Poon, and Abayomi Olabode. It is intended as an opinion paper for *PLoS Pathogens* or similar after it has been finalised with their input. I also contributed to Olabode et al., 2022 which is discussed.

Abayomi S. Olabode, Garway T. Ng, Kaitlyn E. Wade, Mikhail Salnikov, Heather E. Grant, David W. Dick, Art F. Y. Poon 2022. "Revisiting the recombinant history of HIV-1 group M with dynamic network community detection" *PNAS* 119 (19) e2108815119
https://www.pnas.org/doi/10.1073/pnas.2108815119

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Heather E. Grant

Contents

Figures and Tables

Figures

Tables

Chapter 1

General Introduction

1.1 HIV origin and subtypes

There exists a large and diverse reservoir of simian immunodeficiency viruses (SIVs) in wild primate populations in Central and West Africa (Peeters & Delaporte 2012). From this reservoir there have been multiple zoonotic introductions of human immunodeficiency virus (HIV) into people, probably from the consumption of bush meat (Gao et al. 1999, Hahn et al. 2000, Sharp et al. 2001). HIV-1 Group M (M for 'main') is responsible for the overwhelming majority of global infections, whilst HIV-2 or HIV-1 Group O (O for 'outlier') or N (N for 'non-M/O') are distinct zoonotic events (Sharp & Hahn 2011), usually found as part of small epidemics in West Africa. All global Group M diversity can be found nested within modern day diversity in Kinshasa in the Democratic Republic of Congo (Rambaut et al. 2001) where this diversity has been present since the 1960s (Worobey et al. 2008) before the first AIDS cases were reported.

Exports of HIV-1 M strains out of the DRC into new global susceptible populations led to rapid evolution and diversification (Faria et al. 2014), which can be seen in its 'starburst' phylogenetic structure (Archer & Robertson 2007), indicative of rapid expansion. These founder events (Rambaut et al. 2001) created phylogenetically distinct clades, often termed subtypes (A to D, F to H, J and K) (Robertson et al. 2000). Evolution in separate populations for many decades has resulted in between subtype genome nucleotide distances of around 15% (Li et al. 2015) and strong geographical structure. Subtype B for example, is the dominant subtype found in North America and Europe, while subtype C is the most common subtype in Southern Africa. Places with high levels of global immigration such as London have a highly cosmopolitan subtype distribution (Yebra et al. 2018).

1.2 Infection and disease progression

HIV is predominantly transmitted via sexual routes, but can also be transmitted from mother to child, via intravenous drug use, or contaminated blood products (Shaw & Hunter 2012). HIV relies on the CD4 receptor, as well as a secondary chemokine co-receptor to gain entry into T-cells (Wilen et al. 2012). After cell binding and entry, HIV is reverse transcribed into DNA (the provirus), which is then integrated into the host cell genome. This provirus copy is transcribed and translated into viral RNA and proteins which can then be assembled to make new viruses which go on to infect new cells, (see lifecycle overview by Kirchhoff 2013). Once an infection has been initiated, HIV replicates with an exceptionally high mutation rate (Mansky & Temin 1995) and short generation time (Ho et al. 1995) to create large quasi-species diversity within each person. The size of this viral population gives it power to evolve rapidly in response to drugs (Wei et al. 1995) and host immune responses (Bonhoeffer et al. 1995) like cytotoxic T-cells (Poon et al. 2007), or neutralising antibodies (Frost et al. 2005). During the course of infection, T-cells are irreparably lost (Sabin et al. 2000), causing AIDS (defined as fewer than 200 CD4+ T-cells per microL), after which other infections cause death. It is therefore important halt white blood cell decay by starting antiretroviral therapy (ART) as soon as possible (WHO, 2016).

Without ART intervention, each individual patient's path to AIDS is extremely variable, and may depend on a myriad of host genetic factors, including sex and age (Telenti & Johnson 2012), but particularly human leukocyte antigen (HLA) types (Steel et al. 1988, Kaslow et al. 1996). For example, cohort studies in Africa and Europe show that the allele HLA B57 is extremely predictive of viral suppression (Price et al. 2019, Fellay et al. 2009). A very small percentage of people are in a category known as 'elite controllers' because they able to stay AIDS free for longer periods, and the genetics of that groups is of great interest (Deeks & Walker 2007).

Viral load (number of viral copies at some equilibrium during clinical latency) also contributes to the variability in AIDS progression. The clearest cut demonstration of this effect is between HIV-1 and HIV-2, where HIV-2 was shown to be far less virulent in a cohort of Senegalese sex workers (Marlink et al. 1994). HIV-2 infections had a much longer time to AIDS and left higher CD4 counts, whilst eliciting lower viral loads (Kanki et al. 1994). Therefore, a clear difference in viral load and virulence could be demonstrated (Hansmann et al. 2005). Viral load is a key predictor of disease progression (Mellors et al. 1996), but also leads to higher transmission risk (Blaser et al. 2014).

The secondary co-receptor used by the virus (CCR5 or CXCR4) has implications for variability in disease progression. CXCR4 tropism has been shown to more rapidly reduce CD4 counts and the time to AIDS compared with CCR5 tropism (Richman & Bozzette 1994), by more more rapidly depleting T-cells (Penn et al. 1999). This is likely explained by a wider range of T-cells as 'prey' available (Schuitemaker et al. 2010). This wider range includes naive T-cells which are involved in T-cell regeneration, thus further contributing to more drastic declines in CD4 counts (Blaak et al. 2000).

1.3 Co-receptor dynamics

Early on in HIV research, viruses grown in cell-culture could be differentiated as 'syncytium inducing' (SI) or 'non-syncytium inducing' (NSI) (Koot et al. 1993). Syncytium formation is a feature of fast replicating viruses grown in transformed T-cell lines, and ultimately this phenotype was linked to use of the secondary co-receptor CXCR4, while the NSI forming viruses were shown to use CCR5 (Connor et al. 1997).

A viral switch from using the CCR5 co-receptor ('R5 tropic viruses') to CXCR4 ('X4 tropic viruses') is facilitated by changes in the amino acid sequence of the third variable loop (V3) of the envelope protein gp120, notably positively charged amino acids at the 11th and 25th position (the '11/25 rule'; de Wolf et al. 1994). However, this rule is not a particularly accurate predictor in isolation, as the other positions contribute to changes in overall charge and the 3D structure of the protein (Lengauer et al. 2007). There are a multitude of diverse amino acid combinations and pathways that can facilitate a co-receptor switch (Poon et al. 2012), which makes machine learning tools like geno2pheno highly relevant to this complex pattern prediction problem (Lengauer et al. 2007). Other gene regions (such as the V1/V2 region) may also contribute to co-receptor tropism (Pastore et al. 2006), but are often not included in prediction tools.

There is a long-standing notion that R5 variants generally initiate most infections (Zhu et al. 1993, van't Wout et al. 1994). Independent of route of transmission, the mucosal tissues are seemingly the primary site of replication and have a high concentration of memory T cells with high expression of CD4 and CCR5 co-receptors (Lackner et al. 2012). The observation that persons with two copies of a defective CCR5 allele (a 32 base pair deletion) are protected against the initiation of new infections (Wilkinson et al. 1998), also points to the importance of CCR5 at the early stages. Other than the abundance of CCR5 at primary infection sites, several 'gatekeeping' mechanisms have been proposed to explain the apparent scarcity of new infections with X4 variants (Margolis & Shattock 2006), including differences in interactions with heparan sulfate proteoglycans (Moulard et al. 2000) or the integrin $\alpha 4 \beta 7$ (Cicala et al. 2010).

Over the course of infection, dynamic changes in the immunological ecosystem are expected to change the relative fitness of R5 and X4 variants. After immune activation there may be a depletion of memory T-cell populations and a fitness advantage to X4 variants which infect a wider variety of white blood cells (the 'target cell hypothesis'; Davenport et al. 2002). But despite the rapid evolution rate of HIV, there is no inevitability of X4 evolution (Holmes 2001), and not all AIDS patients develop X4 variants. Therefore, there may be an adaptive landscape with a dip in fitness for intermediate forms (Regoes & Bonhoeffer 2005). Furthermore the immune system may actively select against X4 variants (the 'immune-control hypothesis'), which is also likely to change over the duration of infection. For example, there is some evidence that transitional co-receptor forms are more susceptible to certain neutralising antibodies (Bunnik et al. 2007), and dendritic cells may release SDF-1α which inhibit X4 variants (González et al. 2010).

There is therefore evidence that X4 tropic variants can be both a *cause* and a *consequence* of more advanced disease progression. Careful experiments have showed clearly that CXCR4 co-receptor tropism causes more rapid T-cell depletion (for example Penn et al. 1999, Kreisberg et al. 2001), and evidence supporting the 'target cell hypothesis' has also come from a longitudinal study pinpointing certain T-cell population changes as strong predictors for the emergence of X4 variants (Connell et al. 2020).

1.4 Recombination in HIV

Retroviruses co-package two RNA genomes in their virus particles, both of which are required to reverse transcribe a single DNA copy (Panganiban & Fiore 1988), thus HIV is considered 'psuedo-diploid' (Hu & Temin 1990). Recombination mimics sex in that it may serve to bring beneficial mutations together or purge deleterious mutations (Barton & Charlesworth 1998), increasing the variance of fitness and allowing selection to work efficiently (Worobey & Holmes 1999, Burt 2000). Recombination is therefore a powerful tool in HIV evolution which can create better adapted viruses, but also rescue non-functioning viruses from a high mutation rate and error catastrophe (Tripathi et al. 2012). Recombination exceeds the mutation rate, occurring between 2 to 14 times per replication (Jetzt et al. 2000, Zhuang et al. 2002, Cromer et al. 2016), ten times higher than that of other retroviruses (Rhodes et al. 2003). The rate of recombination might also depend on the overall viral load or the cell type where replication is occurring (Levy et al. 2004).

As part of the retroviral lifecycle, HIV is reverse transcribed from RNA to DNA by reverse transcriptase (RT), firstly by making a negative sense DNA copy with an RNA template, then a positive sense DNA copy with the single stranded DNA template (see detailed review by Coffin et al. 1997). If dual infection with two distinct viruses occurs and those RNA genomes are

transcribed in the same cell, it is possible to generate a single particle with two distinct RNA genomes. During reverse transcription in a new cell, distinct RNA genomes may recombine to produce a chimeric provirus, which after integration may later may produce a daughter virus with two identical recombinant RNA genomes.

There are two directional models for recombination: firstly the 'copy choice model' (during the synthesis the minus strand) and secondly the 'strand displacement model' (during synthesis of the positive strand). These models are not mutually exclusive, although it is believed that recombination is more frequent during minus strand synthesis (Coffin et al. 1997). The negative strand 'copy choice' or 'temple switching' hypothesis has more than one proposed mechanism. The early 'forced choice' model arose from observations in Avian Sarcoma Virus that RNA genomes are usually very fragmented, thus RT should be 'forced' to find other RNA templates to continue DNA synthesis (Coffin 1979). However, the degree of RNA fragmentation within the HIV particle is not certain, thus the 'pause model' may be more realistic. In the 'pause model' a slowing of reverse transcription lead to template switching by RT (Destefano et al. 1992, DeStefano et al. 1994, Wu et al. 1995). An extension to this model came from Hwang et al. (2001), who proposed that RNase H can further promote the dissociation or RT by degrading the nascent template (the 'dynamic copy choice model'). Structured region of the RNA genome are thought to be important in encouraging template switching, such as hairpins (for example a kissing hairpin in the dimer initiation sequence; Balakrishnan et al. 2001), and it has been demonstrated experimentally that modulation of the stability of hairpin structures can influence recombination rates (Galetto et al. 2004).

Despite these intricate model, there is no unifying mechanism or model which can explain hotspots of recombination along the genome (Galetto & Negroni 2005), and it seems other factors like increased sequence identity (Baird et al. 2006, Archer et al. 2008), and homopolymeric nucleotide runs (Klarmann et al. 1993) also promote recombination. Onto this background of mechanistic processes is superimposed functional constraints on the virus. A successful HIV recombinant must create a viable virus with functional proteins, able to infect new cells and new hosts, and some recombination events will inevitably bring discordant gene combinations together creating less fit proteins (Galli et al. 2010). Therefore many of the recombinants formed during infection may never leave the person in which they were generated.

Describing recombination is an important first step in any phylogenetic analysis. When sequences that have undergone recombination are used in standard phylogenetics, the branch lengths tend to be overestimated, and there is a breakdown of the molecular clock (Schierup & Hein 2000*a*). Recombination in genomes can be detected with the use of phylogenetics because different sections of the genome represent different evolutionary histories, and will therefore have discordant branching patterns (Robertson et al. 1995, Lemey et al. 2004). Subtyping tools like REGA (de Oliveira et al. 2005) implement a phylogenetic based method with "bootscanning" (Salminen et al. 1995), which involves moving along the genome in 400

base pair windows at 20 base pair increments. The closest reference sequence for each window is reported which can give an indication of where the recombination event took place along the genome because the top references 'switch'. Other subtyping tools such as COMET (Struck et al. 2014) are alignment free and work by building Markov Models over nucleotide frequencies for each given reference sequence. It is therefore 'context based' and rapid, but does not determine the location of any recombination breakpoints. SCUEAL (Kosakovsky Pond et al. 2009) is an HIV specific tool based on the recombination detection program GARD (Kosakovsky Pond et al. 2006) which both work by searching through a multitude of different recombination models using a genetic algorithm. This method is able to provide numerical breakpoint locations and outputs intra-subtype recombination results (discussed more in chapter 2 and appendix A). Ultimately, recombination detection programs are only as good as the reference sequences used, and many recombination events will be undetectable, particularly between highly similar sequences. The rate of recombination is likely under-estimated since a higher pairwise sequence identity makes the physical occurrence of recombination more likely to happen, yet the detection of recombination less likely.

1.5 The HIV genome

The HIV genome is around 9700 base pairs (including the long terminal repeat or LTR regions) see Figure 1.1, (or approximately 8000 bp from *gag* to *nef*). Three structural genes make the essential proteins needed to make an HIV viral particle. *Gag* encodes several proteins including matrix protein (p17), capsid (p24), and nucleocapsid (p7), whilst *pol* encodes the enzymes protease (PR), reverse transcriptase (RT) and integrase (IN) which are all essential for the HIV life cycle. The *env* gene makes proteins SU gp120 and TM gp41 which form a complex trimer spike particle which is essential for CD4 and co-receptor recognition and cellular entry. Gp120 contains 5 conserved regions and 5 variable regions (which directly interact with the immune system), and gp41 makes up the stalk protein, composed of an extracellular and intracellular component, with a hydrophobic transmembrane region between them. Regulatory proteins aid processes like transcription (TAT) and nuclear export (REV) while the LTRs or UTR (untranslated regions) are important for integration into host cell. Among other functions, the accessory genes (*vpr, vpu, vif, nef*) have various roles in downregulating the immune system (Emerman & Malim 1998).

The first HIV sequences were often only few hundred base pairs. The V3 loop was sequenced earlier on e.g. (Bruce et al. 1994) but later researchers started to sequence p17 or gp41 which permitted phylogenetic clustering at a national level without signatures of selection which are often found in V3 e.g. (Leigh Brown et al. 1997). Since the ART era, *pol* sequencing has become common (and particularly the RT region) because it is where many drug resistant

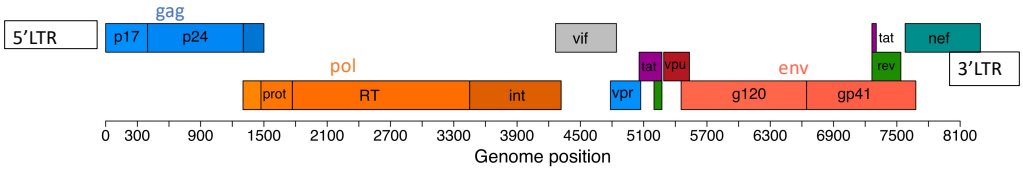

Figure 1.1: The HIV genome

mutations arise (Cane 2011). Working with near full length genomes however, provides over 8000bp of genetic information, shown to give better resolution and accuracy in phylogenetic analysis compared with *pol* alone (Yebra et al. 2016), and more opportunity to look at genotypes that might be of interest across the whole genome.

1.6 The generalized HIV epidemic in Uganda

'Slim disease' as AIDS was colloquially known in Uganda, was first reported in 1982 in the Rakai district (Serwadda et al. 1985), although there were reports of deaths in 1976 which, in retrospect, were AIDS-like (Kuhanen 2010). In the 1980s, particularly high HIV prevalence was seen in lorry drivers and 'barmaids' (female sex workers) found along roadside travel hubs; 35% and 67% respectively (Carswell 1987, Carswell et al. 1989), while phylogenetic evidence points to the importance of the road network in importing HIV into rural southwest Uganda from the DRC (Gray et al. 2009). Unlike the Western epidemic in the 1980s, the Ugandan epidemic was, and still is, overwhelmingly spread heterosexually, with hotspots in urban areas and roadside stops where sex workers are concentrated (Berkley et al. 1989). HIV spread continuously into new populations during the 1980s throughout Uganda, no doubt aided by war in 1979 (Kuhanen 2010), imparting a particularly high incidence and mortality burden e.g. (Sewankambo et al. 2000).

In the late 80s and early 90s Uganda mounted a successful and concerted national effort to encourage large scale behavioural changes, with involvement from the highest levels in government down to the grass roots (Green et al. 2006). A combination of openness to acknowledging the epidemic (Genuis & Genuis 2005), reductions in risky behaviour (Hayes & Weiss 2018) and public health messages like ABC 'abstinence be faithful, use condoms' (Shelton et al. 2004) has helped to reduced incidence. The Ugandan response was one of the best in Africa and provided a good example for other African nations to follow (UNAIDS), although the relative impact of each intervention type on reducing infections during this period was debated (Low-Beer 2002). It appears that incidence peaked in 1987 and prevalence peaked in 1992 (Kirby 2008, Baryarama et al. 2004) but interpretation of incidence and prevalence is extremely difficult (Parkhurst 2002), not least because of rapidly changing birth

rates and AIDS-related deaths (Mbulaiteye et al. 2002). Data from the UN estimate that the population of Uganda has changed from 6 million in 1956, to 15 million in 1986, to 40 million in 2016, despite adult mortality of 438 per 1000 in the years between 1995-2000 (UN Department of Economic and Social Affairs 2019) (age 15-50, both sexes, any cause).

The introduction of ART in 2004 was another big step forward further reducing the burden of disease and also new infection (since ART-initiated individuals are less infectious). Since then, ART use has steadily grown and reduced HIV incidence rates (Grabowski et al. 2017). Despite a reduction in prevalence and incidence, the Ugandan HIV epidemic is highly diverse and highly complex. There is great heterogeneity in infection rate, risk, and prevalence across the country (Chang et al. 2017), and incidence ranges widely according to age, gender, and geographical location (particularly between urban and rural areas). For example, a meta-analysis by Birdthistle et al. (2019) showed young women in a Rakai fishing community might have a 32x greater HIV risk than young women in the rural East Ugandan district Kumi (12.4 vs 0.38, incidence measured in per 100 person years). In the fishing villages in and around Lake Victoria, the prevalence of HIV is as much as 11x higher than the general population in land (Kamali et al. 2016) due to a culture of high-risk behaviours (Kiwanuka et al. 2014). Co-infection or super-infection (multiple infections initiated simultaneously or sequentially) occur (Redd et al. 2014, Ssemwanga et al. 2012), but are difficult to estimate effectively because they often resolve in early infection (Koning et al. 2013, Yang, Daar, Jamieson, Balamurugan, Smith, Pitt, Petropoulos, Richman, Little & Leigh-Brown 2004). There appears to be structured transmission networks within communities, but also substantial connections between adjacent communities (Ratmann et al. 2020, Kiwuwa-Muyingo et al. 2017). Furthermore, these dynamics might not be as obvious as expected. The fishing villages were long thought to be the source of HIV transmission into the general population e.g. (Opio et al. 2013) but it has been recently shown that there is significant movement in the opposite direction, and that these villages may in fact be sinks of transmission (Bbosa, Ssemwanga, Nsubuga, Salazar-Gonzalez, Salazar, Nanyonjo, Kuteesa, Seeley, Kiwanuka, Bagaya, Yebra, Leigh-Brown & Kaleebu 2019).

The Phylogenetics And Networks for Generalised HIV Epidemics in Africa consortium (PANGEA-HIV) (Pillay et al. 2015) funded by the Bill and Melinda Gates Foundation, was set up in order to widen the phylogenetic information available to study HIV transmission in Africa. The first phase of project (2013-2017) was led by Deenan Pillay (UCL), Christophe Fraser (Oxford University), Paul Kellam (EMBL-EBI) and Tulio de Oliveria (Africa Health Research Institute), and Andy Leigh Brown (Edinburgh), and managed by Dr Anne Hoppe. Sequencing was carried out on behalf of several partners including the Medical Research Council/Uganda

Virus Research Institute, the Rakai Health Sciences Program, the Zambart Project, and the Botswana Harvard AIDS Institute Partnership. The PANGEA2 (2017-2021) project will continue this work by producing additional data but mainly focusing on analysis of the existing data.

The Uganda Virus Research Institute in Entebbe directed by Pontiano Kaleebu runs a multitude of cohort studies, the most famous being the 'General Population Cohort', recently described by Deogratius Ssemwanga and collegues in Kyamulibwa in south west Uganda (Ssemwanga et al. 2020), and (Kapaata et al. 2013). This cohort has been run since 1989 and includes local rural communities. Other large cohorts include 'The Good Health for Women Project' and women at high risk to HIV (Kasamba et al. 2019), or cohorts of high-risk sexual behavior e.g. in Masaka (Serwanga et al. 2018), various fisherfolk cohorts e.g. Nsazi Island, and drug resistance cohorts e.g. Ministry of Health Drug Resistance(UVRI-MOHDR). These cohort studies were included as part of the MRC/UVRI and LSHTM Uganda Research Unit contribution to generated PANGEA data.

In 2017 in what was then the Public Health England labs at Porton Down, samples from 1986 Uganda were found during relocation efforts. Pat Cane brought these to the attention of PANGEA, who attempted to sequence them with the PANGEA protocol, but due to the highly degraded nature of the samples, there was limited success. Later, Judy Brueur suggested the use of target capture next generation sequencing on the remaining samples, which vastly improved the outcome of sequencing see (chapter 3).

This thesis presents the most comprehensive characterisation to date of the Ugandan epidemic with full genome data, spanning three decades. The data reflect a very complex epidemic with two HIV subtypes, which have competed in the same population, while dual infections have facilitated the generation of a myriad of unique recombinant forms.

Chapter 2

Pervasive and non-random recombination in near full-length HIV genomes from Uganda

Heather E. Grant[1], Emma B. Hodcroft[2,3], Deogratius Ssemwanga[4,5], John M. Kitayimbwa[6], Gonzalo Yebra[7], Luis Roger Esquivel Gomez[8], Daniel Frampton[9], Astrid Gall[10], Paul Kellam (10), Tulio de Oliveira[11], Nicholas Bbosa[4], Rebecca N. Nsubuga[4], Freddie Kibengo[4], Tsz Ho Kwan [12], Samantha Lycett[7], Rowland Kao[7], David L. Robertson[13], Oliver Ratmann[14], Christophe Fraser[15], Deenan Pillay [10,11], Pontiano Kaleebu[4,5], Andrew J. Leigh Brown[1], on behalf of the PANGEA-HIV consortium

1) Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
2) Biozentrum, University of Basel, Basel, Switzerland
3) Swiss Institute of Bioinformatics, Basel, Switzerland
4) Medical Research Council (MRC)/Uganda Virus Research Institute (UVRI) and London School of Hygiene and Tropical Medicine (LSHTM) Uganda Research Unit, Entebbe, Uganda
5) Uganda Virus Research Institute, Entebbe, Uganda
6) Department of Mathematics, Makerere University, Kampala, Uganda
7) The Roslin Institute, University of Edinburgh, Edinburgh, UK
8) Max Planck Institute for the Science of Human History, Jena, Germany
9) Division of Infection and Immunity, University College London, London, UK
10) European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK
11) Africa Health Research Institute, Nelson R. Mandela School of Medicine, Durban, South Africa
12) Stanley Ho Centre for Emerging Infectious Diseases, The Chinese University of Hong Kong, Shatin, Hong Kong
13) MRC Centre for Virus Research, University of Glasgow, Glasgow, UK

14) Department of Mathematics, Imperial College London, London, UK
15) Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK

Key words: HIV; genome; subtypes; phylogenetics; recombination; breakpoints

2.1 Abstract

Recombination is an important feature of HIV evolution, occurring both within and between the major branches of diversity (subtypes). The Ugandan epidemic is primarily composed of two subtypes, A1 and D, that have been co-circulating for 50 years, frequently recombining in dually infected patients. Here we investigate the frequency of recombinants in this population and the location of breakpoints along the genome. As part of the PANGEA-HIV consortium 1472 consensus genome sequences over 5kb have been obtained from 1857 samples collected by the MRC/UVRI & LSHTM Research unit in Uganda, 465 (31.6%) of which were near-full length sequences (>8kb). Using the subtyping tool SCUEAL we find that of the near-full length dataset, 233 (50.1%) genomes contained only one subtype 30.8% A1 (n=143), 17.6% D (n=82) and 1.7% C (n=8), while 49.9% (n=232) contained more than one subtype (including A1/D (n=164), A1/C (n=13), C/D (n=9); A1/C/D (n=13), and 33 complex types). K-means clustering of the recombinant A1/D genomes revealed a section of envelope (C2gp120-TMgp41) is often inherited intact, whilst a generalized linear model was used to demonstrate significantly fewer breakpoints in the gag-pol and envelope C2-TM regions compared with accessory gene regions. Despite similar recombination patterns in many recombinants, no clearly supported CRF was found, there was limited evidence of the transmission of breakpoints, and the vast majority (153/164; 93%) of the A1/D recombinants appear to be unique recombinant forms (URFs). Thus, recombination is pervasive with clear biases in breakpoint location, but circulating recombinant forms (CRFs) are not a significant feature, characteristic of a complex and diverse epidemic.

2.2 Introduction

Human immunodeficiency virus (HIV) is a highly diverse retrovirus at both the within-individual and population level (Smyth et al. 2012). The HIV reverse transcriptase (RT) is error-prone resulting in a high mutation rate. RT also facilitates recombination via template switching between the two RNA genomes packaged inside the virion (Hu & Hughes 2012). The diversity of HIV allows the virus to evade host defenses, accrue drug resistance mutations, and prevent effective vaccine development (Rambaut et al. 2004).

HIV-1 Group M group contains the greatest genetic diversity. This group likely diversified in Kinshasa (Democratic Republic of Congo or DRC) from the 1920s to the 1960s, before rapidly expanding into global susceptible populations (Korber et al. 2000, Worobey et al. 2008, Faria et al. 2014). Forming phylogenetically distinct clades, the subtypes A to D, F to H, J and K (and sub-subtypes within e.g. A1), are found globally but frequently have broad geographic associations, mainly as the result of founder effects (Rambaut et al. 2001, Archer & Robertson 2007). Meanwhile, the DRC retained as much diversity as the global pandemic (Niama et al. 2006). As they spread, the subtypes almost certainly underwent extensive recombination throughout their evolution including at an early stage (Kalish et al. 2004, Ward et al. 2013, Olabode et al. 2019).

Recombination between different HIV variants occurs in individuals with dual infection (Robertson et al. 1995), either acquired simultaneously (co-infection) or sequentially (super-infection). This gives rise to unique recombinant forms (URFs) especially in regions where more than one subtype is common (Yebra et al. 2015, Bbosa, Ssemwanga, Nsubuga, Salazar-Gonzalez, Salazar, Nanyonjo, Kuteesa, Seeley, Kiwanuka, Bagaya, Yebra, Leigh-Brown & Kaleebu 2019). If three or more recombinant genomes without direct epidemiological linkage are found, they may be defined as a circulating recombinant form (CRF) (Robertson et al. 2000). Additionally, recombination between viruses of the same subtype (intra-subtype) occurs (Kraft et al. 2012), especially where there are high rates of dual infections (Taylor & Korber 2005) although as it is more difficult to detect due to the similarity of the recombining sequences (Yang, Daar, Jamieson, Balamurugan, Smith, Pitt, Petropoulos, Richman, Little & Leigh-Brown 2004) it is therefore less well documented.

HIV-1 subtypes represent major clades that have a lengthy period of distinct identity, thus assigning sequences to subtypes is inherently a phylogenetic problem. Correctly placing sequences into clades of ancestral diversity relies on the availability of representative reference sequences, that themselves are unrecombined and correctly classified. It is made challenging by growing global diversity, the accumulation of drug resistance mutations (essentially equating to convergent evolution), and in particular, widespread recombination. Manual phylogenetics has been seen as a “gold standard” for subtype classification (Pineda-Peña et al. 2013, Fabeni et al. 2017), but a number of automated tools exist e.g. (de Oliveira et al. 2005, Struck et al. 2014) which are particularly useful in subtyping large datasets and databases. Automated subtyping methods have good accuracy compared to manual phylogenetics in the case of the simple ‘pure’ subtype using just the *pol* region (Pineda-Peña et al. 2013, Fabeni et al. 2017), although a similar assessment has not been undertaken for whole-genome tools. Agreement between methods is better for certain subtypes (e.g. B or C), whilst more challenging for others (e.g. A or D), and novel recombinants with sections of different phylogenetic history are a particular source of disagreement (Gifford et al. 2006), highlighting the difficulties in classifying recombinant sequences. The description of new CRFs

for instance, typically involves showing that they form a monophyletic cluster amongst a background of other sequences, a “boot-scanning” sliding window approach (Salminen et al. 1995) to find putative sections of different subtype, followed by more detailed and laborious confirmation by hand e.g. (Carr et al. 1998, Foster et al. 2014).

SCUEAL (Kosakovsky Pond et al. 2009) is an automated tool, which finds the most likely subtype or recombinant mosaic with a model-based evaluation. Briefly, a reference set of pure subtypes and CRF genomes is used to make a reference alignment, tree, and an inferred root sequence which remains constant for each query and model proposal. The query sequence is then aligned to the inferred root sequence, grafted to the reference set to make a three-taxon tree, and the maximum likelihood placement is found. A genetic algorithm acts upon a population of models to create mosaic suggestions for a fixed number of breakpoints. Schwartz’s Bayesian Information Criterion (BIC) is used to assess the fitness of models in the population, which evolve until there is no improvement after several generations (the stopping criteria). Additional breakpoints may be added until there is no further BIC improvement (and a step-down verification). Model averaged support for the best mosaic is found using the sum of Akaike weights of all concordant proposed models. A 95% confidence interval for the breakpoint location is found using a similar principle.

In Uganda, HIV was prevalent by the early 1980s (Serwadda et al. 1985), with two circulating subtypes (A1 and D) present at similar frequencies in the population (Yirrell et al. 2002, 1997), alongside unique A1/D recombinants (Eshleman et al. 2002). These two subtypes are thought to represent independent introductions of HIV diversity into Uganda, with A1 having arrived first via the rural south-west in the 1950s or 60s, followed later by subtype D about 10 years later (Yebra et al. 2015). There were already reports of growing numbers of AIDS cases (then identified as aggressive Kaposi’s sarcoma or slim disease) in the rural Rakai region of south western Uganda in the 1970s, (Serwadda et al. 1986, Kuhanen 2010). Surveillance studies found seropositivity in 1987 in pregnant women attending hospitals in the capital, Kampala, was 24.1% (Carswell 1987). Today the adult prevalence is estimated to be within 5.7% and 6.2% (Joint United Nations Programme on HIV/AIDS, 2019; Ministry of Health Uganda, 2019). Dual infections can be found in female sex workers (Ssemwanga et al. 2012, Redd et al. 2014) but also at substantial levels in general population and low risk rural cohorts (Kiwanuka et al. 2010, Ssemwanga et al. 2012, Redd et al. 2012). Therefore, subtypes A1 and D have been co-circulating in Uganda for perhaps as long as 50 years, with high rates of incidence and dual infection, providing ample opportunity for recombination to occur.

The PANGEA-HIV project (Pillay et al. 2015) was set up with the aims of using phylogenetics to better understand the dynamics and drivers of ongoing transmission in African HIV epidemics and has generated large numbers of near full length genome sequences. The data generated with samples obtained by MRC/UVRI in Uganda presented an opportunity to study the prevalence of recombinants and the distribution of their breakpoint locations along the genome in a population setting.

2.3 Methods

2.3.1 Sample collection

Samples were collected by the MRC/UVRI and LSHTM Uganda Research Unit between 2007 and 2017 from sites and cohorts across southern Uganda. These included the Masaka District in the rural South West, female sex workers from Kampala, and people living in fishing communities on the shores and islands around Lake Victoria. Ethical approval was given by the Uganda Virus Research Institute Research and Ethics Committee (UVRI-REC, Federal Wide Assurance [FWA] No. 00001354), the Uganda National Council for Science and Technology (UNCST FWA No. 00001293) and the University of Edinburgh School of Biological Sciences Ethics Committee (12/06/2018). All participants were recruited voluntarily and provided written informed consent.

2.3.2 Sequencing and alignment

Viral RNA was extracted from plasma by automated extraction. Near full-length HIV-1 genomes were reverse transcribed and amplified in four overlapping amplicons using a one-step RT–PCR protocol and a pan-HIV-1 primer set (Gall et al. 2012). Amplicons were pooled in equimolar amounts and sequenced using Illumina MiSeq 250-bp paired-end technology as described previously (Gall et al. 2014). Consensus sequences were generated from short reads using an in-house de novo assembly pipeline as follows. Trimmomatic (Bolger et al. 2014) was used to trim reads using a mean Phred quality score cut-off of 30. Human reads were removed by mapping to a smalt [https://www.sanger.ac.uk/science/tools/smalt-0] index consisting of HIV genomes [downloaded from Genbank:https://www.ncbi.nlm.nih.gov/genbank] and the hg38 human assembly [downloaded from Ensembl: ftp://ftp.ensembl.org]: read pairs where either or both reads mapped to hg38 were removed. De novo assembly was then performed using Iterative Virus Assembler (IVA) (Hunt et al. 2015), and contigs aligned to their closest viral reference using lastz (Harris 2007). Custom Perl scripts were then used to concatenate contigs into draft genomes and subsequently generate consensus sequences by a process of iterative mapping using smalt and SAMtools (Li et al. 2009). We applied a read depth cut-off of (greater or equal to) 20 reads to these final genomic sequences

before subsequent analyses. In total 1277 consensus genome sequences were produced at the Wellcome Sanger Institute, following the above protocol. In addition, 603 consensus genomes were produced using a similar approach by the Africa Centre (Durban, South Africa). After removal of duplicates the dataset comprised 1857 sequences. Of these, 1472 (79.3%) were over 5000bp, 1218 (65.6%) were over 6000bp, 797 (42.9%) were over 7000bp, and 465 (25.0%) were near-full length at over 8000bp which were used in the breakpoint analyses. Of these last, 371 were sequenced at the Wellcome Sanger Institute and 94 sequenced at the Africa Centre. The consensus sequences were aligned using MAFFT (Katoh & Standley 2013), and where necessary manually edited after visual inspection. The alignment starts from the first codon of *gag* (HXB2, 790) and ends at the last codon of *nef* (HXB2, 9415). Hypervariable loops 1+2, 4, and 5 in *env* (HXB2 6615-6812; 7377-7478; 7599-7637) were removed from the alignment as these can rarely be aligned with confidence (Simmonds et al. 1990). The sequences are submitted to Genbank under the accession numbers MN788736: MN790202.

2.3.3 Subtyping

Preliminary subtyping investigations were carried out with COMET (Struck et al. 2014), REGA (de Oliveira et al. 2005), and SCUEAL (Kosakovsky Pond et al. 2009). Our comparison of these three methods found overall agreement to be only 36.2%, (28.9% of sequences agreed between two methods, and 34.9% had no agreement). Where there was agreement between the 3 methods, these agreements tended to be pure subtypes (82%). Arau et al. (2019) carried out a similar comparison, but found better agreement by transforming the output of the methods to simplify classifications of difficult recombinants, which are much more likely to disagree. Of COMET, REGA and SCUEAL, only SCUEAL outputs breakpoint location numerically. For that reason, subtyping and breakpoint detection were undertaken with SCUEAL implemented locally using 218 full length subtypes and CRFs as references, allowing the program to find recombinant fragments of 300 base pairs and above, with a maximum number of 10 breakpoints. The genetic algorithm population size was set to 128 models and was said to have converged after no score improvements in 50 generations. A validation exercise was undertaken by creating ten random A1/D in-silico recombinants and analysing them 100 times in SCUEAL to test its reliability and accuracy (Supplementary Figure 2.7). The raw SCUEAL output was edited in R (R Core Team 2019) using the packages ape (Paradis & Schliep 2019) and seqinr (Charif & Lobry 2007) to make the following adjustments. Firstly, SCUEAL reports breakpoints at the location in the individual sequence, not the alignment, so these were adjusted to correspond to alignment positions. Secondly, phylogenetic subtyping methods sometimes have difficulty distinguishing subtype B and D in recombinants, owing to their closer common ancestry than other subtypes (Korber et al. 2000). As no pure subtype B sequences have been observed from Uganda (Lihana et al. 2012) and subtype B was only

ever seen as fragments in complex recombinants B calls were changed to D. Similarly, we did not attempt to distinguish A2 fragments from A1, as while A1 has been established in Uganda for decades, other A lineages have not been described. Confidence intervals of individual breakpoints have been stripped for clarity. Intra-subtype breakpoints were also removed.

2.3.4 Identification of transmitted breakpoints

A maximum-likelihood tree of all A1/D recombinant genomes, three A1 sequences, and three D sequences was constructed using IQ-TREE (Nguyen et al. 2015) with fast model selection (Kalyaanamoorthy et al. 2017), in order to identify any obvious CRFs. Similarly, a second tree was also constructed including the non-A1/D recombinants.

In order to distinguish between transmitted breakpoints and independent recombination events, we used a window-based approach to find pairs of sections of the genome linked by a low genetic distance. If a given pair of genomes contained multiple consecutive linked windows and a similar breakpoint was also found inside one of these windows, it was taken as evidence for a transmitted breakpoint.

Custom R scripts (`https://github.com/heathergrant/HIV_recombination`) were used to split genomes into 27 non-overlapping 300 base pair windows and to find linkage with a threshold of 2% divergence using the TN93 nucleotide distance (Tamura & Nei 1993). This is similar to the HIV-TRACE approach (Kosakovsky Pond et al. 2018), but considers multiple windows instead of the whole sequence. This approach was tested with randomly generated recombinants (see Supplementary Figure 2.8), and it was shown that at the 2% level, some references would be linked in some single windows. This 2% threshold was slightly higher than the usual 1.5% threshold often used in studies of transmission clusters using *pol* sequences e.g. Mehta et al. (2015), and there is no set distance that a pair of CRF genomes might be linked to each other: it will depend on the time since recombination and subsequent spread (younger CRFs should have lower thresholds). The purpose of this linkage was not to find recent transmission pairs, but to find sections of the genome that were related and shared a clearly transmitted breakpoint. All of the A1/D recombinant pairs linked by more than two out of 27 windows at the 2% level were examined. Where there was evidence for transmitted breakpoints between pairs of genomes, only one genome was kept in the subsequent GLM analysis to avoid issues of non-independence.

2.3.5 Recombination pattern classification

In order to classify A1/D recombinant genomes, each genome was transformed into binary characters identifying subtype at each nucleotide position (A1 recorded as 0, D recorded as 1). A Euclidean distance matrix was generated from the recoded data and K-means clusters were found using the kmeans function from the package stats v.3.6.0 (part of base R) and the algorithm of Hartigan & Wong (1979), which divides the data into groups by minimizing within-cluster variation. The optimal value of K was judged with the gap statistic (Tibshirani et al. 2001), and the elbow and silhouette methods using the cluster v.2.0.8 (Maechler et al. 2019) and factoextra v.1.0.5 (Kassambara & Mundt 2017) R packages.

2.3.6 Breakpoint and genome location model framework

Breakpoints of all inter-subtype recombinant genomes at different genome positions were analyzed using a generalized linear model in R. The binary response was presence or absence of a breakpoint, aggregated for each window of the genome, transformed with the logit link. Genomes were divided into 27 windows of 300 base pairs in length. The first window did not contain breakpoints (as minimum length to assign a subtype was constrained to 300bp), and the last window was fewer than 300bp. Both were removed from the analysis. Following the genome K-means clustering result, the genome regions were defined into three broad regions of the genome. These were a) windows containing gag-pol (windows 1-13), b) a custom region of envelope (C2-TM, from C2 of gp120 to the transmembrane region of gp41, windows 19-22) and c) accessory gene regions (*vif*, *vpr*, *vpu*, 14-18) and the cytoplasmic tail of gp41 plus *nef* 22-26).

2.4 Results

2.4.1 Subtype distribution

The MRC PANGEA-HIV genome dataset comprised 1857 sequences, of which 1472 were over 5000bp and 465 were over 8000bp. The subtype distribution for the 5000bp dataset was: 411 (27.9%) A1, 235 (16.0%) D, 25 (1.7%) C, 472 (32.1%) A1/D, 63 (4.3%) A1/C, 25 (1.7%) C/D, 54 (3.7%) A1/C/D, and 187 (12.7%) complex. Of the 465 near-full length genomes, 233 (50.1%) were ‘pure’ containing only one subtype (143 A1; 82 D; 8 C), while 232 (49.9%) were inter-subtype recombinants (164 A1/D; 13 A1/C; 9 C/D; 13 A1/C/D; and 33 other complex recombinants Figure 2.1). SCUEAL called more ‘complex’ and ‘other’ subtypes in the 5000bp dataset than the more complete sequences, which may be due to gaps in the sequence. Excluding the ‘complex’ category however, there was no difference in subtype proportions between these two datasets (χ^2= 4.19, df = 6, p=0.65), and the ratio of A1 to D genomes was

similar (1.743 :1 in the 8000bp and 1.748 :1 in the 5000bp dataset), confirming a lack of bias in successful sequencing by subtype or recombinant status. For the remaining analyses we used the near-full length genome dataset where subtype and location of breakpoints could be most accurately determined.

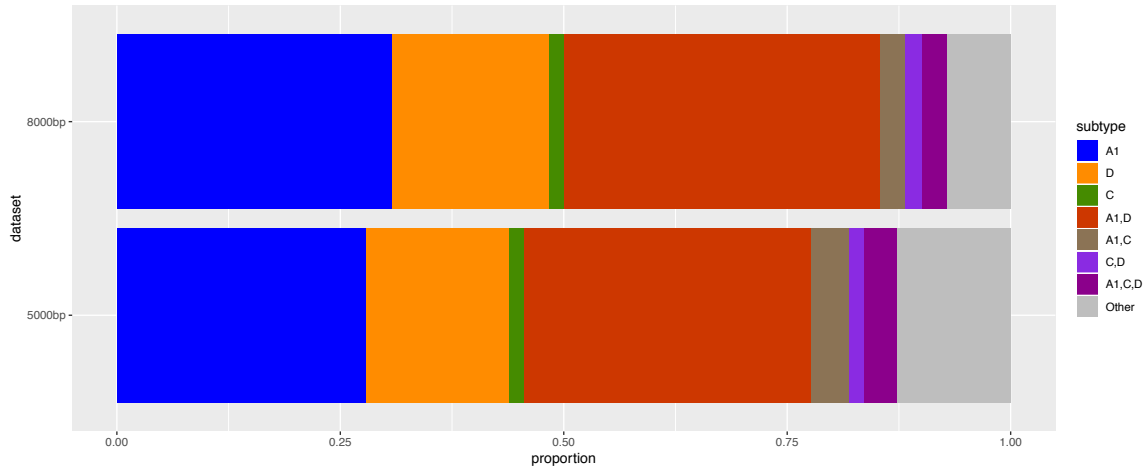

Figure 2.1: Subtype distribution in the 5000bp and above genomes, n=1472, and the near full length 8000bp and above dataset, n=465

2.4.2 Identification of CRFs and transmitted breakpoints

A maximum-likelihood tree of the A1/D recombinants with three A1 and D pure sequences was constructed (Figure 2.2). A similar figure is presented for non-A1/D recombinants, (n=68) in Supplementary Figure 2.9. Although the overall phylogeny is confounded by the violation of the key assumption that there are no recombinants, any CRF should form a clear monophyletic cluster.

Midpoint rooting broadly splits the tree into genomes predominantly containing subtype D, and those predominantly containing subtype A1 (the three references of each subtype fall within these respective groups). There are a few closely related cherries, and one closely related triplet, (Figure 2.2). Notably, some recombinants with a similar recombinant pattern can be found on altogether different parts of the tree, showing a clear evidence of convergent recombination.

We then used a window-based approach to find consecutive genetically linked windows that contained similar breakpoints, in an attempt to distinguish transmitted and unique breakpoints. Of the 164 A1/D recombinants, there were 12 single pairs, linked at a 2% threshold in a minimum of two out of 27 windows (Figure 2.3). There were also pairs forming a triplet (boxed), which had a similar recombination pattern in all three sequences and was tightly linked in multiple windows. However, there is epidemiological linkage of two of these sequences (data not shown) and therefore it does not meet the requirements of a CRF. Pairs 1,2 and 3 were

linked in 27/27 windows and are likely to have been transmitted relatively recently. Pair 2 has an almost identical subtype result and those breakpoints were likely transmitted. Other matching breakpoints outside of linked windows (e.g. in pairs 4 or 6) could represent transmitted breakpoints whose windows have diverged sufficiently to indicate an older common ancestor.

