LUND UNIVERSITY

Inferring transmission dynamics from HIV-1 genealogies

Makau Nduva, George

2022

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Makau Nduva, G. (2022). *Inferring transmission dynamics from HIV-1 genealogies*. [Doctoral Thesis (compilation), Department of Translational Medicine]. Lund University, Faculty of Medicine.

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

LUND UNIVERSITY

Inferring transmission dynamics from HIV-1 genealogies

Makau Nduva, George

2022

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Makau Nduva, G. (2022). *Inferring transmission dynamics from HIV-1 genealogies*. Lund University, Faculty of Medicine.

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Inferring transmission dynamics from HIV-1 genealogies

GEORGE MAKAU NDUVA
TRANSLATIONAL MEDICINE | FACULTY OF MEDICINE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2022 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Translational Medicine

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2022:30
ISBN 978-91-8021-191-8
ISSN 1652-8220

Inferring transmission dynamics from HIV-1 genealogies

George Makau Nduva

LUND UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Medicine, Lund University, Sweden.
To be defended at Lund University at 09:00 AM, 17th March 2022.

Faculty opponent
Professor Christophe Fraser, University of Oxford, UK.

Supervisor
Associate Professor Joakim Esbjörnsson, Lund University, Sweden.

Co-supervisors
Professor Eduard Sanders, University of Oxford, UK.
Dr Amin Hassan, KEMRI-Wellcome Trust Research Program, Kenya.
Professor Patrik Medstrand, Lund University, Sweden.
Dr Kamini Gounder, University of KwaZulu-Natal, South Africa.

Organization
LUND UNIVERSITY

George Makau Nduva

Doctoral dissertation

2022-03-17

Lund University

Title: Inferring transmission dynamics using HIV-1 genealogies

Abstract: With a national prevalence of 4.9% in the adult population, the HIV-1 epidemic in Kenya is the fifth largest in the world. HIV-1 prevalence is more than three-fold higher among HIV key populations – including men who have sex with men (MSM), people who inject drugs (PWID), and female sex workers (FSW) than in the general heterosexual (HET) population. However, the contribution of different risk groups in the propagation of the epidemic has not been investigated. Also, the epidemic is geographically heterogeneous (65% of all new infections occur in nine out of the 47 counties in Kenya). Yet, the rates of HIV-1 transmission between geographic regions have not been described. Also, data are lacking on how levels and trends of HIV drug resistance (HIVDR) in Kenya compare among individuals of different risk groups, with or without antiretroviral therapy (ART) exposure.

The primary objective was to phylogenetically describe virus transmission within and between risk groups (MSM, PWID, FSW, and HET) and geographic locations as well as to determine levels of HIV-1 drug resistance over time within and between risk groups in Kenya. A secondary objective was to phylogenetically characterise transmission patterns in a paediatric HIV-1 outbreak in Pakistan.

In the first objective, clustering patterns in Kenya indicated that HIV-1 transmission between risk groups was rare – where most HIV-1 transmission occurs within-risk groups. In addition, when HIV-1 (infrequently) jumped between risk populations, virus jumps from HET to key populations were more common than vice-versa. There was significant West-to-East transmission (i.e. from high-to-low HIV-1 prevalence regions) in the mixed epidemic. Interestingly, Coast and Nairobi provinces were suggested to be important geographic hubs of HIV-1 dissemination in the MSM-specific HIV-1 sub-epidemic. HIVDR analysis revealed that overall pre-treatment HIVDR increased from 6.9% in 1986-2005 to 24.2% in 2016-2020. This was associated with increased non-nucleoside reverse transcriptase inhibitors (NNRTI) resistance in all risk groups. DRMs of any kind were found in treatment naïve HET (13.9%, 95% CI: 12.7-15.2), FSW (19.9%, 95% CI: 15.8-24.6), MSM (15.1%, 95% CI: 9.7-21.9), PWID (31.0%, 95% CI: 19.5-44.5), and children (41.3%, 95% CI: 30.1-53.3). PWID and children were more likely than HET to have DRMs (aOR, 3.5, 95% CI: 1.7-5.4, $p<0.001$, and aOR, 3.0, 95% CI: 1.8-4.8, $p<0.001$), respectively. No integrase strand transfer inhibitors (INSTI) drug resistance was detected. Hence, current INSTI-based ART regimens may remain effective in controlling HIV-1 in Kenya. In the secondary objective, clustering patterns in the Pakistani paediatric HIV-1 outbreak revealed multiple introductions of HIV-1 and no phylogenetic HIV-1 mixing between children and key populations.

Findings may be relevant for HIV-1 control in Kenya and Pakistan.

Key words: HIV-1; key populations; molecular epidemiology; transmission; HIV-1 drug resistance; phylogenetics

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language: English

ISSN and key title: 1652-8220 Publication 30; Inferring transmission dynamics from HIV-1 genealogies

ISBN: 978-91-8021-191-8

Recipient's notes

Number of pages 84

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature G. Makau

Date 2022-02-10

Inferring transmission dynamics from HIV-1 genealogies

George Makau Nduva

LUND
UNIVERSITY

Cover and back photos by Eger Tiitus

Copyright pp 1-84 George Makau Nduva
Paper 1 © by the Authors (Submitted manuscript)
Paper 2 © by the Authors (Manuscript)
Paper 3 © by the Authors (Published and open Access)
Paper 4 © by the Authors (Published and open Access)
Paper 5 © by the Authors (Published and open Access)
Paper 6 © by the Authors (Published and open Access)

Faculty of Medicine
Department of Translational Medicine, Lund University.

ISBN 978-91-8021-191-8
ISSN 1652-8220
Lund University, Faculty of Medicine Doctoral Dissertation Series 2022:30

Printed in Sweden by Media-Tryck, Lund University
Lund 2022

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se
MADE IN SWEDEN

To Louisa W.

Table of Contents

List of papers **8**

Abbreviations **10**

Aims of this doctoral dissertation **12**

Summary **13**

HIV virology and epidemiology **14**

The epidemic 14
HIV origin and discovery 17
HIV classification 18
HIV genetic variants and global distribution 18
HIV virion 20
HIV-1 genome 21
HIV-1 replication cycle 22
HIV-1 recombination 24
HIV transmission 25
HIV pathogenesis and immunology 26
HIV treatment 27
HIV drug resistance 30
HIV-1 evolutionary dynamics within and between hosts 32

Phylogenetic inference **34**

Introduction 34
Sequence alignment 35
Selecting nucleotide substitution models 36
Methods for constructing phylogenetic trees 37
Distance-based methods 37
Character-based methods 38
Molecular evolution and the molecular clock 41
Population dynamics and the coalescent 41

Materials and methods **44**
Systematic literature review 44
Study population and partial HIV-1 *pol* sequence datasets 44
Partial HIV-1 *pol* sequencing 46
HIV-1 subtype analysis 46
Cluster analysis 47
Bayesian phylodynamic inference 48
Bayesian phylogeographic inference 48
Characterisation of HIV-1 drug resistance 49
Statistical analysis 50
Ethical considerations 50
Data availability 50
Main findings and discussion **51**
Molecular description of the HIV-1 epidemic in Kenya 51
HIV-1 subtype A (sub-subtype A1) dominated the epidemic, with increasing proportions (2004-2019) 51
HIV-1 transmission was compartmentalised by risk groups 53
Evidence of geographic HIV-1 mixing in Kenya 55
The effective population size had stabilised at a high level 56
High levels of pre-treatment and acquired drug resistance in different risk groups 58
Conclusions from a review of studies on HIV-1 phylogenetic linkages between populations in sSA 59
Molecular characterization of a paediatric HIV-1 outbreak in Larkana, Pakistan 59
Limitations, potential solutions, and future perspectives **61**
Acknowledgements **63**
References **65**

List of papers

Original articles included in the dissertation

1. **George M. Nduva**, Frederick Otieno, Joshua Kimani, Elizabeth Wahome, Lyle R. McKinnon, Francois Cholette, Maxwell Majiwa, Moses Masika, Gaudensia Mutua, Omu Anzala, Susan M. Graham, Larry Gelmon, Matt A. Price, Adrian Smith, Robert C. Bailey, Guy Baele, Philippe Lemey, Amin S. Hassan[#], Eduard J. Sanders[#], and Joakim Esbjörnsson[#]. Quantifying rates of HIV-1 flow between risk groups and geographic locations in Kenya: a country-wide phylogenetic study, (*Submitted Manuscript*).
   [#]Equal contribution as senior authors.

2. **George M. Nduva,** Frederick Otieno, Joshua Kimani, Elizabeth Wahome, Lyle R. McKinnon, Francois Cholette, Maxwell Majiwa, Moses Masika, Gaudensia Mutua, Omu Anzala, Susan M. Graham, Larry Gelmon, Matt A. Price, Adrian Smith, Robert C. Bailey, Eduard J. Sanders[#], Joakim Esbjörnsson[#], and Amin S. Hassan[#]. Pre-treatment and acquired HIV-1 drug resistance within and between risk groups in Kenya, (*Manuscript unpublished*).
   [#]Equal contribution as senior authors.

3. **George M. Nduva**, Frederick Otieno, Joshua Kimani, Lyle R. McKinnon, Francois Cholette, Susan M. Graham, Larry Gelmon, Matt A. Price, Adrian Smith, Robert C. Bailey, Amin S. Hassan[#], Eduard J. Sanders[#], and Joakim Esbjörnsson[#]. Phylogeographic reconstruction of HIV-1 spread among MSM populations in Kenya (2006-2019), *Frontiers in Microbiology*, 2022.
   [#]Equal contribution as senior authors.

4. **George M. Nduva**, Amin S. Hassan, Jamirah Nazziwa, Susan M. Graham, Joakim Esbjörnsson[#], and Eduard J. Sanders[#]. HIV-1 Transmission Patterns Within and Between Risk Groups in Coastal Kenya. *Scientific Reports*, 2020.
   [#]Equal contribution as senior authors.

5. **Syed Hani Abidi**[*], **George Makau Nduva**[*], **Dilsha Siddiqui**[*], Wardah Rafaqat, Syed Faisal Mahmood, Amna Rehana Siddiqui, Apsara Ali Nathwani, Aneeta Hotwani, Sharaf Ali Shah, Sikander Memon, Saqib Ali Sheikh, Palwasha Khan, Joakim Esbjörnsson[#], Rashida Abbas Ferrand[#] and Fatima Mir[#]. Phylogenetic and drug-resistance analysis of HIV-1 sequences from an extensive paediatric HIV-1 outbreak in Larkana, Pakistan. *Frontiers in Microbiology*, 2021.
   [*]Equal contribution as first authors. [#]Equal contribution as senior authors.

6. **George M. Nduva*, Jamirah Nazziwa***, Amin S. Hassan, Eduard J. Sanders, and Joakim Esbjörnsson. The Role of Phylogenetics in Discerning HIV-1 Mixing among Vulnerable Populations and Geographic Regions in Sub-Saharan Africa: A Systematic Review. *Viruses*, 2021.
   *Equal contribution as first authors. #Equal contribution as senior authors.

The following paper is not included in this doctoral dissertation but is of relevance to the field:

1. Amin S. Hassan, Joakim Esbjörnsson, Elizabeth Wahome, Alexander Thiong'o, **George M. Nduva**, Matt. A. Price, and Eduard J. Sanders. HIV-1 subtype diversity, transmission networks and transmitted drug resistance amongst acute and early infected MSM populations from Coastal Kenya. *PLoS One*, 2018.

Published papers were printed with permission from the copyright holders.

Abbreviations

Aa	Amino acid
ADR	Acquired drug resistance
AEHI	Acute or early HIV infection
AHI	Acute HIV infection
AIDS	Acquired Immune Deficiency Syndrome
aLRT-SH	Approximate Shimodaira-Hasegawa-like Likelihood Ratio Test
ART	Antiretroviral therapy
ARV	Antiretroviral
AZT	Zidovudine
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
BF	Bayes factors
cART	Combined antiretroviral therapy
CCR5	C-C chemokine receptor type 5
CD4	Cluster of differentiation 4
cDNA	Complementary deoxyribonucleic acid
CRF	Circulating recombinant form
CTMC	Continuous-time Markov chain
CXCR4	C-X-C Motif Chemokine Receptor 4
DNA	Deoxyribonucleic acid
env	Envelope gene
Env	Envelope protein
ESS	Effective sample size
FSW	Female sex workers
gag	Group-specific antigen gene
gp	Glycoprotein
GTR	General time-reversible substitution model
HAART	Highly active antiretroviral therapy
HET	Heterosexual
HIV-1	Human immunodeficiency virus type 1
HIV-2	Human immunodeficiency virus type 2
HIVDR	HIV drug resistance
HPM	Hierarchical phylogenetic model
indel	Insertion or deletion
INSTI	Integrase strand-transfer inhibitor
kb	Kilobases
LASER ART	Long-acting slow effective release antiretroviral therapy
LTR	Long terminal repeat
MCC	Maximum clade credibility
M-group	Major group
MHC	Major histocompatibility complex

ML	Maximum-likelihood
MSM	Men who have sex with men
nef	Negative regulatory factor gene
N group	Non-M, non-O group
NNRTI	Non-nucleoside reverse transcriptase inhibitor
NRTI	Nucleoside reverse transcriptase inhibitor
nt	Nucleotide bases
O group	Outlier group
ORF	Open reading frames
PANGEA	Phylogenetics for generalised epidemics in Africa
PCR	Polymerase chain reaction
PDR	Pre-treatment drug resistance
PIC	Pre-integration complex
PMTCT	Prevention of mother-to-child transmission
pol	Polymerase gene
PR	Protease
PrEP	Pre-exposure prophylaxis
PWID	People who inject drugs
rev	Regulator of expression of virion proteins gene
RNA	Ribonucleic acid
RNAse H	Ribonuclease H
RT	Reverse transcriptase
SIV	Simian immunodeficiency virus
SNP	Single nucleotide polymorphism
ss	Single stranded
STI	Sexually transmitted infection
tar	Transactivation-responsive RNA
tat	Trans-activator of transcription gene
TDR	Transmitted drug resistance
$tRNA_{Lys}$	Transfer ribonucleic acid molecule that binds to lysine
UNAIDS	The Joint United Nations Programme on HIV/AIDS
URF	Unique recombinant form
vif	Virus infectivity factor gene
VL	Virus load
vpr	Virus protein R gene
vpu	Virus protein U gene
WHO	World Health Organization

Aims of this doctoral dissertation

The **overall aim** of this doctoral dissertation was to investigate the molecular epidemiology and reconstruct the evolutionary history of HIV-1 in local and nationwide contexts using phylogenetic approaches. Most of the studies of the thesis relate to the HIV-1 epidemic in Kenya (with a focus on HIV-1 transmission within and between risk groups and geographic regions). In addition, one study focused on characterising an HIV-1 outbreak in 2019 in Larkana, Pakistan that predominantly affected children.

Specific objectives:

Paper I: To investigate the HIV-1 molecular epidemiology in various risk groups and geographic locations in Kenya.

Paper II: To determine levels of HIV-1 drug resistance over time, within and between risk groups in Kenya.

Paper III: To phylodynamically quantify rates of HIV-1 transmission among MSM in various geographic regions in Kenya.

Paper IV: To study patterns of HIV-1 transmission within and between risk groups in Coastal Kenya.

Paper V: To investigate the patterns and source of transmission in the largest HIV-1 outbreak that has been described among children in Pakistan.

Paper VI: To review the role of phylogenetics in discerning HIV-1 mixing between geographies and risk group populations in sub-Saharan Africa (sSA).

Summary

With a national prevalence of 4.9% in the adult population, the HIV-1 epidemic in Kenya is the fifth largest in the world. HIV-1 prevalence is more than three-fold higher among HIV key populations – including men who have sex with men (MSM), people who inject drugs (PWID), and female sex workers (FSW) than in the general heterosexual (HET) population. However, the contribution of different risk groups in the propagation of the epidemic has not been investigated. Also, the epidemic is geographically heterogeneous (65% of all new infections occur in nine out of the 47 counties in Kenya). Yet, the rates of HIV-1 transmission between geographic regions have not been described. Also, data are lacking on how levels and trends of HIV drug resistance (HIVDR) in Kenya compare among individuals of different risk groups, with or without antiretroviral therapy (ART) exposure.

The primary objective was to phylogenetically describe virus transmission within and between risk groups (MSM, PWID, FSW, and HET) and geographic locations as well as to determine levels of HIV-1 drug resistance over time within and between risk groups in Kenya. A secondary objective was to phylogenetically characterise transmission patterns in a paediatric HIV-1 outbreak in Pakistan.

In the first objective, clustering patterns in Kenya indicated that HIV-1 transmission between risk groups was rare – where most HIV-1 transmission occurs within-risk groups. In addition, when HIV-1 (infrequently) jumped between risk populations, virus jumps from HET to key populations were more common than vice-versa. There was significant West-to-East transmission (i.e. from high-to-low HIV-1 prevalence regions) in the mixed epidemic. Interestingly, Coast and Nairobi provinces were suggested to be important geographic hubs of HIV-1 dissemination in the MSM-specific HIV-1 sub-epidemic. HIVDR analysis revealed that overall pre-treatment HIVDR increased from 6.9% in 1986-2005 to 24.2% in 2016-2020. This was associated with increased non-nucleoside reverse transcriptase inhibitors (NNRTI) resistance in all risk groups. DRMs of any kind were found in treatment naïve HET (13.9%, 95% CI: 12.7-15.2), FSW (19.9%, 95% CI: 15.8-24.6), MSM (15.1%, 95% CI: 9.7-21.9), PWID (31.0%, 95% CI: 19.5-44.5), and children (41.3%, 95% CI: 30.1-53.3). PWID and children were more likely than HET to have DRMs (aOR, 3.5, 95% CI: 1.7-5.4, $p<0.001$, and aOR, 3.0, 95% CI: 1.8-4.8, $p<0.001$), respectively. No integrase strand transfer inhibitors (INSTI) drug resistance was detected. Hence, current INSTI-based ART regimens may remain effective in controlling HIV-1 in Kenya. In the secondary objective, clustering patterns in the Pakistani paediatric HIV-1 outbreak revealed multiple introductions of HIV-1 and no phylogenetic HIV-1 mixing between children and key populations.

Findings may be relevant for HIV-1 control in Kenya and Pakistan.

HIV virology and epidemiology

The epidemic

The human immunodeficiency virus infection and acquired immunodeficiency syndrome (HIV/AIDS) is one of the world's most serious health and development challenges[1]. According to the Joint United Nations Programme on HIV and AIDS (UNAIDS), approximately 76 million people have become infected with HIV since the start of the epidemic – and an estimated 38 million people have died from AIDS-related illnesses[2]. Majority (67%) of the individuals infected with HIV are in sub-Saharan Africa (sSA, **Fig. 1**).

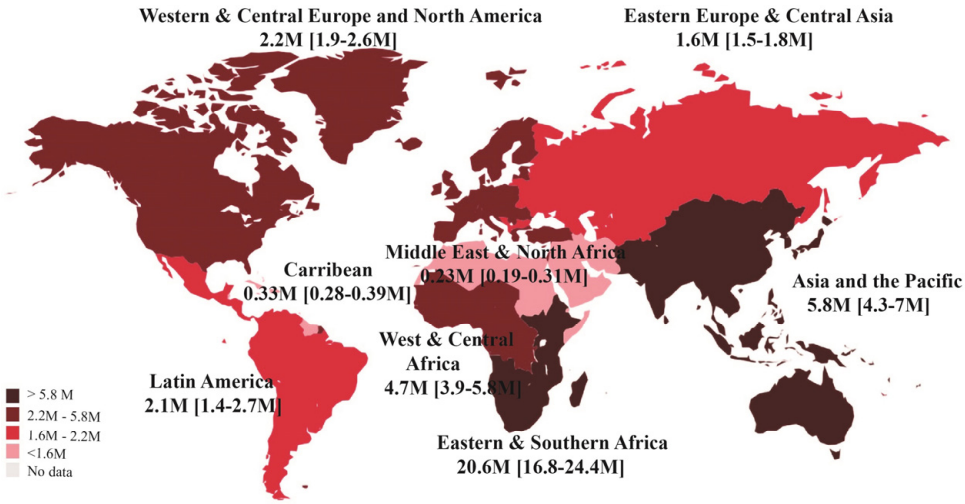


Figure 1. The estimated number of people living with HIV-1 by the end of 2021. Data used in the map were UNAIDS epidemiological estimates available publicly at https://aidsinfo.unaids.org/ as of 10th February 2022.

The number of people with HIV globally has increased consistently over the last three decades (**Fig. 2a**). The reason for this is the high numbers of infections particularly in low-income and middle-income countries (LMICs), and the global increase in treatment rates which have reduced mortality rates[1,2]. Although the number of new infections per year has reduced by 31%, from 2.1 million new infections in 2010 to 1.5 million in 2020 (**Fig. 2b**), the rate of decrease has not been enough to achieve the UNAIDS goal to reduce the global HIV incidence rate to less than half a million new infections by 2020[1]. Likewise, two of three UNAIDS 90–90–90 goals were not achieved as only 84% of people with HIV globally knew their HIV status in 2020, 87% of whom were accessing treatment, and 90% of whom were virally supressed[2,3].