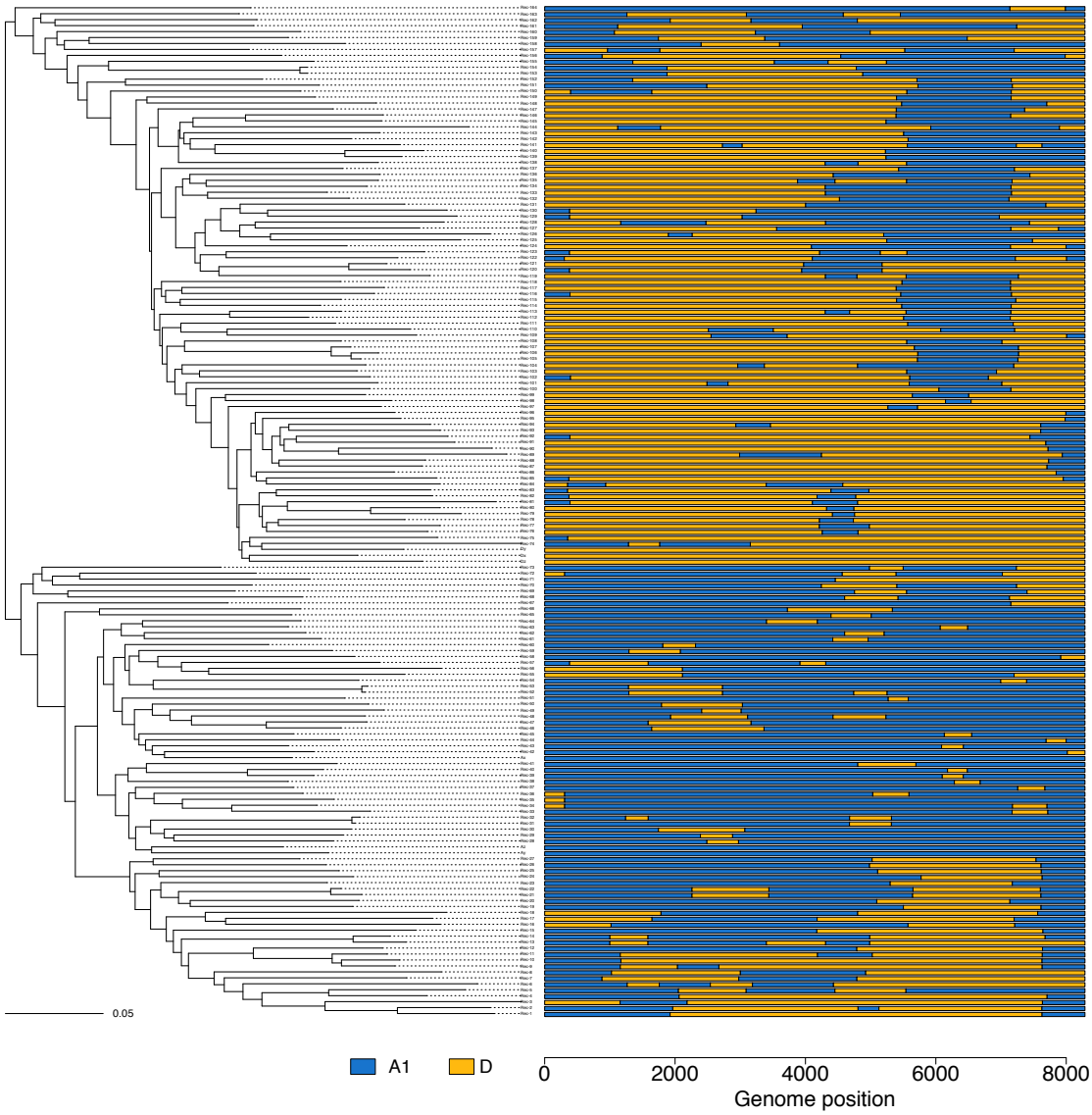

Figure 2.2: Maximum-likelihood reconstruction of the A1/D recombinants using IQ-TREE and their SCUEAL subtype (right). One triplet (Rec-105 to 107), and a few cherries can be seen (e.g. Rec 153-154). Some examples of convergent recombination patterns include Rec-116 & Rec 147, Rec-8 & Rec-160, Rec 29 & Rec-158

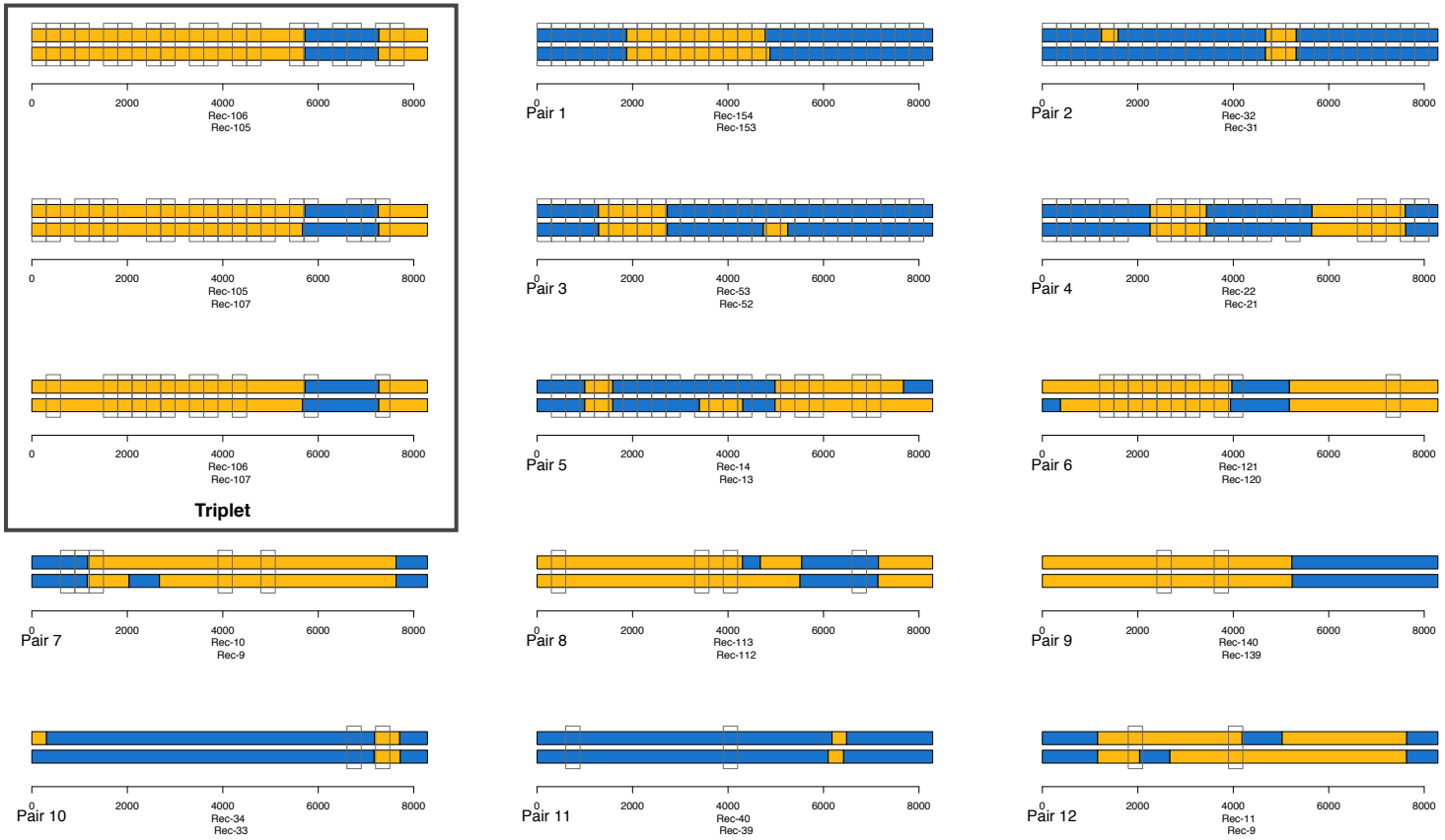

Figure 2.3: Pairs of genomes linked by a distance of less than 2% genetic distance (TN93) in two or more 300 base pair windows along the genome. The matching windows are shown with open clear boxes, and the SCUEAL subtyping result for the genome pairs are in colour (blue for subtype A1 and orange for subtype D)

Assuming there is evidence for transmitted breakpoints in pairs 1-12 (the A1/D pairs) and the triplet, there are 14 A1/D genomes that have evidence for being transmitted wholly or partially, and these pairs and triplet can be found as closely linked tips in the phylogenetic tree (Figure 2.2). Overall, as the vast majority of the A1/D genomes (150/164; 91%) lack linkage with other genomes, we see no evidence for large-scale transmission of individual recombinants such as would be recognised as a CRF, and should be considered unique recombinant forms. Linked windows with non-matching breakpoints (e.g. pairs 1, 3, 5) represent subtle inconsistencies in subtyping results, perhaps in region where divergent subtypes are more similar.

2.4.3 Recombinant groupings

The A1/D recombinants were placed into groups to highlight similarities in recombination patterns. This was done by putting subtype identity at each position along each genome through a K-means clustering algorithm. The optimum number of groups was found to be nine. Figure 2.4 shows a representation of the 164 A1/D recombinant genomes placed into these nine groups, (see Supplementary Figures 2.10, 2.11 and 2.12 for justification of, and alternative values of K). Group 1 contains mostly subtype D (in orange) with small sections of subtype A1 (in blue), whereas group 9 contains mostly subtype A1 with small sections of subtype D. In the remaining groups it is notable that a section of envelope appears to be inherited intact in many A1/D recombinants. This was observed in both directions, where subtype A1 envelope was found on a background of subtype D (groups 3, 4, 5), and subtype D envelope was found on a background of subtype A1 (groups 6 & 7). The part of envelope these groups have in common specifically spans from the C2 part of gp120 through to the transmembrane domain of gp41 (abbreviated C2-TM). In groups 7&8 the intact region of envelope extended into *nef* and there also appeared to be sections of subtype D RT (within *pol*) with A1 subtype either side.

2.4.4 Breakpoint distribution

The distribution of breakpoints along the genome for the A1/D genomes (n=164) and all other inter-subtype recombinants genomes (n=68) is shown in 300 base pair windows in Figure 2.5. The two distributions were strongly positively correlated (Pearson correlation, R^2=0.91, df =25, p<0.001).

Both distributions show a relatively large frequency of breakpoints in the accessory gene region (covering *vif*, *vpu*, *vpr*, *tat1*, *rev1*, and genome positions 4200 to 5700), lower levels of recombination in the gag-pol region and a particularly low level of recombination in the envelope region which was also seen in the K-means clustering result (Figure 2.4). Figure 2.6 shows the distribution within envelope at a finer scale (100 base pair windows) and a lower frequency of recombination within the C2-TM region (windows 20-23).

	Estimate	SE	z	p
Intercept (Gene region = Accessory)	-1.61635	0.05886	-27.462	<0.001
Gene region = gag-pol	-0.92597	0.09147	-10.123	<0.001
Gene region = env C2-TM	-1.40804	0.16688	-8.437	<0.001

Table 2.1: Beta estimates for the GLM on the log-odds scale

Table 2.1 shows the GLM model summary. Regions of the genome containing gag-pol had significantly ($p<0.001$) fewer breakpoints per 300bp window per genome than the accessory gene region, as did the C2-TM region ($p<0.001$). On the data scale the model finds the following estimates of breakpoint per 300bp window per genome: gag-pol 0.073 (95% CI 0.064 – 0.083), env-C2-TM 0.046 (95% CI 0.035 – 0.062), and the accessory regions 0.166 (95% CI 0.150-0.182).

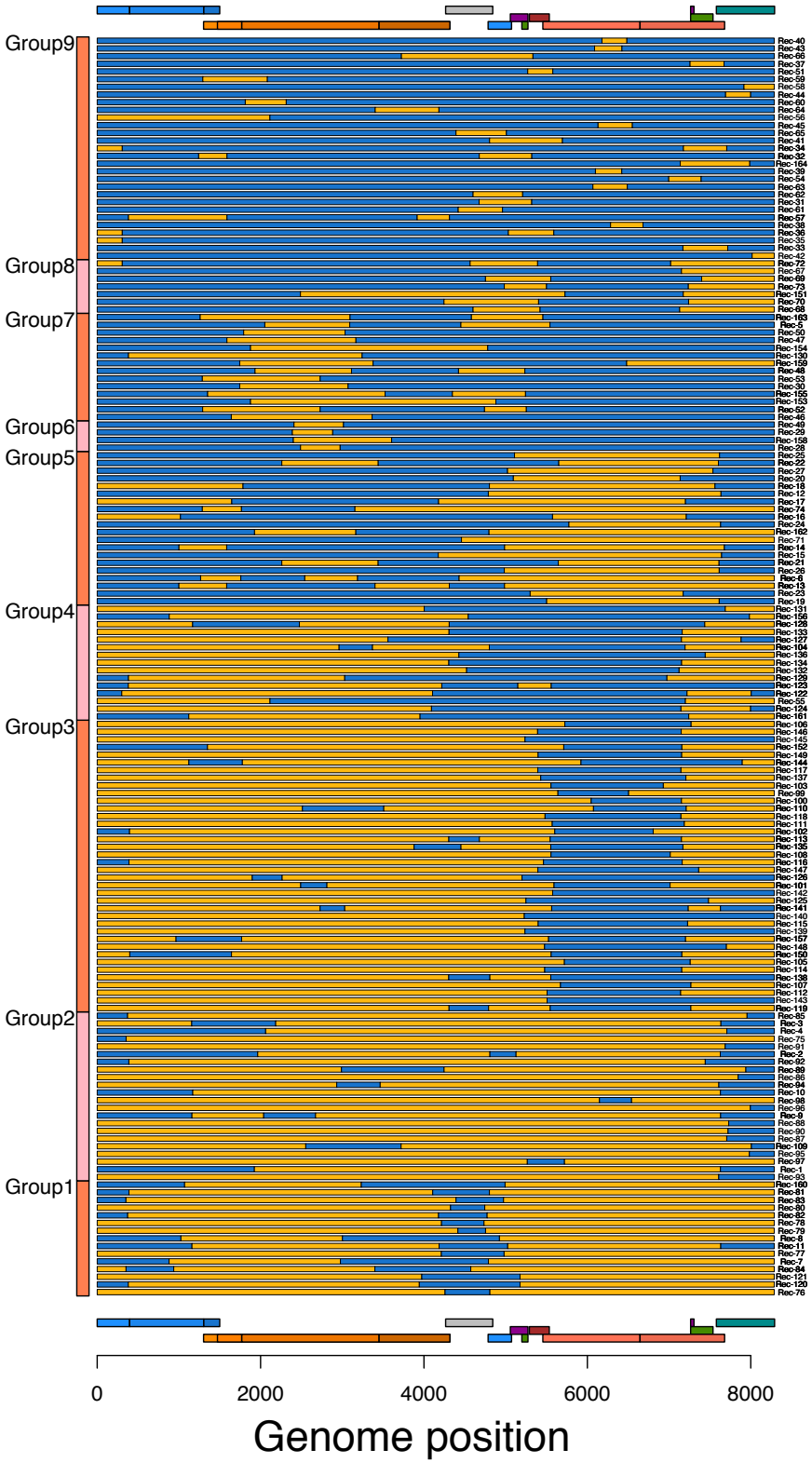

Figure 2.4: Recombination pattern of the A1/D recombinant genomes (n=164). Genome position is on the x-axis and each horizontal bar is an individual genome recombination pattern. Segments of orange colour represent subtype D, while blue colouration represents subtype A1

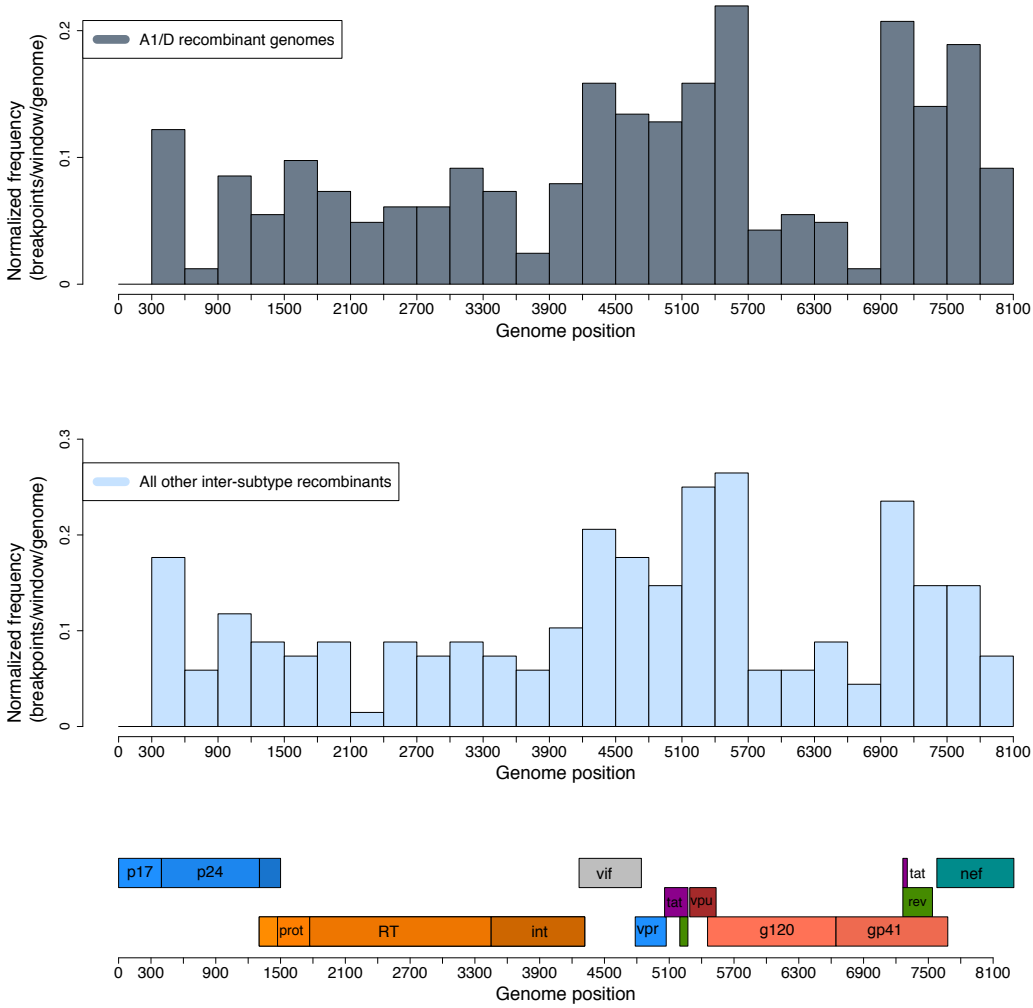

Figure 2.5: Distribution of inter-subtype recombination breakpoints divided into 300 base pair bins in A1/D recombinants (n=164) and all other inter-subtype recombinant genomes (n=68). Genome position numbering corresponds to the first nucleotide of p17 and ends with the last nucleotide of nef

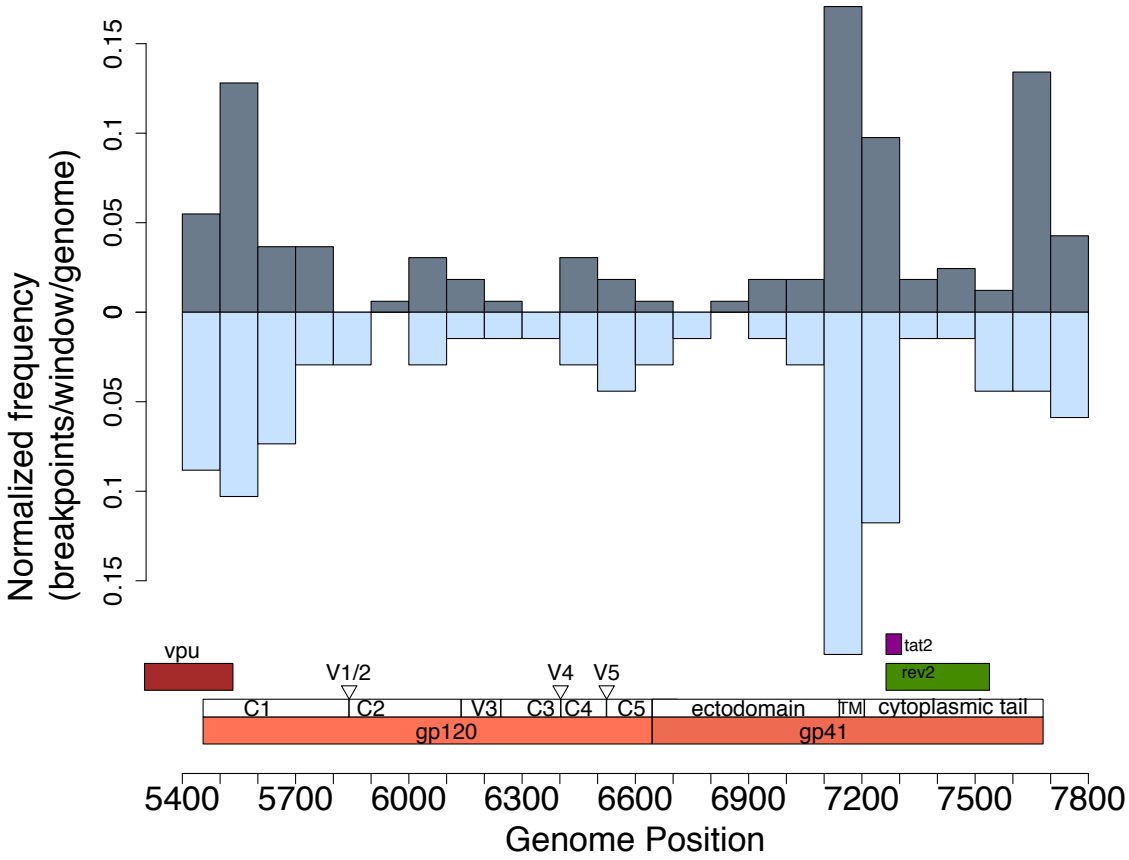

Figure 2.6: Distribution of breakpoints in the envelope region. Breakpoints have been binned into 100 base pair regions and the finer sub-structure of gp120 and gp41 is shown

2.5 Discussion

Multiple studies using single gene regions e.g. (Yirrell et al. 1997, 2002, Kaleebu et al. 2000) have previously described the HIV diversity in Uganda as predominantly subtype A1, subtype D, and A1/D recombinants (including A1/D URFs (Eshleman et al. 2002)). A more recent study suggests that in the *pol* region, around 15% of sequences are detectable inter-subtype recombinants (Bbosa, Ssemwanga, Nsubuga, Salazar-Gonzalez, Salazar, Nanyonjo, Kuteesa, Seeley, Kiwanuka, Bagaya, Yebra, Leigh-Brown & Kaleebu 2019), however, near-full length genomes reveal substantial additional recombination: we observe here that around half (49.9%) of the genomes are inter-subtype recombinants, and that most of these are URFs. Earlier small-scale studies of full-length genomes from Uganda have also shown high numbers of inter-subtype recombinants e.g. (14/46; 30%) (Harris et al. 2002) and (92/200; 46%) (Lee et al. 2017), also predominantly containing A1 and D subtypes.

This dataset, containing large numbers of near-full length sequences from a country already known to contain high numbers of unique recombinants, provided a difficult subtyping challenge. SCUEAL is an automated tool, unique in its ability to find a model-based assessment of recombination, including breakpoint locations. We have tested SCUEAL against *in silico* PANGEA subtype A1 and D recombinant sequences, and found it to perform extremely well (Supplementary Figure 2.7). Further to this, extensive tests were included in the original SCUEAL publication (Kosakovsky Pond et al. 2009), including a test against simulated recombinants, of database sequences, and in a comparison with the boot-scanning tool REGA. While it was shown to perform very well under a wide range of scenarios, accuracy wanes under the most complex scenarios, for instance those with more breakpoints, with closely related recombining sequences, and short fragments. Whilst SCUEAL is an extremely powerful model-based estimation of recombinant history, it is probabilistic, and should be interpreted as such.

There are significantly lower levels of recombination breakpoints in the gag-pol and envelope C2-TM regions compared with the accessory gene regions of the genome. The pattern of breakpoint frequency along the genome is remarkably similar to those in CRFs and URFs from publicly available datasets (Fan et al. 2007). These authors were the first to hypothesize that envelope is often inherited intact, being transferred into new genomes as an integral unit (Archer et al. 2008). Functional constraints of protein and RNA folding could drive these patterns, as has been shown in vitro (Galli et al. 2010), and this appears particularly pertinent in the envelope region, where the K-means clustering and GLMM result showed that the C2-TM region is often inherited intact. The gp120 protein is essential for cellular entry and for outcompeting other strains (Marozsan et al. 2005), and its recombination is likely to come up against functional constraints (Simon-Loriere et al. 2009). The three dimensional structure of envelope shows the interdependence of the gp120 and gp41 proteins, and the disruption

of internal residue contacts is expected to decrease the fitness of recombinants (Woo et al. 2014). The intricate interdependences of *env* proteins have been further demonstrated in vitro (Bagaya et al. 2015), and also by computational simulations of protein folding (Golden et al. 2014).

Sequence identity (Baird et al. 2006, Archer et al. 2008) and RNA structure (Galetto et al. 2004) have been shown to predict recombination frequency along the HIV genome. RNA structures have also been shown to potentially enable the recombination of envelope (Simon-Loriere et al. 2010), and in particular, a hairpin in C2 is identified as a driver of recombination. This mechanistic explanation of recombination in envelope, taken together with the seemingly universal breakpoint pattern and in the global CRF datasets, may suggest the genome recombination pattern and the recombination of C2-TM as an integral unit as observed here, is not unique to Uganda, but may be generalized to other population settings.

Finding potential CRFs among a myriad of recombinant genomes is not straightforward as standard phylogenies are violated by recombination, but sequences that have a more recent common ancestor (such as CRFs) should be identifiable as a cluster. However, independent recombination events with convergent recombination patterns involving the same subtypes and breakpoints will be difficult to distinguish from CRFs that originated years or decades ago. It is also possible that some recombination events are sequential, where recombinant genomes undergo new recombination, creating breakpoints of different ages in the same genome.

We searched all recombinant sequences for shared breakpoints which would suggest recombinants had been transmitted. The error associated with breakpoint assignment in SCUEAL will be related to diversity in the surrounding region. Any case where transmission of a recombinant had occurred would lead to the flanking sequences either side of the breakpoint being homologous even if subsequent recombination caused the descendent sequences to be relocated in the phylogeny. Given the difficulty of applying phylogenetic approaches we estimated simple genetic identity across the breakpoint between putative examples of transmitted recombinants. This revealed a small number which could be assigned to transmission pairs. Overall 91% of these recombinants are unique, as previously seen in *pol* sequences (Yebra et al. 2015), and parallels the general low frequency of transmission pairs in the Ugandan general population (Bbosa, Ssemwanga, Nsubuga, Salazar-Gonzalez, Salazar, Nanyonjo, Kuteesa, Seeley, Kiwanuka, Bagaya, Yebra, Leigh-Brown & Kaleebu 2019). A high prevalence of URFs in Uganda and neighboring Kenya has been seen in earlier studies (Harris et al. 2002, Yang, Li, Shi, Winter, Van Eijk, Ayisi, Hu, Steketee, Nahlen & Lal 2004, Lee et al. 2017) pointing to their continual creation, which would require a relatively high dual infection rate. In

general, this would be expected to be found in transmission networks of higher degree than observed here (we found only 12 linked pairs and a triplet in a pool of 164 A1/D recombinants). It appears from this inconsistency that the HIV transmission network structure in Uganda is more complex than generally thought.

The distinct lack of CRFs in the dataset suggests recombinants are unable to establish in any appreciable way. A recombinant might be transmitted widely if it has some biological advantage (Turk & Carobene 2015) or after going through a bottleneck in a new susceptible population e.g. CRF01_AE (**?**), but neither appears to hold true in this already established and diverse epidemic. However, since the sampling density is low and only a small sample of closely linked pairs of genomes were found, our findings could also be consistent with the presence of circulating recombinants at low frequency.

Recombination is an important evolutionary force, observable at every scale, from within-patient (Song et al. 2018) to deep in HIV evolutionary history, before even the divergence of the subtypes (Olabode et al. 2019). Significant efforts have been made to quantify the general population level of recombination in HIV-1 using coalescent-based estimators (McVean et al. 2002, Taylor & Korber 2005) which concluded that it can be extremely high, particularly in comparison with other viruses with comparable levels of population nucleotide diversity (e.g. HCV). Taylor and Korber extended their analysis to estimate possible levels of superinfection consistent with both the within-individual recombination level they inferred and that of the frequency of recombination inferred at the population level. They suggested that the superinfection level could be as high as 15% in some combinations of parameter values. However, as they pointed out, they did not consider non-random mixing in the population, which generally applies to sexual networks (Liljeros et al. 2001).

Here we have shown pervasive levels of recombination in Uganda, both within and between subtypes. While at the population level some patterns of recombination breakpoints are more prevalent than expected, the effect is not large, and certainly has not given rise to outgrowth of any particular recombinant, or CRF, as the great majority are unique. A major assumption of any phylogenetic analysis is that no recombination between sequences has taken place. The greatest impact of the inferred high level of recombination in the dataset therefore appears to lie on the reconstruction and interpretation of HIV phylogenies. This may be especially true for sequences with overlooked intra-subtype recombination.

Acknowledgements

We are grateful to Dr Jarrod Hadfield for helpful discussions and assistance with the GLMM analysis, and to Dr Anne Hoppe for her skillful management of the PANGEA-HIV programme and Dorothea Seiler Vellame for sharing useful R scripts. We would like to thank the editors and three anonymous reviewers for helpful suggestions and constructive criticism.

Supplementary information / Data Availability

Sequence data analysed in this work have been submitted to Genbank under accession numbers MN788736: MN790202. The whole-genome version of SCUEAL is available on Github (https://github.com/veg/hyphy-analyses).

2.6 Supplementary Information

In silico A1/D inter-subtype recombinants, with a random number of breakpoints from 1 to 3, random breakpoint locations, and a random selection of three “pure” A1 (labelled Ax, Ay, Az) and three “pure” D subtype sequences (labelled Dx, Dy, Dz) taken from the PANGEA dataset (pre-screened with SCUEAL) were created. Each *in silico* recombinant was analysed by SCUEAL 100 times.

The majority of the SCUEAL replicates found inter-subtype breakpoints as expected. Examples 3, 4, 5, and 10 worked particularly consistently (100/100 replicates found breakpoints in the same 100bp region as expected). More difficult scenarios, (examples 1 and 7 with small recombination fragments) still performed well finding breakpoints most of the time (60 or more/100), with other breakpoints close by.

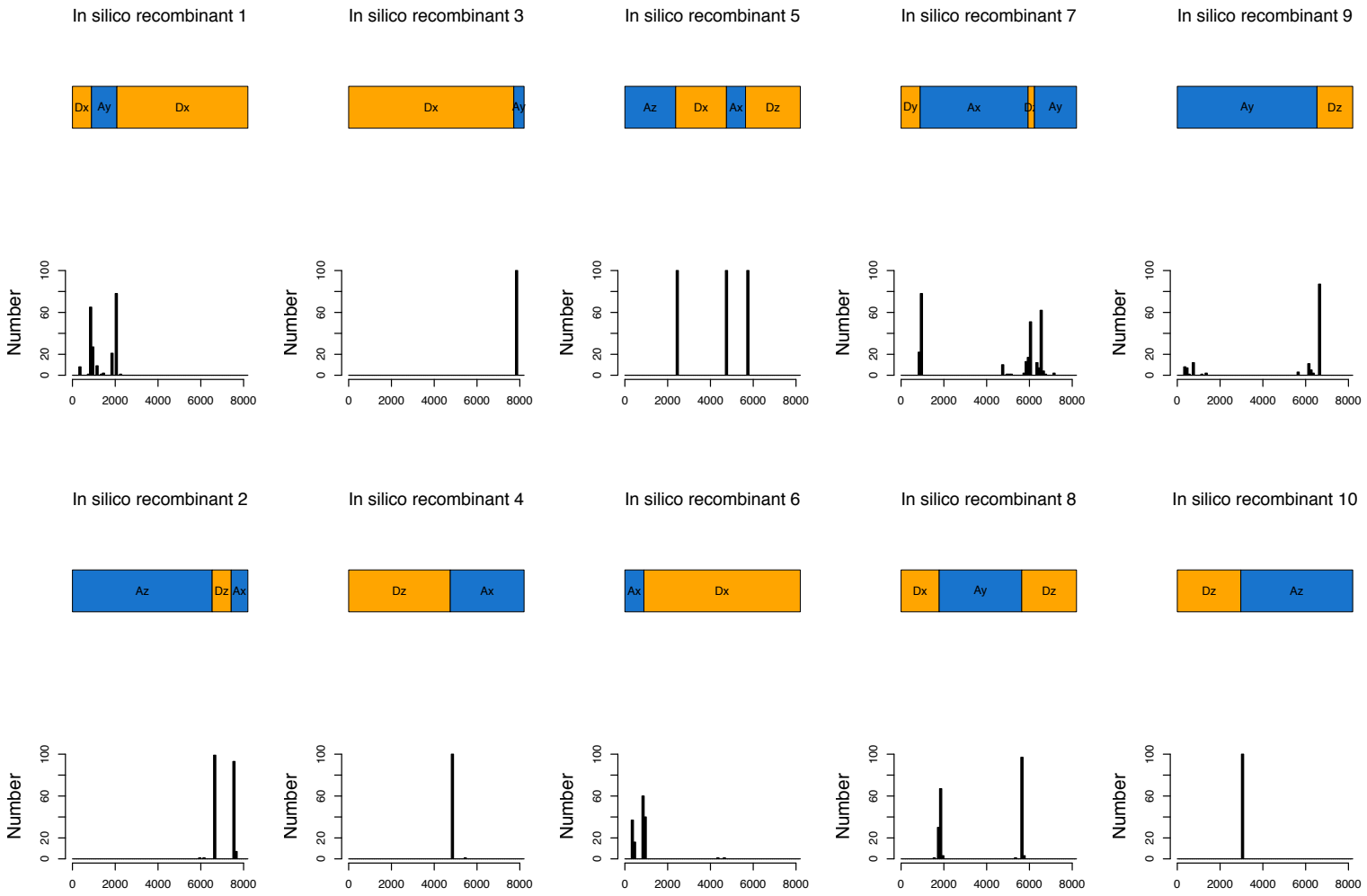

Figure 2.7: For each of ten in-silico inter-subtype recombinants, and the number of inter-subtype breakpoints found for each 100bp-region along the genome in 100 replicate SCUEAL assessments

To test the sliding window pairwise linkage approach described in the methods (section 2.4), dummy recombinants were generated using the SCUEAL reference genomes to show that recombinants with the sections of the same subtype would be linked. Twenty of these recombinants were generated with a random selection of genomes (23 A1 SCUEAL reference genomes and 12 D SCUEAL reference genomes labelled A1:A23, D24:D35) and a random number of breakpoints between 1 and 3. Fig 2.8 shows these recombinants in the pairwise window linkage analysis, where windows containing the same subtype were successfully linked. Note that at the 2% level, some windows will match even with different reference subtypes, (e.g. references A4 and A20 matched in window 13 in pair 5), but this was linkage in a single window only.

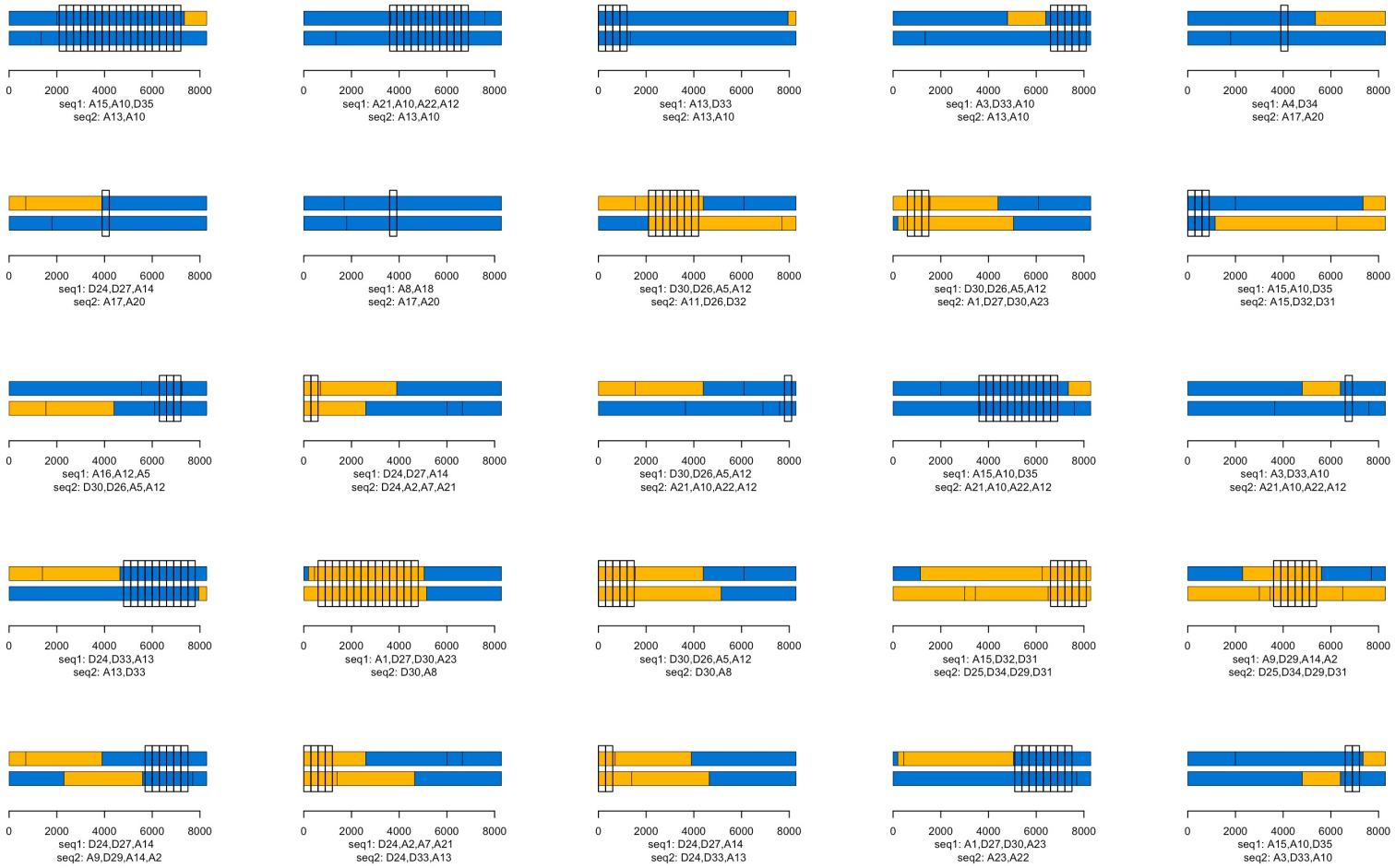

Figure 2.8: Pairs of randomly generated recombinant genomes linked by a distance of less than 2% (TN93) in more than one window along the genomes. The matching windows are shown with open clear boxes, and the SCUEAL subtyping result is in colour (blue for Subtype A1 and orange for Subtype D).

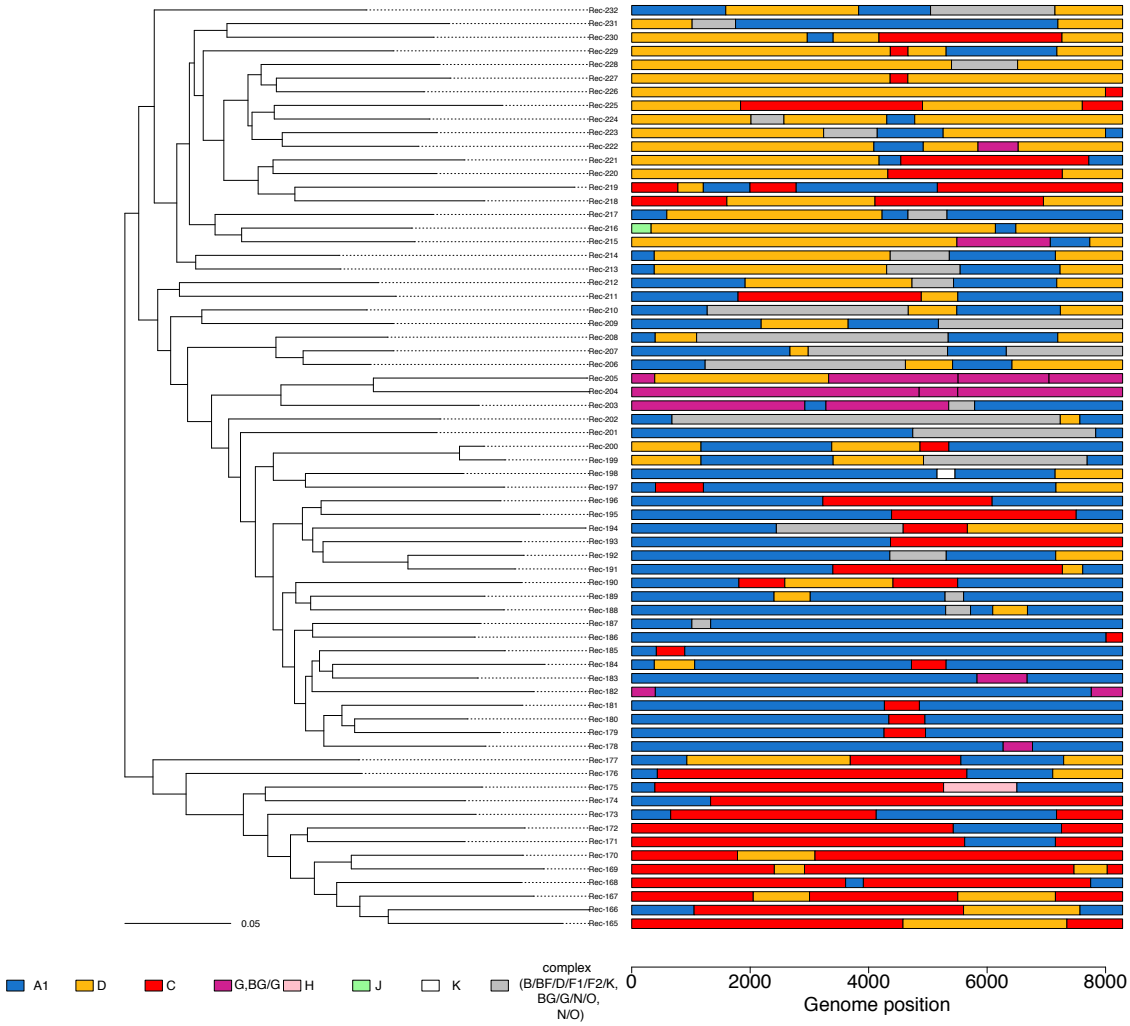

Figure 2.9: Maximum-likelihood reconstruction of the non-A1/D recombinants using IQ-TREE (n=68), midpoint rooted, and their SCUEAL subtype assessment (right). A phylogenetic tree of recombinants clearly violates the assumption of absence of recombination, but is used to show the absence of any clear CRF

Supplementary information: K-means clustering justification

Fig 2.10 shows three methods for determining the optimal number of clusters. The gap statistic (top left) compares the within cluster variation for each value of k against a null reference distribution. Here 9 appears to be most appropriate value of k. The elbow method (top right) plots all within-cluster sum of squares for each value of k, and the bend or elbow in the plot is an indicator of the appropriate number of clusters. There is a curve between k=2 and 5 without an obvious elbow, however there is a kink at k=9. The silhouette method (bottom left) shows the average silhouette width (a measure of how tightly grouped genomes within each cluster are), and k=2, k=3, k=5 or k=9 have higher silhouette width.

Fig 2.11 shows genomes clustered with k=2, k=3, k=5 and k=9. Using k=2 broadly splits the genomes into predominantly subtype A1, and subtype D genomes, k=3 splits the genome into mostly A1, mostly D, and D genomes with an A1 envelope region. Using k=5 adds another group of A1 genomes with a D envelope region, and using k=9 adds a few smaller groups with clustering of subtype D in gag and pol regions. The broad scale pattern of interest (that is the intact inheritance of partial envelope) is evident in k=3, k=5 and k=9.

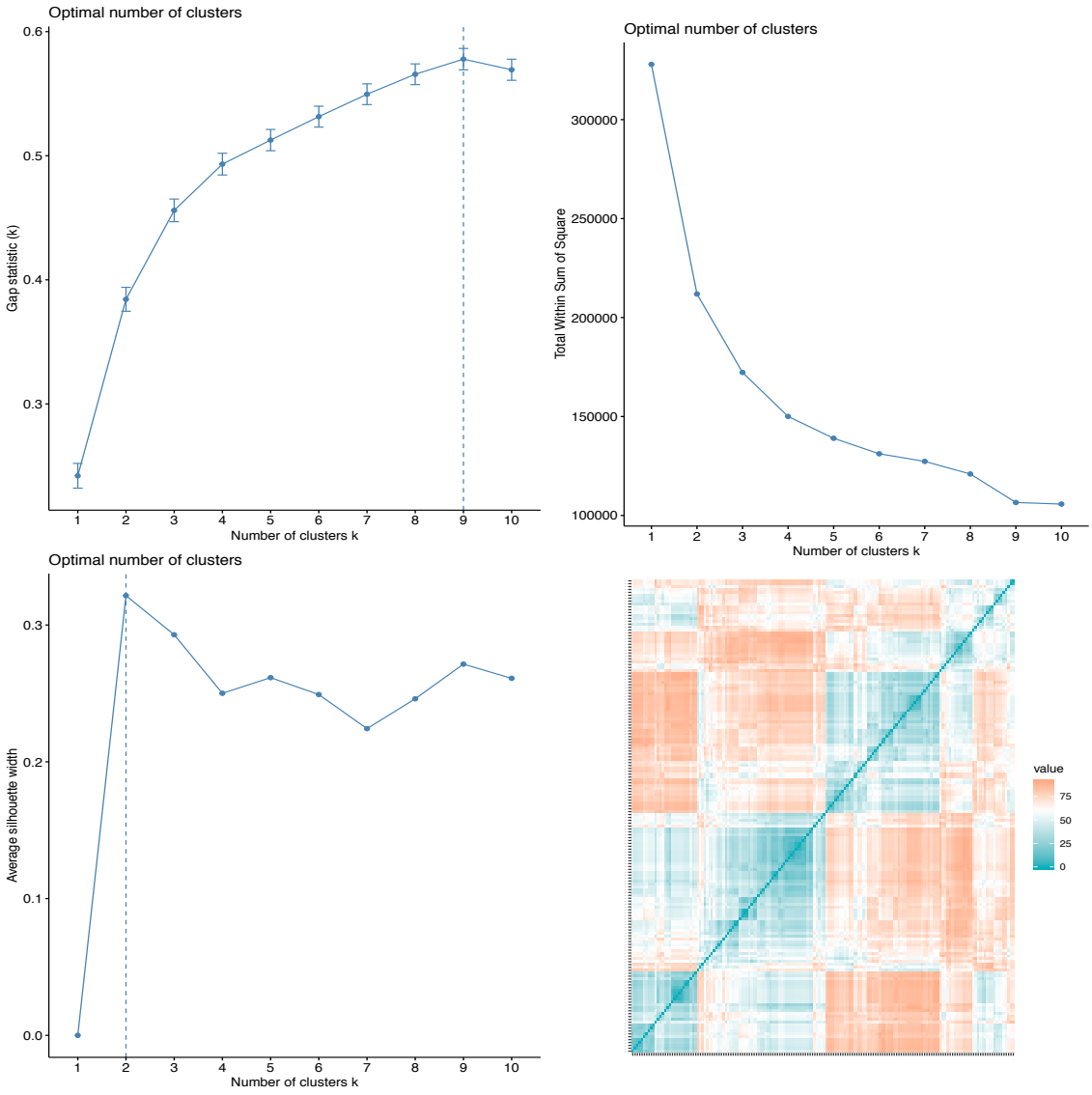

Figure 2.10: top to bottom, left to right, a) Gap statistic b) Elbow method c) Silhouette method and d) visualisation of the raw Euclidean distance matrix

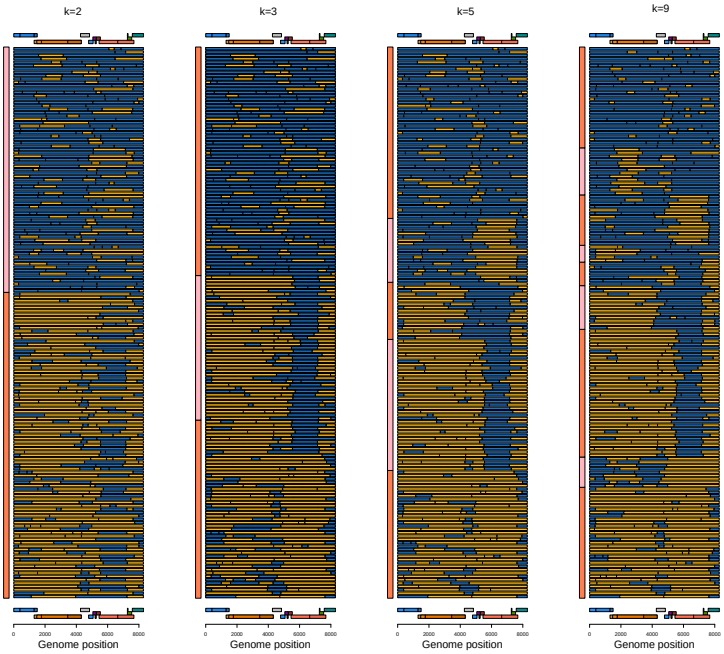

Figure 2.11: Groupings of the A1/D recombinants clustered by kmeans clustering with alternative values of k (2, 3, 5, 9). Segments of orange colour represent Subtype D, while blue colouration represents subtype A1

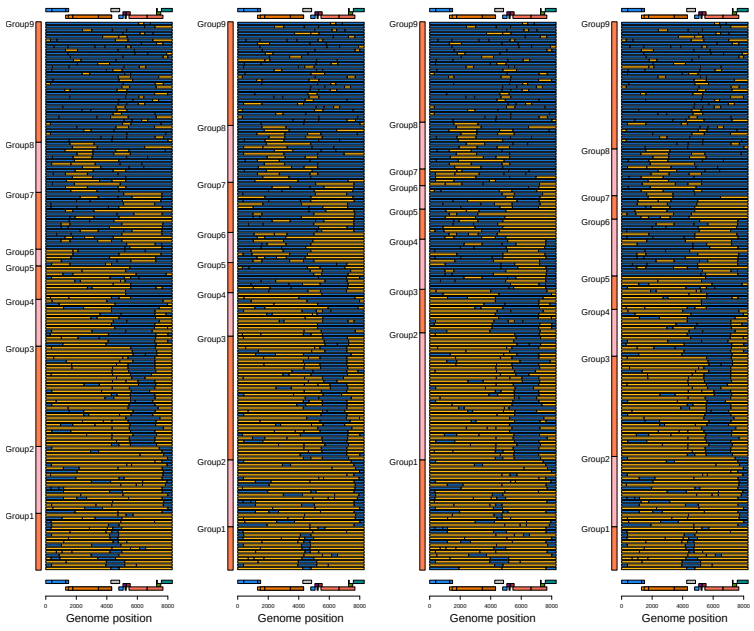

Figure 2.12: Four alternative k=9 groupings. The K-means clustering algorithm is stochastic in that it takes a random selection of k genomes as starting point. There are overall similar patterns albeit with subtle differences

Chapter 3

Historical HIV genomes from 1986 Uganda show a change in subtype frequency driven by co-receptor tropism

Heather E. Grant[1], Sunando Roy[2], Rachel Williams[2], Helena Tutill[2], Bridget Ferns[3], Patricia Cane[4], J. Wilson Carswell, Deogratius Ssemwanga[5], Pontiano Kaleebu[5], Judith Breuer[2], Andrew J. Leigh Brown[1]

1) Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
2) Division of Infection and Immunity, University College London, London, UK
3) Department of Virology, University College London Hospitals NHS foundation Trust, UK
4) High Containment Microbiology, UK Health Security Agency, Porton Down, Salisbury, UK
5) Medical Research Council (MRC)/Uganda Virus Research Institute (UVRI) and London School of Hygiene and Tropical Medicine (LSHTM) Uganda Research Unit, Entebbe, Uganda

3.1 Abstract

We present 109 full-length HIV genomes amplified from blood serum samples obtained during early 1986 from across Uganda, which to our knowledge is the earliest and largest population sample from the initial phase of the HIV epidemic in Africa. Consensus sequences were made from paired-end Illumina reads with a target-capture approach to amplify HIV material following poor success with standard approaches. In comparisons with a smaller 'intermediate' genome dataset from 1998-9 and a 'modern' genome dataset from 2007-2016, the proportion of subtype D was significantly higher initially, dropping from 67% (73/109), down to 57% (26/46) and 17% (82/465) respectively ($p<0.0001$). Subtype D has previously been associated with a faster rate of progression than subtype A1 in Ugandan population

studies, and a higher CXCR4 co-receptor tropism (or 'X4') which is known to decrease the AIDS free period. Here we find significant differences in co-receptor usage between A1 and D subtypes in all three sample periods considered, which is particularly striking in the 1986 sample: 66% (53/80) of subtype D *env* sequences were predicted to be X4 tropic compared with none of the 24 subtype A1. We also analysed the frequency of subtypes in the envelope region of inter-subtype recombinants, and found that subtype A1 is over-represented in *env*, suggesting recombination and selection have acted to remove subtype D from circulation. The reduction of subtype D frequency over three decades appears to be a result of selective pressure against X4 tropism and its higher virulence. Lastly, we find a subtype D specific codon deletion at position 24 in the V3 loop, which may explain a higher propensity for subtype D be X4 tropic.

3.2 Introduction

The main (M) group of HIV-1 viruses that caused the global AIDS epidemic can be categorised into major lineages or "subtypes" (Robertson et al. 2000). Evidence points to the epicentre of the HIV epidemic being Kinshasa in the Democratic Republic of Congo (DRC) (Sharp & Hahn 2011) in the early part of the Twentieth Century (Faria et al. 2014). HIV-1 transmission was largely confined to the DRC for multiple decades where it underwent substantial recombination (Kalish et al. 2004, Ward et al. 2013). However, by the 1960s, strong genetic bottlenecks had created geographically (Faria et al. 2019) and phylogenetically distinct subtypes (Worobey et al. 2008). These subtypes subsequently spread worldwide, resulting in the current global subtype distribution (Bbosa, Kaleebu & Ssemwanga 2019).

There has been much speculation and interest in the possibility of phenotypic differences between subtypes (see review by Geretti 2006) that may have contributed to the relative success of any one subtype over another (Essex 1999). However, viral characteristics such as infectivity or virulence are confounded by a range of factors like mode of transmission, host genetics and health. Subtype comparisons within the same country, population, or cohort are therefore strengthened by the reduction in these factors (Kuritzkes 2008) such as those in Uganda where subtypes A1 and D co-circulate at high enough frequencies. Subtype D has been shown consistently in large population studies to progress to AIDS faster compared with subtype A1 (Kaleebu et al. 2001, 2002, Vasan et al. 2006, Kiwanuka et al. 2008, Bousheri et al. 2009, Ssemwanga et al. 2013, Easterbrook et al. 2010). It has also been reported that individuals infected with subtype D reach higher viral loads more rapidly (Amornkul et al. 2013), and that subtype D clones have a higher replicative capacity in vitro (Baalwa et al. 2013). Other studies have indicated that subtype D viruses are more likely to use CXCR4 co-receptors (Tscherning et al. 1998, Huang et al. 2007, Kaleebu et al. 2007).

Co-receptor tropism (the secondary receptor used alongside CD4) can be distinguished in cell-culture where “fast replicating” syncytium inducing (SI) viruses are CXCR4 (X4) tropic and “slow” non syncytium inducing (NSI) viruses are usually CCR5 (R5) tropic (Connor et al. 1997). Fast replicating X4 tropic viruses have long been associated with faster CD4 decline (Koot et al. 1993), and risk of AIDS progression could be as much as 3.8x higher (Daar et al. 2007) which in real terms translates to several years of reduction in lifespan. Comparisons of R5 and X4 viruses at the V3 loop showed that positive charges at position 11 and 25 are strongly predictive of X4 tropism (the ’11/25 rule’; de Wolf et al. 1994). Currently, more sophisticated machine learning models are used to predict co-receptor tropism based on V3 amino acid training data (e.g. geno2pheno; Lengauer et al. 2007).

Genetic sequencing is an important epidemiological tool for gaining understanding of epidemics which can then be used to inform intervention strategies. HIV sequencing is important for use in testing for drug resistant mutations, but can also provide insights about epidemic size and diversity e.g. (Ssemwanga et al. 2020) or movement between key populations by phylogenetic analysis e.g. (Kiwuwa-Muyingo et al. 2017, Bbosa, Ssemwanga, Nsubuga, Salazar-Gonzalez, Salazar, Nanyonjo, Kuteesa, Seeley, Kiwanuka, Bagaya, Yebra, Leigh-Brown & Kaleebu 2019). Sequencing in East Africa up until 2013 had been limited mostly to consensus Sanger sequences of partial gene sequences of p24 or gp41 (Lamers et al. 2016) although partial *pol* sequencing (where many drug resistance mutations are found), has become more common since the antiretroviral therapy roll out. There is very little genome sequence data from the 20th century, the exception to this being a 46-genome dataset from samples taken in the Rakai district in 1998/1999 (Harris et al. 2002). The PANGEA project (Pillay et al. 2015) aimed to rectify this for the 21st century and has obtained large datasets of full-length sequences from Africa to provide more phylogenetic information (Yebra et al. 2016).

Samples from serological surveys conducted in early 1986 from hospitals and antenatal clinics in Uganda were re-discovered in storage in 2013 during the relocation of what were then the Public Health England laboratories at Porton Down. Standard clinical *pol* sequencing (Cane 2011) had limited success with these historic samples (Yebra et al. 2015), and amplification and sequencing success with the PANGEA full-length genome protocol was limited. To overcome the problems associated with considerable RNA degradation, we used target-capture techniques with RNA baits designed to capture a wide variety of HIV-1 M (Depledge et al. 2011) to recover 109 new full-length (and 37 partial) genomes. This is a unique population dataset from the early African epidemic, shortly after AIDS was discovered, and from a decade where very few HIV genomes are available, particularly from Africa.

3.3 Methods

3.3.1 Sample preparation

Serum samples were collected from across Uganda between January and May 1986, as part of a serological survey of HIV prevalence in different populations (Carswell 1987). Samples were sent to Porton Down in the UK for antibody testing in 1986, which was at the time one of the only UK facilities able to receive such samples. The samples, stored at -80^{o}C since, were brought to the attention of the PANGEA project in 2013 and transferred to UCL. The remaining samples have subsequently been returned to the UVRI in Entebbe, Uganda.

HIV positive samples were identified by ELISA and RNA extracted. A target-capture approach (Depledge et al. 2011) developed for degraded RNA viruses was adopted. Thus 120 base pair capture baits were designed with an in-house pipeline to target the whole HIV genome, using 2635 reference genomes covering global subtype and CRF diversity. cDNA libraries were constructed with SuperScript IV Reverse Transcriptase (Invitrogen) followed by NEB Second Strand cDNA Synthesis before using the SureSelectXT Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library. This included a pre-capture PCR step during library preparation; followed by bait hybridization and a capture step with streptavidin beads to enrich for HIV fragments; and a post-capture indexing PCR. Paired end sequencing was carried out with the Illumina MiSeq v2 500 cycle kit.

3.3.2 Sequence assembly

Trimming, adapter removal, and quality checking of reads was performed with TrimGalore, cutadapt and FastQC (Andrews 2010, Martin 2011, Krueger 2020), using a minimum Phred score of 30. Mapping to reference genomes was done with the Burrow-Wheeler Aligner MEM algorithm (Li 2013) and the samtools and bcftools libraries (Danecek et al. 2021), firstly to 170 reference genomes (encompassing a wide range of subtype and CRF diversity) to identify the best genotype, and then to the best reference for a single reference assembly. A visual assessment in Geneious Prime 2022.0.1 (www.geneious.com) was carried out to check for good coverage across the genome, or any dips that might indicate an inter-subtype recombinant sequence. If this was the case, the multi-reference BAM files were examined, or an alternative de-novo assembly with HAPHPIPE and SPAdes was attempted (Bankevich et al. 2012, Gibson et al. 2020). Either the single reference assembly (or de-novo assembly if improvement could be found) was then fed into the HAPHPIPE framework for fine tuning with three rounds of iterative improvement. Coverage statistics and vcf files were produced for each and finally a consensus sequence with a minimum of 10x coverage at every base pair position was generated using GATK (McKenna et al. 2010) within HAPHPIPE.

Full length consensus genomes (>8000bp from *gag* to *nef*) with minimum 10x coverage were selected for analysis from an extended dataset of partial genomes (> 1000bp per sample). Full-length genome consensus sequences have been deposited in GenBank (numbers OP039379: OP039487), and partial sequences (OP39488: OP039526), and read data are available on request.

3.3.3 'Intermediate' and 'modern' datasets

In addition to the newly generated 'historical' dataset, a collection (n=46) of genomes from the Rakai district (Uganda) in 1998 and 1999 provided an 'intermediate set' (Harris et al. 2002), whilst a 'contemporary set' was taken from the MRC/UVRI PANGEA genome collection (n=465) sampled in Central Uganda between 2007 and 2016 (described fully in Grant et al. 2020).

3.3.4 Subtyping and co-receptor prediction

All genomes were subtyped with the full genome version of SCUEAL (Kosakovsky Pond et al. 2009). All available sequences were subjected to co-receptor prediction using the geno2pheno co-receptor tool (Sing et al. 2007) first by aligning the V3 loop by eye and extracting the amino acid sequence in Geneious Prime 2022.0.1 (www.geneious.com). The inter-subtype recombinant genomes (unique recombinant forms; URF) from all three datasets with a clear majority (over 70% the length of *env* determined as A1 or D by SCUEAL breakpoints and clearly covering the V3 loop) were included. Subtype level consensus amino acid sequences were found and Shannon's Entropy of both were calculated and compared with the Entropy-Two tool from the Los Alamos Database
(https://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html).

3.4 Results

3.4.1 Historical sequences

HIV specific baits were used in a target-capture step to enrich HIV genetic material before Illumina MiSeq sequencing to generate a paired-end read dataset of 109 full length consensus sequences with a minimum of 10x coverage at every position. In addition to these full-length consensus genomes (> 8000bp from gag to nef), 37 partial sequences (> 1000bp) were generated (a 65% genome recovery success from 168 samples, or 87% partial sequence recovery). Table 3.4 shows sequence consensus lengths, SCUEAL subtype assignment, and read depth of the genome assemblies for full-length genomes, and Table 3.5 partial sequences.

Location	Hospital	Subtype D	Subtype A1	Inter-subtype recombinants	Subtype C	Total
Kampala and Central Region		32	12	10	1	55
	Rubaga (Lubaga)	(11)		(1)		
	Mulago	(4)	(2)	(2)		
	Nsambya	(3)	(3)	(1)		
	Antenatal Clinics		(1)	(1)		
	Unknown	(14)	(6)	(5)	(1)	
Masaka District	Kitovu	27	1	3		31
Gulu	Lacor	4				4
Jinja	Unknown			1		1
Other Uganda	Unknown	10	4	4		18
Total		73	17	18	1	109

Table 3.1: Frequency of HIV genomes by subtype recovered from historical samples by sampling location

This method is considerably more sensitive than without the target-capture step; in 2014 some of these samples were subjected to the PANGEA protocol (Gall et al. 2012) with modest success, generating 5 full length genomes and 17 partial genomes, (a success rate of 5% and 22% respectively from a 96-sample plate; data not shown). Average coverage spanned from x27 to x1769, with no significant difference found between subtypes or between subtypes and inter-subtype recombinants (Figure 3.1).

Of the 109 consensus genomes, 90 had some basic location information (Table 3.1). The majority are broadly from the “Central” region (n=53) which includes hospitals within Kampala, Rubaga (n=12), Mulago (n=8) and Nsambya (n=7). There were 31 genomes from Kitovu Hospital (Masaka District), and 4 from Lacor hospital (northern Uganda), one from a hospital in Jinja (80km East of Kampala), and 2 from unidentified antenatal clinics (see map, Figure 3.2).

While subtype distribution varied across Uganda, subtype D predominated in the historical dataset (73/109, 67.0%), followed by subtype A1 (17/109, 15.6%), inter-subtype recombinants composed of A1, D (17/109, 15.6%) and A1, C and D (1/109, 0.9%), and subtype C (1/109, 0.9%). All inter-subtype recombinants had a unique recombination pattern (Figure 3.5).

3.4.2 Temporal change in subtype frequency

The SCUEAL designated subtype distribution for the ‘intermediate’ genome dataset was 26 D (56.5%); 7 A1 (15.2%); and 13 inter-subtype recombinants containing A1, D, and C (28.2%). The ‘modern set’ had the distribution: 82 D (17.6%); 3 C (0.6%); 143 A1 (30.8%) and 232 inter-subtype recombinants (49.9%). The proportional change of subtype over the three periods is illustrated in Figure 3.3. Combining other subtypes with recombinants, the relative frequencies

	Historic 1986	Intermediate 1998/9	Modern 2007-2016	Total
D (genome)	73	26	82	181 (52% of genomes)
D env (URF)	8	7	61	76 (37% of URF)
A1 (genome)	16	5	143	164 (48% of genomes)
A1 env (URF)	8	8	115	131 (63% of URF)

Table 3.2: Number of *env* sequences available from 'pure' genomes and URFs for subtype A1 and D at three sampling time points

Subtype	Historic 1986			Intermediate 1998/9			Modern 2007-2016		
	X4	R5	Proportion X4	X4	R5	Proportion X4	X4	R5	Proportion X4
D (genome)	47	26	53/80 (66%)	9	17	11/33 (33%)	44	38	70/143 (49%)
D env (URF)	6	1		2	5		26	35	
A1 (genome)	0	16	0/24 (0%)	0	5	0/13 (0%)	5	136	13/256 (5%)
A1 env (URF)	0	8		0	8		8	107	
Other	0	1	0%	0	0	0%	5	55	8%
Total	53	52	50%	11	35	24%	88	371	19%

Table 3.3: Co-receptor tropism predictions for subtypes D and A1. Distinction is made between V3 sequences from within "pure" genomes and URFs.

of A1 and D, are significantly different in these three time periods (χ^2 = 122.68, df = 4, p < 0.0001). Specifically, there is evidence for a linear decrease in subtype D frequency over time (Cochran Armitage Z = -10.861, dim = 3, p < 0.0001), with subtype D being replaced with subtype A1 and inter-subtype recombinants.

In addition, the majority subtype within the envelope region of all inter-subtype recombinants from all datasets was assessed. Based on SCUEAL-estimated breakpoints a threshold of 70% over the length of the *env* gene, including the V3 loop, was used to classify *env* as D, A1, or not clearly either. Table 3.2 shows the frequency of *env* subtypes A1 and D from 'pure' genomes and URFs over the three periods. There were many more URFs with subtype A1 *env* (n=131), than URFs with subtype D envelopes (n=76), significantly different to the genome level expectation (χ^2= 45.973, df = 3, p-value < 0.0001).

3.4.3 Co-receptor usage

The machine learning application geno2pheno (Lengauer et al. 2007) was used to predict virus co-receptor tropism of all V3 sequences in the three datasets. Adopting a 5% false positivity rate threshold, there is a significantly higher proportion of CXCR4 coreceptor usage by subtype D compared with A1 at all three sampling times (Table 3.3). Of the historical set, 66% (53/80) of subtype D envelope sequences were predicted to be X4 tropic whilst none (0/24) were predicted to be X4 tropic for the A1 sequences (χ^2= 29.8, df=1, p<0.0001). Of the intermediate genomes, 33% (11/33) of subtype D were X4 tropic compared with none (0/13) of subtype A1 (χ^2=4.01, df=1, p=0.04), and of the modern day 49% (70/143) were X4 tropic, compared with 5% (13/256) for subtype A1 (χ^2=104.5, df=1, p<0.0001).

3.4.4 Subtype specific differences in V3 loop at the amino acid level

The consensus V3 amino acid sequences of Subtypes A1 and D from Uganda are shown with the entropy at each codon position underneath (Table 3.4). We used the Los Alamos Entropy-Two tool which uses randomisation with replacement to test for differences in entropy between subtypes. In total 14 positions were significantly more entropic in Subtype D than A1 (including the crucial positions of 11 and 25), while three sites were more diverse in subtype A1 than D (positions 19, 22, and 24).

The consensus length of subtype A1 was 35 codons, whilst that for Subtype D was 34 codons due to a deletion at position 24. This deletion is found in the vast majority of historic (94%; 68/72), and modern-day (90%; 73/81) subtype D sequences. Whilst the deletion 24 is found in the vast majority of Ugandan subtype D sequences, it is found only in some subtype D outgroup sequences, and not found in the Subtype B consensus (Figure 3.4). This suggests that the deletion happened after the subtype B/D split, but before the appearance of subtype D in Uganda. There are a number of additional sequence changes in the recent PANGEA subtype D V3 loops, including an additional deleted codon at position 23 in many sequences, confirming a distinctive difference in the behaviour of this region of *env* in this subtype. All V3 amino acid sequences and G2P results are available from (github.com/heathergrant/HIVdata).

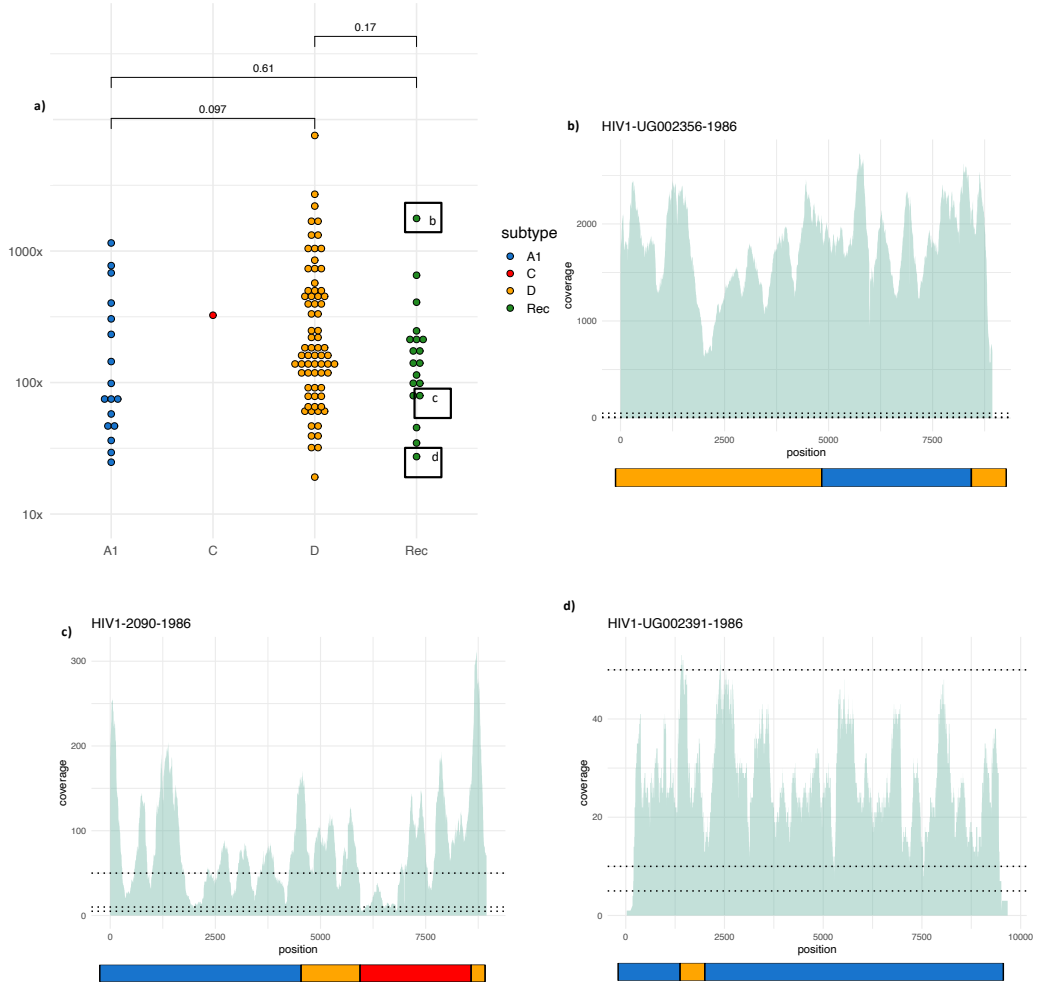

Figure 3.1: Average read coverage across the genome for each genome assembly, and three example coverage plots of inter-subtype recombinant genomes. Panel a) gives the average coverage per genome assembly by subtype (base 10 scale), which shows no significant difference between subtypes A1 and D, or between either subtype and the inter-subtype recombinants. Panel b) shows the highest inter-subtype recombinant coverage plot (x1769), c) the only A1,C,D inter-subtype recombinant genome assembly coverage plot (average x80), and d) the lowest coverage inter-subtype recombinant (average x27). The dotted horizontal lines show 5x, 10x and 50x and the SCUEAL designated breakpoints are shown as bars underneath each plot (subtype A1 in blue, D in orange, and C in red).

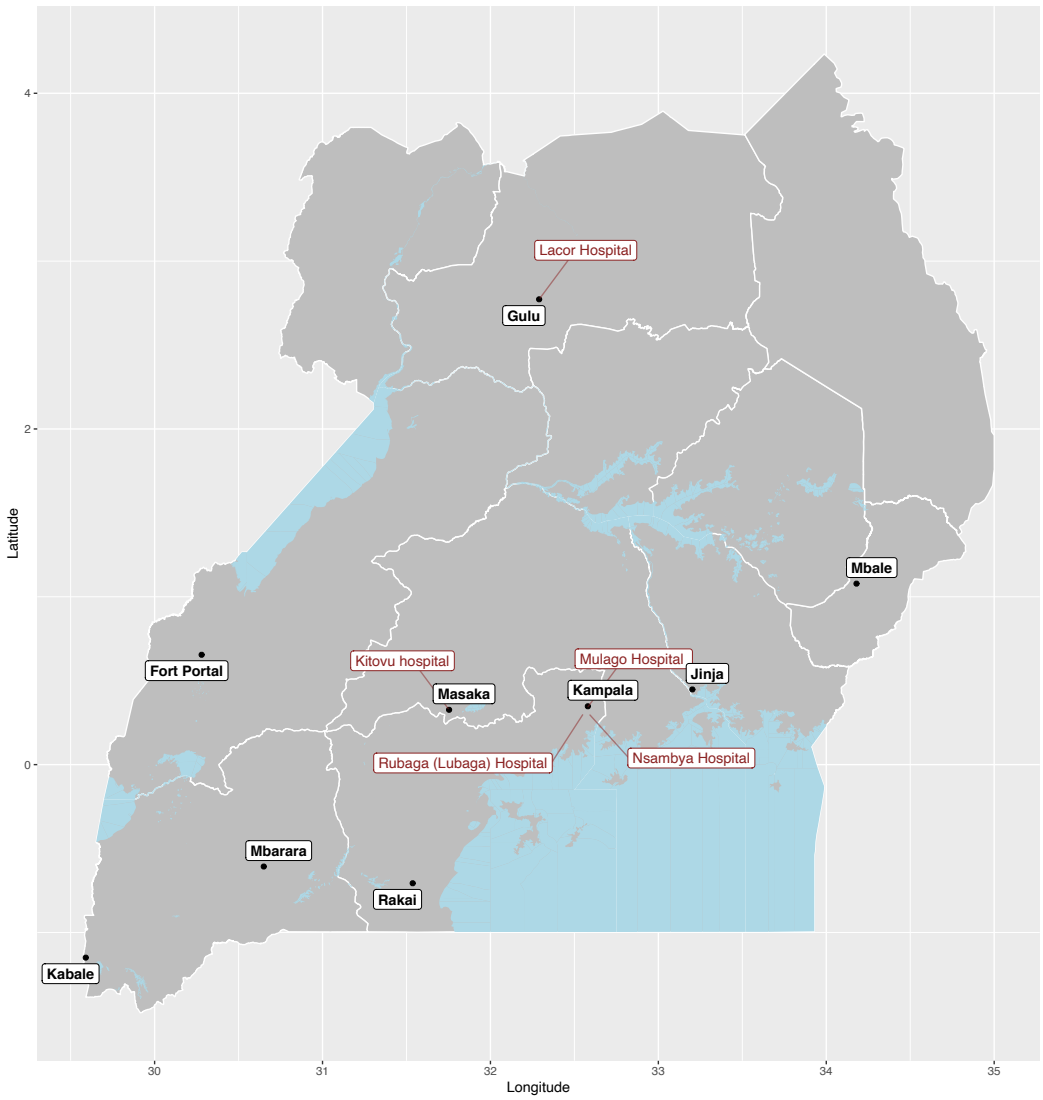

Figure 3.2: Map of Uganda with sampling locations of historical samples, including Kampala, Masaka, Jinja, and Gulu and hospitals Mulago, Nsambya, Rubaga, Kitovu

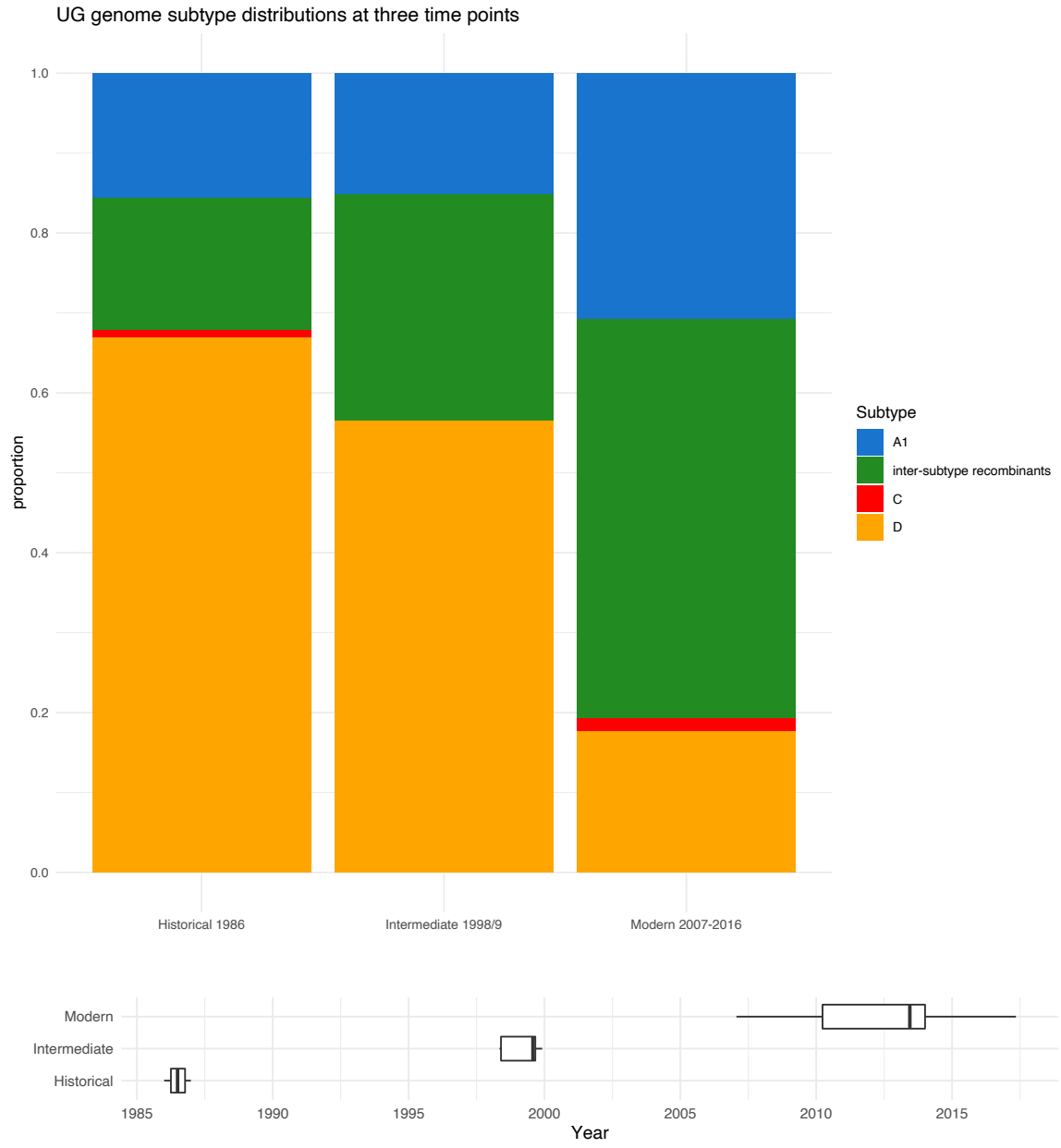

Figure 3.3: Bars show subtype proportions in the 1986 'historical' genome set described here n=109, the 'intermediate set' from Harris et al. 2002 (n=46), and the 'modern' PANGEA dataset (described in Grant et al., 2020, n=465). The sampling dates of each period are shown as box plots underneath.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	G2P prediction
a)	**Subtype A1 and D comparison:**																																				
	Subtype A1 UGANDA (n=164)	C	T	R	P	N	N	N	T	R	K	S	V	H	I	G	P	G	Q	A	F	Y	A	T	G	D	I	I	G	D	I	R	Q	A	H	C	*R5*
	entropy	0.00	1.01	0.00	0.09	0.85	0.09	0.00	0.20	0.09	1.06	0.58	0.86	1.19	0.28	0.19	0.00	0.04	0.60	0.96	0.23	0.26	0.76	0.82	1.47	1.19	0.34	0.33	0.07	0.73	0.16	0.04	0.92	0.00	0.83	0.00	
	Subtype D UGANDA (n=180)	C	T	R	P	Y	N	N	T	R	Q	S	T	H	I	G	P	G	Q	A	L	Y	T	T	-	K	I	I	G	D	I	R	Q	A	H	C	***X4***
	entropy	0.00	0.82	0.00	0.03	1.06	0.40	0.24	0.43	0.37	1.28	0.96	1.12	0.99	0.71	0.07	0.32	0.00	0.95	0.52	1.30	0.70	0.36	1.51	0.55	2.18	0.93	0.94	0.13	0.73	0.26	0.15	0.54	0.00	0.84	0.00	
	Entropy difference Subtype A1 v D	0.0	0.2	0.0	0.1	-0.2	-0.3	-0.2	-0.2	-0.3	-0.2	-0.4	-0.3	0.2	-0.4	0.1	-0.3	0.0	-0.3	0.4	-1.1	-0.4	0.4	-0.7	0.9	-1.0	-0.6	-0.6	-0.1	0.0	-0.1	-0.1	0.4	0.0	0.0	0.0	
	Sites where D more entropic than A1, P-value <0.01					*	*			*		*	*		*		*		*		*		*		*	*	*	*								
	Sites where A1 more entropic than D, P-value <0.01																			*			*		*												
b)	**Subtype D outgroup sequence diversity:**																																				
	Subtype B LANL compendium consensus (n=97)	C	T	R	P	N	N	N	T	R	K	S	I	H	I	G	P	G	R	A	F	Y	A	T	G	D	I	I	G	D	I	R	Q	A	H	C	*R5*
	D.CD.1983.ELI	C	A	R	P	Y	Q	N	T	R	Q	R	T	P	I	G	L	G	Q	S	L	Y	T	T	-	R	S	R	S	I	I	G	Q	A	H	C	***X4***
	D.CD.1985.Z2Z6_Z2_CDC_Z34	C	T	R	P	Y	R	N	I	R	Q	R	T	S	I	G	L	G	Q	A	L	Y	T	T	-	K	T	R	S	I	I	G	Q	A	Y	C	***X4***
	D.CD.2002.LA18ZiAn	C	T	R	P	N	V	Y	T	K	K	G	I	R	T	G	R	G	Q	A	I	L	T	T	-	Q	V	T	G	D	I	R	R	A	H	C	***X4***
	D.CD.2003.LA17MuBo	C	I	R	P	N	N	N	T	R	Q	G	V	G	I	G	P	G	Q	M	F	F	T	T	-	G	I	I	G	D	I	R	Q	A	H	C	*R5*
	D.TD.1999.MN011	C	I	R	P	N	N	N	T	R	R	S	V	H	I	G	P	G	Q	A	L	Y	T	T	-	N	V	I	G	D	I	R	Q	A	H	C	*R5*
	D.SN.1990.SE365	C	T	R	P	Y	N	N	K	R	Q	R	T	P	I	G	L	G	Q	V	L	H	T	T	-	R	V	K	G	D	I	R	Q	A	H	C	***X4***
	D.CD.1984.84ZR085	C	T	R	P	Y	K	K	E	R	Q	R	T	P	I	G	Q	G	Q	A	L	Y	T	T	R	Y	T	T	R	I	I	G	Q	A	Y	C	***X4***
	D.CD.1987.PBS5635	C	T	R	P	Y	N	N	T	R	K	G	I	H	I	G	P	G	Q	A	L	Y	A	S	T	E	I	T	G	D	I	R	Q	A	H	C	*R5*
	D.CM.2001.01CM_4412HAL	C	V	R	P	N	S	N	T	R	K	S	I	N	L	G	P	G	Q	A	F	Y	A	A	T	N	I	I	G	N	I	R	Q	A	H	C	*R5*
	D.ZA.1984.R2	C	T	R	P	Y	K	Y	T	I	Q	K	T	S	I	G	Q	G	Q	A	L	H	T	S	K	R	I	I	G	D	I	R	Q	A	H	C	***X4***
	D.BR.2010.10BR_RJ108	C	T	R	P	Y	N	N	T	R	Q	N	T	Q	I	G	P	G	Q	T	F	Y	T	S	K	R	I	I	G	D	I	R	Q	A	Y	C	***X4***

Figure 3.4: Consensus V3 amino acid sequences of subtypes A1 and D from Uganda with pairwise entropy comparison at each site and b) V3 sequences of the outgroup to subtype D in Uganda. The Entropy-Two tool from the Los Alamos National Laboratory database was used to compare Shannon's Entropy at each codon position (indicating variability at each position). Sites with significantly different ($p<0.01$) entropy between the subtype A1 consensus and the subtype D consensus are highlighted in bold. Positively charged amino acids (K, Lys) and (R, Arg) are shown in blue, while negatively charged amino acids (D, Asp) and (E, Glu) are shown in red, geno2pheno predictions are shown to the right.

3.5 Discussion

Here we describe a population sample of 109 HIV genomes from the early stages of the epidemic in Uganda, a period from which very few genomes are available globally. Two well-known isolates (MAL and ELI; Alizon et al. 1986)) were the first sequences generated from African samples (obtained following culture). However, most of the sequences from the early years of the epidemic are now retrospectively obtained by amplification of material from preserved serum or tissue, or from clones kept in cell culture. The oldest sequence fragment to date is ZR59 (from 1959 DRC) but unfortunately, only a few hundred base pairs were obtained (Zhu et al. 1998). We show here that target capture next generation sequencing can work well on highly degraded serum samples from over 30 years ago. Yamaguchi et al. (2018) also successfully employed similar methods, obtaining genomes from a range of subtypes from 1987 DRC. More recently "jackhammer" techniques recovered a 1966 genome sequence of a subtype C virus preserved in a similar way to ZR59, where target-capture methods failed (Gryseels et al. 2020). New 21st century retrospective sequencing now means that we are increasingly limited by the availability of preserved virus material rather than method sensitivity.

In this historical dataset, most genomes are 'pure' subtypes, consistent with it being an early point in the Ugandan epidemic. However, we also find 18 inter-subtype recombinant forms, all of which have a unique pattern, representing at least 18 independent co-infection or super-infection events with different subtypes before 1986. Dual infection and recombination between the two subtypes was therefore occurring long before 1986. There is now an extremely high prevalence of unique recombinant forms in Uganda (Lee et al. 2017, Capoferri et al. 2020, Grant et al. 2020), without any major circulating recombinant form. This is not unexpected within a generalised epidemic of such large scale and network complexity involving two subtypes at similar prevalence (Bbosa, Ssemwanga, Nsubuga, Salazar-Gonzalez, Salazar, Nanyonjo, Kuteesa, Seeley, Kiwanuka, Bagaya, Yebra, Leigh-Brown & Kaleebu 2019, Ratmann et al. 2020), which has not experienced any obvious bottlenecks.

Ugandan cohort studies have been of wide interest because unusually, the generalised epidemic contains two subtypes at similar proportions, which provides a natural experiment for directly comparing the phenotypes of viral subtypes. These cohort studies have consistently found subtype D to be more virulent than subtype A1, with faster drops in CD4 counts and more rapid progression to AIDS (Kaleebu et al. 2002, Kiwanuka et al. 2008). Since then, a faster rate of progression in Subtype D has been confirmed in Tanzania (Vasan et al. 2006) and in the UK (Easterbrook et al. 2010).

There is an extensive literature on the subject of differences in virulence between viral strains, often framed in terms of viral load (Fraser et al. 2007, Hodcroft et al. 2014, Blanquart et al. 2016). Viral load is a well-known predictor of HIV virulence (Mellors et al. 1996), and some have claimed evidence of higher viral load being achieved faster in subtype D viruses com-

pared with subtype A1 (Amornkul et al. 2013). However, from other cohort studies it appears that differences in viral load cannot explain differences in mortality risk between subtypes A1 and D (Baeten et al. 2007, McPhee et al. 2019), and that subtype contributes to virulence even after accounting for differences in viral load (Eller et al. 2015).

As well as viral load, co-receptor usage is also well known to be associated with virulence of HIV (see review by Schuitemaker et al. 2010), with the presence of X4 variants being associated with a rapid decline in CD4 and progression to AIDS (Koot et al. 1993). We found significant co-receptor usage difference between subtype D and A1, which had been previously reported in small samples (Huang et al. 2007, Kaleebu et al. 2007). In the 1986 genome sample, 66% (53/80) of subtype D envelopes had X4 tropic viruses compared to 0/ 24 subtype A1 viruses. By the time of the 'intermediate' 1998/9 sample (Harris et al. 2002) the frequency of X4 strains in subtype A remained 0 and that in subtype D had fallen to 33%, which coincided with the period of sample collection for the Kaleebu (2002) study. It is not unreasonable to suppose the difference in progression rate might have been up to twice as great if analysed at the time of the historical samples.

We also report a significant drop in subtype D, sustained over a long time period (1986-2016) over the whole of Uganda, in full length genomes. A change in relative proportion of the two subtypes during the 1990s was also suggested previously from sequence fragments of *gag* and gp41 coding regions (Conroy et al. 2010). Extending this to consider the prevalence of the X4 phenotype itself, as well as changes in subtype and subtype within the URF envelope, recombination has played a significant role too. We looked at 207 modern URF genomes containing either A1 or D *env* segments, and found they were significantly more likely to have subtype A1 in their envelope than subtype D, suggesting some degree of 'rescue' of the genome by the replacement of a less virulent V3 sequence.

In the 1990s, Uganda mounted a concerted national effort to encourage large scale behavioural changes. This came from the highest levels in government and was also implemented at the grass roots level (Green et al. 2006). Once the epidemic was no longer growing, HIV variants would have come under a selective pressure to delay the progression to AIDS, thereby increasing reproductive number (R) by expanding exposure window duration (Fraser et al. 2007). However, we propose that instead of modulating viral load, this came about by changes in frequency of the sequences most likely to encode the more virulent X4 phenotype. Differences in co-receptor usage is a very compelling potential explanation for changing subtype dynamics seen and differences in disease progression reported by various cohort subtypes in Uganda in the 20th century.