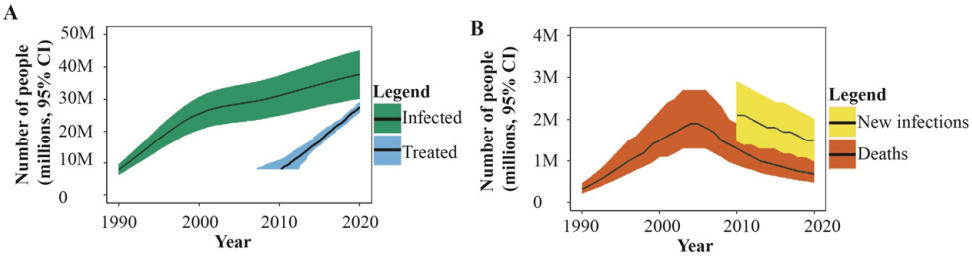


Figure 2. The global HIV epidemic transition metrics 1990-2020. The estimated number of people (in millions) over time are shown as continuous black lines. Graphs represent **(a)** time trends in the number of people living with HIV versus the number of people on treatment with antiretroviral drugs, and **(b)** the number of new infections versus AIDS-related deaths. Data in the plots were UNAIDS epidemiological estimates available publicly at https://aidsinfo.unaids.org/ as of 10th February 2022.

The global human immunodeficiency virus type 1 (HIV-1) distribution also varies between different populations with different HIV-1 transmission risks. The HIV-1 epidemic outside sSA is concentrated to HIV-1 key populations (>90%) – defined by UNAIDs as sex workers (FSW), gay men and other men who have sex with men (MSM), people who inject drugs (PWID), transgender people, prisoners – and their sexual partners[2,4-6]. In contrast, the epidemic in sSA is dominated (>60%) by heterosexual transmission, but with pockets of concentrated sub-epidemics involving key populations[7-9]. In 2020, 65% of the new HIV-1 infections globally, 93% of the new HIV-1 infections outside of sub-Saharan Africa, and 39% of the new HIV-1 infections in sSA were in HIV-1 key populations[1].

The HIV-1 epidemic in Kenya is highlighted in detail in this dissertation. According to UNAIDS epidemiological estimates, the epidemic (which emerged in the mid-1980s) grew exponentially during the mid-1990s and has stabilised during recent years (**Fig. 3**). Mortality due to AIDS-related illnesses has decreased over time, rates of treatment with antiretroviral therapy (ART) have increased exponentially during recent years, and the number of new infections per year has stabilised. In 2020, approximately 1.3 million adults were living with HIV-1 in Kenya, and 36000 became newly infected – making it the fifth largest HIV-1 epidemic in the world[2]. In addition, 96% of people living with HIV-1 knew their HIV-1 status, 86% of people with HIV-1 who knew their HIV-1 status were accessing antiretroviral therapy, and 81% of people on treatment were virally suppressed[10].

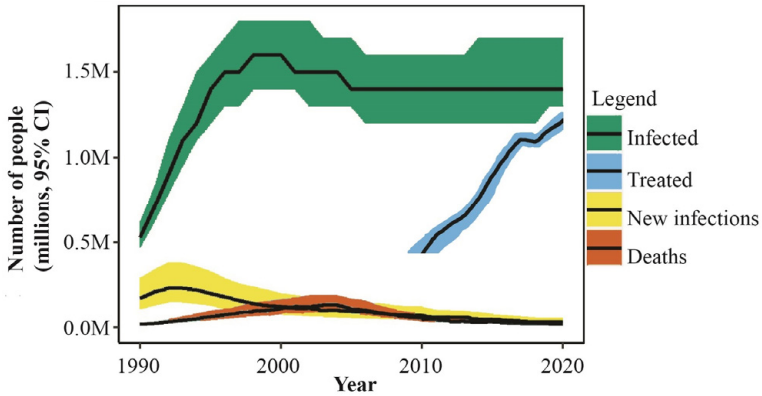


Figure 3. Kenyan HIV-1 epidemic transition metrics between the years 1990 and 2020. The estimated number of people (in millions) over time are shown as continuous black lines. The shaded area represents the 95% confidence interval – coloured Green: infected individuals; Blue: individuals on treatment with antiretroviral drugs; Yellow: new infections; and Orange: deaths. Data are UNAIDS 2021 epidemiological estimates available publicly at https://aidsinfo.unaids.org/ as of 10th February 2022.

The Kenyan Ministry of Health reports high a HIV-1 prevalence among key populations (about 29% among FSW, 18% among MSM, and 18% among PWID, compared to about 5% in the heterosexual (HET) epidemic) – and it has been hypothesised that the ongoing epidemic may be fuelled by HIV-1 key populations (mainly MSM, FSW, and PWID)[11-16]. A national survey on modes of HIV-1 transmission conducted in Kenya suggested that of all new infections, 14% occur in FSW and their clients; 15% in MSM; and 4% in PWID[12,17]. This adds up to that approximately 33% of all new infections in the country would be attributed to key populations. There are reports that MSM sexual networks also involve heterosexual females, which could then be involved in bridging between risk populations[18]. However, the hypothesis of infection flow from high-to-low burden populations is rarely confirmed in practice, in part because it is difficult to measure empirically, and in part, because data from key populations in Africa are extremely scarce[9].

It is well documented that the HIV-1 epidemic in Kenya has extensive geographic heterogeneity, where HIV-1 prevalence ranges from less than 1% in the North Eastern province to more than 20% around the shores of Lake Victoria in the Western regions of the country[11]. The geographic and risk group distributions of HIV-1 in Kenya are summarised in **Fig. 4**.

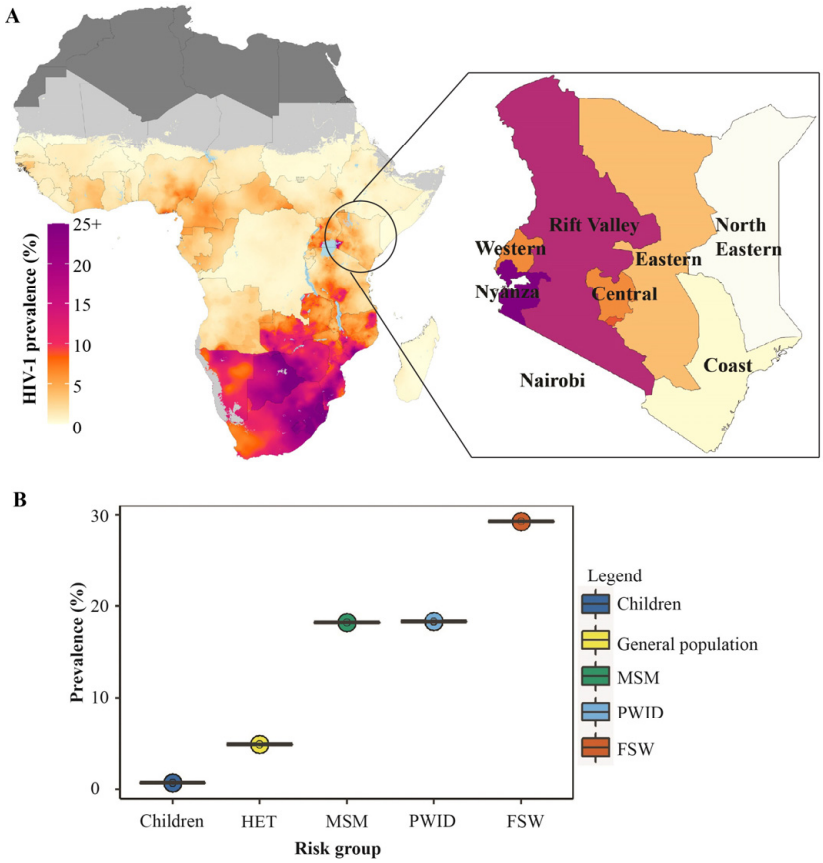


Figure 4. The geographic and risk group distribution of HIV-1 in Kenya in 2019. (A) A map of Africa highlighting HIV-1 prevalence in sub-Saharan Africa, and in Kenya, with the provinces in Kenya coloured per HIV-1 prevalence as inset. The Nyanza province had the highest HIV-1 burden compared to other regions in the country in 2019. **(B)** In Kenya, key populations including men having sex with men (MSM), people who inject drugs (PWID), and female sex workers (FSW) had a three-fold higher prevalence than the relatively lower-at-risk general population (HET) and adolescents/perinatally infected children. The map of Africa was adapted with permission from Laura Dwyer-Lindgren *et al.*, 2019[19], and the HIV-1 estimates in Kenya were adopted from the KENPHIA 2019 report, the 2018 national HIV-1 estimates report, and the 2016 County estimates report[11,13,15].

HIV origin and discovery

Acquired Immune Deficiency Syndrome (AIDS) was first identified as a novel disease in 1981 – following increasing numbers of men who have sex with men (MSM) in New York and California, USA, being infected with *Pneumocystis pneumonia* and Kaposi's sarcoma (a relatively rare and aggressive form of cancer)[20-22]. Subsequently, HIV-1 was acknowledged in causing the emergent infections[23-25]. The identification of HIV-1 was soon followed by isolation of a morphologically similar but antigenically distinct virus causing AIDS in patients in West Africa[26]. The new virus was termed the human immunodeficiency virus type 2 (HIV-2) and was found to be related to both HIV-1 and a simian immunodeficiency virus (SIV) causing immunodeficiency and AIDS-related symptoms in captive macaque[27,28].

Since then, several other simian AIDS viruses have been discovered that are essentially non-pathogenic in their natural non-human primate hosts – but that cluster together with the human and simian AIDS viruses in a phylogenetic lineage within the radiation of lentiviruses (refer to "HIV classification below").

It is now well-established that AIDS likely emerged due to cross-species infections with lentiviruses from different primate species, and that both HIV-1 and HIV-2 originated from zoonotic transmissions of viruses infecting primates in Africa[29,30]. HIV-1 is the pandemic type and likely originated from Cameroon and the Democratic Republic of Congo during the 1920s, from where it expanded globally[31]. On the other hand, HIV-2 likely originated from Guinea-Bissau (together with Cape Verde, Côte d'Ivoire and Senegal) in the 1940s, from where the virus spread locally in West Africa, and to countries sharing socio-historical ties (such as Portugal and France)[32,33]. HIV-2 is thus largely endemic in West Africa, although it is increasingly being replaced by HIV-1 in recent years also in this region[34-39].

HIV classification

HIV belongs to the Lentivirus genus of the family Retroviridae. Lentiviruses are host species-specific, exogenous, and non-oncogenic retroviruses that infect humans, primates, domestic cats, and a variety of livestock (sheep, cattle, horses)[40]. They contain the reverse transcriptase enzyme that converts ribonucleic acid (RNA) into deoxy-ribonucleic acid (DNA) before becoming integrated into the genome of the host (refer to HIV replication cycle). Lentiviruses are tropic for cells of the macrophage lineage *in vivo*[41] – and infections associated with lentiviruses typically result in characteristically long-duration illnesses with a long asymptomatic stage indicative of latent proviruses[42,43].