3.6 Supplementary Information

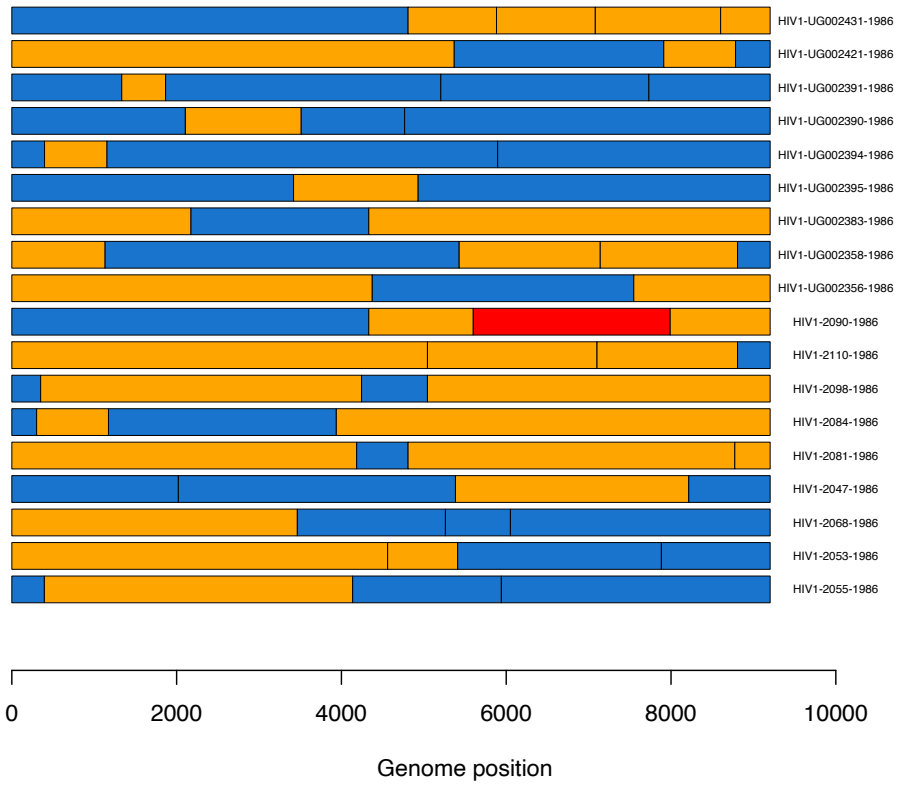

Figure 3.5: SCUEAL assignment of the 18 inter-subtype recombinants from historical samples. All had a unique recombination pattern. Subtype D fragments shown in orange, A1 in blue, subtype C in red

Name	SCUEAL subtype (edited)	Location	Percentage of reads mapped	Average Coverage	Number bp (min 10x each)	Genbank no
HIV1-2034-1986	A1	Mulago	66.69	1151.38	8699	OP039384
HIV1-2037-1986	A1	Mulago	0.54	29.38	8363	OP039386
HIV1-2075-1986	A1	Central	0.83	57.66	8016	OP039407
HIV1-2077-1986	A1		50.3	74.93	8474	OP039408
HIV1-UG002366-1986	A1	Kitovu	1.92	144.38	8106	OP039448
HIV1-UG002370-1986	A1	Central	1.6	47.64	8511	OP039451
HIV1-UG002376-1986	A1	Nsambya	9.96	76.11	8637	OP039454
HIV1-UG002378-1986	A1	Central	33.42	402.13	8650	OP039455
HIV1-UG002399-1986	A1	Antenatal	33.34	774.49	8722	OP039467
HIV1-UG002400-1986	A1		1.27	36.25	8575	OP039468
HIV1-UG002401-1986	A1		0.95	24.78	8178	OP039469
HIV1-UG002408-1986	A1	Central	1.26	45.81	8563	OP039473
HIV1-UG002409-1986	A1	Nsambya	11.43	73.71	8602	OP039474
HIV1-UG002415-1986	A1	Nsambya	32.35	98.53	8604	OP039477
HIV1-UG002419-1986	A1	Central	12.07	232.42	8683	OP039479
HIV1-UG002424-1986	A1	Central	7.45	304.95	8650	OP039481
HIV1-UG002429-1986	A1		51.67	681.4	8687	OP039485
HIV1-2090-1986	A1,C,D	Jinja	0.84	80.63	8461	OP039415
HIV1-2047-1986	A1,D	Mulago	2.81	206.84	8618	OP039396
HIV1-2053-1986	A1,D	Central	2.31	166.72	8511	OP039399
HIV1-2055-1986	A1,D	Central	3.95	221.15	8609	OP039401
HIV1-2068-1986	A1,D	Central	0.37	45.41	8306	OP039406
HIV1-2081-1986	A1,D		2.15	147.75	8596	OP039410
HIV1-2084-1986	A1,D		0.45	34.68	8113	OP039413
HIV1-2098-1986	A1,D	Rubaga	26.25	654.37	8640	OP039417
HIV1-2110-1986	A1,D	Kitovu	4.04	181.23	8618	OP039425
HIV1-UG002356-1986	A1,D	Kitovu	47.05	1769.08	8601	OP039439
HIV1-UG002358-1986	A1,D	Kitovu	0.95	114.09	8627	OP039440
HIV1-UG002383-1986	A1,D	Central	20.04	133.64	8568	OP039457
HIV1-UG002390-1986	A1,D	Antenatal	18.53	204.36	8675	OP039461
HIV1-UG002391-1986	A1,D	Central	19.22	27.32	8187	OP039462
HIV1-UG002394-1986	A1,D	Nsambya	12.1	408.18	8612	OP039463
HIV1-UG002395-1986	A1,D		58.45	247.09	8694	OP039464
HIV1-UG002421-1986	A1,D	Mulago	3.5	93.86	8511	OP039480
HIV1-UG002431-1986	A1,D		1.54	78.34	8343	OP039487
HIV1-UG002385-1986	C	Central	27.71	324.88	8623	OP039458
HIV1-2029-1986	D	Nsambya	9.06	90.32	8558	OP039379
HIV1-2030-1986	D	Rubaga	22.13	153.74	8525	OP039380
HIV1-2031-1986	D	Rubaga	14.73	1593.94	8553	OP039381
HIV1-2032-1986	D	Rubaga	6.89	374.21	8588	OP039382
HIV1-2033-1986	D	Nsambya	57	483.34	8594	OP039383
HIV1-2036-1986	D	Mulago	1.67	65.38	8600	OP039385
HIV1-2038-1986	D	Mulago	1.13	62.45	8587	OP039387
HIV1-2039-1986	D	Rubaga	25.89	385.59	8671	OP039388
HIV1-2040-1986	D	Rubaga	12.06	143.21	8602	OP039389
HIV1-2041-1986	D	Rubaga	77	2197.35	8674	OP039390
HIV1-2042-1986	D	Rubaga	7.73	448.64	8646	OP039391

HIV1-2043-1986	D	Rubaga	88.08	7564.68	8632	OP039392
HIV1-2044-1986	D	Rubaga	10.45	453.74	8618	OP039393
HIV1-2045-1986	D	Mulago	81.83	1779.92	8609	OP039394
HIV1-2046-1986	D	Mulago	1.47	138.57	8443	OP039395
HIV1-2048-1986	D	Central	1.24	76.36	8500	OP039397
HIV1-2052-1986	D	Central	0.66	40.18	8228	OP039398
HIV1-2054-1986	D	Central	9.37	240.65	8599	OP039400
HIV1-2056-1986	D	Central	3.25	184.07	8617	OP039402
HIV1-2057-1986	D	Central	1.03	142.19	8609	OP039403
HIV1-2058-1986	D	Central	75.39	706.32	8649	OP039404
HIV1-2064-1986	D	Central	30.54	2702.45	8685	OP039405
HIV1-2079-1986	D		7.11	90.58	8621	OP039409
HIV1-2082-1986	D		5.79	113.24	8607	OP039411
HIV1-2083-1986	D		22.64	503.26	8600	OP039412
HIV1-2085-1986	D		1.63	157.43	8626	OP039414
HIV1-2097-1986	D	Rubaga	4.57	38.37	8444	OP039416
HIV1-2100-1986	D	Kitovu	8.4	478.59	8619	OP039418
HIV1-2102-1986	D	Kitovu	37.21	764.13	8635	OP039419
HIV1-2103-1986	D	Kitovu	0.89	92.33	8605	OP039420
HIV1-2106-1986	D	Kitovu	3.64	349.84	8635	OP039421
HIV1-2107-1986	D	Kitovu	5.39	32.35	8558	OP039422
HIV1-2108-1986	D	Kitovu	39.81	154.51	8618	OP039423
HIV1-2109-1986	D	Kitovu	1.31	48.41	8611	OP039424
HIV1-2111-1986	D	Kitovu	2.57	151.97	8644	OP039426
HIV1-2112-1986	D	Kitovu	12.72	255.92	8639	OP039427
HIV1-2113-1986	D	Kitovu	2.01	19.09	8054	OP039428
HIV1-2114-1986	D	Kitovu	15.08	227.09	8619	OP039429
HIV1-2115-1986	D	Kitovu	13.4	316.02	8611	OP039430
HIV1-2116-1986	D	Kitovu	0.55	45.05	8593	OP039431
HIV1-UG002345-1986	D	Kitovu	15.37	181.14	8612	OP039432
HIV1-UG002347-1986	D	Kitovu	1.12	103.82	8627	OP039433
HIV1-UG002348-1986	D	Kitovu	7.01	133.46	8637	OP039434
HIV1-UG002349-1986	D	Kitovu	10.54	419.08	8609	OP039435
HIV1-UG002351-1986	D	Kitovu	9.44	143.51	8611	OP039436
HIV1-UG002352-1986	D	Kitovu	5.47	59.16	8604	OP039437
HIV1-UG002354-1986	D	Kitovu	6.16	117.61	8591	OP039438
HIV1-UG002359-1986	D	Kitovu	1.82	80.91	8608	OP039441
HIV1-UG002360-1986	D	Kitovu	20.64	1076.86	8615	OP039442
HIV1-UG002361-1986	D	Kitovu	2.57	179.58	8604	OP039443
HIV1-UG002362-1986	D	Kitovu	4.31	188.1	8601	OP039444
HIV1-UG002363-1986	D	Kitovu	1.13	65.3	8169	OP039445
HIV1-UG002364-1986	D	Kitovu	2.63	170.26	8588	OP039446
HIV1-UG002365-1986	D	Kitovu	12.26	1309.63	8612	OP039447
HIV1-UG002368-1986	D	Lacor	12.6	134.59	8567	OP039449
HIV1-UG002369-1986	D	Lacor	28.12	995.75	8644	OP039450
HIV1-UG002371-1986	D	Lacor	2.5	140.29	8637	OP039452
HIV1-UG002372-1986	D	Lacor	17.6	427.61	8604	OP039453
HIV1-UG002382-1986	D	Central	2.87	115.3	8594	OP039456

HIV1-UG002387-1986	D	Kampala	26.53	65.55	8599	OP039459
HIV1-UG002389-1986	D	Central	9.06	63.09	8629	OP039460
HIV1-UG002396-1986	D	Kampala	0.76	79.79	8628	OP039465
HIV1-UG002397-1986	D	Rubaga	0.43	31.66	8502	OP039466
HIV1-UG002405-1986	D	Central	9.72	516.1	8652	OP039470
HIV1-UG002406-1986	D		62.61	851.56	8609	OP039471
HIV1-UG002407-1986	D		3.72	58.04	8603	OP039472
HIV1-UG002410-1986	D	Central	16.64	570.73	8644	OP039475
HIV1-UG002414-1986	D	Nsambya	3.7	119.62	8583	OP039476
HIV1-UG002417-1986	D	Central	1.14	123.8	8635	OP039478
HIV1-UG002426-1986	D		23.07	1335.02	8651	OP039482
HIV1-UG002427-1986	D		3.48	213.81	8581	OP039483
HIV1-UG002428-1986	D		80.79	1096.89	8637	OP039484
HIV1-UG002430-1986	D		16.39	764.2	8630	OP039486

Table 3.4: Information about each 109 genome sequence including SCUEAL subtype, read depth, location, and Genbank numbers

Name	SCUEAL subtype (edited)	Location	Percentage of reads mapped	Average Coverage	Number bp (min 10x each)	Genbank	Notes about submission
HIV1-2066-1986	A1	Central	0.21	31.05	6832	OP039490	one sequence, large gaps
HIV1-2067-1986	A1	Central	0.28	31.91	7067	OP039491	one sequence, large gaps
HIV1-2071-1986	A1	Central	15.92	12.44	4285	OP039492	one sequence, large gaps
HIV1-2099-1986	A1	Rubaga	0.3	16.32	6161	OP039504	one sequence, large gaps
HIV1-UG002346-1986	A1	Kitovu	0.15	19.59	6360	OP039508	one sequence, large gaps
HIV1-UG002386-1986	A1	Antenatal	0.17	13.09	5192	OP039515	one sequence, large gaps
HIV1-UG002420-1986	A1	Mulago	2.78	18.07	7277	OP039524	one sequence, large gaps
HIV1-2049-1986	A1,D	Central	0.25	14.95	6396	OP039488	one sequence, large gaps
HIV1-2078-1986	A1,D		0.06	12.74	3677	OP039494	one sequence, large gaps
HIV1-2086-1986	A1,D	Jinja	0.07	8.75	2033	OP039497 and OP039498	400 gag + 3000 env-nef
HIV1-UG002379-1986	A1,D	Kampala	33.24	19.34	6701	OP039514	one sequence, large gaps
HIV1-UG002393-1986	A1,D	Antenatal	0.88	16.35	7400	OP039518	one sequence, large gaps
HIV1-UG002404-1986	A1,D	Central	0.31	31.71	6438	OP039520	one sequence, large gaps
HIV1-2072-1986	A1,D,J	Central	0.15	13.73	4981	OP039493	one sequence, large gaps
HIV1-UG002357-1986	A1,K	Kitovu	0.09	12.41	4463	OP039510	one sequence, large gaps
HIV1-UG002412-1986	complex,A1		0.66	7.67	1456	OP039521 and OP039522	1500 gagpol +300 env
HIV1-2093-1986	complex,D	Jinja	0.07	12.2	3506	OP039501	one sequence, large gaps
HIV1-UG002375-1986	complex,D	Kampala	8.44	9.35	2102	OP039512 and OP039513	partial gag 1000 + partial pol 800
HIV1-2065-1986	D	Central	0.12	16.51	4924	OP039489	one sequence, large gaps
HIV1-2080-1986	D		0.22	9.13	3043	OP039495 and OP039496	gagpol 1500 + vif 500
HIV1-2087-1986	D	Jinja	0.11	14.53	6189	OP039499	one sequence, large gaps
HIV1-2092-1986	D	Jinja	0.16	18.36	7522	OP039500	one sequence, large gaps
HIV1-2095-1986	D	Jinja	0.2	16.68	4980	OP039502	one sequence, large gaps

HIV1-2096-1986	D	Jinja	0.02	7.84	1000	OP039503	env-nef 1000
HIV1-2101-1986	D	Kitovu	0.48	18.62	5860	OP039505	one sequence, large gaps
HIV1-2104-1986	D	Kitovu	0.05	7.44	1023	OP039506 and OP039507	gagpol 500 + env 500
HIV1-UG002355-1986	D	Kitovu	0.22	13.16	5063	OP039509	one sequence, large gaps
HIV1-UG002373-1986	D	Lacor	0.17	15.99	5882	OP039511	one sequence, large gaps
HIV1-UG002398-1986	D	Central	0.26	9.08	2640	OP039519	one sequence, large gaps
HIV1-UG002416-1986	D	Central	3.15	12.81	4273	OP039523	one sequence, large gaps
HIV1-UG002422-1986	D	Nsambya	1.28	11.94	5118	OP039525	one sequence, large gaps
HIV1-UG002432-1986	D		0.25	19.18	7499	OP039526	one sequence, large gaps
HIV1-UG002392-1986	D,H	Central	0.59	7.84	1196	OP039516 and OP039517	900bp gagpol + 300 env

Table 3.5: Information about partial genome sequences including SCUEAL subtype, read depth, location, and Genbank numbers

Chapter 4

A tale of two subtypes - Three decades of competitive HIV dynamics in Uganda revealed by full-genome viral sequences

Heather E. Grant[1], Samantha Lycett [2], Deogratius Ssemwanga [3], Judith Breuer [4], Pontiano Kaleebu [3], Andrew J. Leigh Brown [1]

1) Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
2) The Roslin Institute, University of Edinburgh, Edinburgh, UK
3) Medical Research Council (MRC)/Uganda Virus Research Institute (UVRI) and London School of Hygiene and Tropical Medicine (LSHTM) Uganda Research Unit, Entebbe, Uganda
4) Division of Infection and Immunity, University College London, London, UK

4.1 Abstract

While some major global epidemics of HIV-1 such as in the global North or southern Africa are primarily composed of single subtypes (B and C respectively), Uganda has experienced an epidemic composed of two subtypes (A1 and D), co-circulating for many decades. To explore these dynamics, we apply BEAST phylodynamic models to full-length HIV genome data spanning three decades (1986-2016) of the Ugandan HIV epidemic. During the 1960s and 1970s, subtype D expanded extremely rapidly into Uganda, surpassing the pre-existing subtype A1 epidemic to become the dominant subtype in 1986. After the late 1980s and the start of the HIV/AIDS program, subtype D declined much more rapidly than subtype A1, and we see signatures of extinct subtype D lineages. This may be explained by a higher propensity for CXCR4 co-receptor tropism in subtype D, which is known to increase mortality risk and decrease time to AIDS.

4.2 Introduction

HIV-1 group M entered the human population in the early part of the 20th century (Korber et al. 2000), possibly by zoonotic transfer during the butchering of chimpanzee meat (Hahn et al. 2000). After establishing itself in Kinshasa (Democratic Republic of Congo, DRC), its high mutation rate in combination with exponential growth in the local population meant viral diversity rapidly expanded (Faria et al. 2014). Despite appreciable recombination between viruses in the early epidemic (Kalish et al. 2004), by the 1960s, it appears HIV-1 had formed broad-scale lineages (Worobey et al. 2008) which have been categorised into subtypes (Robertson et al. 2000). Furthermore, there is evidence from phylogeographic analysis of *env* sequences that these subtypes were separated geographically within the DRC at an early stage (Faria et al. 2019).

Researchers in Uganda were among the first to document AIDS in Africa, which was typified by cases of aggressive Kaposi's sarcoma and extreme wasting known then as "slim disease" (Serwadda et al. 1985, 1986). Even before it was shown that AIDS was associated with a retrovirus, HIV was already widespread in Ugandan communities. One serological survey in Kampala in February 1987 pointed to 1 in 4 pregnant women being HIV positive (Carswell 1987). In the decade that followed, Uganda mounted a multipronged campaign to tackle HIV/AIDS (Slutkin et al. 2006), which was heralded one of the best responses in Africa (UNAIDS 1998). After peaking in around 1992 (Kirby 2008), prevalence and incidence started to fall steadily (Baryarama et al. 2004) aided further by the start of the antiretroviral therapy roll out in 2004 (Grabowski et al. 2014). Between 2010 and 2018 incidence fell from 3.21 to 1.4 per 1000 population per year (UNAIDS, 2019).

Uniquely, Uganda has had two HIV-1 subtypes (A1 and D) in circulation at similar frequency in the same populations since the beginning of the recorded epidemic (Yirrell et al. 1997, Ssemwanga et al. 2020). Subtype A is widespread globally in different risk groups with a very large intra-subtype diversity (Tongo et al. 2018), while subtype D is mostly confined to heterosexual populations in East Africa. There are however, some key differences between the subtypes; subtype D is less diverse than A1, is confined to heterosexual populations in East Africa, and is thought to have spread into Uganda a decade later (Yebra et al. 2015). Subtype D is also known to use the CXCR4 co-receptor more frequently (Huang et al. 2007, Kaleebu et al. 2007), and cohorts from Uganda, Tanzania, and even UK, have found subtype D to progress faster to AIDS (Kaleebu et al. 2002, Vasan et al. 2006, Easterbrook et al. 2010, Kiwanuka et al. 2008).

In our most recent work, we sequenced 109 full length genomes from samples taken in 1986 from various hospitals across Uganda (chapter 3). This is one of the earliest population samples, both from East Africa and from the early global epidemic. To evaluate longitudinal changes in subtype diversity in Uganda we combined this historical sample to an intermediate (1998-9) sample and a modern (2007-2016) sample. Bioinformatic predictions confirmed a

higher propensity for subtype D to be X4 tropic at each time period, but particularly in the 1986 dataset where 0/24 subtype A1 were X4 tropic compared with 53/80 in subtype D. We also showed the prevalence of subtype D fell continuously from 67% to 57% to 17% over these periods. We proposed therefore, that subtype D had a shorter infectious window and a disadvantage over time due to a higher X4 propensity (since X4 co-receptor tropism is likely to decrease time to AIDS; Daar et al. 2007).

In this study, we combine 620 Ugandan genomes spanning three decades with additional sequences from the Los Alamos database to further analyse and explore the changing HIV dynamics of two subtypes. Using BEAST (Suchard et al. 2018) we examine the subtypes in their wider East African context, the times of most recent common ancestor (tMRCA) of Ugandan subtype diversity, and show subtle differences in the histories of the three major HIV genes. Using the skygrid model we estimate the change in effective population number (N_e) of each subtype independently. Finally, we have incorporated into our phylogenetic analyses gene sequences from unique recombinant forms (URFs) to demonstrate historic and ongoing recombination between subtypes.

4.3 Methods

All plots and graphs were made in R with application of the packages ‘ape’ (Paradis & Schliep 2019), ‘phytools’ (Revell 2012), ‘caper’ (Orme et al. 2018), ‘geiger’ (Pennell et al. 2014), ‘ggtree’ (Yu et al. 2018), and ‘ggplot’ (Wickham 2016). Our dataset was previously subtyped with SCUEAL (Kosakovsky Pond et al. 2009, Grant et al. 2020), including 465 MRC/UVRI PANGEA sequences sampled between 2007 and 2016, and a collection of 46 genomes by (Harris et al. 2002) and 109 HIV genome sequences from 1986 (chapter 3). In order to place these Ugandan genomes in a wider context, full-length genomes were obtained from the Los Alamos National Laboratory database (lanl.hiv.gov) to supplement the dataset. See results and Supplementary Table 4.1 for details of the additional included genome sequences.

4.3.1 Phylogenetic reconstructions of subtype A1 and D

Sequences were separated by subtypes and into three major gene regions ("gene partitions"): *gag* (493 codons), *pol* (939 codons), and *env* (744 codons), and edited by eye to ensure sequences were in a single open reading frame throughout. An initial maximum likelihood tree produced with IQtree (Nguyen et al. 2015) was used as input for Tempest (Rambaut et al. 2016) which identifies outliers in the molecular clock. Alignments were screened with RDP4 (Martin et al. 2015) for individual sequences with evidence of recombination and GARD (Kosakovsky Pond et al. 2006) for evidence of incongruent history in different sections of

each alignment. Where a significant breakpoint was found the majority part of the alignment was kept and the minority discarded. After removing molecular clock outliers, recombinant sequences and recombinant gene regions, genomes which had a representative in every gene partition were retained for analysis.

Bayesian phylogenetic reconstruction was carried out in BEAST 1.10.4 (Suchard et al. 2018) to reconstruct the histories and topology of Ugandan subtypes A1 and D, find their time to most recent common ancestor (tMRCA), and place them in their wider contexts. We chose a relaxed clock model with rates sampled from a lognormal distribution, commonly used for HIV data e.g. (Drummond et al. 2006, Alizon & Fraser 2013). The nucleotide substitution model used was SRD06, in which the HKY85 model is applied to two partitions: codon positions 1+2 and codon position 3, separately (Shapiro et al. 2006). In this model site heterogeneity was modelled using a gamma distribution with 10 discrete categories. This nucleotide model was selected after comparing marginal likelihood estimates of GTR+4, TN93+4, SRD06+4, and SRD06+10 for subtype A1 *pol* (data not shown) with generalised stepping stone path sampling (GSS; (Baele et al. 2016). The Skyride (Minin et al. 2008) was used as the demographic model, which allows the effective population size of the virus to change freely between time points, (irregular intervals determined by coalescent times) with GMRF smoothing between them. Alizon & Fraser (2013) suggest between host evolutionary rates (measured in substitutions per site per year) of ($1.7x10^{-3}$ and $3.1x10^{-3}$) for non-env and *env* regions respectively, and the following evolutionary rate priors were applied: subtype A1 *gag* $1.8x10^{-3}$ with standard deviation (SD) $1.8x10^{-4}$, subtype A1 *pol* $1.5x10^{-3}$ with SD $1.5x10^{-4}$, subtype A1 *env* $2.7x10^{-3}$ with SD $5x10^{-4}$, subtype D *gag* $2x10^{-3}$ with SD $5x10^{-4}$, subtype D *pol* $1.7x\ 10^{-3}$ with SD $5x10^{-4}$, subtype D *env* $2.7x10^{-3}$ with SD $5x10^{-4}$.

MCMC chains were run for a minimum of 300 million generations sampled every 100,000 generations and then extended until deemed converged. This was judged by high (>300) effective sample sizes (ESS) and good mixing of parameter estimates with particular attention paid to a well-mixed evolutionary rate, (a good acceptance rate of MCMC proposals and lack of autocorrelated samples) as seen with Tracer v.1.7 (Rambaut et al. 2018). A Maximum Credibility Clade tree (MCC) was created for each gene partition analysis separately. Finally, the posterior tree states outputs of all gene partitions of each subtype were combined (possible because all runs had the same tip-labels) with equal numbers of tree states (equal partition weight), so that a MCC tree of all three partitions could be constructed to show topology agreement at the genome level (using treeannotator within the BEAST package).

4.3.2 Additional analyses with recombinant sequences

We also examined the dynamics of unique circulating recombinant forms (URFs) in the dataset. From the Ugandan URFs previously described (chapter 2) A1 or D fragments which covered at least 70% of the length of the gene, based on the SCUEAL breakpoints, were extracted. These recombinant fragments were added to gene alignments of Ugandan 'pure' sequences, again checked with RDP4, GARD, and Tempest, and analysed using BEAST in the same way as above. We examined the phylogenetic spread of URF sequences throughout individual gene phylogenies, and counted the number of lineages containing only URF genomes (which represent at least one recombination event with another subtype).

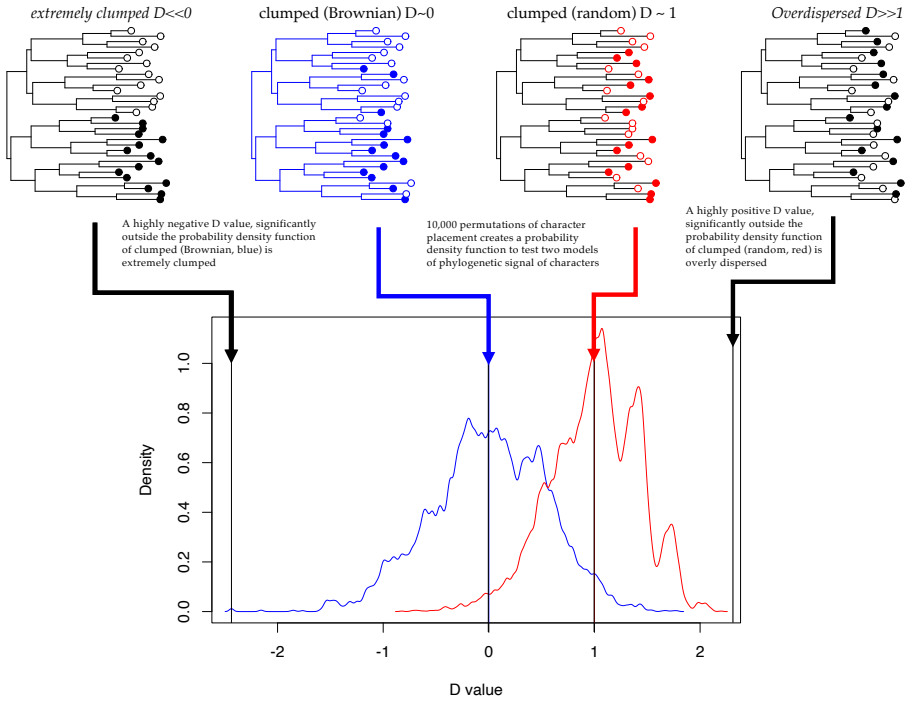

Figure 4.1: The D statistic - adapted from Fritz and Purvis 2010 and the phylo.d() function in caper showing the D-statistic rationale.

Visual inspection of some topologies suggested a level of non-randomness or 'clumping' of sequences from before the year 2000. To formally test this observation we applied the D-statistic (Fritz & Purvis 2010) of phylogenetic signal (using the binary trait of "pre- or post-2000"). The 'D measure' is calculated with weighted sum of sister clade differences. For a given phylogenetic tree and a given number of tips with a binary trait, trait evolution can be simulated under a Brownian random walk (Felsenstein 1985) (expected D=0) or a random distribution (expected D=1). When repeated 1000 times this creates a probability density function for interpreting the significance of the value of D (Figure 4.1). The estimated value of

D is compared against the two null distributions to obtain confidence in phylogenetic signal or clumped-ness of the binary traits. A very high value of D suggests over-dispersal of the trait on the tree, while a very negative value of D suggests extreme clumping or high phylogenetic signal. This was also applied to co-receptor (X4 or R5) tropism in the *env* topologies.

4.3.3 Effective population size estimation

In order to estimate the change in effective population size of the two subtypes over the same time scale, additional BEAST runs were set with the skygrid model (Gill et al. 2012). This model differs from the skyride model in that the time points (where effective population number are estimated) are regular, and therefore a change can be interpreted more easily. With 20 parameters and a cut off of 80 years, each 'grid square' represented a regular 4-year interval. Point estimates were taken for each gene partition separately, and then the log files were combined together (weighted equally) for the overall 95% Bayesian credible interval (BCI) of effective population size using all partitions at the genome level. The same substitution model, clock model, and rate priors as above were used, but only genomes from Uganda were included.

4.4 Results

We estimated the evolutionary history of HIV subtypes A1 and D in Uganda with the skyride model including closely related subtype outgroup genomes (from East Africa, the DRC, or the early subtype B Western epidemic). We combined tree state outputs from separate BEAST analyses for each of the three gene partitions: *gag*, *pol*, and *env*, to make a combined Maximum Clade Credibility tree at the genome level. Thus we were able to examine the gene partitions separately and together, while allowing for differences in evolutionary rates and the effects of inter-gene recombination. At the same time we were able to evaluate the topology and tMRCA over the whole HIV genome. We present the phylogenetic results obtained for each subtype separately, and show the overall dynamics for the two together.

4.4.1 Subtype A1

A total of 167 subtype A1 genomes from Uganda (1986-2016) were complemented with additional full-length genomes from the LANL database (hiv.lanl.gov). Subtype A1 is widely prevalent across East Africa, and therefore a number of genomes were available from surrounding countries. After removing duplicate genomes or those with missing dates, there were 68 from Kenya, 10 from Rwanda, 22 from Tanzania, a single DRC genome, and 11 additional from Uganda. Kenyan genomes were subsampled to 20, a similar number as available from Tanzania, by iteratively removing one from a pair of the most genetically similar genomes from

the same year. The oldest available genomes from the DRC in outgroup subtype G (n=4), A4 (n=4), A5 (n=4) and A2 (n=2 and n=2 from elsewhere) were selected for inclusion. After screening for recombinant sequences with RDP4, 3 sequences from the *gag* alignment, 4 from *pol*, and 10 sequences from the *env* alignment were removed as probable recombinants. There was no evidence according to GARD of recombination in *gag*, but strong evidence for a breakpoint at position 2190 in *pol* (corresponding to mid-integrase) and another breakpoint in *env* at position 1700 (corresponding to the transmembrane region of gp41). The final dataset included 218 genomes each with three gene sequences.

Figure 4.2 shows the genome MCC topology of subtype A1 and its outgroups. All subtypes and sub-subtypes monophyletic and have the overall structure G,(A5,(A2,(A4,A1)))). There is however, low branch support for the placement of A2 and A5. In individal MCC gene trees for *gag* and *pol* subtypes A2 and A5 are sisters, while in *env* A5 is an outgroup to (A2, (A4, A1) (Supplementary Figure 4.6). This gene level incongruence is most likely explained by recombination between A5 and A2 lineages in the DRC in the 1950s and 1960s. The single subtype A1 genome from the DRC falls just outside of the East African diversity, and the tMRCA for the East African/ Ugandan subtype A1 is 1949.4 [1943-1955]. There is some structure within subtype A1 by East African nation. For example, “clade T” has predominantely Tanzanian sequences or “clade K” has the most Kenyan sequences (and also Rwandan, Tanzanian, and Ugandan). These clades have low posterior node support at the genome level, but are more clearly defined in the individual gene trees with better node support (Supplementary Figure 4.6), again reflecting intra-subtype recombination between the gene partitions.

We further explored the evolution of subtype A1 by including A1 gene sequences from URF genomes broken down into ‘pure’ gene sections containing of 70% A1 so that an additional 101 *gag*, 59 *pol*, 118 *env* (total of 278) were included. By including ‘pure’ sections of URFs in the trees, we see the placement of URF fragments is also very well mixed throughout, showing continual generation of unique recombinant forms in all genes. We have picked out some recombinant lineages of different ages, for example, ‘clade u1’ with 3 gag URF sequences with a tMRCA of 1996 or clade u3 with two pol sequences coalescing in 1999, or clade u6 with 5 env A1 URFs with tMRCA 1971 (Supplementary Figure 4.7).

In these analyses, we also see historical subtype A1 sequences spread well throughout the tree, particularly in *pol*. To give a formal measurement of how clumped historical sequences we applied the D-statistic for phylogenetic signal on the extended gene trees of Ugandan A1. There was some clustering of pre-2000 sequences as might be expected under the Brownian random walk model in *gag* and *env* (D= -1.1, D=-0.9) while in *pol* pre-2000 sequences are

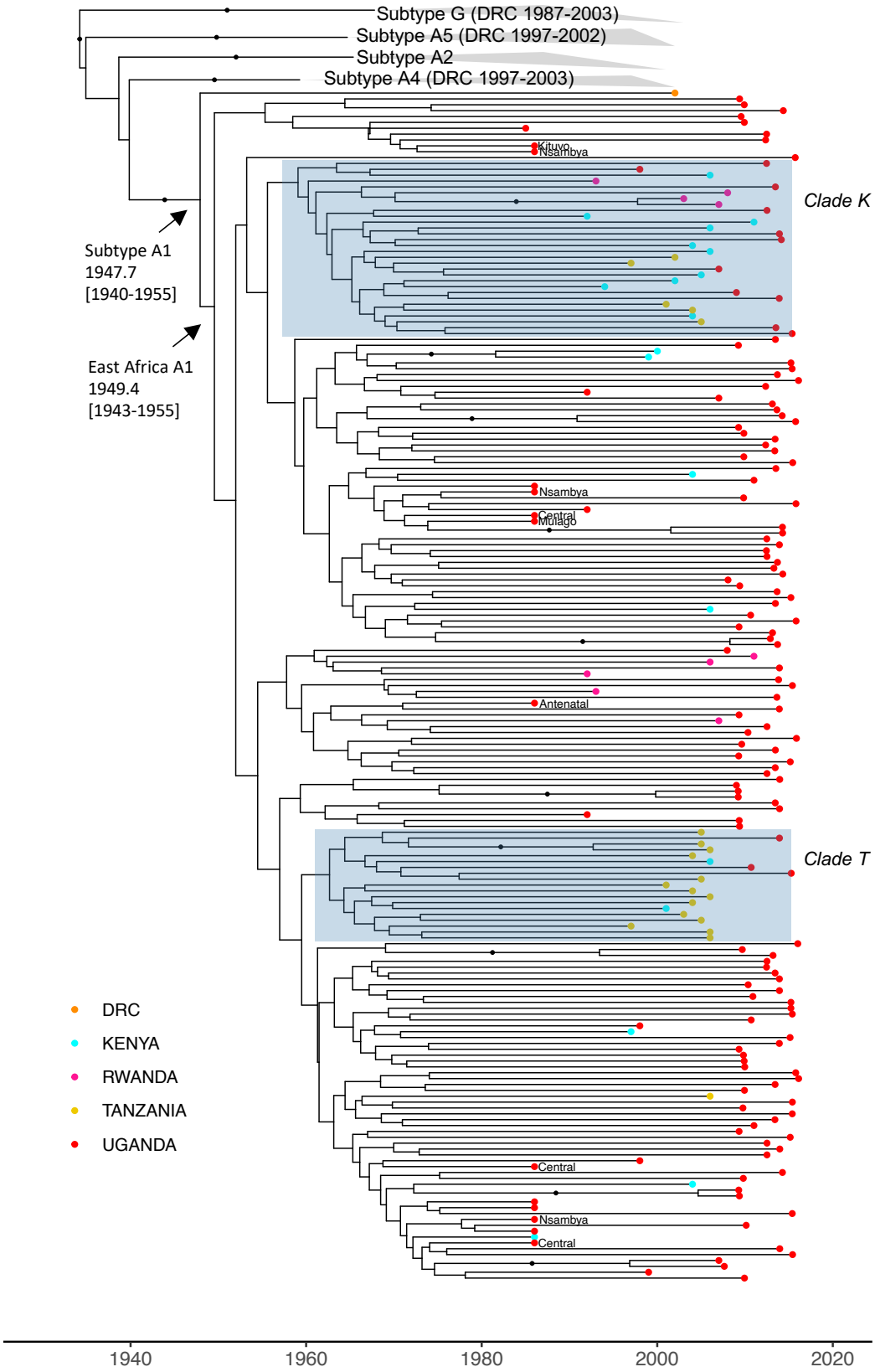

Figure 4.2: Subtype A1 BEAST skyride with outgroups. Maximum Clade Credibility tree from BEAST with median node height for Subtype A1, created with tree states from three independent BEAST runs for *gag*, *pol*, *env* partitions. Outgroups are collapsed for clarity, small black dots at nodes indicate posterior probability of > 0.9. Individual gene MCC trees are shown in Figure 4.6

more dispersed and randomly placed throughout the phylogeny (D=0.9). Lastly, co-receptor tropism associated with faster CD4 decline (X4) is plotted onto the A1 *env* tree, and showed no phylogenetic signal; X4 sequence tips are found under a random expectation (D=1.1) in the phylogeny (see supplementary information for D-statistic details).

4.4.2 Subtype D

Subtype D is much less prevalent outside Uganda, and only 18 additional genomes from East Africa were found (6 from Uganda, 7 from the DRC, 2 from Tanzania, 3 from Kenya) to complement the 181 we have previously described (1986, 1998/9 and 2007-2016). Ugandan sequences therefore make up 187/200 or 94% of East African subtype D genomes available to us. Additionally, we included 16 subtype D genomes from other countries including Cameroon (n=4), Chad (n=4), Senegal (n=1), South Africa (n=2), Yemen (n=2), and Brazil (n=3). Because of the close relationship between subtypes B and D, seven of the oldest (1983-4) subtype B genomes from the early north America epidemic were included as an outgroup (see Supplementary Table 4.1). Subtype D alignments were also assessed for signatures of recombination: RDP4 suggested removal of 1 sequence from *gag*, none from *pol*, and in the *env* alignment 12 sequences were either removed or trimmed. GARD found very similar recombination positions in subtype D as subtype A1 (none in *gag*, position 2300 in *pol* (mid-integrase) and position 1700 in *env* (the transmembrane region of gp41). The final subtype D dataset included 208 genomes with sequence available in all three gene partitions, all screened for recombination.

The genome level MCC tree of subtype D is shown in Figure 4.3. Subtype B forms a clear outgroup to subtype D as expected with tMRCA of 1970.3 [1965-1974]. Subtype D can be broadly separated into three groups each with > 0.9 genome level MCC posterior support. The top clade labelled “D ancestral north” contains genomes from the DRC, from 2 other countries further north (Cameroon, and Chad) and from immigrants in two other countries (Cyprus, Yemen). The second clade, labelled “D ancestral south”, contains genomes from the DRC including ‘ELI’ and ‘NDK’ and from countries from southern Africa (South Africa, Tanzania; also Brazil). The third clade is overwhelmingly Ugandan (97%) with only 6 non-Ugandan genomes (Kenya, Tanzania, and Senegal). Only one East African genome falls outside of this clade (KX907406 from Mbeya, Tanzania, 2004 in ‘ancestral D south’), highlighting a very strong geographical structure. These three clades (ancestral D north, ancestral D south, and D Uganda) can also be found in the individual gene partitions of each of *gag*, *pol*, and *env* (Supplementary Figure 4.8) but with some topological incongruences between them, which again suggests intra-subtype recombination within the DRC of early subtype D lineages.

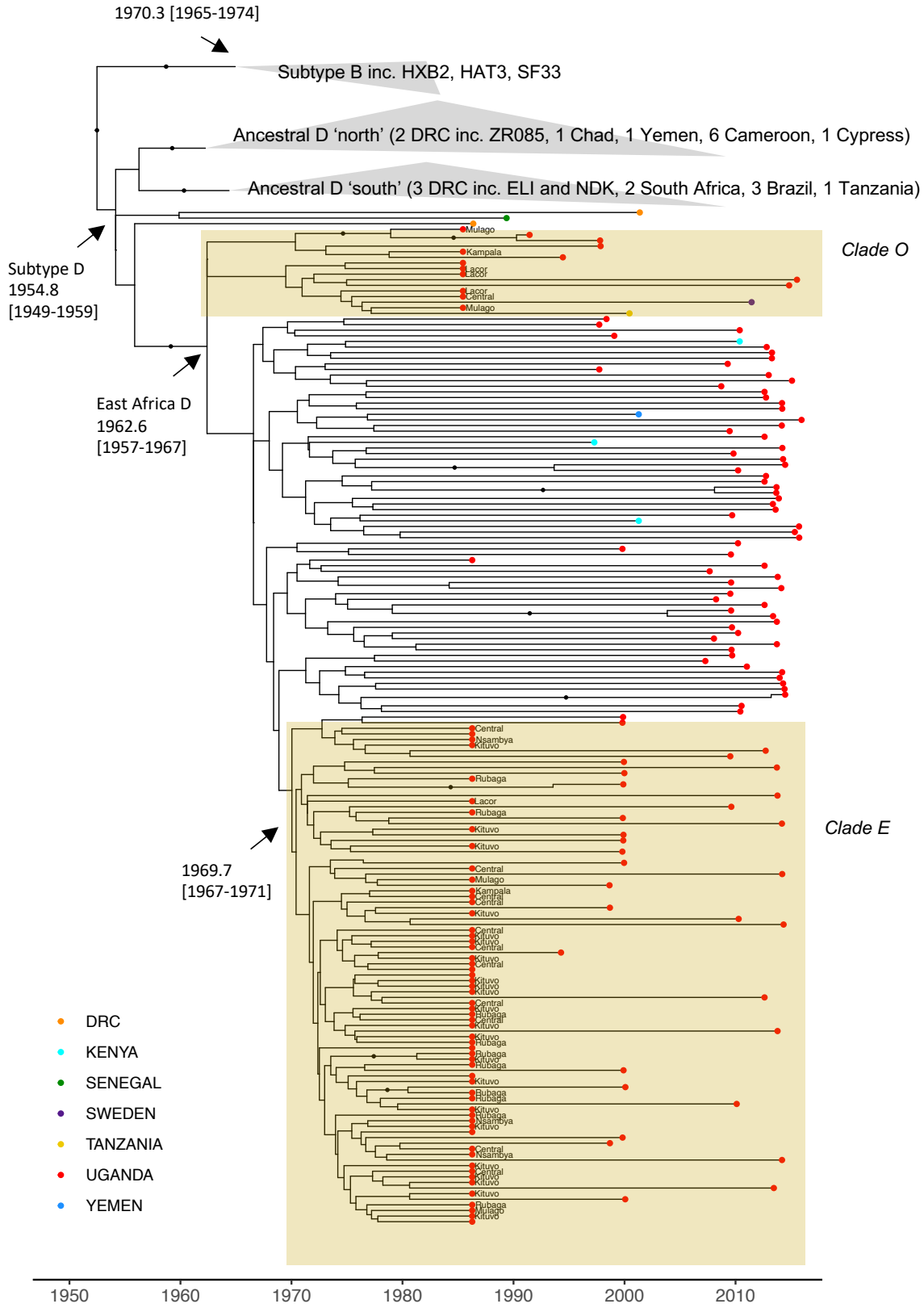

Figure 4.3: Subtype D BEAST skyride with outgroups. MCC tree created with tree states from three independent BEAST runs for *gag*, *pol*, *env* partitions. Outgroups are collapsed for clarity, small black dots at nodes indicate posterior probability of > 0.9. Individual gene MCC trees are shown in Figure 4.6

The historical 1986 subtype D sequences are distributed very unevenly throughout the tree (Figure 4.3), and the majority of modern Ugandan D sequences are not directly related to the genomes sampled in 1986 (sampling location is shown where known). The first bifurcation of Ugandan subtype D appears in 1962.8 (95% BCI: 1958-1967) and forms an older group highlighted as “Clade O”. Clade O contains 12 pre-2000 sequences, but only 3 post-2000 sequences suggesting this clade has severely reduced in size. The clade highlighted “Clade E” with tMRCA 1969.7 [1967-1971] contains the majority of 1986 sequences including all the known Kitovu hospital genomes (from the Masaka district). Although Clade E contains 77 pre-2000 genomes, it does not give rise to many modern descendants, and only includes 14 post-2000 genomes. Therefore, it appears many of these lineages are no longer in circulation.

According to the PhyloD statistic the historical tips in these trees are all extremely clumped (*gag* = -1.5, *pol* = -1.6), and particularly so in *env* (-2.7) (supplementary information). In extended analyses we included an additional 106 *gag*, 93 *pol*, and 65 *env* subtype D gene sequences (264 in total) from URFs. Many recombinant lineages can be seen in Supplementary Figure 4.9. We highlight 3 *gag* sequences that coalesce in 2007 (clade u7), a pair of *pol* sequences in 1992 (clade u8), or 4 *env* sequence in 1994 (clade u9). In previous work we show subtype D has a high propensity to carry X4 co-receptor tropism (66% of 1986 sequences; chapter 3). Here, X4 co-receptor tropism shows some phylogenetic signal in subtype D as its distribution is clumped across the *env* phylogeny (D=0), as might be expected by a Brownian walk (p=0.5), beyond the random null distribution (p<0.001).

4.4.3 Change in effective population size over time

An estimation of N_e over time for the two HIV subtypes was carried out using the skygrid demographic model where the effective population size was estimated at 4-year intervals. Figure 4.4 shows the skygrid effective population size point estimates for each partition for subtype A1 and subtype D with the associated credible intervals. The three genes provide internally consistent estimates for both subtypes. As already established, subtype A1 is present earlier in Uganda (1957 in this analysis), growing exponentially until a peak in the late 1980s and early 1990s, consistent with some estimates (Kirby 2008) of the peak in HIV incidence and prevalence for Uganda. Subtype D is introduced into Uganda about a decade later, but its epidemic grows even more quickly, so that by the early 1990s where it also reaches its peak, the two subtypes have a similar effective population size. Subsequently, N_e for subtype A1 remains fairly constant, with perhaps a slight upward trend toward the present day. Subtype D however, appears less stable, dropping more sharply in the early 1990s and again after 2010, although the credible intervals largely overlap.

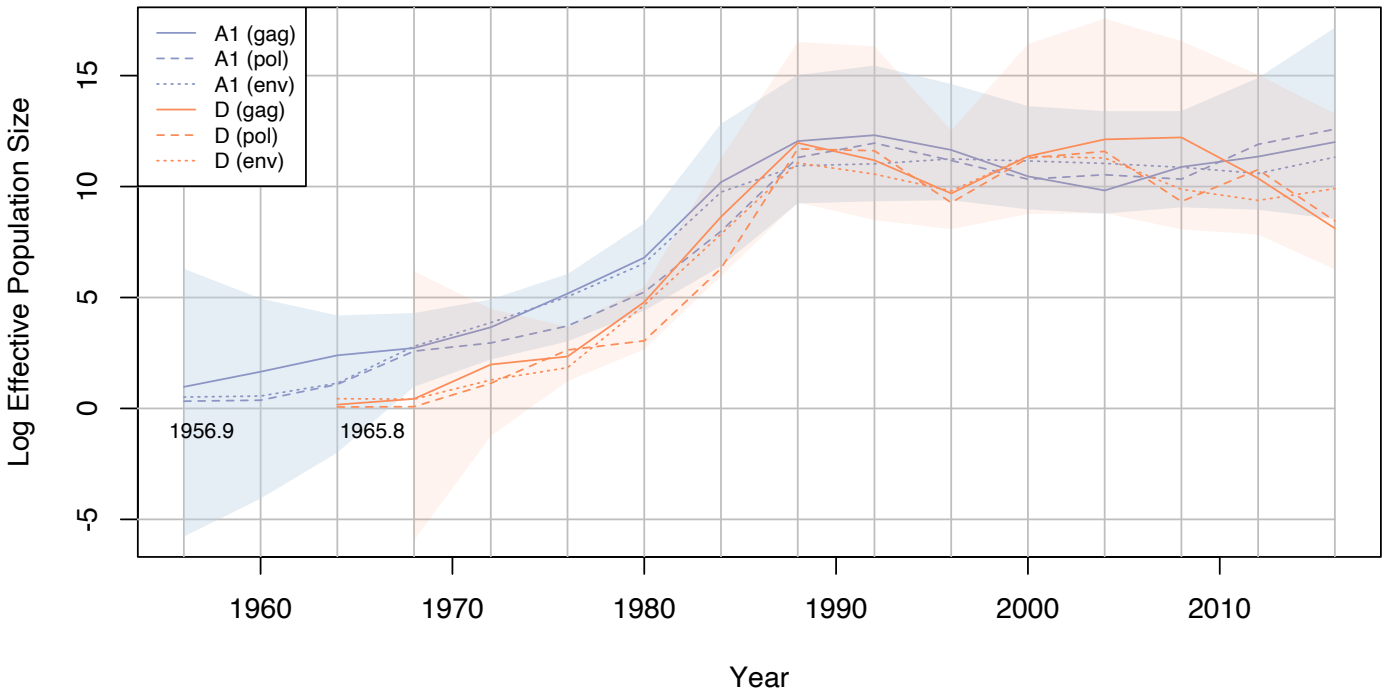

Figure 4.4: BEAST Skygrid estimates of effective population size of subtypes over time, Subtype A1 is shown in blue and subtype D shown in orange. Each line represents the point estimates from each of three genes (*gag, pol, env*) with the 95% BCI based on the three log files combined. Each vertical grid line represents 4 years

4.4.4 Molecular clock rate estimates and tMRCA in Uganda

In the above analyses we have analysed three genes using two demographic models, firstly the skyride with outgroup sequences to look at the placement of the subtypes in their wider context (Figures 4.2 and 4.3) and secondly the skygrid model with only Ugandan sequences to examine changes in N_e over time (Figure 4.4). We find subtype D to have a higher evolutionary rate than subtype A1 in all comparisons of gene and demographic model (Figure 4.5. Evolutionary rates in subtype A1 are lowest in *pol* ($1.51x10^{-3}$), followed by a marginally faster rate in *gag* (mean $1.73x10^{-3}$), whilst *env* rates are considerably higher (mean $2.53x10^{-3}$). The same pattern with respect to the three genes is seen in subtype D (*pol* mean = $1.74x10^{-3}$, *gag* mean = $1.99x10^{-3}$, *env* mean = $2.93x10^{-3}$). The clock estimates for the skygrid model were much the same as the skyride model if slightly higher (*pol*: 1.58 vs 1.51) (*gag* :1.77 vs 1.73), and (*env*: 2.65 vs 2.53) for subtype A1 and (*pol*: 1.83 vs 1.74) (*gag*: 2.10 vs 1.99) and (*env*: both 2.93) $x10^{-3}$ for subtype D. Accordingly, the tMRCA ages as estimated with the skygrid model were slightly later due to a faster rate (compare panels c and d in Figure 4.5). The individual gene skyride tMRCA estimates were remarkably consistent in subtype D skygrid (Figure 4.5d) with distinct peaks in the posterior around the date 1966, while there was more variation between genes in the subtype A1 skyride analysis (Figure 4.5c).

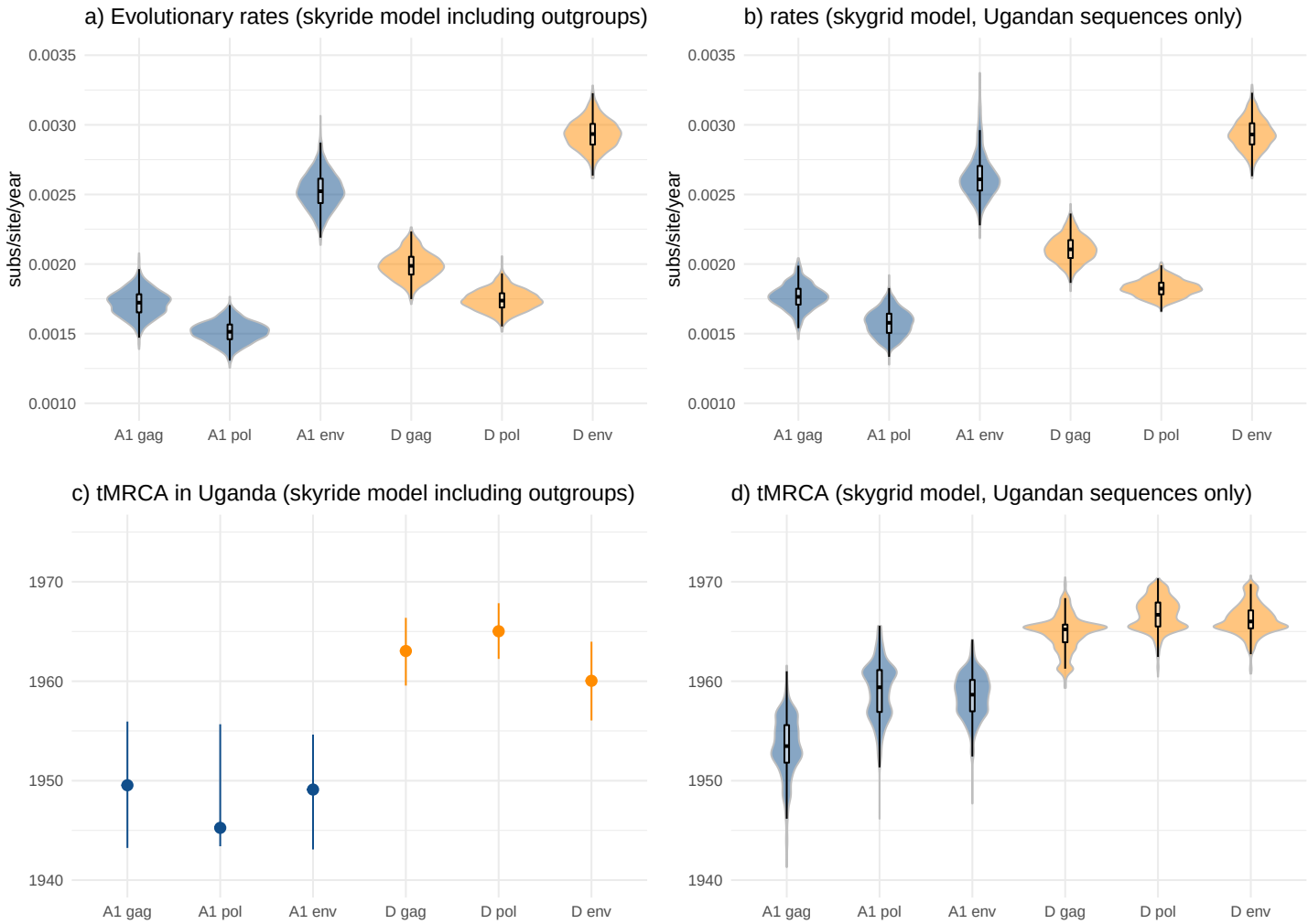

Figure 4.5: Rate and date estimates from BEAST analyses, a) evolutionary rates estimated with the skyride model and outgroups, b) evolutionary rates with skygrid and only Ugandan sequences, c) time to most recent common Ugandan ancestor estimated with the skyride model (age height and BCI of the node is shown) d) root age estimate with the Skygrid model and Ugandan sequences only.

4.5 Discussion

Using full-length HIV genomes spanning three decades of the Ugandan epidemic (1986-2016) we have shown that the two predominant subtypes, A1 and D, entered the country sequentially. Subtype A1 preceded subtype D, as (Yebra et al. 2015) first suggested using *pol* fragment sequence data (A:1960 vs D:1973). Here we have confirmed and refined this difference using full genome data from 1986. We now estimate the age of the Ugandan epidemics to be between 1949 and 1956 for subtype A1, and 1962 and 1965 for subtype D, (estimates from skyride with outgroups and skygrid with Uganda only models respectively). After its later appearance, the Ugandan subtype D clade grew substantially faster than subtype A1 and both subtypes peaked with a similar N_e in the 1988, and thereafter remaining similar. By 1986, subtype D was the predominant (67%) subtype in that sample. However, these early subtype D sequences are not randomly distributed in the subtype D phylogeny. There is clearly visible (and statistically significant) clumping in the distribution of historical sequences in the tree, many of which do not give rise to modern day descendants and we are therefore seeing the extinction of many early subtype D lineages.

As known for some time (e.g. Yirrell et al. 1997), subtype D is well established in Uganda, and was infact the predominant subtype in 1986, the date of the historical samples. This was despite an already widespread subtype A1 epidemic already in place since around 1950. A national epidemic of two HIV subtypes at similar prevalence is unusual. A comparable example might be the early Thai epidemic where the B and CRF01 subtypes were clearly evident (Mccutchan et al. 1992). However, these were found in different geographical areas and were associated with different risk groups (Ou et al. 1992). In Uganda, subtypes A1 and D have been found within the general population, and have both been associated with heterosexual transmission, and diverse URF genomes containing both subtypes confirms that these epidemics were in the same population early on, since dual infection is a prerequisite for their creation. East African Subtype D genomes available from our sample and in the Los Alamos Database are 97% Ugandan, and so it appears that while subtype A spread into the rest of Africa (and indeed across the whole world), subtype D remained endemic to Uganda.

From the phylodynamic reconstruction, exponential growth is inferred for both subtypes until the late 1980s. The slowing of the growth corresponds to the beginning of a strong and multipronged HIV public education campaign in Uganda. After this point there seems to be a change in dynamics, and subtype D decreases in proportion from 67% in 1986 to 17% in the modern day (chapter 3); a decline which other large Ugandan cohorts have also reported (Conroy et al. 2010, Lamers et al. 2020). Subtype D infections progress relatively quickly to AIDS (Kaleebu et al. 2002, Ssemwanga et al. 2013, Vasan et al. 2006, Kiwanuka et al. 2008, Baeten et al. 2007, McPhee et al. 2019, Easterbrook et al. 2010). We and others have shown that subtype D has a higher propensity for CXCR4 tropism (Huang et al. 2007, Kaleebu et al. 2007) which might be the cause of increase in virulence. Under many models a reduction in

risk through behavioural change on a population level, might be expected to favour a variant with a lower virulence and consequently a longer exposure window (Blanquart et al. 2016, Fraser et al. 2007), and in Uganda that may have favoured subtype A1 over subtype D in the later period.

In our analysis we have adopted a novel approach to making genome level MCC trees by analysing genes with different evolutionary rates separately and then combining them to reflect the genome history as a whole. We have taken particular care in our analyses to remove intra-subtype recombination from each of our gene alignments, since the clock rate will be particularly sensitive to recombination (Schierup & Hein 2000*b*). The gene level results were consistent enough that we were able to combine the topologies into one consensus tree, whilst allowing for uncertainty introduced by intra-gene recombination via incongruent branching patterns. This method seems to be robust in recovering similar key tMRCA dates at the subtype level. Thus, the MCC skyride genome tMRCA of all subtype D ($\approx$1955) is consistent with earlier date estimates (Faria et al. 2014, 2019, Korber et al. 2000), as were those of subtypes B and A (Worobey et al. 2016, Tongo et al. 2018).

We used two different population growth models in our BEAST phylogeny reconstructions. Firstly, we used the skyride with a wide range of outgroup sequences to look at the placement of the subtypes in their wider context. Secondly we used the skygrid with sequences solely from Uganda to look at the change in effective population over time. We observed that the skygrid (with single subtype) evolutionary rate estimates are consistently slightly higher and the dates are consistently slightly earlier than those from the skyride (with multiple subtype) model. Previously, Wertheim et al. (2012) described the phenomenon of heterotachy in different subtype clock rate estimations. Specifically, they showed that subtype evolutionary rates are not all the same, and analyses using single subtypes produce earlier tMRCAs than combined subtype analyses, which is consistent with our observations.

The evolutionary rate in subtype D appears higher than that for subtype A with all models and for all genes. Yebra et al. also previously reported a faster rate of subtype D than A1 using just *pol* data, but with a larger difference (2.4 $x10^{-3}$ vs $1.7x10^{-3}$). The use of historical genome sequences from 1986 has significantly reduced this gap, and improved the accuracy of evolutionary rate estimates. Subtype D is approximately 10 years younger than subtype A (depending on the model examined), therefore we believe that the faster rate of subtype D is a subtle effect of the time dependent clock rates observed in viruses e.g. (Duchêne et al. 2014, Aiewsakun & Katzourakis 2015). Others have suggested that using mixed clock models (Bletsa et al. 2019) aid with evolutionary rate estimates, but having full-length genomes spanning 3 decades should add more of an improvement to estimates than the clock model itself,

as others have found (Wilkinson et al. 2015, Worobey et al. 2016). Similarly, Patino-Galindo & Gonzalez-Candelas (2017) suggest the clock rate for subtype D is slower than other subtypes, but the subtype D genome data presented here doubles what has been previously available, and so has provided a better estimate.

We suggest that differences in tMRCA in each gene might not necessarily reflect errors in the BEAST model, but that there may be real differences in evolutionary processes occurring in each gene, particularly since the same full genomes were used for all models and partitions. For instance, the 5-year discrepancy between Skyride *pol* D (1965) and skyride *env* D (1960) might indicate an intra-subtype gene conversion event in *pol* that might that makes it appear younger than other genes. Also of note, is that when examining the individual gene tree in *pol* D (Supplementary Figure 4.6) a well-supported coalescent event at 1973.4 is observed. This is where the majority of modern sequences meet, and is consistent with our previous estimate of 1973 based on pol data alone and without the historical sequences (Yebra et al. 2015). This highlights the importance of the sampling of the sequences as well as the estimation procedure.

We present a tale of two subtypes in Uganda. Firstly, subtype A1 enters Uganda between 1949 and 1956, growing exponentially. Secondly, subtype D enters between 1962 and 1965 and rapidly overtakes to become the dominant subtype (67%) in 1986. After the AIDS education era begins in the late 80s, there is a reversal of fitness as subtype D declines rapidly. We propose that many historical subtype D lineages went extinct and failed to infect new persons because a faster progression to AIDS gave the virus a shorter window to do so.

4.6 Supplementary Information

Subtype	Country	Sampling Year	Accession	Name
B	FRANCE	1983	K03455	HXB2
B	UNITED STATES	1983	AY835748	5096
B	UNITED STATES	1983	AY835754	5084
B	UNITED STATES	1983	AY835770	5082
B	UNITED STATES	1983	AY835777	5018
B	UNITED STATES	1983	AY835781	5157
B	UNITED STATES	1983	M17451	HAT3
D	DEM REP OF CONGO	1983	K03454	ELI
D	DEM REP OF CONGO	1983	M27323	NDK
D	DEM REP OF CONGO	1984	U88822	84ZR085
D	DEM REP OF CONGO	1985	M22639	Z2Z6_Z2_CDC_Z34
D	DEM REP OF CONGO	1987	MH705152	PBS5635
D	DEM REP OF CONGO	2002	KU168272	LA18ZiAn
D	DEM REP OF CONGO	2003	KU168271	LA17MuBo
D	BRAZIL	2010	KJ787684	10BR_RJ095

D	BRAZIL	2010	KJ787683	RJ108
D	BRAZIL	2010	KU749394	DEMD10BR034
D	CAMEROON	2001	AY371155	01CM_0009BBY
D	CAMEROON	2001	AY371156	01CM_0175BA
D	CAMEROON	2001	AY371157	01CM_4412HAL
D	CAMEROON	2008	MN153488	08CMBDSH25
D	CAMEROON	2010	JX140670	DEMD10CM009
D	CAMEROON	2010	KP109501	DEMD10CM018
D	CHAD	1999	AJ488927	MN012
D	CYPRESS	2006	FJ388945	CY163
D	SENEGAL	1990	AB485649	SE365
D	KENYA	2001	AF457090	NKU3006
D	KENYA	1997	AY322189	ML415_2
D	KENYA	2011	KF716476	DEMD11KE003
D	TANZANIA	2004	KX907406	CO6405V4
D	TANZANIA	2001	AY253311	A280
D	UGANDA	1991	AB485650	UG270
D	UGANDA	1991	AB485651	UG270
D	UGANDA	1992	AJ320484	92UG001
D	UGANDA	1993	AY713418	93UG_065
D	UGANDA	1994	U88824	94UG114
D	UGANDA	1995	MH705143	42-877
D	UGANDA	2007	JX236670	p191647
D	SOUTH AFRICA	1984	AY773338	R2
D	SOUTH AFRICA	1985	AY773339	R214
D	YEMEN	2001	AY795903	01YE386
D	YEMEN	2002	AY795907	02YE516
D	CHAD	1999	AJ488926	MN011
D	CHAD	1999	AJ519488	MN011
D	CHAD	1999	AJ519489	MN012

Table 4.1: Additional genomes from the Los Alamos Database (for subtype D)

Subtype	Country	Sampling Year	Accession	Name
G	DEM REP OF CONGO	2003	KU168277	LA23LiEd
G	DEM REP OF CONGO	1987	MH705134	PBS1191
G	DEM REP OF CONGO	1987	MH705155	P406
G	DEM REP OF CONGO	1987	MH705162	87-2580
A2	CYPRUS	1994	AF286237	94CY017_41
A2	DEM REP OF CONGO	1997	AF286238	97CDKTB48
A2	CAMEROON	2001	GU201516	01CM_1445MV
A2	DEM REP OF CONGO	1987	MH705163	PBS1195
A4	DEM REP OF CONGO	1997	AM000053	97CD_KCC2
A4	DEM REP OF CONGO	1997	AM000054	97CD_KTB13
A4	DEM REP OF CONGO	2002	AM000055	02CD_KTB035
A5	DEM REP OF CONGO	1997	FM877777	97CD_KTB119

A5	DEM REP OF CONGO	2002	FM877780	02CD_KS069
A5	DEM REP OF CONGO	2002	FM877781	02CD_LBTB084
A5	DEM REP OF CONGO	2002	FM877782	02CD_MBTB047
A1	DEM REP OF CONGO	2002	KU168256	LA01AlPr
A1	KENYA	1994	AF004885	Q23_17
A1	KENYA	2000	AF457055	KER2012
A1	KENYA	1999	AF457063	KNH1088
A1	KENYA	1986	AF539405	ML170_1986
A1	KENYA	1997	AY322185	ML013_2
A1	KENYA	2001	EU110085	ML752
A1	KENYA	2002	EU110092	ML1990
A1	KENYA	2006	FJ623475	06KECst_006
A1	KENYA	2006	FJ623480	06KECst_009
A1	KENYA	2006	FJ623483	06KECst_016
A1	KENYA	2006	FJ623488	06KECst_017
A1	KENYA	2011	KF716474	DEMA111KE002
A1	KENYA	2004	KT022360	04KE169579V3
A1	KENYA	2004	KT022361	04KE263806V2
A1	KENYA	2004	KT022364	04KE378531V2
A1	KENYA	2004	KT022365	04KE406723V2
A1	KENYA	2005	KT022370	05KE185405V4
A1	KENYA	2006	KT022378	06KE196199V6
A1	TANZANIA	1997	AF361872	97TZ02
A1	TANZANIA	1997	AF361873	97TZ03
A1	TANZANIA	2001	AY253305	A173
A1	TANZANIA	2001	AY253314	A341
A1	TANZANIA	2005	KX907341	CO0260V5
A1	TANZANIA	2004	KX907343	CO0330V4
A1	TANZANIA	2006	KX907346	CO0434V7
A1	TANZANIA	2005	KX907347	CO0439V5
A1	TANZANIA	2003	KX907348	CO0543V2
A1	TANZANIA	2002	KX907352	CO0783V0
A1	TANZANIA	2005	KX907364	CO3083V4
A1	TANZANIA	2004	KX907372	CO3365V2
A1	TANZANIA	2006	KX907378	CO3504V7
A1	TANZANIA	2004	KX907383	CO3718V3
A1	TANZANIA	2004	KX907389	CO3878V2
A1	TANZANIA	2005	KX907401	CO6161V5
A1	TANZANIA	2005	KX907412	CO6592V5
A1	TANZANIA	2006	KX907414	CO6637V7
A1	TANZANIA	2006	KX907423	CO6830V7
A1	TANZANIA	2006	KX907431	CO6974V7
A1	RWANDA	1992	AB253421	92RW008
A1	RWANDA	1992	AB287376	92RW025A
A1	RWANDA	1993	AB287378	93RW037A
A1	RWANDA	1993	AY713406	93RW_024
A1	RWANDA	2011	KF716472	DEMA111RW002
A1	RWANDA	2003	KF716499	DEMA03RW001

A1	RWANDA	2007	KP109528	DEMA07RW002
A1	RWANDA	2007	KP223844	R880_MPL_C11
A1	RWANDA	2006	KU749423	DEMA106RW003
A1	RWANDA	2008	KU749424	DEMA108RW010
A1	UGANDA	1992	AB098332	UG029
A1	UGANDA	1992	AB253428	92UG037_A35
A1	UGANDA	1992	AY713407	92UG_029
A1	UGANDA	2007	JX236669	p191084
A1	UGANDA	2007	JX236671	p191845
A1	UGANDA	2007	JX236676	p9004SDM
A1	UGANDA	2009	KF716478	DEMA109UG001
A1	UGANDA	2011	KF716486	DEMA110UG009
A1	UGANDA	2011	KF859745	DEMA110UG001
A1	UGANDA	2009	KP109490	DEMA109UG017
A1	UGANDA	1985	M62320	U455_U455A

Table 4.2: Additional genomes from Los Alamos (for subtype A1)

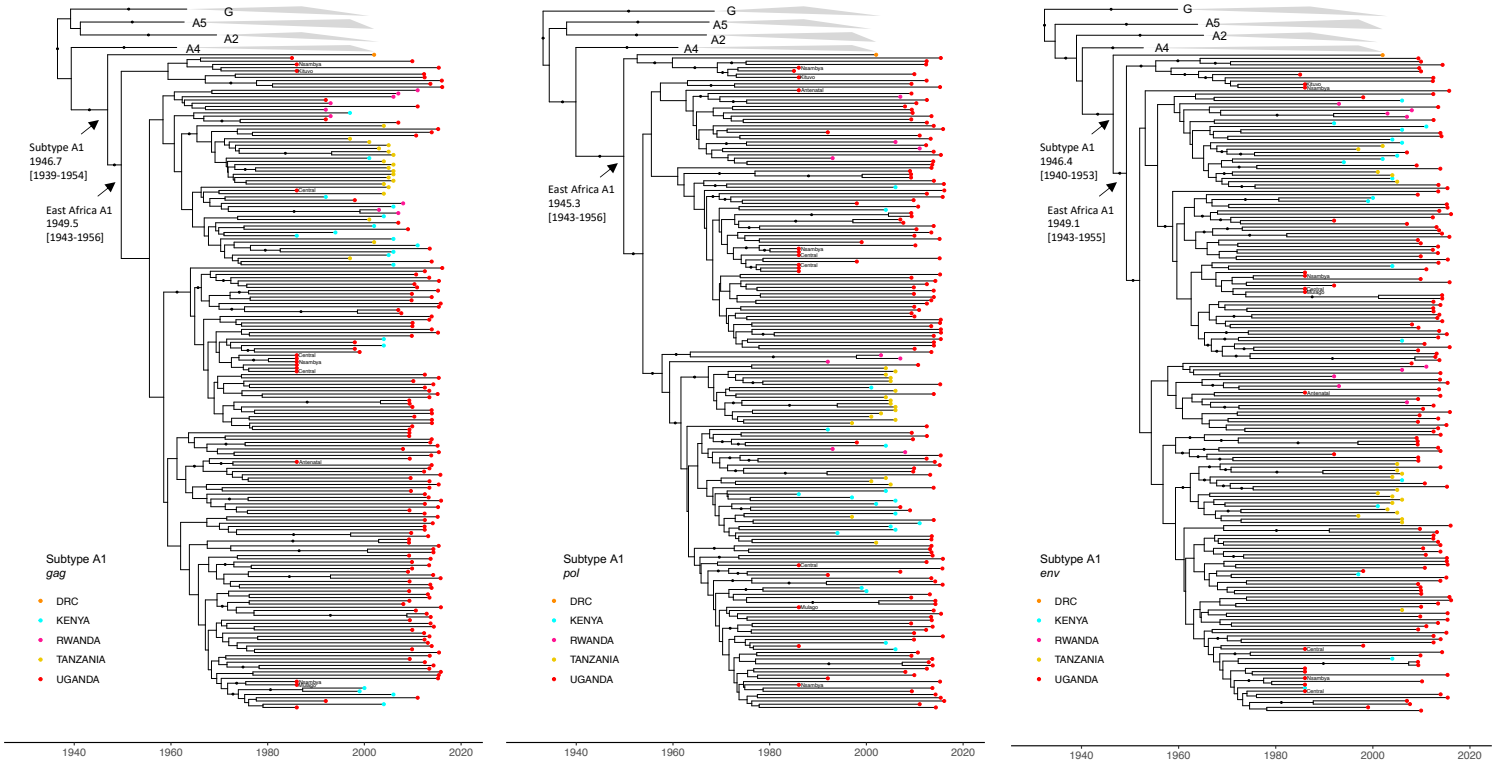

Figure 4.6: Subtype A1 individual gene Maximum Clade Credibility tree from BEAST with median node height *gag* (left) and *pol* (middle) and *env* (right). Outgroups are collapsed for clarity, small black dots at nodes indicate posterior probability of > 0.9

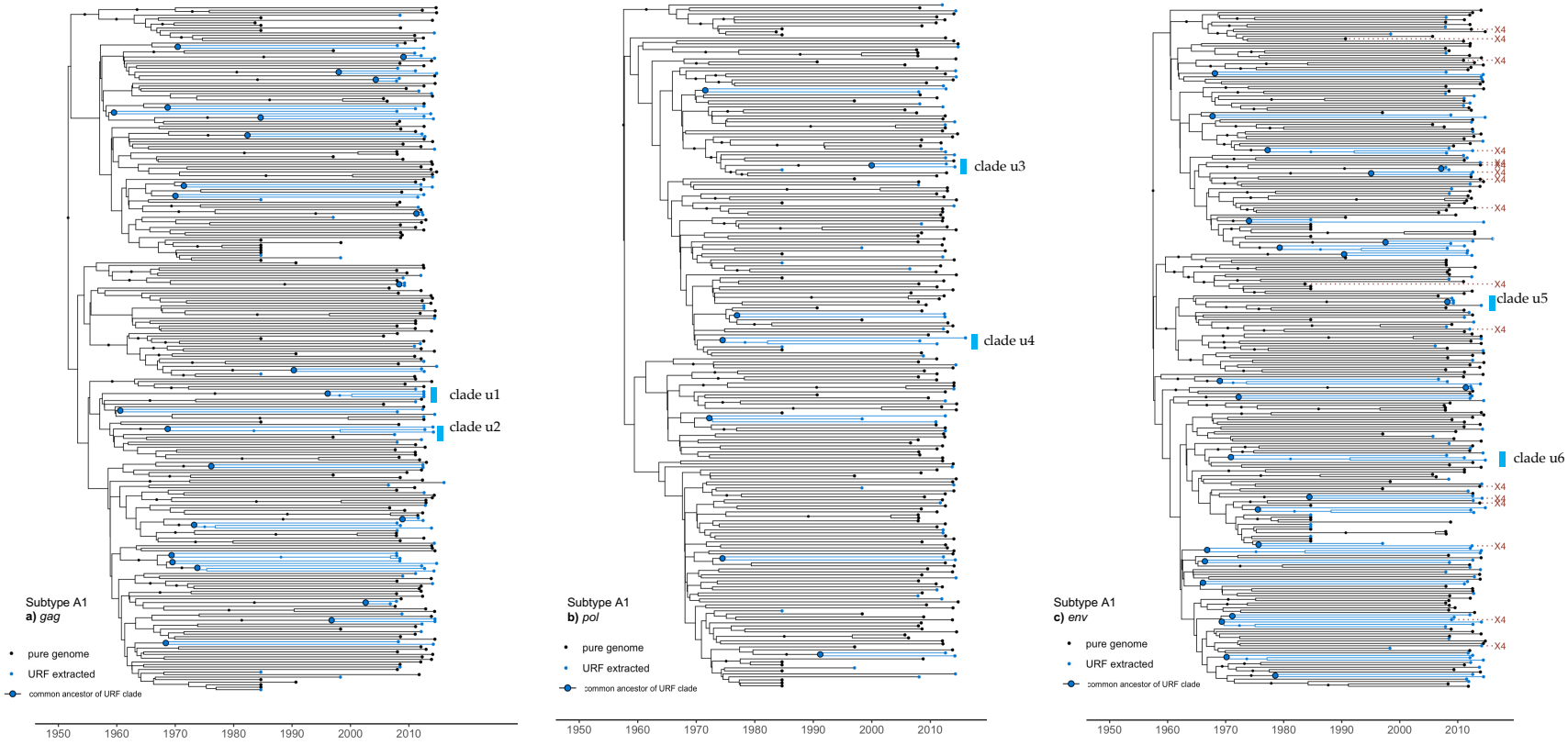

Figure 4.7: Subtype A1 individual gene Maximum Clade Credibility trees with URF extracted gene sequences (blue tips) included alongside 'pure' genomes (black tips). Some circulating recombinant clades are highlighted. The *env* phylogeny indicates X4 tropic sequences.

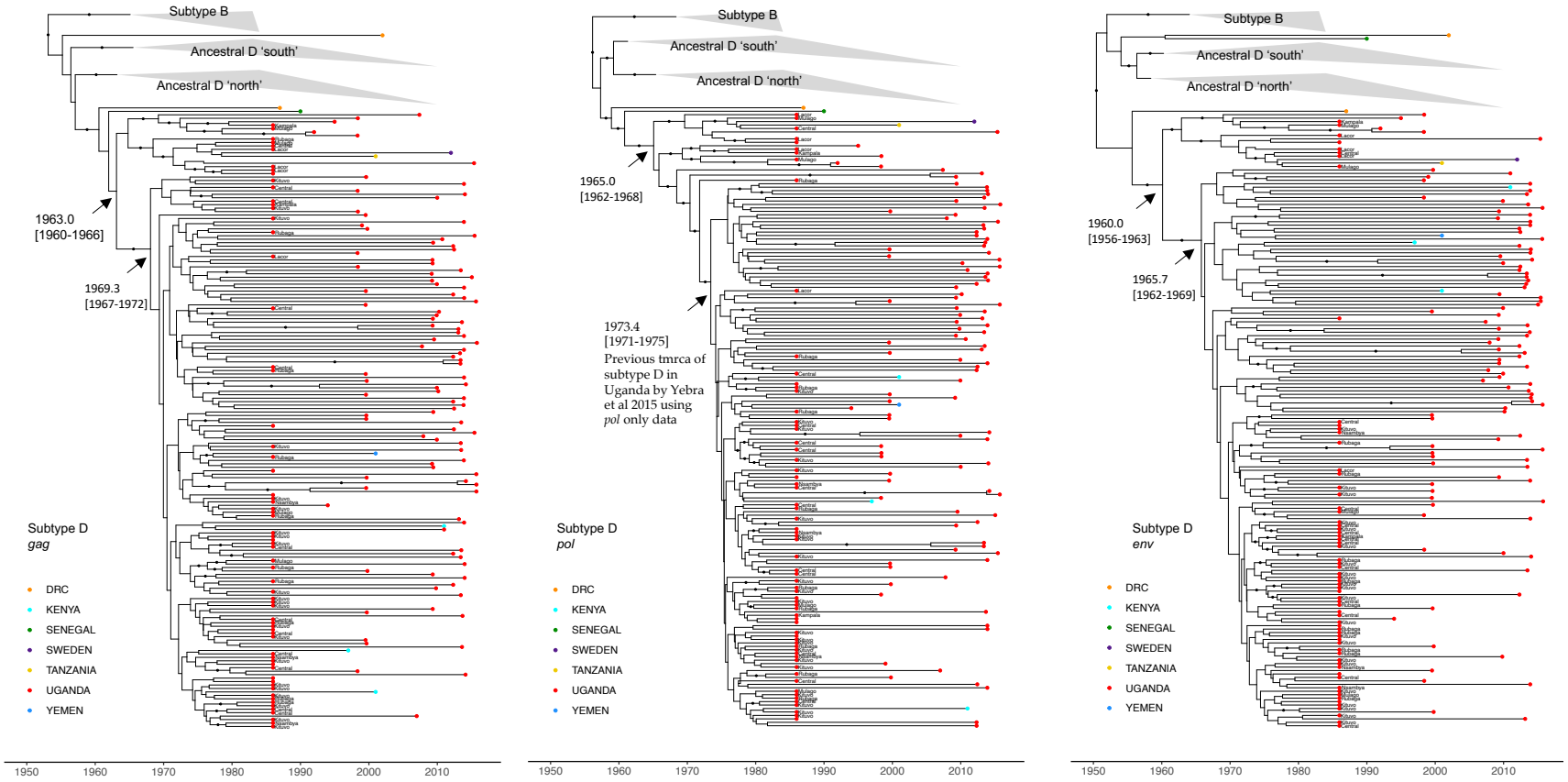

Figure 4.8: Subtype D individual gene Maximum Clade Credibility trees from BEAST with median node height *gag* (left) and pol (middle) and *env* (right). Outgroups are collapsed for clarity, small black dots at nodes indicate posterior probability of > 0.9

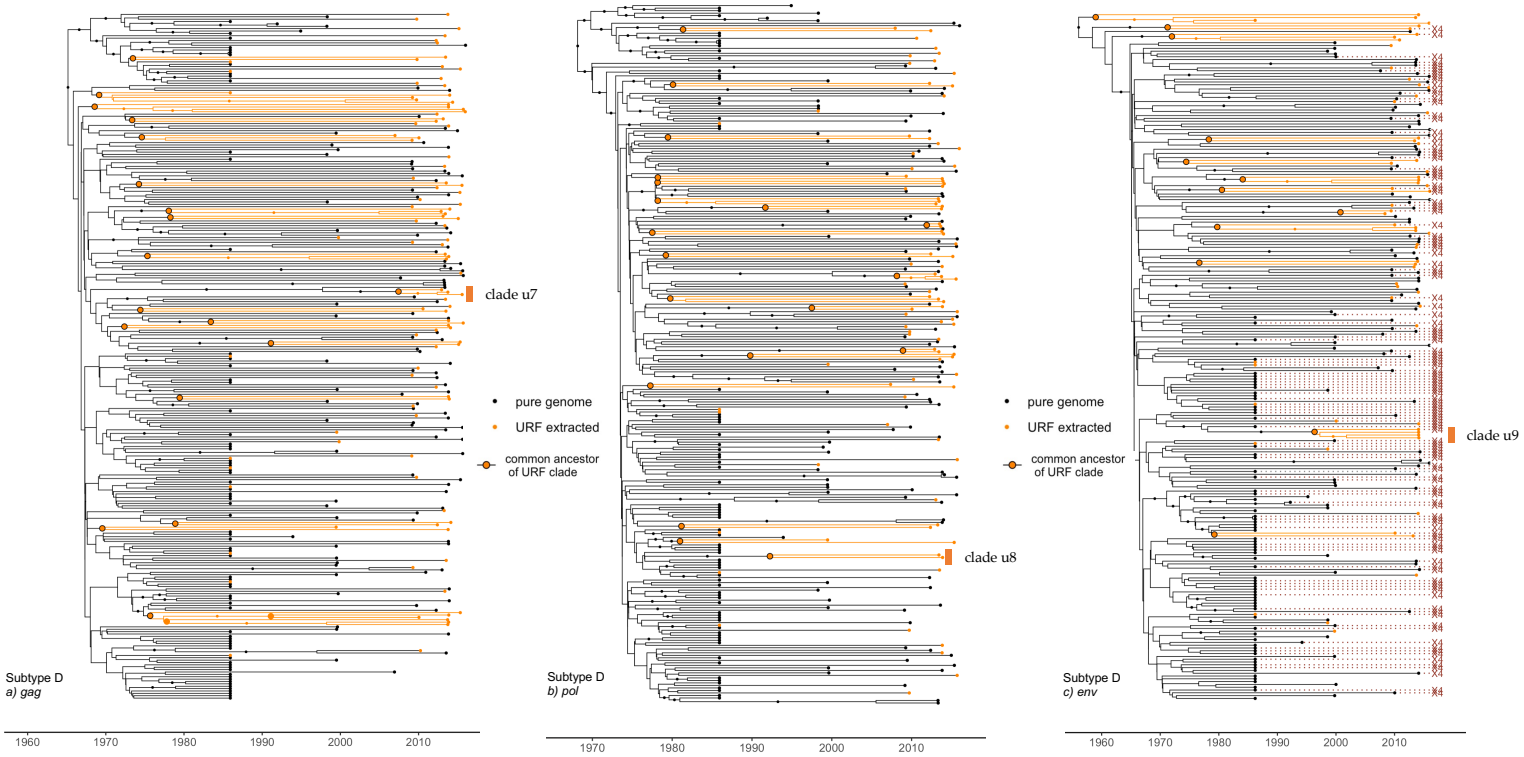

Figure 4.9: Subtype D individual gene Maximum Clade Credibility trees with URF sequences (orange) included alongside 'pure' genomes (black). Some circulating recombinant clades are highlighted. The *env* phylogeny indicates X4 tropic sequences.

Partition	post-2000	pre-2000	D	p(clumped Brownian)	p(clumped random)	Interpretation
Gag A1	239	32	-1.119	0.019	0.000	clumped
Pol A1	31	197	0.925	0.033	0.408	random
Env A1	256	29	-0.934	0.945	0.000	clumped
Gag D	116	172	-1.535	1.000	0.000	extremely clumped
Pol D	112	166	-1.635	1.000	0.000	extremely clumped
Env D	109	135	-2.791	1.000	0.000	extremely clumped

Table 4.3: The Fritz and Purvis "D-statistic" for phylogenetic distribution of age of sequences (pre-2000/ post-2000)

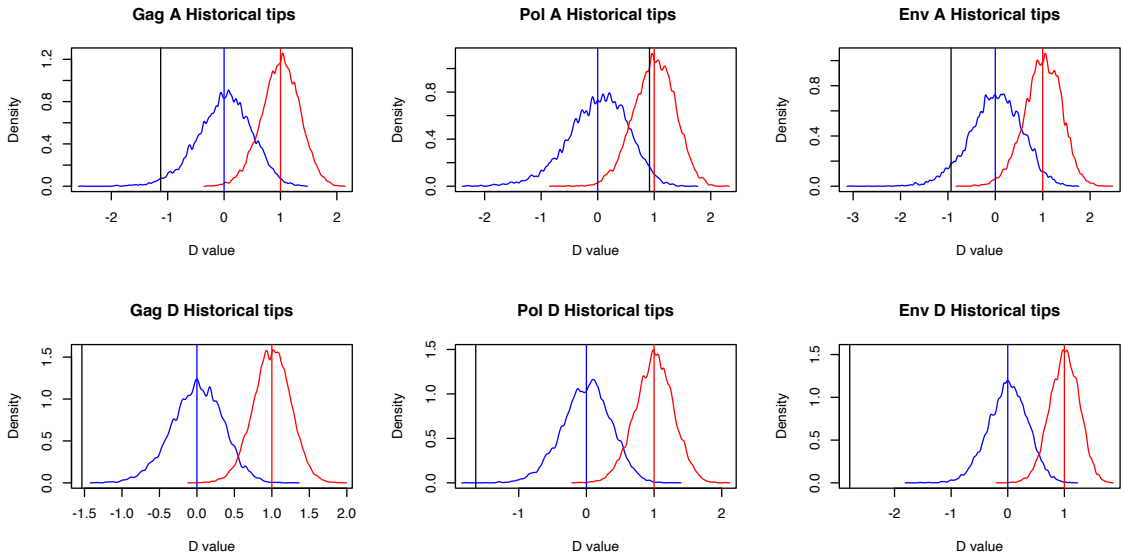

Figure 4.10: Prob density functions of phyloD statistic applied to historical (pre-2000) tips

Partition	R5	X4	D	p(clumped Brownian)	p(clumped random)	Interpretation
Env A1	268	17	1.084	0.051	0.545	random
Env D	113	131	0.040	0.464	0.000	clumped

Table 4.4: The Fritz and Purvis "D-statistic" for phylogenetic distribution of binary character X4 tropism

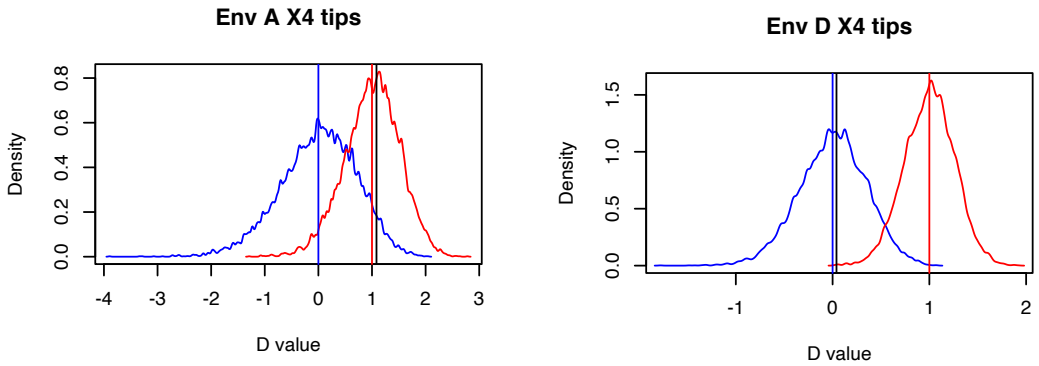

Figure 4.11: Prob density functions of phyloD statistic applied to tips with X4 tropism

Chapter 5

Opinion: HIV-1 circulating recombinant forms are biologically distracting and misleading

Heather E. Grant[1], Abayomi S. Olabode [2], Art F. Poon [2], Andrew J. Leigh Brown[1], and David L. Robertson[3]

1) Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
2) Department of Pathology & Laboratory Medicine, University of Western Ontario, London, Canada
3) MRC Centre for Virus Research, University of Glasgow, Glasgow, UK

5.1 Introduction

The division of viruses into groups that meaningfully reflect their evolutionary history is an important pursuit (Kuhn et al. 2019), but difficult, and as such there are no universally agreed methodologies for virus classification (Kuhn 2021). The International Committee on Taxonomy of Viruses (ICTV) states that virus taxonomy should serve the community in facilitating international agreement, provide stability, avoid confusion, and avoid the unnecessary creation of names. HIV diversity is currently partitioned into subtypes A-D and F-H, J and K, with sub-subtypes A1-A6, F1-F2 and 117 circulating recombinant forms composed of more than one subtype, see lanl.gov/crfs.

The major HIV subtypes have gone through large bottlenecks resulting in genetically distinct subtypes (around 15% divergence; (Li et al. 2015), with distinct patterns with respect to geography and risk groups. Subtype B for example is associated with men who have sex with men in high income countries (Vermund & Leigh-Brown 2012), subtype C with heterosexual populations in South Africa (Wilkinson et al. 2015), or subtype A6 in people who inject drugs in the countries of the former Soviet Union (Abidi et al. 2021).

The ancestral sequence of COVID-19 was known before the major lineages had started to emerge (Rambaut et al. 2020). By contrast, when HIV was first sequenced in 1985, it was already the cause of a large pandemic with geographically distinct lineages that had accrued around 60 years of mutation and recombination (Worobey et al. 2008). It was not until sequences from Africa became available to researchers, that the scale of the genetic diversity (Alizon et al. 1986, Potts et al. 1993) or age (Zhu et al. 1998) of the pandemic became apparent. Sequencing and sampling at the time was not able to capture the breadth and depth of global diversity, so that when the first subtypes (A-E) were designated in 1992, it was not by full-genome sequences, but by single gene sequences from *gag* or *env*.When more data became available, it was became increasingly clear that recombination was a significant factor in the evolution of HIV-1 (Robertson et al. 1995). Therefore, when 'subtype E' was sequenced in full, the *gag* and *pol* regions appeared more like subtype A1 (Gao et al. 1996, Carr et al. 1996) and so the lineage became known as CRF01_AE. In 1999, a committee met and promulgated a new rule. Circulating recombinant forms (CRF) were defined as three or more epidemiologically unlinked genomes with a distinct recombination pattern with more than one subtype (Robertson et al. 2000).

5.2 CRF frequencies

The Los Alamos HIV Database (accessed 24 February 2022) was searched for CRF sequences of any length. Of the 117 CRF classifications, 51 (44%) had fewer than ten sequences, and the majority (92 or 78%) had fewer than a hundred representative sequences, see Figure 5.1. An example is CRF 23, a subtype B and G chimeric genome described from Cuba (Sierra et al. 2007), which has not been documented since, either as a partial gene sequence or as a full-length genome. Partial gene sequences are much more common in the database, and when we specifically looked for full-length sequences, a large majority 98 or (83%) of CRFs had ten or fewer genome representatives.

Therefore, most CRFs have poor representation in the database. The two main exceptions to this are CRF01and CRF02 with 66,368 and 23,999 sequences (of any length) respectively, and they alone make up 71.3% of the CRF sequences available. In fact, both of these CRFs have more numerous representations than the subtypes F, G, H, J, and K combined (n=16,185). This is in part explained by their age; CRF01 originated in Africa, before seeding epidemics in Thailand and China, and is a large enough clade that there have multiple proposals to further stratify it into sub-lineages (Feng et al. 2013, Li et al. 2017, An et al. 2020). Similarly, CRF02 (composed of subtypes A and G) is a very old and diverse lineage found widely in West and Central West Africa (Faria et al. 2012, Mir et al. 2016).