HIV genetic variants and global distribution

Phylogenetic analysis comparing the genetic relationship between HIV-1, HIV-2 and SIVs indicate that HIV-1 is closely related to SIV from chimpanzees (SIV_{CPZ}) and gorillas (SIV_{GOR}) while HIV-2 is closely related to SIV from sooty mangabeys (SIV_{SM}, **Fig. 5**). Thus, HIV-1 and HIV-2 comprise various groups, each having resulted from a separate cross-species transmission to humans of SIV from non-human primates[44]. HIV-1 group M (Main group) subtypes (A-D, F, H, J, and K) and 118 associated inter-subtype circulating recombinants forms (CRFs), are spread globally[45,46]. The global co-circulation of multiple HIV-1 subtypes – and co-infection or super-infection of individuals with multiple subtypes has resulted in the emergence of recombinant forms (refer to "HIV-1 recombination"). HIV recombinants are classified into either circulating recombinant forms (CRFs) – defined as characteristic full-length or near full-length HIV sequences that are found in three epidemiologically unlinked individuals, or unique recombinant forms (URFs) when these criteria are not met[44,47]. Other HIV-1 groups such as group O (Outlier), group N (Not-M, Not-O) and P (pending the identification of further

human cases) are rare and have been identified mostly in West- Africa where they co-circulate with the HIV-2 epidemic group (A and B) and non-epidemic groups (C-G)[32,47].

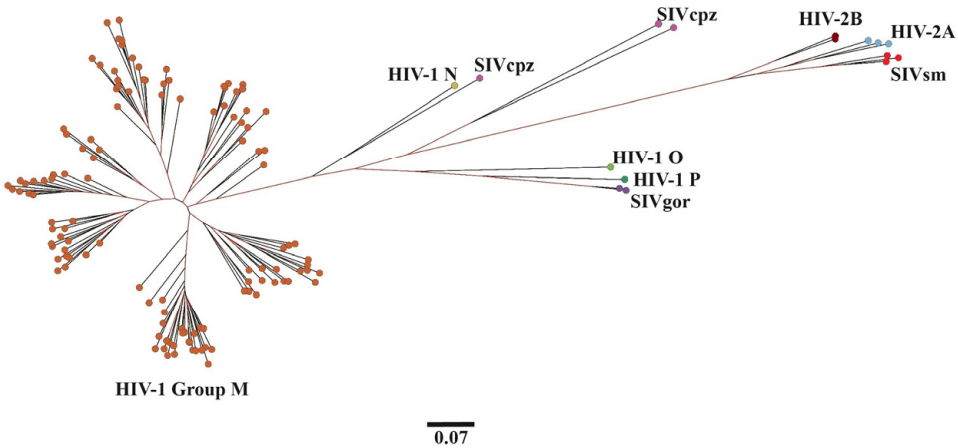


Figure 5. The phylogenetic relatedness between human and simian immunodeficiency viruses: A molecular phylogeny depicting the genetic relationship between HIV-1, HIV-2 and SIV. Branch tips are coloured to represent the different strains. Orange: HIV-1 group M (Main group subtypes and circulating recombinant forms, CRFs – responsible for the global epidemic); Light green: HIV-1 group O (Outlier group); Yellow: HIV-1 group N (Not-M, Not-O group); Dark green: HIV-1 group P (pending the identification of further human cases); Purple: SIV_{GOR} (from gorillas); Magenta: SIV_{CPZ} (from chimpanzees); Sky Blue: HIV-2A; Maroon: HIV-2B; and Red: SIV_{SM} (from sooty mangabeys). The tree is rooted at mid-point and the scale bar indicates 7% nucleotide sequence substitutions per site.

A recent global survey of the distribution of HIV-1 subtypes and CRFs showed that HIV-1 subtype C represents the largest proportion of all HIV-1 infections worldwide (46%), followed by different recombinants (CRFs & URFs, 23%), subtype B (12%), subtype A (10%), subtype G (5%), and subtype D (3%) and the subtypes F, H, J, K (all <1%, respectively)[45]. HIV-1 distribution also reveals distinct distribution patterns, where some strains are dominant in distinct geographic regions (**Fig. 6**). For instance, whereas all HIV-1 subtypes, many CRFs and URFs co-circulate in Central Africa, the epidemic in Southern Africa, Ethiopia, and South Asia (India) is dominated by subtype C. On the other hand, subtype B is dominant in the Middle East and North Africa, Western and Central Europe, North America, Caribbean, Latin America, and Oceania. Interestingly, some regions are dominated by CRFs. For instance, the CRF01_AE dominates in Southeast Asia and East Asia, and the CRF02_AG is dominant in West Africa where it co-circulates with subtype G.

Figure 6. Global and regional distributions of major HIV-1 subtypes, circulating recombinant forms (CRFs), and unique recombinant forms (URFs): Subtypes C, A1, D are dominant in Africa, whereas subtype B infections dominate North America, Latin America, and Europe. CRFs are more common in Africa and Asia. The map of the world was obtained from Wikimedia Commons, the free media repository, and modified to depict HIV-1 subtype diversity worldwide.

HIV virion

The mature HIV-1 virion is spherical (approximately 100 nm in diameter) and is enveloped by an outer lipid membrane[48]. The HIV-1 virion has trimers of gp120 (surface proteins) on the surface and these are anchored to the lipid membrane by trimers of the gp41 transmembrane protein (**Fig. 7**)[49]. The lipid membrane encapsulates a symmetrical layer of matrix proteins, protecting the capsid and the core of the virion. The HIV-1 core houses the replication enzymes reverse transcriptase (RT) and integrase (IN) as well as the virus genomic RNA and is encased by the cone-shaped HIV-1 capsid (CA). The HIV-1 genomic RNA exists as a non-covalent dimer, with a 5′ cap and 3′ polyadenylated tail, and is complexed with a human transfer RNA ($tRNA_{Lys}$) molecule and the virus nucleocapsid protein (NP), a nucleic acid chaperone[48,49].

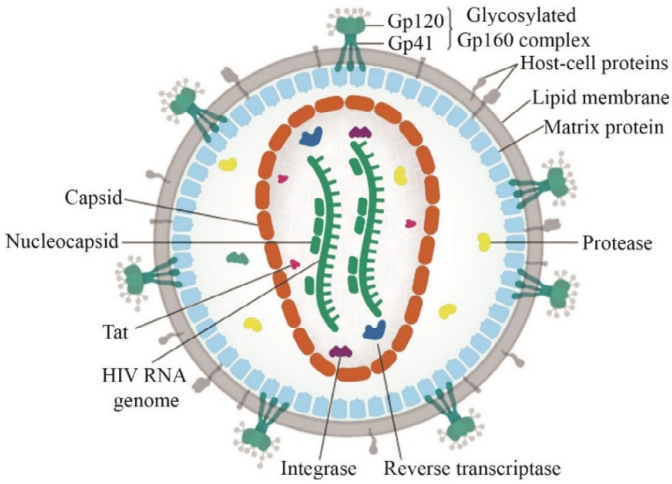


Figure 7. Cross-sectional schematic diagram of the HIV-1 virion: The HIV-1 virion expresses around 35-70 glycoproteins composed of gp120 and gp41 and are embedded with carbohydrate molecules (green). The gp41 transmembrane protein associates non-covalently with the gp120 surface protein, and both are important for virus entry into host cells. The virus envelope has several host cell membrane proteins (Grey) such as class I and class II MHC molecules that are obtained as the virion buds off. Beneath the lipid membrane is a symmetrical layer of matrix proteins (Light blue), protecting the capsid (Orange) and the virus core/nucleocapsid. The nucleocapsid comprises the HIV-1 genome (two copies of positive-sense ssRNA) linked with two molecules of reverse transcriptase (Dark blue) and nucleoid proteins p10 (Green), a protease (Yellow), and integrase (Magenta). Tat, an activator of transcription of virus genes is also found in the virion core (Pink). Source: Thomas Splettstoesser (www.scistyle.com).

HIV-1 genome

The HIV-1 genome (**Fig. 8**) is a coding RNA having nine genes that encode for various virus proteins (**Table 1**). The protein-coding genes are flanked at the ends by the virus long terminal repeats (LTRs), containing transcriptional regulatory elements, RNA processing signals, packaging sites, and integration sites.

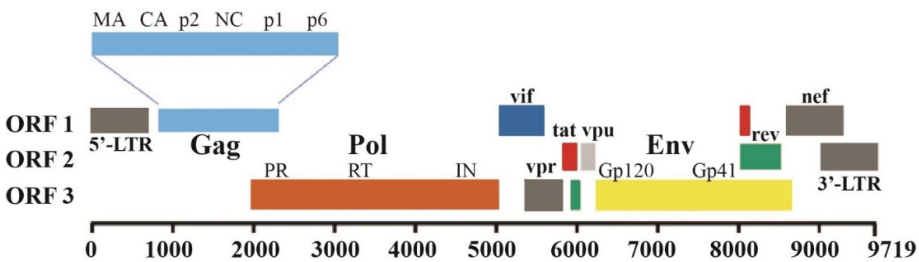


Figure 8. A schematic summary of the HIV-1 genome: Open reading frames are shown as coloured rectangles representing three structural genes (*gag, pol*, and *env*), four accessory genes (*vif, vpr, vpu*, and *nef*), and two regulatory genes (*tat*, and *rev*). The proteins encoded by HIV-1 *gag (MA, CA, p2, NC, p1, and p6)*, *pol (PR, RT, IN)* and *env* (gp120 and gp41) are also shown. The diagram was redrawn to simplify the HIV-1 genome available at www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html.

The Gag polyprotein precursor is proteolytically processed to generate the matrix (MA), capsid (CA), nucleocapsid (NC) and p6 proteins. The Gag-Pol polyprotein is cleaved to produce protease (PR), reverse transcriptase (RT) and integrase (IN).

Gene	Size *	Protein	Function
gag	p24	Capsid protein (CA)	Formation of conical capsid
	p17	Matrix protein (MA)	Formation of the inner membrane layer
	p7	Nucleoprotein (NC)	Formation of the nucleoprotein/RNA complex
	p6	Core protein	Involved in virus particle release
pol	p10	Protease (PR)	Proteolytic cleavage of Gag and Gag-Pol precursor protein
	p51	Reverse transcriptase (RT)	Transcription of HIV-1 RNA in provirus DNA
	p15 (66)	RNase H	Degradation of virus RNA in the virus RNA/DNA replication complex
	p32	Integrase (IN)	Integration of provirus DNA into the host genome
env	gp120	Gp120	Attachment of virus to the target cell
	gp41	Gp41	Anchorage of gp120 and the fusion of virus and cell membrane
tat	p14	Trans-activator protein (Tat)	Activator of transcription of virus genes
rev	p19	RNA splicing regulator (Rev)	Regulates the export of non-spliced and partially spliced virus mRNA
nef	p27	Negative regulatory factor (Nef)	Alters virus replication - downregulates CD4 and MHC I
vif	p23	Virus infectivity protein (Vif)	For infectious virus production in vivo
vpr	p15	Virus protein r (Vpr)	Facilitates virus infectivity, inhibits cell division
vpu	p16	Virus protein unique (Vpu)	Enhances virion release from cells, downregulates CD4 and MHC class I expression

Table 1. Functions of the different HIV-1 proteins. The major HIV-1 genes are the structural genes *gag, pol,* and *env*, but the genome also comprises regulatory (*tat* and *rev*) and accessory genes – *nef, vif, vpr, and vpu*. *Numbers corresponds to the size of the proteins (p) or glycoproteins (gp) in kilo Daltons (kDa).

A key difference between the HIV-1 genome and the HIV-2 genome is that the HIV-1 *env* gene encodes the gp120 and gp41 proteins, whilst the HIV-2 *env* encodes the gp125 and gp36 proteins[50]. Also, the HIV-2 *vpx* gene encodes the Vpx protein which plays a role in nuclear translocation of the virus pre-integration complex (PIC) and is therefore required for the virus to infect non-dividing cells.

HIV-1 replication cycle

The HIV-1 life cycle begins when the gp120 trimer binds non-covalently to CD4, the primary receptor on CD4+ T-lymphocytes, macrophages, dendritic cells, monocytes, and brain microglia **(Fig. 9)**. For the first step (virus entry), gp120 interacts with the surface receptor CD4 and triggers exposure and binding of a second co-receptor – classically the chemokine co-receptor CCR5, although some viruses can bind alternatively to the C-X-C chemokine receptor CXCR4[51-53]. Viruses that use CCR5 are designated R5 viruses, CXCR4-using viruses, X4, and viruses able to utilize both receptors, R5X4[41]. Additional minor co-receptors have

been described including CCR2b[54], CCR3[55], CCR8[56], and CXCR6[57] that may be used – albeit in combination with CCR5 and/or CXCR4[58]. The binding of the co-receptor prompts fusion of the virus envelope and host cell envelope which is mediated by the gp41-fusion peptide, resulting in entry of the virus particle into the cell[59,60]. Upon entry, partial capsid uncoating occurs[61]. Concurrent with uncoating, reverse transcription is primed by a host transfer RNA ($tRNA_{Lys}$) bound to the virus positive-strand RNA (+ssRNA) genome, allowing the virus RT to replicate the virus RNA genome into a complementary DNA (cDNA) molecule with long terminal repeats (LTRs) in the 5' and 3' ends of the genome. Unlike DNA polymerases, the HIV RT is error-prone and lacks proofreading capacity, thus allowing for the accumulation of mutations in the genome[62].

After reverse transcription, a pre-integration complex (PIC) is formed consisting of the HIV-1 DNA and virus proteins RT, IN, MA, NC, and Vpr[63]. However, HIV-2/SIV PIC differs from that of HIV-1 by having an additional factor, virus protein X (Vpx), which is critical for nuclear import[64]. Formation of PIC allows import into the cell nucleus where the PIC-associated virus integrase orchestrates insertion of the virus DNA into the host cell genome forming a provirus that can remain transcriptionally latent[65] or can be transcribed by a cellular RNA polymerase II[61]. The unspliced or partially spliced RNAs are transported from the nucleus into the cytoplasm for translation into proteins[66]. Env proteins are glycosylated and trimerized in the endoplasmic reticulum and the Golgi apparatus. Genomic RNA and virus proteins are positioned on the plasma membrane for virion assembly. Virions then bud off from the cell, acquiring a lipid envelope containing the gp160 virus protein complex trimers and other host cell membrane proteins[67].

Shortly after budding, the virus life cycle ends with maturation involving cleavage of the Gag-Pol polyprotein by the virus protease resulting in functional proteins. The average duration of the HIV-1 life cycle *in vivo* is estimated to be 1.2 days, and the average HIV-1 generation time (the time from the release of a virus particle until it infects another cell and causes the release of a new generation of virus particles) is 2.6 days[68].

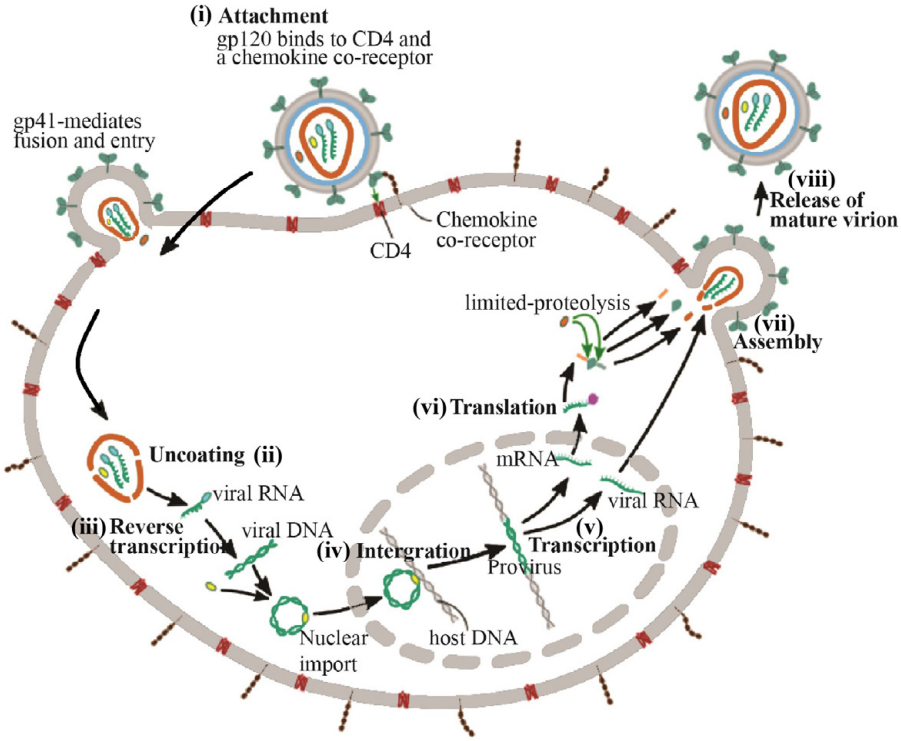


Figure 9. A schematic overview of the HIV-1 replication cycle. The main stages of the HIV-1 life cycle include: (i) virus binding to CD4 primary receptor mediated by virus gp160 surface protein and virus envelope and host cell membrane fusion mediated by the gp41 virus transmembrane protein; (ii) partial capsid disassembly; (iii) reverse transcription mediated by the virus reverse transcriptase enzyme; (iv) formation of pre-initiation complex, nuclear import, and integration into host cell genome mediated by virus integrase enzyme; (v) replication and transcription to produce HIV-1 genomic RNA and HIV-1 mRNA; (vi) translation of virus mRNA to produce virus proteins; (vii) proteolysis of proteins and packaging of the virus genome and proteins into virions at the host cell inner membrane; and (viii) virus particle budding, release from the cell, and cleavage of Gag-Polypeptide by virus protease, resulting in a mature and infectious virus particle. Source: Jmarchn, Wikimedia Commons, https://commons.wikimedia.org/w/index.php?curid=58188472. The original image has been modified for clarity.

HIV-1 recombination

Co-infection or super-infection of individuals with multiple subtypes may result in the emergence of recombinant forms (refer to "HIV genetic variants and global distribution"). A recombinant is thus a genetic sequence that carries regions from two or more genetically distinct parental strains[69]. Given that two RNA copies of the HIV genome are found in each virion, and because of low processivity, the RT enzyme jumps between the two RNA templates during reverse transcription. In case the two RNA templates are of different strains, recombinant genomes may be produced[62].

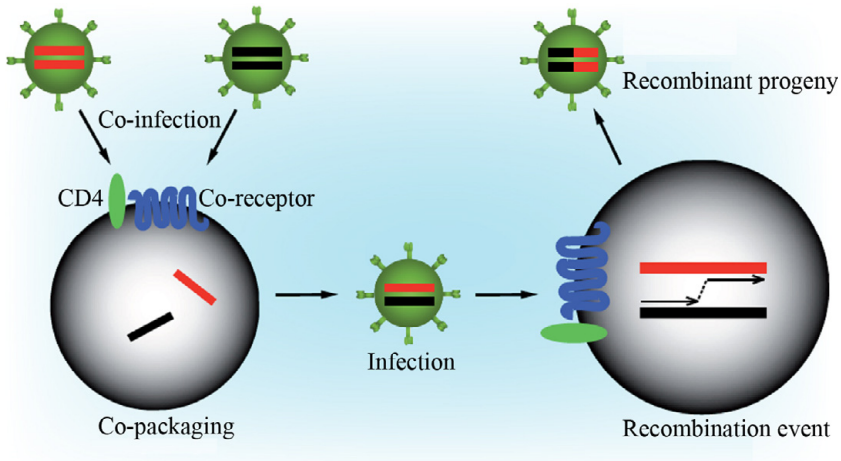


Figure 10. HIV-1 recombination. Two genetically distinct virus particles (red and black) infect the same target cell. Both viruses start to replicate within the cell, and during assembly, one genome from each virus variant are co-packaged into the same budding particle. Second, this new particle infects a new target cell, and during reverse transcription the viral reverse transcriptase jumps between the RNA strands, resulting in recombination events. Consequently, a recombinant progeny, with a mosaic genome from both the parental strains is generated. The figure was adopted with permission from Joakim Esbjörnsson[70].

HIV transmission

HIV is most commonly transmitted through sexual contact across mucosal surfaces, maternal-infant exposure, or percutaneous inoculation[71]. Globally, approximately 80% of HIV infections result from sexual transmission across anal and genital mucosal surfaces (approximately 70% of which are heterosexual transmissions); whilst the remaining proportion results from MSM, injection drug use, and perinatal infections[1]. Some risk factors for HIV transmission include: having unprotected vaginal[72] or anal sex[73]; sharing contaminated needles and syringes when injecting drugs[74]; and parenteral exposure[75] among others[71,76]. The HIV transmission probability per exposure event depends on transmission route – and is estimated to range between 1/200-1/2000 for exposures across the female genital tract (but could be as high as 95/100 for exposures through direct blood contacts, e.g. injection drug use and parenteral inoculations)[71].

The HIV viral load (VL) is defined as the number of copies of HIV RNA per millilitre of blood and influences the transmission probability in sexual transmission. Transmission is rare among individuals with levels of less than 1000 copies of HIV RNA per millilitre, and a 2.5-fold increase in transmission rate has been reported for every 10-fold increase in VL[77-79]. The treatment status of the transmitting partner also influences HIV transmissibility. That is, the risk that individuals on successful treatment with ART (i.e. maintaining undetectable viral loads) should transmit the infection to others is negligable[80]. Medical male circumcision has also been shown to decrease HIV-1 acquisition in the circumcised

male by a factor of 60%[81,82]. Overall, effective prevention strategies include implementing behaviour change programs[83], serodiscordant couples counselling (about condoms, sexual risk, fertility, and contraception)[84,85] harm reduction efforts for injecting drug users[86], male circumcision[82], and adherence to antiretroviral drugs that suppress viremia in infected individuals[80,87,88].