There has been much speculation that recombination between multiple subtypes might bring together favourable mutations to bring forth some biological advantage which allows them to outgrow their parent subtypes (e.g. Turk & Carobene 2015). With some exceptions however, it appears that the majority of CRFs are not reported after their initial designation. These events therefore appear to be more lost to genetic drift than confer any selective advantage.

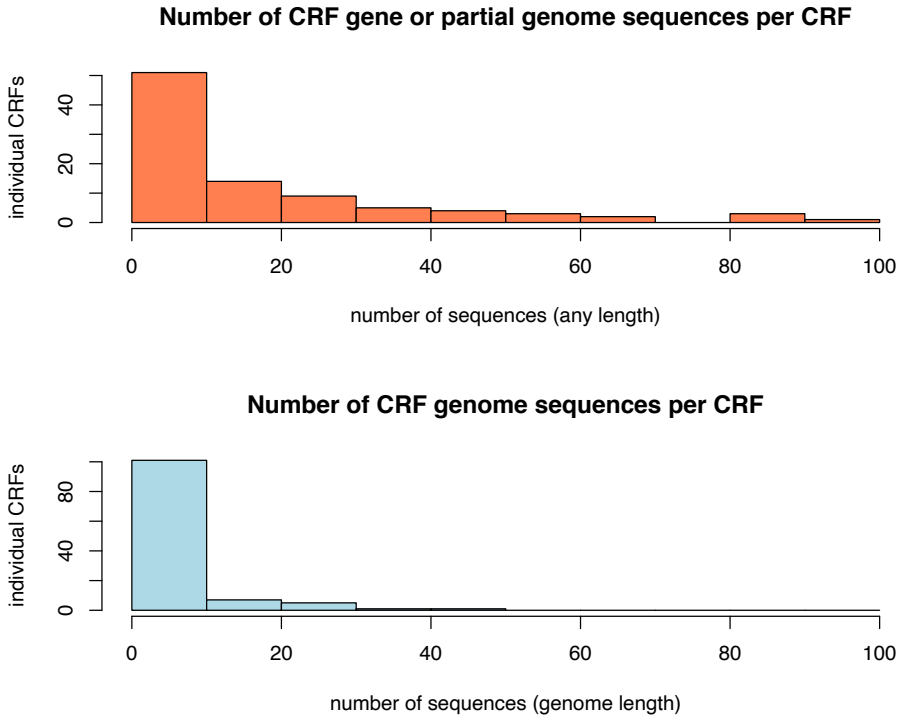

Figure 5.1: Los Alamos Database was queried for full length genomes

5.3 Recombination is a continuous process

Mutation and recombination are not independent processes (Schlub et al. 2014), because they are both a consequence of the reverse transcription process (Coffin et al. 1997). Template switching is an obligate part of reverse transcription (Temin 1993) and additional template switches can happen about 5 to 14 times per replication event (Cromer et al. 2016), making recombination more frequent than mutation (Hu & Temin 1990). Recent advances in single genome amplification have meant recombination within a patient can be well documented and described (Song et al. 2018). Recombination is not only a mechanism for increasing diversity, but also a mechanism to avoid error catastrophe (Tripathi et al. 2012) and maintaining genome structure and integrity (Rawson et al. 2018). We would argue that a single recombination event is therefore not remarkable in of itself.

Signatures of recombination can be detected in both distant and recent histories. In distant history, the SIV in the chimpanzee (from which HIV-1 M was descended) is a chimera of two SIVs from different monkey prey species of chimpanzees (Bailes et al. 2003). Within Group 1-M, a deep recombination event is evidenced by differences in the time to most recent common ancestors across the genome (Olabode et al. 2019). It is therefore not surprising that global CRF lineages represent a spectrum of ancient and modern recombination events (Zhang et al. 2010). In the same way, breakpoints within the same genome will have different ages as sequential recombination events take place (Gao et al. 2021). Another example, CRFs 79, 102, 104, 109, 113, 117, which are all found in China, are composed of CRFs 01 and 07, and will therefore have breakpoints between CRF 01 and CRF 07 parents, but also older breakpoints between subtypes C and B in the CRF07 genomic regions.

5.4 Recombination patterns along the genome

Fan et al., in 2007 described the distribution of 354 unique breakpoints amongst 80 CRFs and URFs Figure 5.2 a). They reported moderate levels of recombination in structural genes *gag* and *pol*, very low levels within envelope, and much higher levels at the accessory gene regions (*vif, vpr, tat*, and *nef*). Whilst there are a number of factors that determine breakpoint frequency and location along the genome, like RNA structure (Simon-Loriere et al. 2009, 2010), and sequence similarity (Baird et al. 2006), recombination in the *env* gene seems to be highly suppressed. This may be because the envelope proteins, which are essential for cellular entry, have an extremely complex trimer structure, providing a significant functional constraint (Bagaya et al. 2015, Woo et al. 2014, Golden et al. 2014) against their disruption, thereby being recombined intact like a 'cassette tape' in and between viruses (Archer et al. 2008).

In Uganda, two subtypes co-circulate at high frequency, leading to myriad unique recombinant forms (Grant et al. 2020). Previously, we described the distribution of breakpoints of A1 and D URFs in Uganda using the phylogenetic method SCUEAL (Kosakovsky Pond et al. 2009) (Figure 5.2b). We found a remarkably similar distribution to that of Fan et al. (Figure 5.2a). We then extended our SCUEAL analyses to a large curated global dataset (Olabode et al. 2019) with 2558 URFs, Figure 5.2c), and again find peaks of recombination in the accessory gene regions either side of envelope. In addition, SCUEAL, which is capable of identifying within subtype recombination, also finds the same pattern within subtype B (n=434, Figure 5.2d). Finally, we recover similar breakpoint distributions using two other non-phylogenetic methods using JPHMM (a hidden Markov Model based tool, n=1394 genomes, Figure 5.2e) and using dynamic stochastic block modelling from (Olabode et al. 2022), Figure 5.2f, demonstrating that this is not a result of using standard phylogenetics or SCUEAL specifically.

We therefore demonstrate a reproducible pattern of recombination along the HIV genome using phylogenetic and non-phylogenetic methods, that can be seen both within and between subtypes. Given that there are clear hotspots of recombination, genomes may have the same inter-subtype recombination pattern by chance, despite having independent evolutionary origins. This is indeed what has been observed in Uganda, and as was the case with CRF84 which was retracted some time after its publication for this reason.

5.5 'Pure' subtypes

The concept of 'pure' subtypes is misleading, as all subtypes originated in the Democratic Republic of Congo and were likely the result of recombination (Kalish et al. 2004). The lack of clarity as to what comprises a 'parent' subtype is exemplified by the unknown "Subtype E" parent of CRF01 (Anderson et al. 2000) and indeed, it was shown that CFR02 is the parent of subtype G, not the other way around (Abecasis et al. 2007).

Recently we have applied a new clustering method inspired by network science, to detect 'communities' of HIV diversity (Olabode et al. 2022). A network is constructed based on genetic distances (TN93) across sliding windows between the genomes, and an optimal number of clusters determined with ICL criterion. Using this method, HIV global diversity can be categorised into 25 optimal 'communities' of diversity, with only 5% of the genomes in the analysis belonging to the same 'community' in every window along their genome, therefore deemed 95% of the genomes recombinant.

Furthermore, when this analysis was applied to the subtype reference genomes of A-D,F-H,J,and K, the communities which were found were not clearly delineated in all windows of their genome, showing some evidence of recombination between the subtypes, particularly subtypes F,G, J and K. Therefore, even the most 'pure' HIV subtypes cannot be strictly delineated.

5.6 Misleading classification

The subtype classification is useful where there have been strong bottlenecks and strong geographic structure is present, like for example the clade A6, which is found in the countries of the former Soviet Union, or subtype B in the global North. Where there are multiple subtypes circulating in the same region however, myriad recombinant forms will arise, as is seen today in cosmopolitan centres with high levels of immigration such as London (Yebra et al. 2018).

CRFs and pure subtypes classifications, particularly in the same geographical regions, may therefore create some confusing and misleading subtyping results, and when we include new CRF references in the subtyping process, it becomes difficult to decipher new CRF sections from parts of the genome that resemble the parent. Figure 5.3 shows an example of

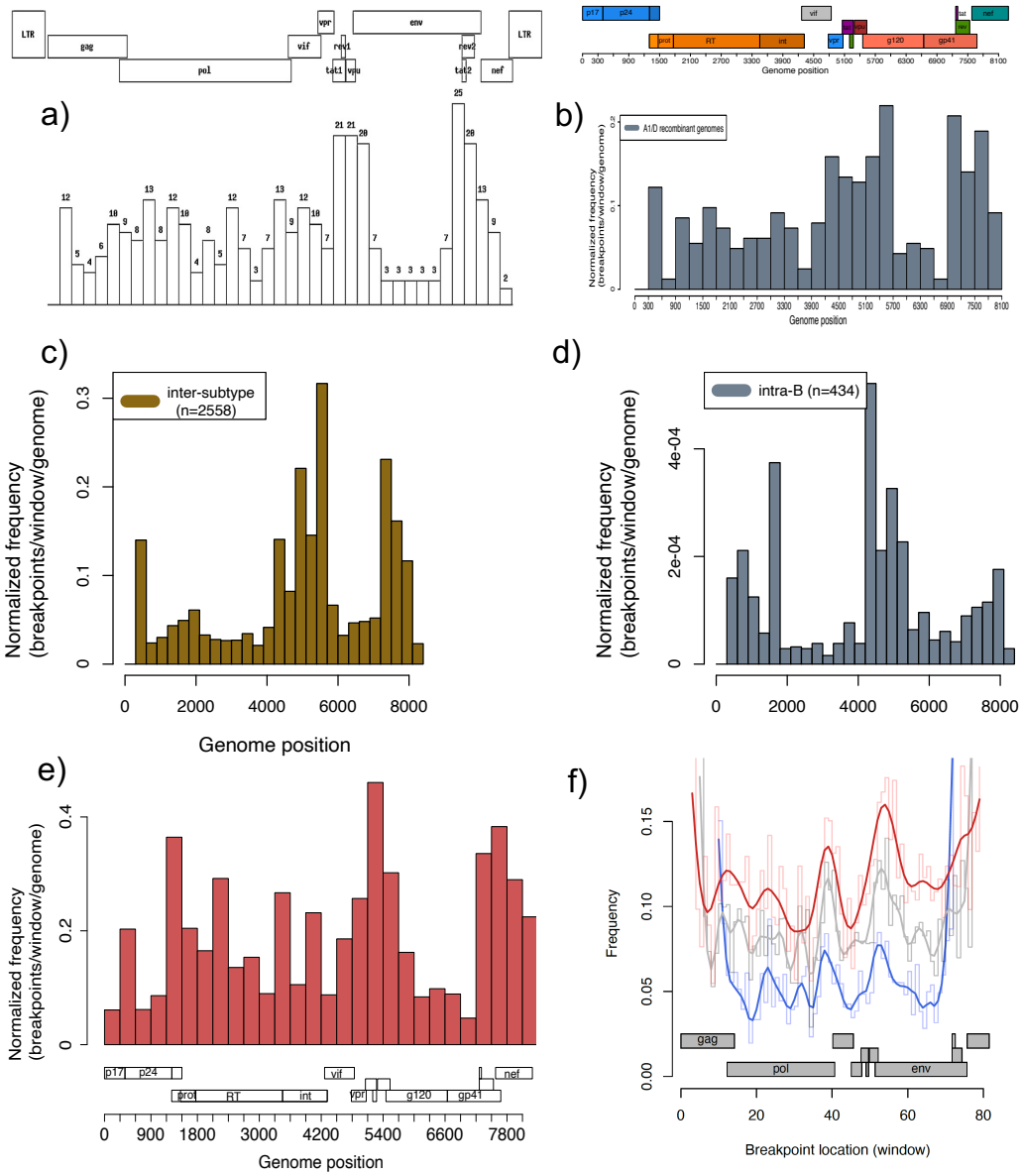

Figure 5.2: Patterns of recombination along the genome as determined by a) (Fan et al. 2007) b) (Grant et al. 2020), c) SCUEAL and global HIV divesity d) SCUEAL within subtype B, e) JPHMM inter-subtype breakpoint distribution on the same genomes f) the dynamic stochastic block model (DBM) breakpoint distribution as found by (Olabode et al. 2022)

a genome from Spain in 2014 [KT276261] subtyped by SCUEAL, with the recombination pattern G/CRF25/G/CRF43. Examining the structure of CRF25 and CRF43 however, the picture becomes more confusing as the section of the genome predicted to be CRF25 has parents subtype A and G, and the parents of CRF43 are CRF02 and subtype G. A strict

interpretation would suggest this genome is an inter-subtype recombinant with three parents, G, CRF25 and CRF43. It could also be the case that KT276261 is a subtype G genome with genome regions that are similar to parent sections of those CRFs (also subtype G). Moreover, the fact that CRFs25 and 43 are from Saudi Arabia might suggest a priori they are related to the parent rather than the CRF.

Some subtyping programs, for example JPHMM, do not include CRFs in their reference set. This would mean that CRF03, for example would be reported as a C/B recombinant by JPHMM. For the automated subtyping tools which do include CRFs however, it becomes cumbersome to include new CRF references as they arise, and they may face the kind of misleading taxonomy presented in Figure 5.3.

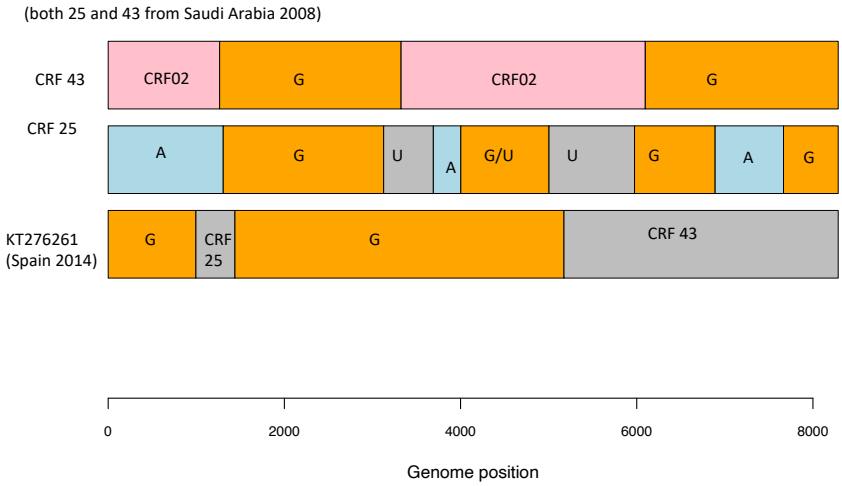

Figure 5.3: CRFs make describing diversity more difficult. KT276261 is subtyped with SCUEAL and appears related to CRFs 25 and 43, but may just be phylogenetically similar to the parent sections of those CRFs.

5.7 Conclusion

Both subtypes and CRFs have undergone a multitude of recombination events in recent and ancient times, and are far from uniform in terms of relative frequency or age. Incomplete sampling of diversity in the early part of the epidemic meant that certain subtypes were 'grandfathered' in (Foley et al. 2016), but the subtype system, whilst imperfect, is widely understood and reflects founder effects that happened before the 1960s (Worobey et al. 2008).

We argue that while some CRFs seed significant epidemics, most are never documented again and therefore represent small clusters transmission clusters which confuse our understanding of HIV diversity. While retiring the current subtype classification may well be 'throwing the baby out with the bathwater', we must accept that recombination is a pervasive, ongoing, and even predictable evolutionary process in HIV-1. Because recombination is an ongoing process, the use of highly curated subtype references means that our understanding of HIV diversity is fixed in a single point in time. Genome references, whilst they may have been carefully selected, will have undoubtedly have undergone recombination, and will undoubtedly give rise to new recombinant forms. We therefore make two suggestions.

Firstly, we suggest the use of unsupervised clustering methods in finding reference sequences, such as the DSBM (Olabode et al. 2022). Analyses that require large numbers of sequences to explore the diversity and evolution of HIV (like phylodynamic or clustering studies) would certainly benefit from this approach, as it would include the most appropriate references, whilst also reducing the burden of inter and intra-subtype recombination. Furthermore, classification of diversity can be updated as more data becomes available, and reflect the frequency of new sequences and the dynamic process of recombination.

Secondly, we propose that the new naming of CRFs is counterproductive, because many do not have large enough numbers of sequences available (and therefore epidemiological significance), because the presence of shared breakpoints among three genomes alone is not an interesting feature enough in of itself, and because of the taxonomic confusion they add.

Chapter 6

General Discussion

6.1 HIV phylogenetics with full-length genomes

This thesis presents two large datasets of genomes from modern Uganda by the PANGEA-HIV consortium, and a historic sample of 109 genomes from 1986, providing new depth and detail about the HIV epidemic in Uganda. With longer genome sequences however the challenges posed by recombination (Posada & Crandall 2002) become more severe, particularly in this well-established generalised epidemic with two subtypes and high levels of superinfection (Redd et al. 2012, Ssemwanga et al. 2012). In the modern dataset (chapter 2), over half of the genomes were distinct, unique, inter-subtype recombinant forms (Figure 2.4), in chapter 3, a range of distinct, unique, inter-subtype recombinant forms are found by the year 1986 (Figure 3.5), and in chapter 4, by including sections of genes from URFs into phylodynamic analyses, the diversity of ages of URF clades from the 1970s to 2000s is demonstrated (Figures 4.7 and 4.9). Recombinant forms are therefore being continually created.

If dual infections with different HIV subtypes occur, so too must dual infections with two viruses of the same subtype, and intra-subtype breakpoints of a range of ages should also be present. During the preparation of chapter 2, I also described the intra-subtype recombination distribution along the genome, (included here as Appendix A). SCUEAL is the only program able to detect intra-subtype recombination because it uses multiple reference genomes from the same subtype that have been carefully selected and screened with GARD for recombination (Kosakovsky Pond et al. 2009). Using SCUEAL I showed that almost all (91.8%) of the genomes exhibit either intra- or inter-subtype recombination, and that the distribution of intra-subtype breakpoints mirrored that of inter-subtype recombination along the genome (Figure A.1), but with slightly higher levels of intra-subtype recombination in C2-TM *env* region (Figure A.2). This pattern might be explained by greater levels of protein or RNA mis-folding between diverse subtypes compared to the same subtype, where proteins are more likely to fold and function as required where there has been recombination within the same subtype. However, after an *in silico* validation of the intra-subtype detection, it became clear that intra-subtype detection was not perfectly reliable (Figures A.3 and A.4), and the analysis was removed from the publication. The unreliability of intra-subtype recombination detection was particularly high the subtype A1 experiments, where certain sections were conflated with other

subtype A sub-subtypes and CRFs containing subtype A parents. This further highlights the problems with referenced based taxonomy (discussed in chapter 5), and the struggle to finding references that are truly “recombination-free”. The intra-subtype recombination in Uganda (and the world) therefore remains largely undocumented because it is so hard to detect with increasing pairwise similarity.

Intra-subtype recombination may also artificially inflate branch lengths (Schierup & Hein 2000*a*). The classic ‘starburst’ (Archer et al. 2008) shape of HIV therefore may be partially a consequence of intra-subtype recombination. There seems to be no easy solution to the problem of intra-subtype recombination in phylogenetics, since recombination is a fundamental part of the HIV lifecycle, and absence of recombination is a fundamental assumption of phylogenetics, and many phylogenetic studies must exclude large numbers of recombinant sequences from analyses (e.g. studies using *pol* data in Uganda). In chapter 4, I extracted sections of ’pure’ gene sequence from recombinant genomes to decrease the proportion of excluded data. I also allowed each gene to have a separate phylogenetic history to allow for recombination between genes, whilst strictly screening for inter- and intra-subtype recombination within each gene with GARD and RDP4.

In chapter 5, we suggest the use of dynamic stochastic block models (Olabode et al. 2022) for selecting reference sequences from the database, an approach inspired by network theory which is highly suited to HIV biology and recombination. This method however will require some more testing before it can be more widely used. For example, when I applied the DSBM to Ugandan genome data (not shown), the results of the DSBM were highly dependent on the down sampling process and the quality of the curated alignment. The number of communities also tends to increase with increasing numbers of sequences. Therefore, finding the appropriate diversity and number of sequences to apply the DSBM approach will be the next important consideration.

6.2 Virulence in subtype D

There is a correlation between faster disease progression and subtype D viruses which has been shown in several studies both in Uganda and elsewhere. Chapter 3 presents evidence for a high propensity for subtype D to use the CXCR4 co-receptor using the bioinformatic tool geno2pheno. Bioinformatic co-receptor prediction is imperfect, and there are other parts of the genome outside of V3 which contribute to co-receptor tropism (such as V1/V2). Raymond et al. (2011) suggest some improvements to subtype D specific predictions which might be further explored, but ultimately the most accurate determination of co-receptor tropism is through cell culture. With that caveat aside, I think we can be very certain that subtype D viruses are more

likely to be X4 tropic, because this was also found by others using cell culture (Huang et al. 2007, Kaleebu et al. 2007), and viruses from the same 1986 survey samples sequenced in chapter 3 were also found to be syncytium forming when they were examined in cell culture in 1991 (Oram et al. 1991).

In chapter 3, I compare the V3 amino acid sequence of subtype D to some outgroups such as subtype B and some older subtype D references. We should be wary of interpretation of some of the older samples (like MAL and HXB2) that would have been grown in cell culture before sequencing (as was the practice in 1986), because X4 tropic mutations commonly evolve in cell culture and might not necessarily reflect the virus extracted from the patient. However, there is a conspicuous deletion in codon position 24 in subtype D, which was probably present during the bottleneck of the expansion of subtype D into Uganda. This may have facilitated the increased entropy at positions 23 and 25, and help predispose subtype D to increased X4 tropism. Poon et al. (2012) show that, during the course of infection, there are a multitude of mutational pathways that V3 loops can take to switch from R5 to X4 form. The inherited "sequence space" of V3 may therefore alter the mutational pathways available and predispose subtype D to X4. Furthermore, in chapter 4, I show that the X4 tropic sequences in subtype D *env* are clumped in their phylogenetic distribution, providing some evidence that propensity to be X4-tropic is a heritable trait within subtype D. Interestingly the opposite effect has been observed in subtype C which has lower propensity to be X4 (Pollakis et al. 2004, Ping et al. 1999), possibly because it requires additional mutations to reach X4 tropism (Coetzer et al. 2011).

R5 viruses are said to generally initiate infection (Zhu et al. 1993) while during later stages, X4 viruses may become prominent (Connor et al. 1997). The 'switch' from R5 to X4 is often said to happen in half of AIDS patients (Koot et al. 1992), (although I would argue many of these observations are from the global North and may be a subtype B specific observation). In theory, during the course of infection, the T-cell population undergo changes which encourage the evolution of X4 tropism (reviewed by Regoes & Bonhoeffer 2005). X4 viruses are sometimes regarded as 'dead ends' because of several 'gatekeeping' mechanisms in the mucosa, which should prohibit the transmission of X4 viruses (reviewed by Margolis & Shattock 2006). In chapter 3, however, I report that 66% of all subtype D envelopes from 1986 were X4 tropic, while all the subtype A viruses were R5 tropic. This large difference between subtypes is surprising given that many of the samples (of *both* subtypes) came from late-stage AIDS patients in hospitals. Subsequent studies have provided clear evidence that X4 tropic transmission can occur sexually (Chalmet et al. 2012), and from mother to child (Church et al. 2008) leading to Hedskog et al. (2012) to assert that there is no evidence against a 'random transmission hypothesis' and the detection of R5 viruses in early infection may simply be a sampling bias. More evidence for early infection with X4 variants come from Wambui et al. (2012) who show no association between disease stage and X4-tropism in

a Kenyan cohort. Although an extremely small sample size, it is remarkable that recently infected patients (>500 CD4), were just as likely to have X4 tropic viruses (n=4) as R5 (n=5) (as determined by cell culture). In this thesis I show that subtype D viruses from 1986-2016 have a higher X4 frequency compared with subtype A1, but this data does not address the 'random transmission hypothesis' or the 'gatekeeping hypothesis'. Either way, my conclusions are consistent with one of a few possibilities, 1) X4 receptors are transmitted in the form observed, 2) they are transmitted in an R5 form which has the predisposition to evolve very quickly to X4, or 3) they exist as dual-tropic (R5/X4) viruses .

Viral load is also an important factor in virulence, (Mellors et al. 1996), and Eller et al. (2015) claim it to be the single largest contributor to virulence. However, the relationship between viral load, transmissibility, and disease progression is subject to an evolutionary trade off (Fraser et al. 2007) and the relationships between these factors quickly become very complicated (see outlined in Figure 6.1), not least because viral load increase and co-receptor switches may be correlated as the disease progresses to AIDS. Compared to the large difference in viral loads between HIV-1 and HIV-2 (Hansmann et al. 2005), differences of viral load within Group M are less obvious, and estimates of heritability of viral load vary considerably within and between subtypes (Hodcroft et al. 2014). While single point mutations might arise almost immediately in response to drugs (Coffin 1995), genetic factors determining viral load in Group M are presumed to be multiple loci with small-scale contributions to replication capacity or T-cell-activation activities (Fraser et al. 2014). A more virulent subtype B strain with higher viral load was clearly identified recently in Europe, but still, it remains difficult to attribute virulence effects at the genetic level (Wymant et al. 2022). The evolution of virulence, particularly with respect to the "transmission virulence trade-off" is a fascinating area of theoretical research e.g. (Alizon et al. 2009, Alizon & Michalakis 2015). However, without fully understanding the underlying mechanisms for virulence it is difficult to predict the direction of evolution (Read 1994). This issue has also been of great interest with respect to Covid-19 (Wertheim 2022).

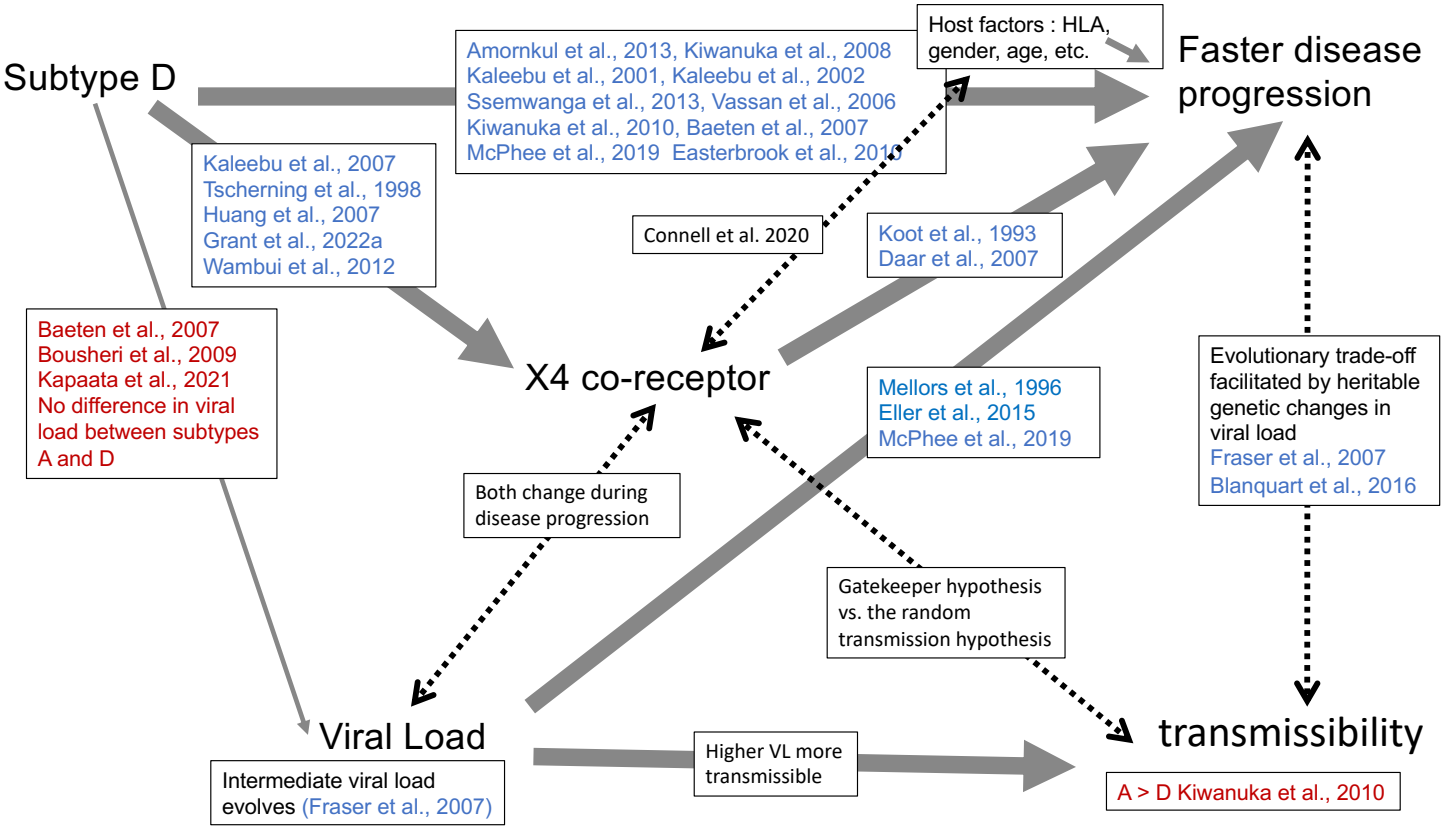

Figure 6.1: The complicated relationship between subtype D, disease progression, viral load, transmission, and co-receptor

Many authors find no significant difference in viral load between subtypes A1 and D in Uganda, at least in the modern era (Baeten et al. 2007, Bousheri et al. 2009, Kapaata et al. 2021). The historic samples had no accompanying viral load measurements or information about disease stage, so we could not compare the set point viral loads of the two subtypes. Even if more information was available, it's likely many of patients were in a late stage of the disease, which would confound the comparison. However, read depth in NGS assemblies are sometimes used as a proxy for viral load (Bonsall et al. 2020, Frampton et al. 2021), which we might be able to use to give an indication about differences in viral load between subtypes in 1986. Figure 3.1 from chapter 3, shows no significant difference in read depth between A1 and D genome assemblies, which suggests there are no differences in viral load between the subtypes, although this must be a tentative conclusion because of the lack of meta data (crucially disease stage) and the possibility of unequal degradation of RNA samples (e.g. different time to freezer). Amornkul et al. (2013) report that subtype D reached higher viral loads faster than other subtypes, but this may be a consequence faster disease progression due to differences in propensity for the X4 co-receptor.

6.3 Subtype specific differences

Phenotypic difference between the Group M subtypes has been of great interest, particularly with respect to vaccine design, drug resistance, virulence or infectivity which may give one subtype an advantage over another (Essex 1999, Geretti 2006, Lessells et al. 2012, Taylor et al. 2008, Buonaguro et al. 2007). The principle that one subtype confers a selective advantage over another is the crux of the argument made by (Turk & Carobene 2015); that combination of phenotypic differences from different subtypes would be significant in creating a new CRF with some selective advantage.

Subtype D virulence seems to be well documented by many authors, both within Ugandan cohorts, and in Western cohorts. Another interesting subtype effect might be that subtype C seems to have lower propensity to form X4 viruses (Bjorndal et al. 1999, Peeters et al. 1999, Tscherning et al. 1998). However, many subtype specific phenotype reports are not replicated, and it may be particularly difficult to clarify this because any virulence or transmissibility factors heritable by the virus are confounded by host genetic factors, ethnicity, geography, and route of transmission (Butler et al. 2007, Peeters 2000). Some studies might be able to find subtype specific motifs of interest (for example Tenzer et al. 2014). However, studies which do not investigate underlying genetic mechanisms of subtype-specific definitions show only correlations. Moreover, many of the studies that compare subtypes are based on single gene classification (e.g. (Price et al. 2019) only sequences pol, or (Palm et al. 2014) only sequence V3), which therefore rely on an assumed linkage between parts of the genome which contain theoretical loci of interest. In this thesis, whole genomes were examined, and two subtypes

were compared within the same human population, reducing some of these confounding factors. We found a deletion in position 24 of the V3 loop of subtype D: a subtype specific motif already associated with a known regulator of virulence. Furthermore, we showed that this subtype D specific V3 loop appears to have been removed by selection and recombination in URF sequences.

6.4 HIV diversity in Uganda

Sampling biases are always important to consider in phylogenetic studies (Hills 1998). Uganda is one of the best sampled countries in East Africa, but even so, the data here represents only a tiny fraction of the Ugandan epidemic. The modern PANGEA data was collected by the UVRI which is based in Entebbe, and many of the cohorts therefore represent the central region of Uganda. Some subtype heterogeneity between Ugandan regions can be seen in the map presented by Poon et al. (2019), which also highlights that many areas are under-sampled. It is important to note that studies from the Rakai Health Sciences Program in South West Uganda appear to sample more subtype D than the central region. Some Rakai studies even find geographical patterns of subtype within the Rakai region (Collinson-Streng et al. 2009).

Although the relative subtype frequencies in Uganda have shifted, the Ugandan epidemic has been overwhelmingly driven by subtypes A1 and D for three decades. The founder events that seeded the epidemics East Africa seem to have occurred early on. Considering the proximity of Uganda to the DRC, it was surprising not to see other DRC subtype exports in the 1986 sample. Similarly, subtype C, whilst common in Kenya is not seen in Uganda in any abundance, even though Uganda sits between the DRC and Kenya. The sub-structure of subtype A1 by East African country (Figure 4.2) suggests that the movement of people between countries was fairly constrained, particularly in the earlier decades.

We can provide no clear explanation as to why subtype D invaded Uganda so successfully and not the surrounding countries like Kenya or Tanzania. The period before 1986 period is an undocumented part of the epidemic and we can only speculate based on the coalescence patterns from BEAST. One explanation might be that subtype D had a higher viral load in the early epidemic (and therefore was more infectious per sexual contact) but we find no evidence to support a higher viral load in subtype D (discussed above). Therefore, the spread of subtype D may just simply be a very strong founder effect, where subtype D established itself very rapidly in Uganda in the 20 years (approx 1965-1987) before the national “education era” began to curb infections. Faria et al. (2019) discuss the introduction of subtype D into Uganda. They suggest that the nearest cities Bwamanda and Kisangani in the DRC are known to have high subtype D incidence (Vidal et al. 2005, 2000) and subtype D may have entered into northern Uganda from these northern DRC cities. Interestingly then, Clade O (the oldest clade in Uganda, Figure 4.3) contains 3 of the 4 genomes from Lacor hospital in Gulu (northern

Uganda), so subtype D may have entered Uganda from the north side. Northern Uganda is not included in the PANGEA dataset however, so it seems possible that more descendants might be found with more sampling in the region. Faria et al. (2019) tentatively suggest that blood transfusions may have played a role in amplifying the founder effects of subtype D in Uganda. However, at that time blood donations were usually from family members and one-time events, and there were no financial incentives for repeat donations (personal communication Dr JW Carswell, December 2021), so while this route of infection may have happened occasionally it is unlikely to have been a large factor.

Appendix A

SCUEAL intra-subtype recombination

Because SCUEAL is a phylogenetic based method with multiple references for each subtype, it is also capable of detecting intra-subtype recombination. During preparation of Chapter 2, I had initially included the intra-subtype breakpoint distributions of subtype A and subtype D genomes in Uganda (Figure A.1 and Figure A.2). I reported that subtype A1 had more intra-subtype breakpoints per genome (mean 3.5) than subtype D (mean 2.5), (Mann-Whitney U-test, $p<0.0001$),

The reviewers asked for a validation of the SCUEAL breakpoint detection. Using real Ugandan genomes, I created in silico A1/D inter-subtype recombinants, and A1 and D intra-subtype recombinants to test the sensitivity of SCUEAL. A random number of breakpoints from 1 to 3, with random breakpoint locations, and a random selection of three “pure” A1 (labelled A1, A2, A3) and three “pure” D subtype sequences (labelled D4, D5, D6) taken from the PANGEA dataset (already designated as ’pure’ by SCUEAL) were used to make the recombinants. Each in silico recombinant was analysed by SCUEAL 100 times.

For the A1/D recombinants, the majority of the SCUEAL replicates found inter-subtype breakpoints as expected (as seen in Chapter 2 supplementary information Fig 2.7) which demonstrated high SCUEAL sensitivity to A1/D breakpoints. However, the SCUEAL intra-subtype detection was not as reliable (Figures A.3 and A.4). SCUEAL *was not* able to accurately describe the intra-subtype breakpoint placements in many cases, and a number of additional inter-subtype breakpoints were found, especially in subtype A1 simulations Figure A.4, where the correct call (pure A1) was made only between 38-92% of the time.

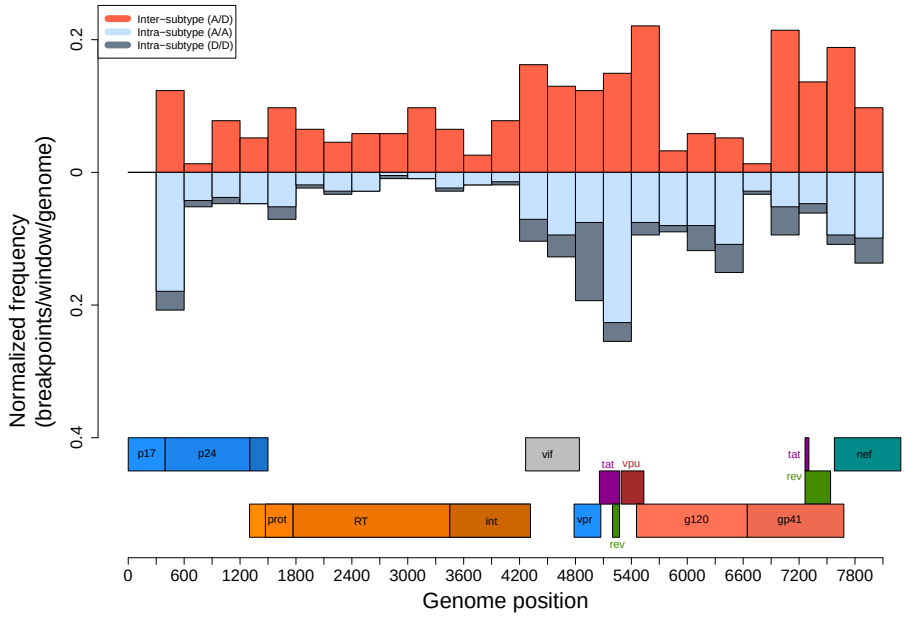

Figure A.1: Intra-subtype breakpoints from the 164 A1/D genomes are shown in red above the axis and inter-subtype breakpoints are stacked in light blue (A1) and dark blue (D)

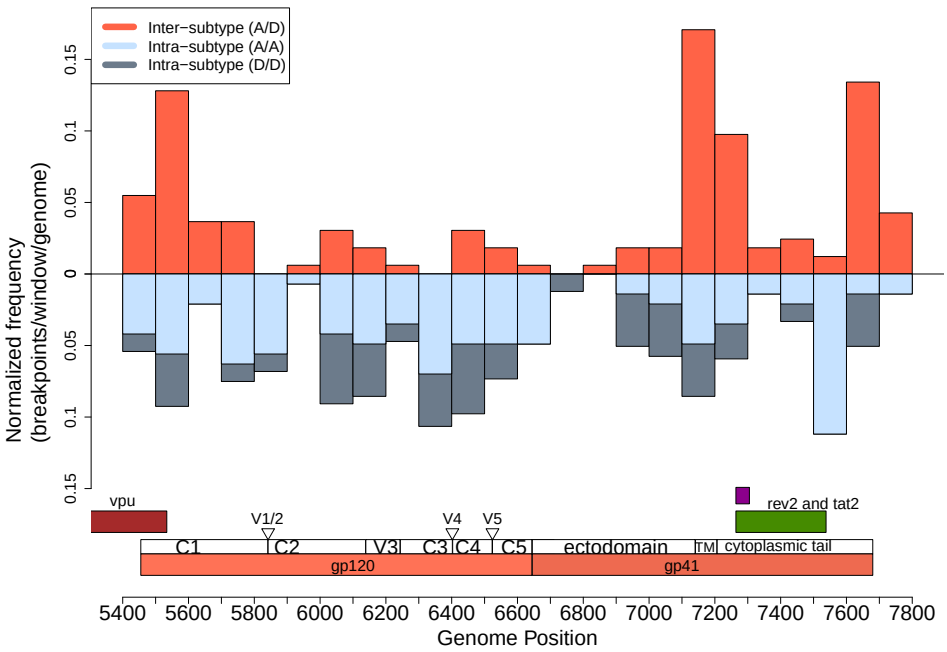

Figure A.2: Breakpoints have been binned into 100 base pair regions and the finer sub-structure of gp120 and gp41 is shown

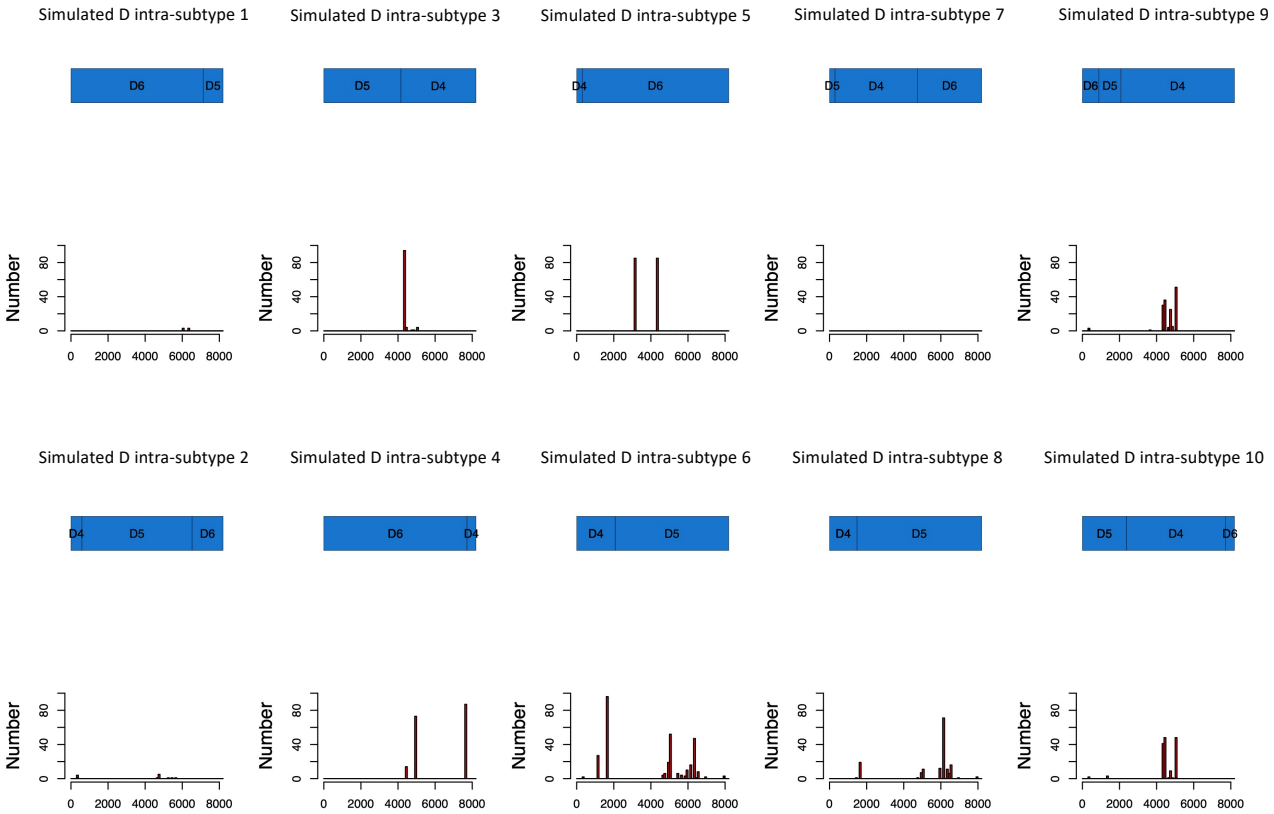

Figure A.3: For each of ten in-silico subtype D intra-subtype recombinants, the number of inter-subtype (black) and intra-subtype recombinant breakpoints (red) for each 100bp-region along the genome, found in 100 replicate SCUEAL assessments. The in-silico recombination pattern shown above

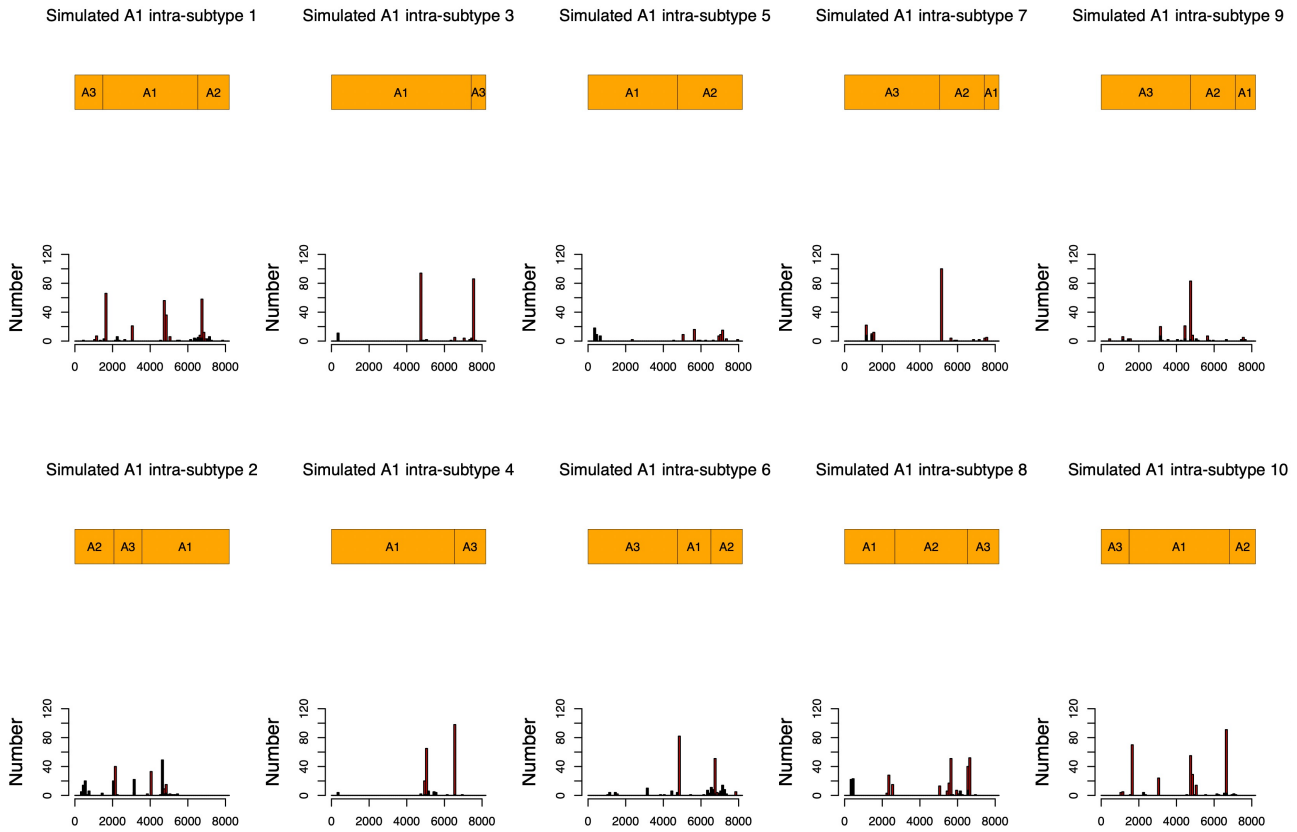

Figure A.4: For each of ten in-silico subtype A1 intra-subtype recombinants, the number of inter-subtype (black) and intra-subtype recombinant breakpoints (red) for each 100bp-region along the genome, found in 100 replicate SCUEAL assessments. The in-silico recombination pattern shown above

Bibliography

Abecasis, A. B., Lemey, P., Vidal, N., de Oliveira, T., Peeters, M., Camacho, R., Shapiro, B., Rambaut, A. & Vandamme, A.-M. (2007), 'Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: Subtype G is a circulating recombinant form', *Journal of Virology* **81**(16), 8543–8551.