HIV pathogenesis and immunology

As summarised in **Fig. 11**, the natural (untreated) course of an HIV infection may be divided into three stages: the acute stage, the asymptomatic stage, and the AIDS stage[89,90].

The acute HIV stage occurs during the first three to six weeks of infection and is characterised by elevated viremia levels and a rapid drop in CD4+ T cell counts (the concentration of CD4+ T cells in peripheral blood) because of increasing HIV-1 replication in CD4+ T cells. During this stage, HIV-1 is widely disseminated in the host's lymphoid tissues[91]. An initial innate immune response is mounted by cytotoxic CD8+ T cells, causing a decline in viremia, albeit without full suppression of virus replication (HIV expression can persist in lymph nodes even in the absence of detectable viremia in plasma, or HIV mRNA in peripheral blood mononuclear cells)[92]. At this point, seroconversion, (i.e. the development of antibodies, occurs and the CD4+ T cell counts begin to recover as the immune system attempts to fight the virus.

During the asymptomatic stage, viral loads fluctuate around a steady set-point value (i.e. set-point viral load; spVL), and spVL varies up to 1,000-fold between patients[93]. Yet, HIV-infected individuals with higher spVL typically progress faster to AIDS without ART[94,95]. HIV disease progression is affected by the HIV-1 subtype – for instance, subtype D may be associated with faster disease progression than subtype A[96]. Likewise, some subtype-specific variants may vary in virulence, as was recently shown in the Netherlands where individuals infected with a subtype-B variant (termed VB variant) had higher viral loads (than those with non-VB variants) and would have experienced a faster decline in CD4+ T cells count without treatment initiaion[97].

Interestingly, one study has found no synergy between spVL and the infecting subtype in determining progression to clinical AIDS, suggesting that both may act independent of each other[98]. It has been hypothesised that spVL values cluster around 4.52 $\log_{10}$ copies per millilitre and that this value may be a possible outcome of natural selection acting on HIV-1 to maximize opportunities for onwards transmission[93]. Overall, the asymptomatic infection stage lasts several years (average, 10 years)[99] and is characterized by low but persistent viral replication and a slow, continuous loss of CD4+ T-cells and high CD8+ anti-HIV responses[92]. Also,

X4 or R5/X4 virus populations may emerge, permitting entry into alternative host cells, increasing CD4+ T cell depletion rates[100,101].

Eventually, the number of effector T cells needed to mount an adequate immune response can no longer be maintained. Individuals become susceptible to lethal opportunistic infections such as tuberculosis[102], and Pneumocystis Pneumonia[103]. This marks the AIDS stage and unless treated with appropriate antiretroviral drugs, an infected person has 2-3 years of life expectancy[92].

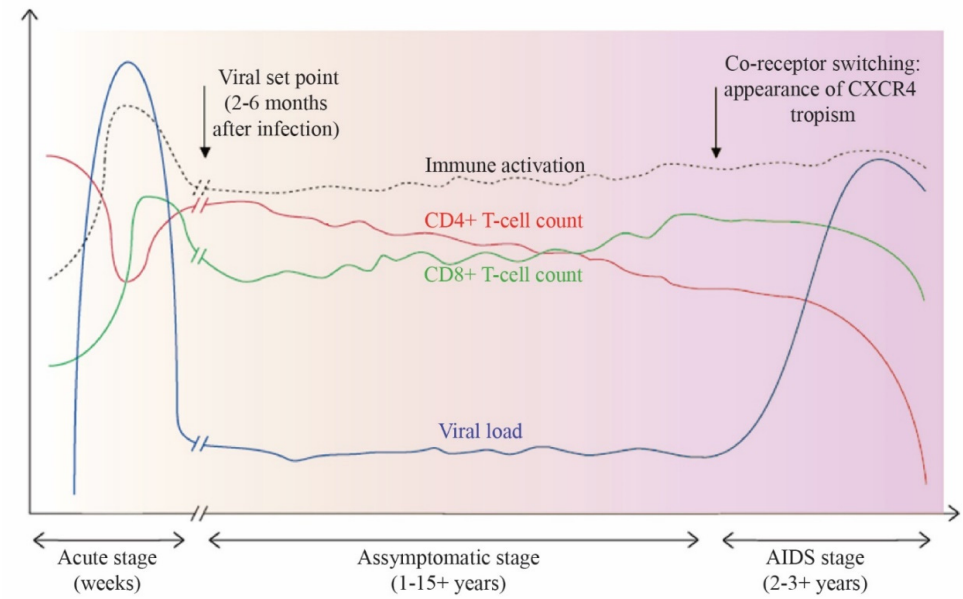


Figure 11. Natural history of HIV-1 infection. A schematic representation of the characteristics of HIV-1 disease progression course in the absence of treatment. Changing dynamics during the acute stage include a rapid increase in viremia, a decrease in CD4+ T cell count, and an increase in CD8+ T cell count. After the acute stage, a relatively stable (setpoint) viral load is achieved, which lasts throughout the asymptomatic stage. During the early AIDS stage, there is a rapid increase in viremia, a sharp decline in CD4+ T cells, and a decrease in CD8+ T cell counts. During the disease course, close to the onset of AIDS, the virus population may switch or broaden its coreceptor use to include CXCR4 (instead of, or in addition to CCR5). The figure was adapted with permission from Joakim Esbjörnsson[70].

HIV treatment

HIV treatment comprises the use of antiretroviral drugs that target different stages of the HIV life cycle. These drugs are grouped into several classes depending on their properties and mechanism of interference with the HIV life cycle (**Fig. 12** and **Table 2**).

CCR5-inhibitors block the CCR5 coreceptor on the surface target cells to prevent virus attachment[104]. Fusion inhibitors and post-attachment inhibitors stop the fusion of the HIV envelope with the host cell lipid membrane, effectively blocking entry[105].

Nucleoside reverse transcriptase inhibitors (NRTIs) prevent the formation of a 3'-5'-phosphodiester bond in growing DNA chains to prevent virus replication[106-110]. Non-nucleoside reverse transcriptase inhibitors (NNRTIs) interfere with the reverse transcriptase enzyme by binding directly to it, blocking the reverse transcription process)[111-115]. Protease inhibitors (PIs) bind selectively to the virus protease enzyme to block proteolytic cleavage of protein precursors that are necessary to produce infectious virus particles[116-119]. Integrase strand transfer inhibitors (INSTIs) block the integration of virus DNA into host DNA by virus integrase[120-123]. A capsid inhibitor (Lenacapavir) which disrupts the functions of HIV capsid protein is currently in early clinical trials, and has shown high potency and promises to be useful as a long-acting drug[124,125].

Drug class	**Mechanism of action**	**Examples**
CCR5 inhibitors	Block the CCR5 coreceptor on the surface target cells to prevent HIV attachment	Maraviroc[104]
Fusion inhibitors	Stops the fusion of the HIV envelope protein with the CD4+ T cell lipid membrane	Enfuvirtide[105]
NRTI	Prevent the formation of a 3'-5'-phosphodiester bond in growing DNA chains to prevent viral replication	Abacavir, emtricitabine, lamivudine, tenofovir, and zidovudine[106-110]
NNRTI	Bind to reverse transcriptase enzyme to block reverse transcription	Doravirine, efavirenz, etravirine, nevirapine, and rilpivirine[111-115]
INSTI	Inhibit the viral enzyme integrase to block integration of viral DNA into the host cell DNA	Elvitegravir, dolutegravir, bictegravir, and raltegravir[120-123]
PI	Block the viral protease enzyme necessary to produce mature virions	Lopinavir, ritonavir, indinavir, nelfinavir, amprenavir, darunavir, and atazanavir[116-119]
Capsid inhibitors	Disrupt HIV capsid	Lenacapavir (an experimental drug)[124,125]

Table 2. Common antiretroviral drug classes with their mechanisms of action. Abbreviations: CCR5, C-C chemokine receptor type 5; NRTI, nucleoside reverse transcriptase inhibitors; NNRTI, non-nucleoside reverse transcriptase inhibitors; INSTI, integrase inhibitors; and PI, protease inhibitors.

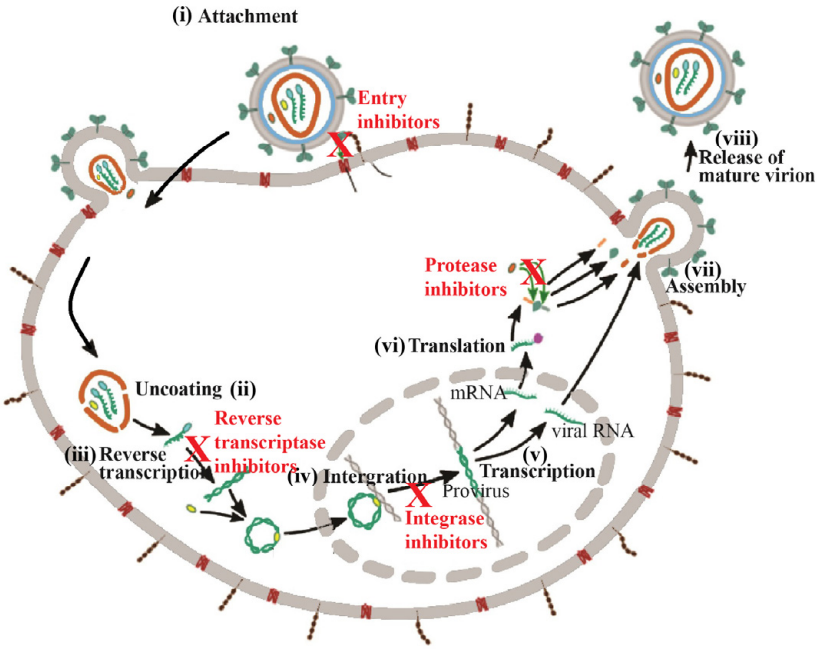


Figure 12. Schematic description of the mechanism of the four classes of available antiretroviral drugs against HIV. Antiretroviral drugs are broadly classified by the phase of the retrovirus life-cycle that the drug inhibits. Fusion inhibitors (interfere with the binding, fusion, or entry of an HIV virion), reverse-transcriptase inhibitors (interfere with the translation of viral RNA into DNA), integrase inhibitors (block the viral enzyme integrase, that inserts the viral genome into the DNA of the host cell), and protease inhibitors (block proteolytic cleavage of protein precursors that are necessary for the production of infectious viral particles). Source: Thomas Splettstoesser (www.scistyle.com). The original image has been modified for clarity.

In the late 1980s and early 1990s, management of HIV/AIDS using the NRTI drug zidovudine (AZT) as a mono-therapy offered sub-optimal HIV-1 control due to the rapid emergence of viruses carrying mutations that permitted replication in the presence of the drug[126]. In 1995, the U.S. Food and Drug Administration (FDA) approved the first protease inhibitor, called Invirase (saquinavir)[127]. In 1996, the combined use of a PI with two NRTIs (an approach named highly active antiretroviral therapy, denoted HAART) resulted in better control of viremia, an increase in CD4+ T cell counts, and reduced mortality[128-131].

Although highly active, HAART had some disadvantages – some drugs of the time were toxic and caused potentially severe metabolic effects[132]. Also, there were concerns about the development of drug resistance if treatment adherence was not maintained – especially linked to the use of NNRTI[133]. In addition, HAART was associated with a high pill burden – for instance, patients were required to take three capsules of Invirase every 8 hours, and this schedule was difficult to sustain over the long term[127]. To maximize on risk-benefit ratio, HAART was delayed until the immune function dropped below a CD4+ T cell count of less than 350. In 2001, the

NRTI tenofovir disoproxil fumarate was introduced – and this drug had fewer side effects, required one pill daily, and could overcome drug resistance[134].

Combination antiretroviral therapy (cART, also generally referred to as ART) has since become the recommended treatment regimen, and modern ART regimens can suppress viral load to undetectable levels with a low pill burden and minimal side-effects[135,136]. Other key developments include so-called "long-acting slow effective release ART" (LASER ART) such as the long-acting cabotegravir and rilpivirine combinations, which reduce the pill-burden, improve patient adherence, and effectively maintain HIV-1 suppression[137].

ART can also be taken as pre-exposure prophylaxis (PrEP) to reduce the risk of getting infected[138]. In high-income countries, treatment is often supplemented with plasma HIV-1 RNA quantification to monitor treatment efficacy and prognosis for patients on treatment, and drug resistance testing is part of routine care and is essential for maintaining treatment regimen efficacy[135,139,140].

Overall, HIV-1 treatment has become widely available and has contributed to preventing early mortality, improving lifelong survival, increasing quality of life, and preventing new HIV infections following studies showing a reduced risk of onward transmission in virologically suppressed patients[141,142].

HIV drug resistance

Modern cART combination regimens are effective in blocking virus replication and maintaining long-term virus suppression – if no baseline pre-treatment HIV drug resistance is present and drug concentrations are maintained at an optimal concentration[143]. Yet, increased use of ART at the population level is known to contribute to HIV drug resistance (HIVDR) – and transmission of resistant viruses can also further compromise therapy[128,144-146].

HIVDR emergence is a consequence of within-host virus evolution which results in the generation of virus quasispecies (a swarm of closely related but genetically diverse virus populations) including numerous potentially drug-resistant mutant variants[147]. Such drug-resistant variants may have limited replicative fitness (presumably due to mutation-induced structural changes in the binding of the natural substrate and in the catalytic activity of the virus RT and PR)[148,149] – and are in most cases outgrown by the more fit and drug-susceptible wild-type viruses. However, when drug concentration is sub-optimal (e.g. when patients fail to adhere to medication), selective pressure is exerted that encourages the growth of pre-existing drug-resistant mutants resulting in acquired drug resistance (ADR)[150,151].

Drug-resistant viruses may be transmitted to other individuals – this is referred to as transmitted drug resistance (TDR, resistance detected among antiretroviral drug-

naïve people with no history of antiretroviral drug exposure). Pre-treatment drug resistance (PDR) comprises drug resistance (mutations) detected in HIV-infected persons before starting ART, and such resistance could be due to TDR or prior exposure to antiretroviral drugs (e.g. short-course prophylaxis for prevention of mother-to-child transmission (PMTCT), pre or post-exposure prophylaxis, or first-line ART restarters after earlier treatment interruption[152].

Overall, multiple factors contribute to HIVDR. First, a drug-related factor is the drug's genetic barrier to resistance – defined as the number of mutations needed for the virus to overcome the drug-selective pressure. Whereas some drugs (e.g. ritonavir-boosted PIs and the INSTI dolutegravir (DTG), bictegravir (BIC) and cabotegravir) have a high genetic barrier and their inhibitory ability is not easily overcome by virus mutations, drugs with a low genetic barrier – e.g. the NNRTIs efavirenz (EFV) and nevirapine (NVP), the NRTIs emtricitabine (FTC) and lamivudine (3TC), and the INSTIs raltegravir (RAL) and elvitegravir (EVG) may lose their effectiveness even in the presence of only a single mutation[143,153].

Second, some factors are virus-related[154-157]. For instance, when transmitted, the M184V mutation (where methionine replaces valine at position 184 in reverse transcriptase) reduces the virus replicative capacity in the absence of drug pressure[154,155]. Hence a variant with the M184V mutation is less likely to be transmitted compared to a variant harbouring a low-fitness-cost mutation (e.g. the K103N or L90M mutations)[154,155,158]. The infecting virus subtype may also influence the propensity for the development of resistance – this has been suggested to be linked to differences in virus polymorphisms that influence drug binding affinity/kinetics[156,157].

Third, patient-related factors comprise adherence levels (i.e. sustaining high-level adherence to self-administered medication ensures optimal drug concentration and reduce the risk of the emergence of resistant HIV-1 strains)[159]. Likewise, perinatally infected infants have been suggested to have a higher risk of harbouring PDRs due to prior exposure to suboptimal maternal ART during breastfeeding[160].

From a broader perspective, in high-income settings, the ART regimens are often guided by baseline assessment of drug-resistance mutations and then individually designed and monitored to minimize pill burden, therapeutic side-effects, and virological failure. In contrast, most low- and middle-income countries (LMICs) follow WHO-standardised ART guidelines – in 2021, these comprised the use of dolutegravir (DTG)-based regimens in combination with NRTIs as the preferred first-line treatment for adults and children[152]. In addition, baseline HIVDR testing is not common, virological monitoring is sub-optimal, and treatment history is not always known, altogether resulting in high levels of both PDR and ADR on the population level[152,161-163].

HIV-1 evolutionary dynamics within and between hosts

HIV-1 is characterised by a high evolutionary rate resulting from accumulated mutations because of an error-prone reverse transcriptase[164], a high virus turnover[68], and high genetic recombination rates[165]. It has been established that sexual transmission of HIV-1 is characterised by a genetic bottleneck where in most cases, systemic infection is derived from a single variant – the transmitted/founder virus (TF)[166,167]. The early within-host virus population is relatively homogenous and is targeted by both cellular and humoral immune responses[168,169]. This causes accumulation of immune-escape mutations and a rapid increase in HIV-1 within-host diversity[168,170,171].

Paradoxically, adaptive within-host evolution may decrease virus transmissibility[172]. Yet, it has been proposed that a fraction of ancestral (TF-like) variants with limited within-host evolution (and thus higher transmission potential) may persist as archived provirus sequences in long-lived memory CD4+ T cells[173-175]. In a 'store-and-retrieve' fashion, these TF-like variants are likely resurrected and preferentially transmitted compared to non-TF-like variants[176,177]. Whilst within-host evolutionary dynamics are influenced by selective forces and competitive fitness[176,178-180], between-host HIV-1 evolution is mostly influenced by neutral processes – and depends on the social dynamics of the host population and the biological processes defining the transmission event[180-182].

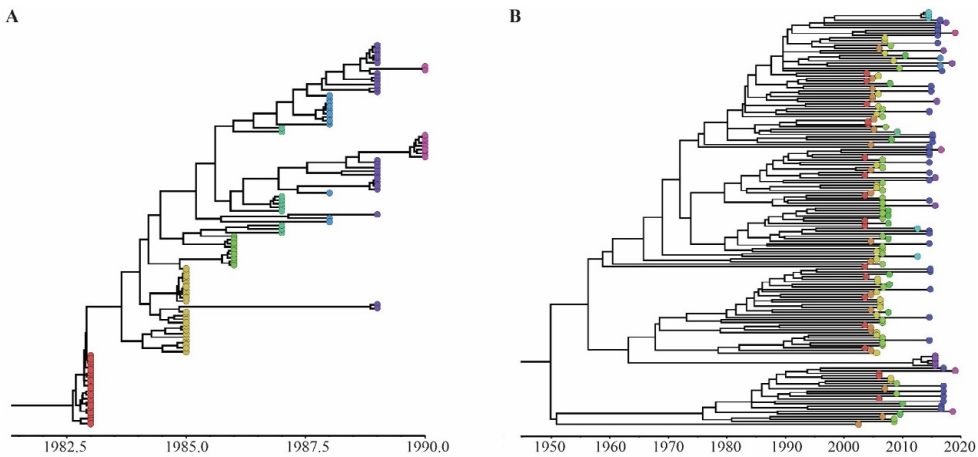


Figure 13. Phylogenetic visualisation of intra-host and inter-host HIV-1 diversity. The phylogenies represent (A) HIV-1 within-host phylogeny with branch lengths in time units partial *env* gene longitudinally sampled from a single patient over 80 months (subtype B, 106 sequences, 516 bp; patient 3)[171]; and (B) HIV-1 population phylogeny with branch lengths in time units: full-length *pol* gene sampled from Coastal Kenya (subtypes A (A1), C, D and recombinants, 163 sequences, 967 bp)[183].

The rapid rates of HIV-1 evolution allow for the reconstruction of phylogenetic trees that could be used to infer patterns in within-host or between-host relatedness of viruses. When visualised on a phylogeny, within-hosts heterochronous HIV-1 sequences exhibit an asymmetrical or "ladder-like" profile with little diversity at any given time point (**Fig. 13a**)[179,184]. In comparison, a phylogeny of heterochronous inter-host HIV-1 sequences typically reflects the persistence of multiple lineages over time and display less of a ladder-like conformation **(Fig. 13b)**[179,183].

Phylogenetic inference

Introduction

Phylogenetic inference applies to the study of evolutionary relationships among organisms based on genetic information. An evolutionary relationship can be thought of as a branching process – where organisms are altered over time through speciation (differentiation into separate branches and develop into novel species)[185], hybridization (fusion into a new population by the mating of two former distinct populations)[186], and extinction (disappearance of an entire species)[187].

Traditionally, relationships between organisms were estimated by comparing their morphological features[188]. However, in recent years, advances in genetic sequencing technology have increased the availability of sequence data, allowing the use of gene sequence data to infer genetic relationships. For instance, the genetic relationship between five hypothetical taxa is illustrated in **Fig. 14.** This is a typical bifurcating phylogenetic tree – where taxa (i.e. A-E) are placed on the tips of the tree and every two branches are linked at a node representing the most recent common ancestor (MRCA). Based on this tree, all taxa are descended from one common ancestor (MRCA 4) at the root. In the rectangular tree layout, the genetic relatedness between organisms is represented by the lengths of the horizontal (not vertical) branches.