Abidi, S. H., Aibekova, L., Davlidova, S., Amangeldiyeva, A., Foley, B. & Ali, S. (2021), 'Origin and evolution of HIV-1 subtype A6', *PLoS ONE* **16**(December), 1–13.

Aiewsakun, P. & Katzourakis, A. (2015), 'Time dependency of foamy virus evolutionary rate estimates', *BMC Evolutionary Biology* **15**(1), 1–15.

Alizon, M., Wain-Hobson, S., Montagnier, L. & Sonigo, P. (1986), 'Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients.', *Cell* **46**(1), 63–74.

Alizon, S. & Fraser, C. (2013), 'Within-host and between-host evolutionary rates across the HIV-1 genome', *Retrovirology* **10**(1), 49.

Alizon, S., Hurford, A., Mideo, N. & Van Baalen, M. (2009), 'Virulence evolution and the trade-off hypothesis: History, current state of affairs and the future', *Journal of Evolutionary Biology* **22**(2), 245–259.

Alizon, S. & Michalakis, Y. (2015), 'Adaptive virulence evolution: The good old fitness-based approach', *Trends in Ecology and Evolution* **30**(5), 248–254.

Amornkul, P. N., Karita, E., Kamali, A., Rida, W. N., Sanders, E. J., Lakhi, S., Price, M. A., Kilembe, W., Cormier, E., Anzala, O., Latka, M. H., Bekker, L. G., Allen, S. A., Gilmour, J. & Fast, P. E. (2013), 'Disease progression by infecting HIV-1 subtype in a seroconverter cohort in sub-Saharan Africa', *AIDS* **27**(17), 2775–2786.

An, M., Han, X., Zhao, B., English, S., Frost, S. D., Zhang, H. & Shang, H. (2020), 'Cross-continental dispersal of major HIV-1 CRF01_AE clusters in China', *Frontiers in Microbiology* **11**.

Anderson, J. P., Rodrigo, A. G., Learn, G. H., Madan, A., Delahunty, C., Coon, M., Girard, M., Osmanov, S., Hood, L. & Mullins, J. I. (2000), 'Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E', *Journal of Virology* **74**(22), 10752–10765.

Andrews, S. (2010), 'FastQC: A Quality Control Tool for High Throughput Sequence Data'. **URL:** *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*

Arau, P. M. M., Martins, J. S. & Oso, N. S. (2019), 'SNAPPy: A snakemake pipeline for scalable HIV-1 subtype by phylogenetic pairing', *Virus Evolution* **5**(2), 1–8.

Archer, J., Pinney, J. W., Fan, J., Simon-Loriere, E., Arts, E. J., Negroni, M. & Robertson, D. L. (2008), 'Identifying the important HIV-1 recombination breakpoints', *PLoS Computational Biology* **4**(9).

Archer, J. & Robertson, D. L. (2007), 'Understanding the diversification of HIV-1 groups M and O', *AIDS* **21**(13), 1693–1700.

Baalwa, J., Wang, S., Parrish, N. F., Decker, J. M., Keele, B. F., Learn, G. H., Yue, L., Ruzagira, E., Ssemwanga, D., Kamali, A., Amornkul, P. N., Price, M. A., Kappes, J. C., Karita, E., Kaleebu, P., Sanders, E., Gilmour, J., Allen, S., Hunter, E., Montefiori, D. C., Haynes, B. F., Cormier, E., Hahn, B. H. & Shaw, G. M. (2013), 'Molecular identification, cloning and characterization of transmitted/founder HIV-1 subtype A, D and A/D infectious molecular clones', *Virology* **436**(1), 33–48.

Baele, G., Lemey, P. & Suchard, M. A. (2016), 'Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty', *Systematic Biology* **65**(2), 250–264.

Baeten, J. M., Chohan, B., Lavreys, L., Chohan, V., McClelland, R. S., Certain, L., Mandaliya, K., Jaoko, W. & Overbaugh, J. (2007), 'HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads', *Journal of Infectious Diseases* **195**(8), 1177–1180.

Bagaya, B. S., Vega, J. F., Tian, M., Nickel, G. C., Li, Y., Krebs, K. C., Arts, E. J. & Gao, Y. (2015), 'Functional bottlenecks for generation of HIV-1 intersubtype Env recombinants', *Retrovirology* **12**(1), 1–17.

Bailes, E., Gao, F., Bibollet-Ruche, F., Courgnaud, V., Peeters, M., Marx, P. A., Hahn, B. H. & Sharp, P. M. (2003), 'Hybrid origin of SIV in chimpanzees', *Science* **300**(5626), 1713.

Baird, H. A., Gao, Y., Galetto, R., Lalonde, M., Anthony, R. M., Giacomoni, V., Abreha, M., Destefano, J. J., Negroni, M. & Arts, E. J. (2006), 'Influence of sequence identity and unique breakpoints on the frequency of intersubtype HIV-1 recombination', *Retrovirology* **3**, 1–17.

Balakrishnan, M., Fay, P. J. & Bambara, R. A. (2001), 'The kissing hairpin sequence promotes recombination within the HIV-1 5' leader region', *Journal of Biological Chemistry* **276**(39), 36482–36492.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. (2012), 'SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing', *Journal of Computational Biology* **19**(5), 455–477.

Barton, N. H. & Charlesworth, B. (1998), 'Why Sex and Recombination?', *Science* **281**(5385), 1986–1990.

Baryarama, F., Bunnell, R. E., Ransom, R. L., Ekwaru, J. P., Kalule, J., Tumuhairwe, E. B. & Mermin, J. H. (2004), 'Using HIV voluntary counseling and testing data for monitoring the Uganda HIV epidemic, 1992-2000', *Journal of Acquired Immune Deficiency Syndromes* **37**(SUPPL. 1), 1180–1186.

Bbosa, N., Kaleebu, P. & Ssemwanga, D. (2019), 'HIV subtype diversity worldwide', *Current opinion in HIV and AIDS* **14**(3), 153–160.

Bbosa, N., Ssemwanga, D., Nsubuga, R. N., Salazar-Gonzalez, J. F., Salazar, M. G., Nanyonjo, M., Kuteesa, M., Seeley, J., Kiwanuka, N., Bagaya, B. S., Yebra, G., Leigh-Brown, A. & Kaleebu, P. (2019), 'Phylogeography of HIV-1 suggests that Ugandan fishing communities are a sink for, not a source of, virus from general populations', *Scientific Reports* **9**(1), 1–8.

Berkley, S. F., Widy-wirski, R., Okware, S. I., Downing, R., Linnan, M. J., White, K. E., Sempala, S., The, S., Diseases, I., Jul, N., Berkley, S. F., Widy-wirski, R., Okware, S. I., Downing, R., Linnan, M. J. & White, K. E. (1989), 'Risk Factors Associated with HIV Infection in Uganda', *Journal of Infectious Diseases* **160**(1), 22–30.

Birdthistle, I., Tanton, C., Tomita, A., de Graaf, K., Schaffnit, S. B., Tanser, F. & Slaymaker, E. (2019), 'Recent levels and trends in HIV incidence rates among adolescent girls and young women in ten high-prevalence African countries: a systematic review and meta-analysis', *The Lancet Global Health* **7**(11), e1521–e1540.

Bjorndal, A., Sonnerborg, A., Tscherning, C., Albert, J. & Fenyo, E. M. (1999), 'Phenotypic characteristics of human immunodeficiency virus type 1 subtype C isolates of ethiopian AIDS patients', *AIDS Research and Human Retroviruses* **15**(7), 647–653.

Blaak, H., Van't Wout, A., Brouwer, M., Hooibrink, B., Hovenkamp, E. & Schuitemaker, H. (2000), 'In vivo HIV-1 infection of CD45RA CD4 T cells is established primarily by syncytium-inducing variants and correlates with the rate of CD4 T cell decline', *Proceedings of the National Academy of Sciences* **97**(3), 1269–1274.

Blanquart, F., Grabowski, M. K., Herbeck, J., Nalugoda, F., Serwadda, D., Eller, M. A., Robb, M. L., Gray, R., Kigozi, G., Laeyendecker, O., Lythgoe, K. A., Nakigozi, G., Quinn, T. C., Reynolds, S. J., Wawer, M. J. & Fraser, C. (2016), 'A transmission-virulence evolutionary trade-off explains attenuation of HIV-1 in Uganda', *eLife* **5**, 1–32.

Blaser, N., Wettstein, C., Estill, J., Vizcaya, L. S., Wandeler, G., Egger, M. & Keiser, O. (2014), 'Impact of viral load and the duration of primary infection on HIV transmission: Systematic review and meta-analysis', *AIDS* **28**(7), 1021–1029.

Bletsa, M., Suchard, M. A., Ji, X., Gryseels, S., Vrancken, B., Baele, G., Worobey, M. & Lemey, P. (2019), 'Divergence dating using mixed effects clock modelling: An application to HIV-1', *Virus Evolution* **5**(2), 1–11.

Bolger, A. M., Lohse, M. & Usadel, B. (2014), 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics* **30**(15), 2114–2120.

Bonhoeffer, S., Holmes, E. C. & Nowak, M. A. (1995), 'Causes of HIV diversity', *Nature* **376**(6536), 125.

Bonsall, D., Golubchik, T., de Cesare, M., Limbada, M., Kosloff, B., MacIntyre-Cockett, G., Hall, M., Wymant, C., Azim Ansari, M., Abeler-Dörner, L., Schaap, A., Brown, A., Barnes, E., Piwowar-Manning, E., Eshleman, S., Wilson, E., Emel, L., Hayes, R., Fidler, S., Ayles, H., Bowden, R. & Fraser, C. (2020), 'A comprehensive genomics solution for HIV surveillance and clinical monitoring in low-income settings', *Journal of Clinical Microbiology* **58**(10).

Bousheri, S., Burke, C., Ssewanyana, I., Harrigan, R., Martin, J., Hunt, P., Bangsberg, D. R. & Cao, H. (2009), 'Infection with different HIV subtypes is associated with CD4 activation-associated dysfunction and apoptosis', *Journal of Acquired Immune Deficiency Syndromes* **52**(5), 548–552.

Bruce, C., Clegg, C., Featherstone, A., Smith, J., Biryahawaho, B., Downing, R. & Oram, J. (1994), 'Presence of Multiple Genetic Subtypes of Human Immunodeficiency Virus Type 1 Proviruses in Uganda', *AIDS Research and Human Retroviruses* **10**(11), 1543–1550.

Bunnik, E. M., Quakkelaar, E. D., van Nuenen, A. C., Boeser-Nunnink, B. & Schuitemaker, H. (2007), 'Increased neutralization sensitivity of recently emerged CXCR4-using human immunodeficiency virus type 1 strains compared to coexisting CCR5-using variants from the same patient', *Journal of Virology* **81**(2), 525–531.

Buonaguro, L., Tornesello, M. L. & Buonaguro, F. M. (2007), 'Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: Pathogenetic and therapeutic implications', *Journal of Virology* **81**(19), 10209–10219.

Burt, A. (2000), 'Sex, recombination and the efficacy of selection - was Weismann right?', *Evolution* **54**(2), 337–351.

Butler, I., Pandrea, I., Marx, P. & Apetrei, C. (2007), 'HIV Genetic Diversity: Biological and Public Health Consequences', *Current HIV Research* **5**(1), 23–45.

Cane, P. (2011), HIV drug resistance testing, *in* 'Methods in Molecular Biology', Springer, chapter 665, pp. 123–132.

Capoferri, A. A., Lamers, S. L., Grabowski, M. K., Rose, R., Wawer, M. J., Serwadda, D., Gray, R. H., Quinn, T. C., Kigozi, G., Kagaayi, J., Laeyendecker, O., Abeler-Dörner, L., Ayles, H., Bonsall, D., Bowden, R., Calvez, V., Cohen, M., Denis, A., Frampton, D., de Oliveira, T., Essex, M., Fidler, S., Fraser, C., Golubchik, T., Hayes, R., Herbeck, J. T., Hoppe, A., Kaleebu, P., Kellam, P., Kityo, C., Leigh-Brown, A., Lingappa, J. R., Novitsky, V., Paton, N., Pillay, D., Rambaut, A., Ratmann, O., Seeley, J., Ssemwanga, D. & Tanser, F. (2020), 'Recombination analysis of near full-length HIV-1 sequences and the identification of a potential new circulating recombinant form from Rakai, Uganda', *AIDS Research and Human Retroviruses* pp. 1–26.

Carr, J. K., Salminen, M. O., Albert, J., Sanders-Buell, E., Gotte, D., Birx, D. L. & McCutchan, F. E. (1998), 'Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants', *Virology* **247**(1), 22–31.

Carr, J. K., Salminen, M. O., Koch, C., Gotte, D., Artenstein, A. W., Hegerich, P. A., St Louis, D., Burke, D. S. & McCutchan, F. E. (1996), 'Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand', *Journal of Virology* **70**(9), 5935–5943.

Carswell, J. W. (1987), 'HIV infection in healthy persons in Uganda', *AIDS (London, England)* **1**(4), 223–7.

Carswell, J. W., Lloyd, G. & Howells, J. (1989), 'Prevalence of HIV-1 in east African lorry drivers', *AIDS* **3**(11), 759–761.

Chalmet, K., Dauwe, K., Foquet, L., Baatz, F., Seguin-Devaux, C., Van Der Gucht, B., Vogelaers, D., Vandekerckhove, L., Plum, J. & Verhofstede, C. (2012), 'Presence of CXCR4-Using HIV-1 in patients with recently diagnosed infection: Correlates and evidence for transmission', *Journal of Infectious Diseases* **205**(2), 174–184.

Chang, L. W., Health, R., Program, S., Program, B. I., Hopkins, J., Grabowski, M. K., Health, R., Program, S., Ssekubugu, R., Health, R., Program, S., Nalugoda, F., Health, R., Program, S., Kigozi, G., Health, R., Program, S., Nantume, B., Health, R., Program, S. & Lessler, J. (2017), 'Heterogeneity of the HIV epidemic: an observational epidemiologic study of agrarian, trading, and fishing communities in Rakai, Uganda', *Lancet HIV* **3**(8), 1–20.

Charif, D. & Lobry, J. R. (2007), SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis, *in* U. Bastolla, M. Porto, H. E. Roman & M. Vendruscolo, eds, 'Structural Approaches to Sequence Evolution: Molecules, Networks, Populations', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 207–232.

Church, J. D., Huang, W., Mwatha, A., Toma, J., Stawiski, E., Donnell, D., Guay, L. A., Mmiro, F., Musoke, P., Jackson, J. B., Parkin, N. & Eshleman, S. H. (2008), 'HIV-1 tropism and survival in vertically infected Ugandan infants', *Journal of Infectious Diseases* **197**(10), 1382–1388.

Cicala, C., Arthos, J. & Fauci, A. S. (2010), 'HIV-1 envelope, integrins and co-receptor use in mucosal transmission of HIV', *Journal of Translational Medicine* **9**(Suppl 1), 1–10.

Coetzer, M., Nedellec, R., Cilliers, T., Meyers, T., Morris, L. & Mosier, D. E. (2011), 'Extreme genetic divergence is required for coreceptor switching in HIV-1 subtype C', *Journal of Acquired Immune Deficiency Syndromes* **56**(1), 9–15.

Coffin, J., Hughes, S. & Varmus, H. (1997), Reverse Transcription of the Viral Genome in vivo, *in* 'Retroviruses', Cold Spring Harbor Laboratory Press, New York.
URL: *https://www.ncbi.nlm.nih.gov/books/NBK19380/*

Coffin, J. M. (1979), 'Structure, replication, and recombination of retrovirus genomes: Some unifying hypotheses', *Jounral of General Virology* **42**, 1–26.

Coffin, J. M. (1995), 'HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy', *Science* **267**(5197), 483–489.

Collinson-Streng, A. N., Redd, A. D., Sewankambo, N. K., Serwadda, D., Rezapour, M., Lamers, S. L., Gray, R. H., Wawer, M. J., Quinn, T. C. & Laeyendecker, O. (2009), 'Geographic HIV type 1 subtype distribution in Rakai district, Uganda', *AIDS Research and Human Retroviruses* **25**(10), 1045–1048.

Connell, B. J., Hermans, L. E., Wensing, A. M., Schellens, I., Schipper, P. J., van Ham, P. M., de Jong, D. T., Otto, S., Mathe, T., Moraba, R., Borghans, J. A., Papathanasopoulos, M. A., Kruize, Z., Venter, F. W., Kootstra, N. A., Tempelman, H., Tesselaar, K. & Nijhuis, M. (2020), 'Immune activation correlates with and predicts CXCR4 co-receptor tropism switch in HIV-1 infection', *Scientific Reports* **10**(1), 1–10.

Connor, R. I., Sheridan, K. E., Ceradini, D., Choe, S. & Landau, N. R. (1997), 'Change in coreceptor use correlates with disease progression in HIV-1 infected individuals', *The Journal of Experimental Medicine* **185**(2), 621–628.

Conroy, S. A., Laeyendecker, O., Redd, A. D., Collinson-Streng, A., Kong, X., Makumbi, F., Lutalo, T., Sewankambo, N., Kiwanuka, N., Gray, R. H., Wawer, M. J., Serwadda, D. & Quinn, T. C. (2010), 'Changes in the Distribution of HIV Type 1 Subtypes D and A in Rakai District, Uganda Between 1994 and 2002', *AIDS Research and Human Retroviruses* **26**(10), 1087–1091.

Cromer, D., Grimm, A. J., Schlub, T. E., Mak, J. & Davenport, M. P. (2016), 'Estimating the in-vivo HIV template switching and recombination rate', *AIDS* **30**(2), 185–192.

Daar, E. S., Kesler, K. L., Petropoulos, C. J., Huang, W., Bates, M., Lail, A. E., Coakley, E. P., Gomperts, E. D. & Donfield, S. M. (2007), 'Baseline HIV type 1 coreceptor tropism predicts disease progression', *Clinical Infectious Diseases* **45**(5), 643–649.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. & Li, H. (2021), 'Twelve years of SAMtools and BCFtools', *GigaScience* **10**(2), 1–4.

Davenport, M., Zaunders, J., Hazebverg, M., Schuitemaker, H. & van Rij, R. (2002), 'Cell turnover and cell tropism in HIV-1 infection', *Trends in Microbiology* **10**(6), 275–278.

de Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E. J., Wensing, A. M., van de Vijver, D. A., Boucher, C. A., Camacho, R. & Vandamme, A. M. (2005), 'An automated genotyping system for analysis of HIV-1 and other microbial sequences', *Bioinformatics* **21**(19), 3797–3800.

de Wolf, F., Hogervorst, E., Goudsmit, J., Fenyö, E. M., Rübsamen-Waigmann, H., Holmes, H., Galvao-Castro, B., Karita, E., Wasi, C., Sempala, S. D., Baan, E., Zorgdrager, F., Lukashov, V., Osmanov, S., Kuiken, C. & Cornelissen, M. (1994), 'Syncytium-inducing and non-syncytium-inducing capacity of human immunodeficiency virus type 1 subtypes other than B: Phenotypic and genotypic characteristics', *AIDS Research and Human Retroviruses* **10**(11), 1387–1400.

Deeks, S. G. & Walker, B. D. (2007), 'Human Immunodeficiency Virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy', *Immunity* **27**(3), 406–416.

Depledge, D. P., Palser, A. L., Watson, S. J., Lai, I. Y. C., Gray, E. R., Grant, P., Kanda, R. K., Leproust, E., Kellam, P. & Breuer, J. (2011), 'Specific capture and whole-genome sequencing of viruses from clinical samples', *PLoS ONE* **6**(11).

DeStefano, J. J., Bambara, R. A. & Fay, P. J. (1994), 'The mechanism of human immunodeficiency virus reverse transcriptase-catalyzed strand transfer from internal regions of heteropolymeric RNA templates', *Journal of Biological Chemistry* **269**(1), 161–168.

Destefano, J. J., Mallaber, L. M., Rodriguez-Rodriguez, L., Fay, P. J. & Bambara1, R. A. (1992), 'Requirements for strand transfer between internal regions of heteropolymer templates by human immunodeficiency virus reverse transcriptase', *Journal of Virology* **66**(11), 6370–6378.

Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. (2006), 'Relaxed phylogenetics and dating with confidence', *PLoS Biology* **4**(5), 699–710.

Duchêne, S., Holmes, E. C. & Ho, S. Y. (2014), 'Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates', *Proceedings of the Royal Society B: Biological Sciences* **281**(1786).

Easterbrook, P. J., Smith, M., Mullen, J., O'Shea, S., Chrystie, I., De Ruiter, A., Tatt, I. D., Geretti, A. M. & Zuckerman, M. (2010), 'Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy', *Journal of the International AIDS Society* **13**(1), 1–9.

Eller, M. A., Opollo, M. S., Liu, M., Redd, A. D., Eller, L. A., Kityo, C., Kayiwa, J., Laeyendecker, O., Wawer, M. J., Milazzo, M., Kiwanuka, N., Gray, R. H., Serwadda, D., Sewankambo, N. K., Quinn, T. C., Michael, N. L., Wabwire-Mangen, F., Sandberg, J. K. & Robb, M. (2015), 'HIV type 1 disease progression to AIDS and death in a rural ugandan cohort is primarily dependent on viral load despite variable subtype and T-cell immune activation levels', *Journal of Infectious Diseases* **211**(10), 1574–1584.

Emerman, M. & Malim, M. H. (1998), 'HIV-1 regulatory/accessory genes: Keys to unraveling viral and host cell biology', *Science* **280**(5371), 1880–1884.

Eshleman, S. H., Gonzales, M. J., Becker-Pergola, G., Cunningham, S. C., Guay, L. a., Jackson, J. B. & Shafer, R. W. (2002), 'Identification of Ugandan HIV type 1 variants with unique patterns of recombination in pol involving subtypes A and D.', *AIDS Research and Human Retroviruses* **18**(7), 507–511.

Essex, M. (1999), 'Human immunodeficiency viruses in the developing world', *Advances in Virus Research* **53**(C), 71–88.

Fabeni, L., Berno, G., Fokam, J., Bertoli, A., Alteri, C., Gori, C., Forbici, F., Takou, D., Vergori, A., Pennica, A., Mastroianni, M., Montella, F., Mussini, C., Andreoni, M., Antinori, A. & Perno, F. (2017), 'Comparative Evaluation of Subtyping Tools for Surveillance of Newly Emerging HIV-1 Strains', *Journal of Clinical Microbiology* **55**(9), 2827–2837.

Fan, J., Negroni, M. & Robertson, D. L. (2007), 'The distribution of HIV-1 recombination breakpoints', *Infection, Genetics and Evolution* **7**(6), 717–723.

Faria, N. R., Nicole Vidal, Sigaloff, E., Tatem, A. J., Vijver, D. A. M. V. D., Rose, R., Wallis, C. L., Ahuka-mundeke, S., Pybus, O. G., Lemey, P. & Dellicour, S. (2019), 'Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa', *PLoS Pathogens* pp. 1–23.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O. G. & Lemey, P. (2014), 'The early spread and epidemic ignition of HIV - 1 in human populations', *Science* **346**(6205), 56–61.

Faria, N. R., Suchard, M. A., Abecasis, A., Sousa, J. D., Ndembi, N., Bonfim, I., Camacho, R. J., Vandamme, A. M. & Lemey, P. (2012), 'Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon', *Infection, Genetics and Evolution* **12**(2), 453–460.

Fellay, J., Ge, D., Shianna, K. V., Colombo, S., Ledergerber, B., Cirulli, E. T., Urban, T. J., Zhang, K., Gumbs, C. E., Smith, J. P., Castagna, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Günthard, H. F., Mallal, S., Mussini, C., Dalmau, J., Martinez-Picado, J., Miro, J. M., Obel, N., Wolinsky, S. M., Martinson, J. J., Detels, R., Margolick, J. B., Jacobson, L. P., Descombes, P., Antonarakis, S. E., Beckmann, J. S., O'Brien, S. J., Letvin, N. L., McMichael, A. J., Haynes, B. F., Carrington, M., Feng, S., Telenti, A. & Goldstein, D. B. (2009), 'Common genetic variation and the control of HIV-1 in humans', *PLoS Genetics* **5**(12).

Felsenstein, J. (1985), 'Phylogenies and the comparative method', *American Naturalist* **125**(1), 1–15.

Feng, Y., He, X., Hsi, J. H., Li, F., Li, X., Wang, Q., Ruan, Y., Xing, H., Lam, T. T. Y., Pybus, O. G., Takebe, Y. & Shao, Y. (2013), 'The rapidly expanding CRF01-AE epidemic in China is driven by multiple lineages of HIV-1 viruses introduced in the 1990s', *AIDS* **27**(11), 1793–1802.

Foley, B. T., Leitner, T., Paraskevis, D. & Peeters, M. (2016), 'Primate immunodeficiency virus classification and nomenclature: Review', *Infection, Genetics and Evolution* **46**, 150–158.

Foster, G. M., Ambrose, J. C., Hué, S., Delpech, V. C., Fearnhill, E., Abecasis, A. B., Leigh Brown, A. J. & Geretti, A. M. (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), 1–10.

Frampton, D., Rampling, T., Cross, A., Bailey, H., Heaney, J., Byott, M., Scott, R., Sconza, R., Price, J., Margaritis, M., Bergstrom, M., Spyer, M. J., Miralhes, P. B., Grant, P., Kirk, S., Valerio, C., Mangera, Z., Prabhahar, T., Moreno-Cuesta, J., Arulkumaran, N., Singer, M., Shin, G. Y., Sanchez, E., Paraskevopoulou, S. M., Pillay, D., McKendry, R. A., Mirfenderesky, M., Houlihan, C. F. & Nastouli, E. (2021), 'Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study', *The Lancet Infectious Diseases* **21**(9), 1246–1256.

Fraser, C., Hollingsworth, T. D., Chapman, R., De Wolf, F. & Hanage, W. P. (2007), 'Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis', *Proceedings of the National Academy of Sciences of the United States of America* **104**(44), 17441–17446.

Fraser, C., Lythgoe, K., Leventhal, G. E., Shirreff, G., Hollingsworth, T. D., Alizon, S. & Bonhoeffer, S. (2014), 'Virulence and pathogenesis of HIV-1 infection: An evolutionary perspective', *Science* **343**(6177).

Fritz, S. A. & Purvis, A. (2010), 'Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits', *Conservation Biology* **24**(4), 1042–1051.

Frost, S. D. W., Wrin, T., Smith, D. M., Kosakovsky Pond, S. L., Liu, Y., Paxinos, E., Chappey, C., Galovich, J., Beauchaine, J., Petropoulos, C. J., Little, S. J. & Richman, D. D. (2005), 'Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection.', *Proceedings of the National Academy of Sciences of the United States of America* **102**(51), 18514–9.

Galetto, R., Moumen, A., Giacomoni, V., Véron, M., Charneau, P. & Negroni, M. (2004), 'The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot in vivo', *Journal of Biological Chemistry* **279**(35), 36625–36632.

Galetto, R. & Negroni, M. (2005), 'Mechanistic features of recombination in HIV', *AIDS Reviews* **7**(2), 92–102.

Gall, A., Ferns, B., Morris, C., Watson, S., Cotten, M., Robinson, M., Berry, N., Pillay, D. & Kellam, P. (2012), 'Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes', *Journal of Clinical Microbiology* **50**(12), 3838–3844.

Gall, A., Morris, C. & Kellam, P. (2014), 'Complete genome sequence of the WHO international standard for HIV-1 RNA determined by deep sequencing', *Genome Announcements* **2**(1), 10–11.

Galli, A., Kearney, M., Nikolaitchik, O. A., Yu, S., Chin, M. P. S., Maldarelli, F., Coffin, J. M., Pathak, V. K. & 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', *Journal of Virology* **84**(15), 7651–7661.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M. & Hahn, B. H. (1999), 'Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes', *Nature* **397**(6718), 436–441.

Gao, F., Robertson, D. L., Morrison, S. G., Hui, H., Craig, S., Decker, J., Fultz, P. N., Girard, M., Shaw, G. M., Hahn, B. H. & Sharp, P. M. (1996), 'The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin', *Journal of Virology* **70**(10), 7013–7029.

Gao, Y., He, S., Tian, W., Li, D., An, M., Zhao, B., Ding, H., Xu, J., Chu, Z., Shang, H. & Han, X. (2021), 'First complete-genome documentation of HIV-1 intersubtype superinfection with transmissions of diverse recombinants over time to five recipients', *PLoS Pathogens* **17**(2), 1–17.

Genuis, S. J. & Genuis, S. K. (2005), 'HIV/AIDS prevention in Uganda: Why has it worked?', *Postgraduate Medical Journal* **81**(960), 615–617.

Geretti, A. M. (2006), 'HIV-1 subtypes: epidemiology and significance for HIV management', *Current Opinion in Infectious Diseases* **19**, 1–7.

Gibson, K. M., Steiner, M. C., Rentia, U., Bendall, M. L., Pérez-Losada, M. & Crandall, K. A. (2020), 'Validation of variant assembly using haphpipe with next-generation sequence data from viruses', *Viruses* **12**(7).

Gifford, R., De Oliveira, T., Rambaut, A., Myers, R. E., Gale, C. V., Dunn, D., Shafer, R., Vandamme, A. M., Kellam, P. & Pillay, D. (2006), 'Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity', *AIDS* **20**(11), 1521–1529.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B. & Suchard, M. A. (2012), 'Improving Bayesian population dynamics inference : A coalescent-based model for multiple loci', **30**(3), 713–724.

Golden, M., Muhire, B. M., Semegni, Y. & Martin, D. P. (2014), 'Patterns of recombination in HIV-1M are influenced by selection disfavouring the survival of recombinants with disrupted genomic RNA and protein structures', *PLoS ONE* **9**(6), 1–8.

González, N., Bermejo, M., Calonge, E., Jolly, C., Arenzana-Seisdedos, F., Pablos, J. L., Sattentau, Q. J. & Alcamí, J. (2010), 'SDF-1/CXCL12 production by mature dendritic cells inhibits the propagation of X4-tropic HIV-1 isolates at the dendritic cell T-cell infectious synapse', *Journal of Virology* **84**(9), 4341–4351.

Grabowski, M. K., Lessler, J., Redd, A. D., Kagaayi, J., Laeyendecker, O., Ndyanabo, A., Nelson, M. I., Cummings, D. A., Bwanika, J. B., Mueller, A. C., Reynolds, S. J., Munshaw, S., Ray, S. C., Lutalo, T., Manucci, J., Tobian, A. A., Chang, L. W., Beyrer, C., Jennings, J. M., Nalugoda, F., Serwadda, D., Wawer, M. J., Quinn, T. C. & Gray, R. H. (2014), 'The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: Evidence from spatial clustering, phylogenetics, and egocentric transmission models', *PLoS Medicine* **11**(3).

Grabowski, M. K., Serwadda, D. M., Gray, R. H., Nakigozi, G., Kigozi, G., Kagaayi, J., Ssekubugu, R., Nalugoda, F., Lessler, J., Lutalo, T., Galiwango, R. M., Makumbi, F., Kong, X., Kabatesi, D., Alamo, S. T., Wiersma, S., Sewankambo, N. K., Tobian, A. A., Laeyendecker, O., Quinn, T. C., Reynolds, S. J., Wawer, M. J. & Chang, L. W. (2017), 'HIV prevention efforts and incidence of HIV in Uganda', *New England Journal of Medicine* **377**(22), 2154–2166.

Grant, H. E., Hodcroft, E. B., Ssemwanga, D., Kitayimbwa, J. M., Yebra, G., Roger, L., Gomez, E., Frampton, D., Gall, A., Kellam, P., Oliveira, T. D., Bbosa, N., Nsubuga, R. N., Kibengo, F., Kwan, T. H., Lycett, S., Kao, R., Robertson, D. L., Ratmann, O., Fraser, C., Pillay, D., Kaleebu, P. & Leigh Brown, A. J. (2020), 'Pervasive and non-random recombination in near full-length HIV genomes from Uganda', *Virus Evolution* **6**(1), 1–12.

Gray, R. R. R. H., Tatem, A. J. A., Lamers, S., Hou, W., Laeyendecker, O., Serwadda, D., Sewankambo, N., Gray, R. R. R. H., Wawer, M., Quinn, T. C. T., Goodenow, M. M. M. & Salemi, M. (2009), 'Spatial phylodynamics of HIV-1 epidemic emergence in east Africa', *AIDS* **23**(14), 1–14.

Green, E. C., Halperin, D. T., Nantulya, V. & Hogle, J. A. (2006), 'Uganda's HIV prevention success: The role of sexual behavior change and the national response', *AIDS and Behavior* **10**(4), 335–346.

Gryseels, S., Watts, T. D., Mpolesha, J. M. K., Larsen, B. B., Lemey, P., Muyembe-Tamfum, J. J., Teuwen, D. E. & Worobey, M. (2020), 'A near full-length HIV-1 genome from 1966 recovered from formalin-fixed paraffin-embedded tissue', *Proceedings of the National Academy of Sciences of the United States of America* **117**(22).

Hahn, B. H., Shaw, G. M., De Cock, K. M. & Sharp, P. M. (2000), 'AIDS as a zoonosis: Scientific and public health implications', *Science* **287**(January), 607–614.

Hansmann, A., Van Der Loeff, M. F., Kaye, S., Awasana, A. A., Sarge-Njie, R., O'Donovan, D., Ariyoshi, K., Alabi, A., Milligan, P. & Whittle, H. C. (2005), 'Baseline plasma viral load and CD4 cell percentage predict survival in HIV-1- and HIV-2-infected women in a community-based cohort in The Gambia', *Journal of Acquired Immune Deficiency Syndromes* **38**(3), 335–341.

Harris, M., Serwadda, D., Sewankambo, N., Kim, B., Kigozi, G., Kiwanuka, N., Phillips, J. B., Wabwire, F., Meehen, M., Lutalo, T., Lane, J. R., Merling, R., Gray, R., Wawer, M., Birx, D. L., Robb, M. L. & McCutchan, F. E. (2002), 'Among 46 near full length HIV type 1 genome sequences from Rakai District, Uganda, subtype D and AD recombinants predominate', *AIDS Research and Human Retroviruses* **18**(17), 1281–1290.

Harris, R. (2007), Improved pairwise alignment of genomic DNA, PhD thesis, The Pennsylvania State University.

Hartigan, M. & Wong, J. (1979), 'Algorithm AS 136 : A K-Means Clustering Algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1), 100–108.

Hayes, R. & Weiss, H. (2018), 'Understanding HIV Epidemic Trends in Africa', *Science* **311**(February), 620–621.

Hedskog, C., Mild, M. & Albert, J. (2012), 'Transmission of the X4 phenotype of HIV-1: Is there evidence against the "random transmission" hypothesis?', *Journal of Infectious Diseases* **205**(2), 163–165.

Hills, D. (1998), 'Taxonomic sampling, phylogenetic accuracy, and investigator bias', *Systematic Biology* **47**(1), 3–9.

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

Hodcroft, E., Hadfield, J. D., Fearnhill, E., Phillips, A., Dunn, D., O'Shea, S., Pillay, D., Leigh Brown, A. J., Celia, A., David, A., Anton, P., Patricia, C., Hannah, C., David, D., Esther, F., Kholoud, P., David, C., Duncan, C., Duncan, C., Simon, C., Valerie, D., Samuel, D., Anna, M. G., Antony, H., Stéphane, H., Steve, K., Paul, K., Linda, L., Andrew, L. B., Tamyo, M., Nicola, M., Chloe, O., Eleni, N., Deenan, P., Andrew, P., Caroline, S., Erasmus, S., Kate, T., Peter, T., Daniel, W., Ian, W., Hongyi, Z., Mark, Z., Jonathan, A., Sris, A., Jane, A., Abdel, B., David, C., Valerie, D., David, D., Martin, F., Brian, G. C., Richard, G., Mark, G., Phillip, H., Teresa, H., Margaret, J., Stephen, K., Clifford, L., Fabiola, M., Mark, N., Chloe, O., Adrian, P., Andrew, P., Deenan, P., Jillian, P., Frank, P., Caroline, S., Memory, S., Achim, S., Anjum, T. & John, W. (2014), 'The contribution of viral genotype to plasma viral set-point in HIV infection', *PLoS Pathogens* **10**(5).

Holmes, E. C. (2001), 'On the origin and evolution of the human immunodeficiency virus (HIV)', *Biological Reviews of the Cambridge Philosophical Society* **76**(2), 239–254.

Hu, W. S. & Hughes, S. H. (2012), 'HIV-1 reverse transcription', *Cold Spring Harbor Perspectives in Medicine* **2**(a006882), 1–22.

Hu, W. S. & Temin, H. M. (1990), 'Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination.', *Proceedings of the National Academy of Sciences of the United States of America* **87**(4), 1556–60.

Huang, W., Eshleman, S. H., Toma, J., Fransen, S., Stawiski, E., Paxinos, E. E., Whitcomb, J. M., Young, A. M., Donnell, D., Mmiro, F., Musoke, P., Guay, L. A., Jackson, J. B., Parkin, N. T. & Petropoulos, C. J. (2007), 'Coreceptor Tropism in Human Immunodeficiency Virus Type 1 Subtype D: High Prevalence of CXCR4 Tropism and Heterogeneous Composition of Viral Populations', *Journal of Virology* **81**(15), 7885–7893.

Hunt, M., Gall, A., Ong, S. H., Brener, J., Ferns, B., Goulder, P., Nastouli, E., Keane, J. A., Kellam, P. & Otto, T. D. (2015), 'IVA: Accurate de novo assembly of RNA virus genomes', *Bioinformatics* **31**(14), 2374–2376.

Hwang, C. K., Svarovskaia, E. S. & Pathak, V. K. (2001), 'Dynamic copy choice: Steady state between murine leukemia virus polymerase and polymerase-dependent RNase H activity determines frequency of in vivo template switching', *Proceedings of the National Academy of Sciences of the United States of America* **98**(21), 12209–12214.

Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D. & Dougherty, J. P. (2000), 'High rate of recombination throughout the human immunodeficiency virus type 1 genome', *Journal of Virology* **74**(3), 1234–1240.

Kaleebu, P., French, N., Mahe, C., Yirrell, D., Watera, C., Lyagoba, F., Nakiyingi, J., Rutebemberwa, A., Morgan, D., Weber, J., Gilks, C. & Whitworth, J. (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', *The Journal of Infectious Diseases* **185**(9), 1244–1250.

Kaleebu, P., Nankya, I. L., Yirrell, D. L., Shafer, L. A., Kyosiimire-Lugemwa, J., Lule, D. B., Morgan, D., Beddows, S., Weber, J. & Whitworth, J. A. (2007), 'Relation between chemokine receptor use, disease stage, and HIV-1 subtypes A and D: Results from a rural Ugandan cohort', *Journal of Acquired Immune Deficiency Syndromes* **45**(1), 28–33.

Kaleebu, P., Ross, A., Morgan, D., Yirrell, D., Oram, J., Rutebemberwa, A., Lyagoba, F., Hamilton, L., Biryahwaho, B. & Whitworth, J. (2001), 'Relationship between HIV-1 Env subtypes A and D and disease progression in a rural Ugandan cohort', *AIDS* **15**(3), 293–299.

Kaleebu, P., Whitworth, J., Hamilton, L., Rutebemberwa, A., Lyagoba, F., Morgan, D., Duffield, M., Biryahwaho, B., Magambo, B. & Oram, J. (2000), 'Molecular epidemiology of HIV type 1 in a rural community in southwest Uganda', *AIDS Research and Human Retroviruses* **16**(5), 393–401.

Kalish, M. L., Robbins, K. E., Pieniazek, D., Schaefer, A., Nzilambi, N., Quinn, T. C., St. Louis, M. E., Youngpairoj, A. S., Phillips, J., Jaffe, H. W. & Folks, T. M. (2004), 'Recombinant viruses and early global HIV-1 epidemic', *Emerging Infectious Diseases* **10**(7), 1227–1234.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A. & Jermiin, L. S. (2017), 'ModelFinder: Fast model selection for accurate phylogenetic estimates', *Nature Methods* **14**(6), 587–589.

Kamali, A., Nsubuga, R. N., Ruzagira, E., Bahemuka, U., Asiki, G., Price, M. A., Newton, R., Kaleebu, P. & Fast, P. (2016), 'Heterogeneity of HIV incidence: A comparative analysis between fishing communities and in a neighbouring rural general population, Uganda, and implications for HIV control', *Sexually Transmitted Infections* **92**(6), 447–454.

Kanki, P. J., Travers, K. U., MBoup, S., Hsieh, C. C., Marlink, R. G., Gueye-NDiaye, A., Siby, T., Thior, I., Hernandez-Avila, M., Sankalé, J. L., NDoye, I. & Essex, M. (1994), 'Slower heterosexual spread of HIV-2 than HIV-1', *The Lancet* **343**(8903), 943–946.

Kapaata, A., Balinda, S. N., Xu, R., Salazar, M. G., Herard, K., Brooks, K., Laban, K., Hare, J., Dilernia, D., Kamali, A., Ruzagira, E., Mukasa, F., Gilmour, J., Salazar-Gonzalez, J. F., Yue, L., Cotten, M., Hunter, E. & Kaleebu, P. (2021), 'HIV-1 gag-pol sequences from Ugandan early infections reveal sequence variants associated with elevated replication capacity', *Viruses* **13**(2), 1–14.

Kapaata, A., Lyagoba, F., Ssemwanga, D., Magambo, B., Nanyonjo, M., Levin, J., Mayanja, B. N., Mugasa, C., Parry, C. M., Kaleebu, P., Nsubuga Mayanja, B., Mugasa, C., Parry, C. M. & Kaleebu, P. (2013), 'HIV-1 subtype distribution trends and evidence of transmission clusters among incident cases in a rural clinical cohort in southwest Uganda, 2004-2010', *AIDS Research and Human Retroviruses* **29**(3), 520–7.

Kasamba, I., Nash, S., Shahmanesh, M., Baisley, K., Todd, J., Kamacooko, O., Mayanja, Y., Seeley, J. & Weiss, H. A. (2019), 'Missed study visits and subsequent HIV incidence among women in a predominantly sex worker cohort attending a dedicated clinic service in Kampala, Uganda', *Journal of Acquired Immune Deficiency Syndromes* **82**(4), 343–354.

Kaslow, R., Carrington, M., Apple, R., Park, L., Munoz, A., Saah, A., Goedert, J., Winkler, C., O'Brien, S., Rinald, C., DetelsI, S., Blattner, W., Phair, J., Erlich, H. & Mann, D. (1996), 'Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection', *Nature medicine* **2**(4), 405–411.

Kassambara, A. & Mundt, F. (2017), 'factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5'.

Katoh, K. & Standley, D. M. (2013), 'MAFFT multiple sequence alignment software version 7: Improvements in performance and usability', *Molecular Biology and Evolution* **30**(4), 772–780.

Kirby, D. (2008), 'Changes in sexual behaviour leading to the decline in the prevalence of HIV in Uganda: Confirmation from multiple sources of evidence', *Sexually Transmitted Infections* **84**(2).

Kirchhoff, F. (2013), HIV Life Cycle: Overview, *in* 'Encyclopedia of AIDS', Springer New York, pp. 1–9.

Kiwanuka, N., Laeyendecker, O., Robb, M., Kigozi, G., Arroyo, M., McCutchan, F., Eller, L., Eller, M., Makumbi, F., Birx, D., Wabwire-Mangen, F., Serwadda, D., Sewankambo, N., Quinn, T., Wawer, M. & Gray, R. (2008), 'Effect of human immunodeficiency virus type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection', *The Journal of Infectious Diseases* **197**(5), 707–713.

Kiwanuka, N., Robb, M., Laeyendecker, O., Kigozi, G., Wabwire-mangen, F., Fredrick, E., Nalugoda, F., Kagaayi, J., Eller, L. A., Serwadda, D., Nelson, K., Reynolds, S. J., Quinn, T. C., Ronald, H., Wawer, M. J. & Whalen, C. C. (2010), 'HIV-1 viral subtype differences in the rate of CD4+ T-Cell decline', *Journal of Acquired Immune Deficiency Syndromes* **54**(2), 180–184.

Kiwanuka, N., Ssetaala, A., Nalutaaya, A., Mpendo, J., Wambuzi, M., Nanvubya, A., Sigirenda, S., Kitandwe, P. K., Nielsen, L. E., Balyegisawa, A., Kaleebu, P., Nalusiba, J. & Sewankambo, N. K. (2014), 'High incidence of HIV-1 infection in a general population of fishing communities around Lake Victoria, Uganda', *PLoS ONE* **9**(5), 1–9.

Kiwuwa-Muyingo, S., Nazziwa, J., Ssemwanga, D., Ilmonen, P., Njai, H., Ndembi, N., Parry, C., Kitandwe, P. K., Gershim, A., Mpendo, J., Neilsen, L., Seeley, J., Seppälä, H., Lyagoba, F., Kamali, A. & Kaleebu, P. (2017), 'HIV-1 transmission networks in high risk fishing communities on the shores of Lake Victoria in Uganda: A phylogenetic and epidemiological approach', *PLoS ONE* **12**(10), 1–23.

Klarmann, G. J., Schauber, C. A. & Preston, B. D. (1993), 'Template-directed pausing of DNA synthesis by HIV-1 reverse transcriptase during polymerization of HIV-1 Sequences in vitro', *The Journal of Biological Chemistry* **268**(13), 9793–9802.

Koning, F. A., Badhan, A., Shaw, S., Fisher, M., Mbisa, J. L. & Cane, P. A. (2013), 'Dynamics of HIV type 1 recombination following superinfection', *AIDS Research and Human Retroviruses* **29**(6), 963–970.

Koot, M., Keet, I. P., Vos, A. H., De Goede, R. E., Roos, M. T. L., Coutinho, R. A., Miedema, F., Schellekens, P. T. A. & Tersmette, M. (1993), 'Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS', *Annals of Internal Medicine* **118**(9), 681–688.

Koot, M., Vos, A. H., Keet, R. P., de Goede, R. E., Dercksen, M. W., Terpstra, F. G., Coutinho, R. A., Miedema, F. & Tersmette, M. (1992), 'HIV-1 biological phenotype in long-term infected individuals evaluated with an MT-2 cocultivation assay.', *AIDS (London, England)* **6**(1), 49–54.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S. & Bhattacharya, T. (2000), 'Timing the ancestor of the HIV-1 pandemic strains', *Science* **288**(5472), 1789–1796.

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. (2006), 'GARD: A genetic algorithm for recombination detection', *Bioinformatics* **22**(24), 3096–3098.

Kosakovsky Pond, S. L., Posada, D., Stawiski, E., Chappey, C., Poon, A. F., Hughes, G., Fearnhill, E., Gravenor, M. B., Brown, A. J. & Frost, S. D. (2009), 'An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1', *PLoS Computational Biology* **5**(11), 1–21.

Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J. & Wertheim, J. O. (2018), 'HIV-TRACE (Transmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens.', *Molecular Biology and Evolution* **35**(7), 1–16.

Kraft, C. S., Basu, D., Hawkins, P. A., Hraber, P. T., Chomba, E., Mulenga, J., Kilembe, W., Khu, N. H., Derdeyn, C. A., Allen, S. A., Manigart, O. & Hunter, E. (2012), 'Timing and source of subtype-C HIV-1 superinfection in the newly infected partner of Zambian couples with disparate viruses', *Retrovirology* **9**, 1–14.

Kreisberg, J. F., Kwa, D., Schramm, B., Trautner, V., Connor, R., Schuitemaker, H., Mullins, J. I., van't Wout, A. B. & Goldsmith, M. A. (2001), 'Cytopathicity of human immunodeficiency virus type 1 primary isolates depends on coreceptor usage and not patient disease status', *Journal of Virology* **75**(18), 8842–8847.

Krueger, F. (2020), 'TrimGalore'.
URL: *github.com/FelixKrueger/TrimGalore*

Kuhanen, J. (2010), 'Sexualised space, sexual networking & the emergence of AIDS in Rakai, Uganda', *Health and Place* **16**(2), 226–235.

Kuhn, J. H. (2021), Virus Taxonomy, *in* 'Virus Taxonomy', Elsevier Inc., pp. 28–37.

Kuhn, J. H., Wolf, Y. I., Krupovic, M., Zhang Yong-Zhen, Maes, P., Dolja Valerian V. & Koonin, E. V. (2019), 'Classify viruses - the gain is worth the pain', *Nature* **566**, 318–320.

Kuritzkes, D. R. (2008), 'HIV-1 subtype as a determinant of disease progression', *Journal of Infectious Diseases* **197**(5), 638–639.

Lackner, A. A., Lederman, M. M. & Rodriguez, B. (2012), 'HIV pathogenesis: The host', *Cold Spring Harbor Perspectives in Medicine* **2**(9), 1–24.

Lamers, S. L., Barbier, A. E., Ratmann, O., Fraser, C., Rose, R., Laeyendecker, O. & Grabowski, M. K. (2016), 'HIV-1 sequence data coverage in Central East Africa from 1959 to 2013', *AIDS Research and Human Retroviruses* **32**(9), 904–908.

Lamers, S. L., Rose, R., Cross, S., Rodriguez, C. W., Redd, A. D., Quinn, T. C., Serwadda, D., Kagaayi, J., Kigozi, G., Galiwango, R., Gray, R. H., Grabowski, M. K. & Laeyendecker, O. (2020), 'HIV-1 subtype distribution and diversity over 18 years in Rakai, Uganda', *AIDS Research and Human Retroviruses* **36**(6), 522–526.

Lee, G. Q., Bangsberg, D. R., Mo, T., Lachowski, C., Brumme, C. J., Zhang, W., Lima, V. D., Boum, Y., Mwebesa, B. B., Muzoora, C., Andia, I., Mbalibulha, Y., Kembabazi, A., Carroll, R., Siedner, M. J., Haberer, J. E., Mocello, A. R., Kigozi, S. H., Hunt, P. W., Martin, J. N. & Harrigan, P. R. (2017), 'Prevalence and clinical impacts of HIV-1 intersubtype recombinants in Uganda revealed by near-full-genome population and deep sequencing approaches', *AIDS* **31**(17), 2345–2354.

Leigh Brown, A., Lobidel, D., Wade, C., Rebus, S., Phillips, A., Brettle, R., France, A., Leen, C., McMenamin, J., McMillan, A., Maw, R., Mulcahy, F., Robertson, J. R., Sankar, K., Scott, G., Wyld, R. & Peutherer, J. (1997), 'The molecular epidemiology of human immunodeficiency virus type 1 in six cities in Britain and Ireland.', *Virology* **235**(1), 166–77.

Lemey, P., Pybus, O. G., Rambaut, A., Drummond, A. J., Robertson, D. L., Roques, P., Worobey, M. & Vandamme, A. M. (2004), 'The molecular population genetics of HIV-1 group O', *Genetics* **167**(3), 1059–1068.

Lengauer, T., Sander, O., Sierra, S., Thielen, A. & Kaiser, R. (2007), 'Bioinformatics prediction of HIV coreceptor usage', *Nature Biotechnology* **25**(12), 1407–1410.

Lessells, R. J., Katzenstein, D. K. & Oliveira, T. D. (2012), 'Are subtype differences important in HIV drug resistance?', *Current Opinion in Virology* **2**(5), 636–643.

Levy, D. N., Aldrovandi, G. M., Kutsch, O. & Shaw, G. M. (2004), 'Dynamics of HIV-1 recombination in its natural target cells', *Proceedings of the National Academy of Sciences* **101**(12), 4204–4209.

Li, G., Piampongsant, S., Faria, R. N., Voet, A., Pineda-Peña, A. C., Khouri, R., Lemey, P., Vandamme, A. M. & Theys, K. (2015), 'An integrated map of HIV genome-wide variation from a population perspective', *Retrovirology* **12**(1).

Li, H. (2013), Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009), 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics* **25**(16), 2078–2079.

Li, X., Liu, H., Liu, L., Feng, Y., Kalish, M. L., Ho, S. Y. & Shao, Y. (2017), 'Tracing the epidemic history of HIV-1 CRF01-AE clusters using near-complete genome sequences', *Scientific Reports* **7**(1), 1–11.

Lihana, R., Ssemwanga, D., Abimiku, A. & Ndembi, N. (2012), 'Update on HIV-1 diversity in Africa: A decade in review', *AIDS Reviews* **14**(2), 83–100.

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E. & Åberg, Y. (2001), 'The web of human sexual contacts', *Nature* **411**(6840), 907–908.

Low-Beer, D. (2002), 'HIV-1 incidence and prevalence trends in Uganda', *The Lancet* **360**, 1788–1792.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. (2019), 'cluster: Cluster Analysis Basics and Extensions. R package version 2.0.8'.

Mansky, L. M. & Temin, H. M. (1995), 'Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase', *Journal of Virology* **69**(8), 5087–5094.

Margolis, L. & Shattock, R. (2006), 'Selective transmission of CCR5-utilizing HIV-1: The 'gatekeeper' problem resolved?', *Nature Reviews Microbiology* **4**(4), 312–317.

Marlink, R., Kanki, P., Thior, I., Travers, K., Eisen, G., Siby, T., Traore, I., Hsieh, C. C., Dia, M. C., Gueye, E. H., Hellinger, J., Guèye-Ndiaye, A., Sankalé, J. L., Ndoye, I., Mboup, S. & Essex, M. (1994), 'Reduced rate of disease development after HIV-2 infection as compared to HIV-1', *Science* **265**(5178), 1587–1590.

Marozsan, A. J., Moore, D. M., Lobritz, M. A., Fraundorf, E., Abraha, A., Reeves, J. D. & Arts, E. J. (2005), 'Differences in the fitness of two diverse wild-type human immunodeficiency virus type 1 isolates are related to the efficiency of cell binding and entry', *Journal of Virology* **79**(11), 7121–7134.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. (2015), 'RDP4: Detection and analysis of recombination patterns in virus genomes', *Virus Evolution* **1**(1), 1–5.

Martin, M. (2011), 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet* **17**(1), 10–12.

Mbulaiteye, S. M., Mahe, C., Whitworth, J. A., Ruberantwari, A., Nakiyingi, J. S., Ojwiya, A. & Kamali, A. (2002), 'Declining HIV-1 incidence and associated prevalence over 10 years in a rural population in south-west Uganda: A cohort study', *Lancet* **360**(9326), 41–46.

Mccutchan, F. E., Hegerich, P. A., Brennan, T. P., Phanuphak, P., Singharaj, P., Jugsudee, A., Berman, P. W., Gray, A. M., Fowler, A. K. & Burke, D. S. (1992), 'Genetic Variants of HIV-1 in Thailand', *AIDS Research and Human Retroviruses* **8**(11), 1887–1895.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. (2010), 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Research* **20**(9), 1297–1303.

McPhee, E., Grabowski, M. K., Gray, R. H., Ndyanabo, A., Ssekasanvu, J., Kigozi, G., Makumbi, F., Serwadda, D., Quinn, T. C. & Laeyendecker, O. (2019), 'Short communication: The interaction of HIV set point viral load and subtype on disease progression', *AIDS Research and Human Retroviruses* **35**(1), 49–51.

McVean, G., Awadalla, P. & Fearnhead, P. (2002), 'A coalescent-based method for detecting and estimating recombination from gene sequences.', *Genetics* **160**(3), 1231–41.

Mehta, S. R., Wertheim, J. O., Brouwer, K. C., Wagner, K. D., Chaillon, A., Strathdee, S., Patterson, T. L., Rangel, M. G., Vargas, M., Murrell, B., Garfein, R., Little, S. J. & Smith, D. M. (2015), 'HIV Transmission Networks in the San Diego-Tijuana Border Region', *EBioMedicine* **2**(10), 1456–1463.

Mellors, J. W., Rinaldo, C. R., Gupta, P., White, R. M., Todd, J. A. & Kingsley, L. A. (1996), 'Prognosis in HIV-1 infection predicted by the quantity of virus in plasma', *Science* **272**(5265), 1167–1170.

Minin, V. N., Bloomquist, E. W. & Suchard, M. A. (2008), 'Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics', *Molecular Biology and Evolution* **25**(7), 1459–1471.

Mir, D., Jung, M., Delatorre, E., Vidal, N., Peeters, M. & Bello, G. (2016), 'Phylodynamics of the major HIV-1 CRF02_AG African lineages and its global dissemination', *Infection, Genetics and Evolution* **46**, 190–199.

Moulard, M., Lortat-Jacob, H., Mondor, I., Roca, G., Wyatt, R., Sodroski, J., Zhao, L. U., Olson, W., Kwong, P. D. & Sattentau, Q. J. (2000), 'Selective interactions of polyanions with basic surfaces on human immunodeficiency virus type 1 gp120', *Journal of Virology* **74**(4), 1948.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. (2015), 'IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular Biology and Evolution* **32**(1), 268–274.

Niama, F. R., Toure-Kane, C., Vidal, N., Obengui, P., Bikandou, B., Ndoundou Nkodia, M. Y., Montavon, C., Diop-Ndiaye, H., Mombouli, J. V., Mokondzimobe, E., Diallo, A. G., Delaporte, E., Parra, H. J., Peeters, M. & Mboup, S. (2006), 'HIV-1 subtypes and recombinants in the Republic of Congo', *Infection, Genetics and Evolution* **6**(5), 337–343.

Olabode, A. S., Avino, M., Ng, G. T., Abu-Sardanah, F., Dick, D. W. & Poon, A. F. Y. (2019), 'Evidence for a recombinant origin of HIV-1 Group M from genomic variation', *Virus Evolution* **5**(1), 1–8.

Olabode, A. S., Ng, G. T., Wade, K. E., Salnikov, M., Grant, H. E., Dick, D. W. & Poon, A. F. Y. (2022), 'Revisiting the recombinant history of HIV-1 group M with dynamic network community detection', *Proceedings of the National Academy of Sciences* **119**(19).

Opio, A., Muyonga, M. & Mulumba, N. (2013), 'HIV Infection in Fishing Communities of Lake Victoria Basin of Uganda - A Cross-Sectional Sero-Behavioral Survey', *PLoS ONE* **8**(8).

Oram, J. D., Downing, R. G., Roff, M., Serwankambo, N., Clegg, J. C., Featherstone, A. S. & Booth, J. C. (1991), 'Sequence analysis of the V3 loop regions of the env genes of Ugandan human immunodeficiency proviruses', *AIDS Research and Human Retroviruses* **7**(7), 605–614.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N. & Pearse, W. (2018), 'caper: Comparative Analyses of Phylogenetics and Evolution in R'.

Ou, C. Y., Takebe, Y., Luo, C. C., Kalish, M., Auwanit, W., Bandea, C., de la Torre, N., Moore, J. L., Schochetman, G., Yamazaki, S., Gayle, H. D., Young, N. L. & Weniger, B. G. (1992), 'Wide distribution of two subtypes of HIV-1 in Thailand', *AIDS Research and Human Retroviruses* **8**(8), 1471–1472.

Palm, A. A., Esbjörnsson, J., Månsson, F., Kvist, A., Isberg, P. E., Biague, A., Da Silva, Z. J., Jansson, M., Norrgren, H. & Medstrand, P. (2014), 'Faster progression to AIDS and AIDS-related death among seroincident individuals infected with recombinant HIV-1 A3/CRF02-AG compared with sub-subtype A3', *Journal of Infectious Diseases* **209**(5), 721–728.

Panganiban, A. T. & Fiore, D. (1988), 'Transfer During Reverse Transcription', *Science* **241**, 1064–1069.

Paradis, E. & Schliep, K. (2019), 'Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R', *Bioinformatics* **35**(3), 526–528.

Parkhurst, J. O. (2002), 'The Ugandan success story? Evidence and claims of HIV-1 prevention', *Lancet* **360**(9326), 78–80.

Pastore, C., Nedellec, R., Ramos, A., Pontow, S., Ratner, L. & Mosier, D. E. (2006), 'Human immunodeficiency virus type 1 coreceptor switching: V1/V2 gain-of-fitness mutations compensate for V3 loss-of-fitness mutations', *Journal of Virology* **80**(2), 750–758.

Patino-Galindo, J. A. & Gonzalez-Candelas, F. (2017), 'The substitution rate of HIV-1 subtypes: a genomic approach', *Virus Evolution* **3**(2), 1–7.

Peeters, M. (2000), 'Recombinant HIV sequences: Their role in the global epidemic', *HIV Sequence Compendium 2000* pp. 39–54.

Peeters, M. & Delaporte, E. (2012), 'Simian retroviruses in African apes', *Clinical Microbiology and Infection* **18**(6), 514–520.

Peeters, M., Vincent, R., Perret, J. L., Lasky, M., Patrel, D., Liegeois, F., Courgnaud, V., Seng, R., Matton, T., Molinier, S. & Delaporte, E. (1999), 'Evidence for differences in MT2 cell tropism according to genetic subtypes of HIV-1: Syncytium-inducing variants seem rare among subtype C HIV- 1 viruses', *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **20**(2), 115–121.

Penn, M. L., Grivel, J.-C., Schramm, B., Goldsmith, M. A. & Margolis, L. (1999), 'CXCR4 utilization is sufficient to trigger CD4 T cell depletion in HIV-1-infected human lymphoid tissue', *Proceedings of the National Academy of Sciences* **96**, 663–668.

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., Fitzjohn, R. G., Alfaro, M. E. & Harmon, L. J. (2014), 'Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees', *Bioinformatics* **30**(15), 2216–2218.

Pillay, D., Herbeck, J., Cohen, M. S., Oliveira, T. d., Fraser, C., Ratmann, O., Leigh-Brown, A. & Kellam, P. (2015), 'PANGEA-HIV: Phylogenetics for generalised epidemics in Africa', *The Lancet Infectious Diseases* **15**(3), 259–261.

Pineda-Peña, A. C., Faria, N. R., Imbrechts, S., Libin, P., Abecasis, A. B., Deforche, K., Gómez-López, A., Camacho, R. J., De Oliveira, T. & Vandamme, A. M. (2013), 'Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools', *Infection, Genetics and Evolution* **19**(100), 337–348.

Ping, L.-H., Nelson, J. A. E., Hoffman, I. F., Schock, J., Lamers, S. L., Goodman, M., Vernazza, P., Kazembe, P., Maida, M., Zimba, D., Goodenow, M. M., Eron, J. J., Fiscus, S. A., Cohen, M. S. & Swanstrom, R. (1999), 'Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: Underrepresentation of X4 variants', *Journal of Virology* **73**(8), 6271–6281.

Pollakis, G., Abebe, A., Kliphuis, A., Chalaby, M. I. M., Bakker, M., Mengistu, Y., Brouwer, M., Goudsmit, J., Schuitemaker, H. & Paxton, W. A. (2004), 'Phenotypic and genotypic comparisons of CCR5- and CXCR4-tropic human immunodeficiency virus type 1 biological clones isolated from subtype C-infected individuals', *Journal of Virology* **78**(6), 2841–2852.

Poon, A. F., Kosakovsky Pond, S. L., Bennett, P., Richman, D. D., Leigh Brown, A. J. & Frost, S. D. (2007), 'Adaptation to human populations is revealed by within-host polymorphisms in HIV-1 and hepatitis C virus', *PLoS Pathogens* **3**(3).

Poon, A. F., Swenson, L. C., Bunnik, E. M., Edo-Matas, D., Schuitemaker, H., van 't Wout, A. B. & Harrigan, P. R. (2012), 'Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data', *PLoS Computational Biology* **8**(11).

Poon, A. F. Y., Avino, M., Kityo, C., Nankya, I., ARTS, E. J., Quiñones-Mateu, M. E., Gibson, R., Ndashimye, E. & Kyeyune, F. (2019), 'First-line HIV treatment failures in non-B subtypes and recombinants: a cross-sectional analysis of multiple populations in Uganda', *AIDS Research and Therapy* **16**(1), 1–10.

Posada, D. & Crandall, K. (2002), 'The effect of recombination on the accuracy of phylogeny estimation', *Journal of Molecular Evolution* **54**(April), 396–402.

Potts, K. E., Kalish, M. L., Bandea, C. I., Orloff, G. M., Louis, M. S., Brown, C., Malanda, N., Kavuka, M., Schochetman, G., Ou, C. Y. & Heyward, W. L. (1993), 'Genetic diversity of human immunodeficiency virus type 1 strains in Kinshasa, Zaire', *AIDS Research and Human Retroviruses* **9**(7), 613–618.

Price, M. A., Rida, W., Kilembe, W., Karita, E., Inambao, M., Ruzagira, E., Kamali, A., Sanders, E. J., Anzala, O., Hunter, E., Allen, S., Edward, V. A., Wall, K. M., Tang, J., Fast, P. E., Kaleebu, P., Lakhi, S., Mutua, G., Bekker, L. G., Abu-Baker, G., Tichacek, A., Chetty, P., Latka, M. H., Maenetje, P., Makkan, H., Kibengo, F., Priddy, F. & Gilmour, J. (2019), 'Control of the HIV-1 load varies by viral subtype in a large cohort of African adults with incident HIV-1 infection', *Journal of Infectious Diseases* **220**(3), 432–441.

R Core Team (2019), 'R: A Language and Environment for Statistical Computing'.
URL: *https://www.r-project.org/*

Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. (2018), 'Posterior summarization in Bayesian phylogenetics using Tracer 1.7', *Systematic Biology* **67**(5), 901–904.

Rambaut, A., Holmes, E. C., OToole, A., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L. & Pybus, O. G. (2020), 'A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology', *Nature Microbiology* **5**(11), 1403–1407.

Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. (2016), 'Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)', *Virus Evolution* **2**(1).

Rambaut, A., Posada, D., Crandall, K. A. & Holmes, E. C. (2004), 'The causes and consequences of HIV evolution', *Nature Reviews Genetics* **5**(1), 52–61.

Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M. & Holmes, E. C. (2001), 'Phylogeny and the origin of HIV-1', *Nature* **410**(6832), 1047–1048.

Ratmann, O., Kagaayi, J., Hall, M., Golubchick, T., Kigozi, G., Xi, X., Wymant, C., Nakigozi, G., Abeler-Dörner, L., Bonsall, D., Gall, A., Hoppe, A., Kellam, P., Bazaale, J., Kalibbala, S., Laeyendecker, O., Lessler, J., Nalugoda, F., Chang, L. W., de Oliveira, T., Pillay, D., Quinn, T. C., Reynolds, S. J., Spencer, S. E., Ssekubugu, R., Serwadda, D., Wawer, M. J., Gray, R. H., Fraser, C., Grabowski, M. K., Ayles, H., Bowden, R., Calvez, V., Cohen, M., Dennis, A., Essex, M., Fidler, S., Frampton, D., Hayes, R., Herbeck, J., Kaleebu, P., Kityo, C., Lingappa, J., Novitsky, V., Paton, N., Rambaut, A., Seeley, J., Ssemwanga, D., Tanser, F., Lutalo, T., Galiwango, R., Makumbi, F., Sewankambo, N. K., Nabukalu, D., Ndyanabo, A., Ssekasanvu, J., Nakawooya, H., Nakukumba, J., Kigozi, G. N., Nantume, B. S., Resty, N., Kambasu, J., Nalugemwa, M., Nakabuye, R., Ssebanobe, L., Nankinga, J., Kayiira, A., Nanfuka, G., Ahimbisibwe, R., Tomusange, S., Galiwango, R. M., Nakalanzi, M., Otobi, J. O., Ankunda, D., Ssembatya, J. L., Ssemanda, J. B., Kato, E., Kairania, R., Kisakye, A., Batte, J., Ludigo, J., Nampijja, A., Watya, S., Nehemia, K., Anyokot, S. M., Mwinike, J., Kibumba, G., Ssebowa, P., Mondo, G., Wasswa, F., Nantongo, A., Kakembo, R., Galiwango, J., Ssemango, G., Redd, A. D., Santelli, J., Kennedy, C. E., Wagman, J. & Tobian, A. (2020), 'Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda', *The Lancet HIV* **7**(3), e173–e183.

Rawson, J. M., Nikolaitchik, O. A., Keele, B. F., Pathak, V. K. & Hu, W. S. (2018), 'NAR Breakthrough Article: Recombination is required for efficient HIV-1 replication and the maintenance of viral genome integrity', *Nucleic Acids Research* **46**(20), 10535–10545.

Raymond, S., Delobel, P., Chaix, M. L., Cazabat, M., Encinas, S., Bruel, P., Sandres-Sauné, K., Marchou, B., Massip, P. & Izopet, J. (2011), 'Genotypic prediction of HIV-1 subtype D tropism', *Retrovirology* **8**(1), 56.

Read, A. F. (1994), 'The evolution of virulence', *Trends in Microbiology* **2**(3), 73–76.

Redd, A. D., Mullis, C. E., Serwadda, D., Kong, X., Martens, C., Ricklefs, S. M., Tobian, A. A., Xiao, C., Grabowski, M. K., Nalugoda, F., Kigozi, G., Laeyendecker, O., Kagaayi, J., Sewankambo, N., Gray, R. H., Porcella, S. F., Wawer, M. J. & Quinn, T. C. (2012), 'The rates of HIV superinfection and primary HIV incidence in a general population in Rakai, Uganda', *Journal of Infectious Diseases* **206**(2), 267–274.

Redd, A. D., Ssemwanga, D., Vandepitte, J., Wendel, S. K., Ndembi, N., Bukenya, J., Nakubulwa, S., Grosskurth, H., Chris, M., Martens, C., Bruno, D., Porcella, S. F., Quinn, T. C. & Kaleebu, P. (2014), 'The rates of HIV-1 superinfection and primary HIV-1 infection are similar in female sex workers in Uganda', *AIDS* **28**(14), 2147–2152.

Regoes, R. R. & Bonhoeffer, S. (2005), 'The HIV coreceptor switch: A population dynamical perspective', *Trends in Microbiology* **13**(6), 269–277.

Revell, L. (2012), 'Phytools: An R package for phylogenetic comparative biology (and other things).', *Methods in Ecology and Evolution* **3**, 217–223.

Rhodes, T., Wargo, H. & Hu, W.-S. (2003), 'High rates of Human Immunodeficiency Virus Type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication', *Journal of Virology* **77**(20), 11193–11200.

Richman, D. D. & Bozzette, S. A. (1994), 'The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression', *Journal of Infectious Diseases* **169**(5), 968–974.

Robertson, D. L., Anderson, J., Bradac, J., Carr, J., Foley, B., Funkhouser, R., Gao, F., Hahn, B., Kuiken, C., Learn, G., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Kalish, M., Salminen, M., Sharp, P., Wolinsky, S. & Korber, B. (2000), 'HIV-1 Nomenclature Proposal', *Science* **288**(5463), 55.

Robertson, D. L., Hahn, B. H. & Sharp, P. M. (1995), 'Recombination in AIDS viruses', *Journal of Molecular Evolution* **40**(3), 249–259.

Sabin, C. A., Devereux, H., Phillips, A. N., Hill, A., Janossy, G., Lee, C. A. & Loveday, C. (2000), 'Course of viral load throughout HIV-1 infection', *JAIDS Journal of Acquired Immune Deficiency Syndromes* **23**(2), 172–177.

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

Schierup, M. H. & Hein, J. (2000*a*), 'Consequences of recombination on traditional phylogenetic analysis', *Genetics* **156**(2), 879–891.

Schierup, M. H. & Hein, J. (2000*b*), 'Recombination and the Molecular Clock', *Molecular Biology and Evolution* **17**, 1578–1579.

Schlub, T. E., Grimm, A. J., Smyth, R. P., Cromer, D., Chopra, A., Mallal, S., Venturi, V., Waugh, C., Mak, J. & Davenport, M. P. (2014), 'Fifteen to twenty percent of HIV substitution mutations are associated with recombination', *Journal of Virology* **88**(7), 3837–3849.

Schuitemaker, H., Van't Wout, A. B. & Lusso, P. (2010), 'Clinical significance of HIV-1 coreceptor usage', *Journal of Translational Medicine* **9**(1), 1–17.

Serwadda, D., Carswell, W., Ayuko, W. O., Wamukota, W., Madda, P. & Downing, R. G. (1986), 'Further experience with kaposi's sarcoma in Uganda', *British Journal of Cancer* **53**(4), 497–500.

Serwadda, D., Sewankambo, N. K., Carswell, J. W., Bayley, A. C., Tedder, R. S., Weiss, R. A., Mugerwa, R. D., Lwegaba, A., Kirya, G. B., Downing, R. G., Clayden, S. A. & Dalgleish, A. G. (1985), 'Slim disease: a new disease in Uganda and its association with HTLV-III infection', *The Lancet* **326**(8460), 849–852.

Serwanga, J., Ssemwanga, D., Muganga, M., Nakiboneka, R., Nakubulwa, S., Kiwuwa-Muyingo, S., Morris, L., Redd, A. D., Quinn, T. C., Kaleebu, P., Mayanja, Y., Hermanus, T., Ilmonen, P., Jonathan, L. & Porcella, S. F. (2018), 'HIV-1 superinfection can occur in the presence of broadly neutralizing antibodies', *Vaccine* **36**(4), 578–586.

Sewankambo, N. K., Gray, R. H., Ahmad, S., Serwadda, D., Wabwire-Mangen, F., Nalugoda, F., Kiwanuka, N., Lutalo, T., Kigozi, G., Li, C., Meehan, M. P., Brahmbatt, H. & Wawer, M. J. (2000), 'Mortality associated with HIV infection in rural Rakai District, Uganda', *AIDS* **14**(15), 2391–2400.

Shapiro, B., Rambaut, A. & Drummond, A. (2006), 'Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences', *Molecular Biology and Evolution* **23**(1), 7–9.

Sharp, P. M., Bailes, E., Chaudhuri, R. R., Rodenburg, C. M., Santiago, M. O. & Hahn, B. H. (2001), 'The origins of acquired immune deficiency syndrome viruses: where and when?', *Philosophical Transactions of the Royal Society B: Biological Sciences* **356**(1410), 867–876.

Sharp, P. M. & Hahn, B. H. (2011), 'Origins of HIV and the AIDS pandemic', *Cold Spring Harbor Perspectives in Medicine* **1**(1), 1–22.

Shaw, G. M. & Hunter, E. (2012), 'HIV transmission', *Cold Spring Harbor Perspectives in Medicine* **2**(11), 1–24.

Shelton, J. D., Halperin, D. T., Nantulya, V., Potts, M., Gayle, H. D. & Holmes, K. K. (2004), 'Partner reduction is crucial for balanced "ABC" approach to HIV prevention', *British Medical Journal* **328**(7444), 891–893.

Sierra, M., Thomson, M. M., Posada, D., Pérez, L., Aragonés, C., González, Z., Pérez, J., Casado, G. & Nájera, R. (2007), 'Identification of 3 phylogenetically related HIV-1 BG intersubtype circulating recombinant forms in Cuba', *Journal of Acquired Immune Deficiency Syndromes* **45**(2), 151–160.

Simmonds, P., Balfe, P., Ludlam, C., Bishop, O. & Leigh Brown, A. J. (1990), 'Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1.', *Journal of virology* **64**(12), 5840–5850.

Simon-Loriere, E., Galetto, R., Hamoudi, M., Archer, J., Lefeuvre, P., Martin, D. P., Robertson, D. L. & Negroni, M. (2009), 'Molecular mechanisms of recombination restriction in the envelope gene of the human immunodeficiency virus', *PLoS Pathogens* **5**(5).

Simon-Loriere, E., Martin, D. P., Weeks, K. M. & Negroni, M. (2010), 'RNA structures facilitate recombination-mediated gene swapping in HIV-1', *Journal of Virology* **84**(24), 12675–12682.

Sing, T., Low, A. J., Beerenwinkel, N., Sander, O., Cheung, P. K., Domingues, F. S., Büch, J., Däumer, M., Kaiser, R., Lengauer, T. & Harrigan, P. R. (2007), 'Predicting HIV coreceptor usage on the basis of genetic and clinical covariates', *Antiviral Therapy* **12**(7), 1097–1106.

Slutkin, G., Okware, S., Naamara, W., Sutherland, D., Flanagan, D., Carael, M., Blas, E., Delay, P. & Tarantola, D. (2006), 'How Uganda reversed its HIV epidemic', *AIDS and Behavior* **10**(4), 351–360.

Smyth, R. P., Davenport, M. P. & Mak, J. (2012), 'The origin of genetic diversity in HIV-1', *Virus Research* **169**(2), 415–429.

Song, H., Giorgi, E. E., Ganusov, V. V., Cai, F., Athreya, G., Yoon, H., Carja, O., Hora, B., Hraber, P., Romero-Severson, E., Jiang, C., Li, X., Wang, S., Li, H., Salazar-Gonzalez, J. F., Salazar, M. G., Goonetilleke, N., Keele, B. F., Montefiori, D. C., Cohen, M. S., Shaw, G. M., Hahn, B. H., McMichael, A. J., Haynes, B. F., Korber, B., Bhattacharya, T. & Gao, F. (2018), 'Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection', *Nature Communications* **9**(1).

Ssemwanga, D., Bbosa, N., Nsubuga, R. N., Ssekagiri, A., Kapaata, A., Nannyonjo, M., Nassolo, F., Karabarinde, A., Mugisha, J., Seeley, J., Yebra, G., Brown, A. L. & Kaleebu, P. (2020), 'The molecular epidemiology and transmission dynamics of HIV type 1 in a general population cohort in Uganda', *Viruses* **12**(11), 1–17.

Ssemwanga, D., Ndembi, N., Lyagoba, F., Bukenya, J., Seeley, J., Vandepitte, J., Grosskurth, H. & Kaleebu, P. (2012), 'HIV type 1 subtype distribution, multiple infections, sexual networks, and partnership histories in female sex workers in Kampala, Uganda', *AIDS Research and Human Retroviruses* **28**(4), 357–365.

Ssemwanga, D., Nsubuga, R. N., Mayanja, B. N., Lyagoba, F., Magambo, B., Yirrell, D., van der Paal, L., Grosskurth, H. & Kaleebu, P. (2013), 'Effect of HIV-1 Subtypes on Disease Progression in Rural Uganda: A Prospective Clinical Cohort Study', *PLoS ONE* **8**(8).

Steel, C. M., Beatson, D., Cuthbert, R. J., Morrison, H., Ludlam, C. A., Peutherer, J. F., Simmonds, P. & Jones, M. (1988), 'HLA haplotype A1 B8 DR3 as a risk factor for HIV-related disease', *The Lancet* **331**(8596), 1185–1188.

Struck, D., Lawyer, G., Ternes, A. M., Schmit, J. C. & Bercoff, D. P. (2014), 'COMET: Adaptive context-based modeling for ultrafast HIV-1 subtype identification', *Nucleic Acids Research* **42**(18), 1–11.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J. & Rambaut, A. (2018), 'Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10', *Virus Evolution* **4**(1), 1–5.

Tamura, K. & Nei, M. (1993), 'Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpazees', *Molecular Biology and Evolution* **10**(3), 512–526.

Taylor, B., Sobieszczyk, M., McCutchan, F. & Hammer, S. M. (2008), 'The challenge of HIV-1 subtype diversity', *The New England Journal of Medicine* **358**(15), 1590–1602.

Taylor, J. E. & Korber, B. T. (2005), 'HIV-1 intra-subtype superinfection rates: Estimates using a structured coalescent with recombination', *Infection, Genetics and Evolution* **5**(1), 85–95.

Telenti, A. & Johnson, W. E. (2012), 'Host genes important to HIV replication and evolution', *Cold Spring Harbor Perspectives in Medicine* **2**(4).

Temin, H. M. (1993), 'Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation.', *Proceedings of the National Academy of Sciences* **90**(15), 6900–6903.

Tenzer, S., Crawford, H., Pymm, P., Gifford, R., Sreenu, V. B., Weimershaus, M., deOliveira, T., Burgevin, A., Gerstoft, J., Akkad, N., Lunn, D., Fugger, L., Bell, J., Schild, H., vanEndert, P. & Iversen, A. K. (2014), 'HIV-1 adaptation to antigen processing results in population-level immune evasion and affects subtype diversification', *Cell Reports* **7**(2), 448–463.

Tibshirani, R., Walther, G. & Hastie, T. (2001), 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society. Series B (Methodological)* **63**(2), 411–423.

Tongo, M., Harkins, G. W., Dorfman, J. R., Billings, E., Tovanabutra, S., de Oliveira, T. & Martin, D. P. (2018), 'Unravelling the complicated evolutionary and dissemination history of HIV-1M subtype A lineages', *Virus Evolution* **4**(1), 1–13.

Tripathi, K., Balagam, R., Vishnoi, N. K. & Dixit, N. M. (2012), 'Stochastic simulations suggest that HIV-1 survives close to its error threshold', *PLoS Computational Biology* **8**(9).

Tscherning, C., Alaeus, A., Fredriksson, R., Bjorndal, A., Deng, H., Littman, D. R., Fenyo, E. M. & Albert, J. (1998), 'Differences in chemokine coreceptor usage between genetic subtypes of HIV-1', *Virology* **241**(2), 181–188.

Turk, G. & Carobene, M. G. (2015), 'Deciphering how HIV-1 intersubtype recombination shapes viral fitness and disease progression', *EBioMedicine* **2**(3), 188–189.

UN Department of Economic and Social Affairs (2019), 'World Population Prospects'.

UNAIDS (1998), 'A measure of success in Uganda: the value of monitoring both HIV prevalence and sexual behaviour. UNAIDS Case Study. Best Practice Collection 8 Geneva: UNAIDS', *Best Practice Collection* **8**.

van't Wout, A. B., Kootstra, N. A., Mulder-Kampinga, G. A., Albrecht-van Lent, N., Scherpbier, H. J., Veenstra, J., Boer, K., Coutinho, R. A., Miedema, F. & Schuitemaker, H. (1994), 'Macrophage-tropic variants initiate human immunodeficiency virus type 1 infection after sexual, parenteral, and vertical transmission', *Journal of Clinical Investigation* **94**, 2060–2067.

Vasan, A., Renjifo, B., Hertzmark, E., Chaplin, B., Msamanga, G., Essex, M., Fawzi, W. & Hunter, D. (2006), 'Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype', *Clinical Infectious Diseases* **42**(6), 843–852.

Vermund, S. H. & Leigh-Brown, A. J. (2012), 'The HIV epidemic: High-income countries', *Cold Spring Harbor Perspectives in Medicine* **2**(5), 1–24.

Vidal, N., Mulanga, C., Bazepeo, S. E., Mwamba, J. K., Tshimpaka, J. W., Kashi, M., Mama, N., Laurent, C., Lepira, F., Delaporte, E. & Peeters, M. (2005), 'Distribution of HIV-1 variants in the Democratic Republic of Congo suggests increase of subtype C in Kinshasa between 1997 and 2002', *Journal of Acquired Immune Deficiency Syndromes* **40**(4), 456–462.

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B. & Delaporte, E. (2000), 'Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa', *Journal of Virology* **74**(22), 10498–10507.

Wambui, V., Kiptoo, M., Kinyua, J., Odera, I., Muge, E., Muiruri, P., Lihana, R., Kinyanjui, P. & Songok, E. M. (2012), 'Predicted HIV-1 coreceptor usage among Kenya patients shows a high tendency for subtype D to be CXCR4 tropic', *AIDS Research and Therapy* **9**, 1–7.

Ward, M. J., Lycett, S. J., Kalish, M. L., Rambaut, A. & Leigh Brown, A. J. (2013), 'Estimating the Rate of Intersubtype Recombination in Early HIV-1 Group M Strains', *Journal of Virology* **87**(4), 1967–1973.

Wei, X., Ghosh, S., Taylor, M., Johnson, A., Emini, E., Deutsch, P., Lifson, J., Bonhoeffer, S., Nowak, M., Hahn, B., Saag, M. & Shaw, G. (1995), 'Viral dynamics in human immunodeficiency virus type 1 infection', *Infectious Diseases in Clinical Practice* **4**(3), 180–181.

Wertheim, J. O. (2022), 'When viruses become more virulent.', *Science (New York, N.Y.)* **375**(6580), 493–494.

Wertheim, J. O., Fourment, M. & Kosakovsky Pond, S. L. (2012), 'Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy', *Molecular Biology and Evolution* **29**(2), 451–456.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York. **URL:** *https://ggplot2.tidyverse.org*

Wilen, C. B., Tilton, J. C. & Doms, R. W. (2012), 'HIV: Cell binding and entry', *Cold Spring Harbor Perspectives in Medicine* **2**(8), 1–14.

Wilkinson, D. A., Operskalski, E. A., Busch, M. P., Mosley, J. W. & Koup, R. A. (1998), 'A 32-bp Deletion within the CCR5 Locus Protects against Transmission of Parenterally Acquired Human Immunodeficiency Virus but Does Not Affect Progression to AIDS-Defining Illness', *Journal of Infectious Diseases* **178**, 1163–1166.

Wilkinson, E., Engelbrecht, S. & De Oliveira, T. (2015), 'History and origin of the HIV-1 subtype C epidemic in South Africa and the greater southern African region', *Scientific Reports* **5**(November).

Woo, J., Robertson, D. L. & Lovell, S. C. (2014), 'Constraints from protein structure and intra-molecular coevolution influence the fitness of HIV-1 recombinants', *Virology* **454-455**(1), 34–39.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J. J., Kabongo, J. M. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T. P. & Wolinsky, S. M. (2008), 'Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960', *Nature* **455**(7213), 661–664.

Worobey, M. & Holmes, E. C. (1999), 'Evolutionary aspects of recombination in RNA viruses', *Journal of General Virology* **80**(May), 2535–2543.

Worobey, M., Watts, T. D., McKay, R. A., Suchard, M. A., Granade, T., Teuwen, D. E., Koblin, B. A., Heneine, W., Lemey, P. & Jaffe, H. W. (2016), '1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America', *Nature* **539**(7627), 98–101.

Wu, W., Blumbergj, B. M., Fay, P. J. & Bambara, R. A. (1995), 'Strand transfer mediated by Human Immunodeficiency Virus reverse transcriptase in vitro is promoted by pausing and results in misincorporation', *The Journal of Biological Chemistry* **270**(1), 325–332.

Wymant, C., Bezemer, D., Blanquart, F., Ferretti, L., Hall, M., Golubchik, T., Bakker, M., Ong, S. H., Zhao, L. & Bonsall, D. (2022), 'A highly virulent variant of HIV-1 circulating in the Netherlands', *Science* **545**(February), 540–545.

Yamaguchi, J., Olivo, A., Laeyendecker, O., Forberg, K., Ndembi, N., Mbanya, D., Kaptue, L., Quinn, T. C., Cloherty, G. A., Rodgers, M. A. & Berg, M. G. (2018), 'Universal target capture of HIV sequences from NGS libraries', *Frontiers in Microbiology* **9**(SEP), 1–13.

Yang, C., Li, M., Shi, Y.-P., Winter, J., Van Eijk, A., Ayisi, J., Hu, D. J., Steketee, R., Nahlen, B. L. & Lal, R. B. (2004), 'Genetic diversity and high proportion of intersubtype recombinants among HIV type 1-infected pregnant women in Kisumu, Western Kenya', *AIDS Research and Human Retroviruses* **20**(5), 565–574.

Yang, O., Daar, E., Jamieson, B., Balamurugan, A., Smith, D., Pitt, J., Petropoulos, C., Richman, D., Little, S. & Leigh-Brown, A. (2004), 'Human immunodeficiency virus type 1 clade B superinfection: Evidence for differential immune containment of distinct clade B strains', *Journal of Virology* **79**(2), 860–868.

Yebra, G., Frampton, D., Gallo Cassarino, T., Raffle, J., Hubb, J., Ferns, R. B., Waters, L., Tong, C. Y. W., Kozlakidis, Z., Hayward, A., Kellam, P., Pillay, D., Clark, D., Nastouli, E., Leigh Brown, A. J., Consortium, o. b. o. t. I., Cassarino, T. G., Raffle, J., Hubb, J., Ferns, R. B., Waters, L., Tong, C. Y. W., Kozlakidis, Z., Hayward, A., Kellam, P., Pillay, D., Clark, D., Nastouli, E. & Leigh Brown, A. J. (2018), 'A high HIV-1 strain variability in London, UK, revealed by full-genome analysis: Results from the ICONIC project', *PLoS ONE* **13**(2), 1–18.

Yebra, G., Hodcroft, E. B., Ragonnet-Cronin, M. L., Pillay, D., Leigh Brown, A. J., Fraser, C., Kellam, P., De Oliveira, T., Dennis, A., Hoppe, A., Kityo, C., Frampton, D., Ssemwanga, D., Tanser, F., Keshani, J., Lingappa, J., Herbeck, J., Wawer, M., Essex, M., Cohen, M. S., Paton, N., Ratmann, O., Kaleebu, P., Hayes, R., Fidler, S., Quinn, T., Novitsky, V., Haywards, A., Nastouli, E., Morris, S., Clark, D. & Kozlakidis, Z. (2016), 'Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic', *Scientific Reports* **6**(December), 1–6.

Yebra, G., Ragonnet-Cronin, M., Ssemwanga, D., Parry, C. M., Logue, C. H., Cane, P. A., Kaleebu, P. & Leigh Brown, A. J. (2015), 'Analysis of the history and spread of HIV-1 in Uganda using phylodynamics', *Journal of General Virology* **96**(7), 1890–1898.

Yirrell, D. L., Kaleebu, P., Morgan, D., Watera, C., Magambo, B., Lyagoba, F. & Whitworth, J. (2002), 'Inter-and intra-genic intersubtype HIV-1 recombination in rural and semi-urban Uganda', *AIDS* **16**(2), 279–286.

Yirrell, D. L., Pickering, H., Palmarini, G., Hamilton, L., Rutebemberwa, A. & Biryahwaho, B. (1997), 'Molecular epidemiological analysis of HIV in sexual networks in Uganda', *AIDS* **12**, 285–290.

Yu, G., Lam, T. T. Y., Zhu, H. & Guan, Y. (2018), 'Two methods for mapping and visualizing associated data on phylogeny using GGTree', *Molecular Biology and Evolution* **35**(12), 3041–3043.

Zhang, M., Foley, B., Schultz, A. K., Macke, J. P., Bulla, I., Stanke, M., Morgenstern, B., Korber, B. & Leitner, T. (2010), 'The role of recombination in the emergence of a complex and dynamic HIV epidemic', *Retrovirology* **7**, 1–15.

Zhu, T., Korber, B., Nahmias, A. J., Hooper, E., Sharp, P. M. & Ho, D. D. (1998), 'An African HIV-1 sequence from 1959 and implications for the origin of the epidemic', *Nature* **391**(February), 594–597.

Zhu, T., Mo, H., Wang, N., Nam, D. S., Cao, Y., Koup, R. A. & Ho, D. D. (1993), 'Genotypic and phenotypic characterization of HIV-1 in patients with primary infection', *Science* **261**(5125), 1179–1181.

Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B. D. & Dougherty, J. P. (2002), 'Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots.', *Journal of Virology* **76**(22), 11273–82.