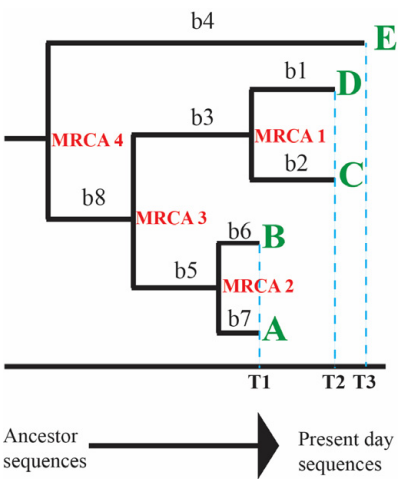


Figure 14. A typical bifurcating phylogenetic tree showing the genetic relationship between five taxa A-E. Horizontal lines represent genetic distance depicting divergence from their most recent common ancestors (MRCAs). The phylogeny depicts that sequences A and B are more closely related, but more distantly related to sequences C and D, and least related to sequence E. Nodes represent the MRCA of the branches leading up to that node.

Typically, organisms that are more genetically similar cluster more closely together in a phylogenetic tree compared to more distantly related organisms. For instance, the branching process in **Fig. 14**, shows that sequence A is more closely related to sequence B, but more distantly related to sequence C and D.

In addition to revealing the pattern of evolutionary relationships itself, contemporary interests include generating phylogenies to derive information regarding the processes responsible for the observed pattern of evolutionary relationships. In this case, the tree topology is itself the framework upon which further inference is drawn. Thus, phylogenetic trees also facilitate inference of rates of evolution and population demographics[189-191]. In this thesis, phylogenetic trees have been used to determine the HIV-1 subtype, the underlying transmission network (based on HIV-1 sequence clustering patterns), and to infer populations dynamics.

Sequence alignment

One of the first steps in constructing phylogenetic trees involves aligning two or more biological sequences. In this process, gaps are added to a matrix of sequences such that nucleotides or amino acid residues in one column of the matrix are related to each other based on the assumption that they are descendants of a common ancestral residue[192]. Aligning multiple sequences allows the identification and localisation of specific evolutionary alterations, such as single nucleotide polymorphisms (SNPs), or insertions or deletions (indels) that have been accumulated by the studied lineages since they diverged from a shared ancestor. A match occurs when the corresponding nucleotide base or amino acid is encountered at a given position, a mismatch occurs where at least one substitution has occurred since the divergence event, and a gap indicates that an indel has occurred in one or more of the compared sequences (**Fig 15**).

Several programs employing sequence alignments algorithms exist, the most widely used being ClustalX2 and MAFFT[193,194]. ClustalX2 implements a scoring system where base matches or mismatches are assigned a positive score, whereas gaps are assigned negative scores. The severity of the gap penalty varies when either a gap is introduced or extended. A heuristic search is then performed to select the best alignment, i.e. the alignment with the highest score[193]. The MAFFT alignment program involves a trade-off between accuracy and speed[194]. MAFFT aligns a large number of sequences by employing a fast group-to-group alignment algorithm based on Fast Fourier Transform (FFT) and an approximate distance calculation method (the so-called “6mer method”) to facilitate rapid alignments[195].

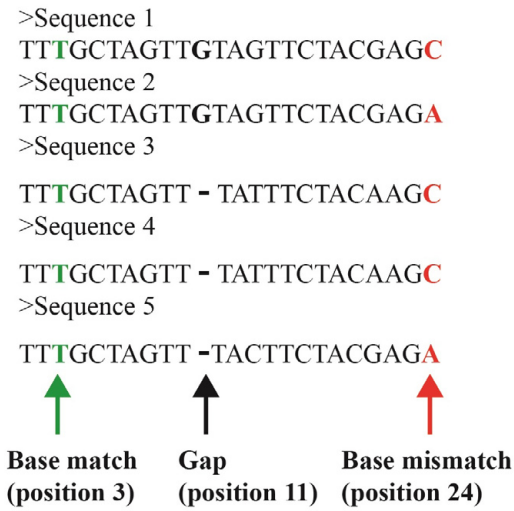


Figure 15. A schematic multiple sequence alignment showing different types of bases impairments. A match occurs when the same base is encountered at a given position, a mismatch is found when at least one substitution occurred since the two sequences diverged from each other, and a gap indicates that one or more deletions or insertions have occurred in one or more of the compared sequences.

Selecting nucleotide substitution models

The second step in reconstructing phylogenetic trees involves selecting an appropriate nucleotide substitution model. Pairwise distance (a measure of divergence between two taxa from a common ancestor) can be estimated by comparing the observed distance between the taxa. However, divergence may not reflect the true number of point mutations that happened at a specific nucleotide site during the evolutionary process. Therefore, this is often addressed by modelling the evolutionary process using nucleotide substitution models that define the rates of change of fixed mutations among sequences[196]. Several substitution models are available, all of which use a matrix that stipulates the rates of nucleotide changes across sites (substitution rates, i.e. A-C, A-G, A-T, C-G, C-T, and G-T) along the alignment with the assumption that the relative frequencies of the nucleotide bases (πA, πC, πG, πT) are at equilibrium. The Jukes and Cantor (JC or JC69) model is the simplest model, and assumes equal equilibrium frequencies for all bases (i.e. $\pi A=\pi C=\pi G=\pi T=0.25$) and that the substitution rates are equal (i.e. $a=b=c=d=e=f$)[197]. In genetics, transitions (A↔G or C↔T) are generally more common than transversions (A↔C, A↔T, C↔G, or G↔T, **Fig. 16**) – therefore, too simplified substitution models may not be satisfactory as they do not account for the transition/transversion ratio and/or unequal base frequencies[198].

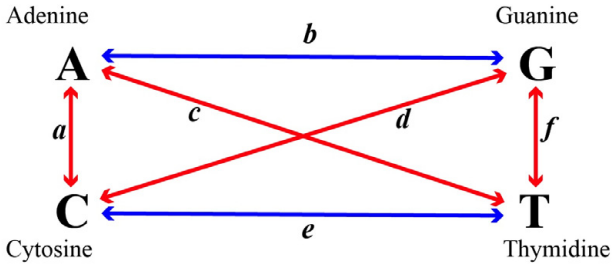


Figure 16. Illustration of nucleotide substitutions. Transitions (A↔G or C↔T, coloured blue) are more common in the evolutionary process than transversions (A↔C, A↔T, C↔G, or G↔T, coloured red). Six possible rates of nucleotide substitution (labelled a-f) are also shown, assuming a reversible substitution process.

Other models that are more parameter-rich, and that are often preferred over the JC69 model, include the Felsenstein model (F81, which allows the frequencies for all bases to change)[199], the Kimura two-parameter model (K2P or K80, which assumes equal equilibrium frequencies for all bases whilst accommodating transitions/transversions)[200], the Hasegawa, Kishino, and Yano model (HKY85, that allows both free base frequencies whilst accommodating transitions/transversions)[201], and the general time-reversible model (GTR, that assumes that each nucleotide base has a separate rate and that transitions/transversions are reversible at similar rates)[192]. In addition, the different models can be complemented by allowing for rate variations in nucleotide distribution across the alignment (e.g. through a discrete gamma "Γ" distribution), sometimes acknowledging that a proportion of sites are invariant and hence have a zero rate of change. To increase accuracy in phylogenetic inference, some tree construction programs now allow for automatic model-selection and flexible rate heterogeneity across sites model, resulting in substantial improvements in the fit between tree, model and data, and with significant gains in computing time[198,202].

Methods for constructing phylogenetic trees

Distance-based methods

Tree-building methods are classified into distance methods or character-based methods. Distance-based methods exploit a matrix of pairwise genetic distances to infer a phylogenetic tree and include the unweighted pair-group method with arithmetic means (UPGMA) and the neighbour-joining (NJ) methods[203,204]. The UPGMA method finds the pair of sequences/taxa with the smallest genetic distance between them, where the distance from each of the two sequences to the MRCA is half the distance between the two sequences. These two taxa are then joined to form one cluster, reducing the number of the sequences in the alignment by one. The model then recalculates a new distance matrix considering the genetic distance from the cluster to each of the remaining taxa, and then adds the taxa with the shortest distance to the sequences in the cluster (unless there is a new pair with a shorter genetic distance). The reiterative process of generating new matrices and adding up taxa in the phylogeny based on genetic distance is repeated until culminating in a

tree. A limitation with the UPGMA method is that it (unrealistically) assumes that substitutions accumulate at the same rate in all lineages diverging from a common ancestor and that all taxa are equally distant from the root. These assumptions, therefore, reduce the reliability of UPGMA trees[205].

The more commonly used NJ method rejects the assumption of a constant molecular clock and does not construct intermediate clusters with nodes at the midpoint (unlike the UPGMA). The NJ method first assumes a star-like tree topology, converting all sequences into a distance matrix representing the genetic distance between all sequences in the alignment. The two taxa with the shortest distance are paired up and connected to a node representing their MRCA. A new matrix is then re-computed in which the two taxa have been replaced with their MRCA. Internal branches are progressively inserted iteratively by calculating the genetic distance between the successive MRCAs and the most similar taxa among the remaining sequences in the alignment until all nodes are bound into one tree. The total length of the tree is then calculated by summing up the distances between each external node where the branches minimizing the total tree length are retained and the shortest tree is selected. However, the shortest tree may not be the "true" tree, but NJ provides a good starting tree for the generally more computer-intensive character-based methods.

Character-based methods

Character-based methods use discrete character states (such as amino acid or nucleotide positions) to infer a phylogenetic tree and include maximum parsimony, maximum likelihood (ML), and Bayesian inference methods. The maximum parsimony approach has been shown to result in inaccurate results with an infinite amount of data and is thus less often favoured[206].

The ML approach is the most used method, and requires a sequence alignment, a user-defined model of nucleotide substitution, and an initial tree topology[199]. The approach exploits the concept of likelihood, i.e. the probability P of observing the data D given the hypothesis H, denoted $L=P(D/H)$ – where D is the sequence alignment of interest, and H is the given phylogenetic tree. In other words, ML searches for the tree that maximises the probability P(Data/Tree)[199].

Trees are searched by cutting off subtrees and reassembling them in different positions on the original tree using branch swapping. The three most commonly used branch swapping methods are the nearest-neighbour interchange (NNI), the subtree pruning and regrafting (SPR), and the tree bisection and reconnection (TBR) methods[192]. During the branch-swapping procedure, the likelihood of (almost) all possible (unrooted) trees for the specified alignment is calculated, and the tree(s) with the highest compound probabilities (maximum likelihood) of character distribution is selected as the ML tree. The higher the number of sequences in the

alignment, the more the expected number of possible trees, and hence the more computationally intensive it is to carry out a tree search (i.e. to investigate all possible trees). However, the iterative tree search process is terminated after many trees have been assessed without any increase in ML score during the iterations. To reduce computation time, a popular approach to tree construction is to use an NJ tree as a starting topology for an ML search, and this has been incorporated in most automated tree building algorithms[207,208].

Another common practice in tree construction is to place some measure of statistical confidence on the inferred phylogeny. Statistical support of a phylogenetic tree is a measure of the robustness of associations between the sequences within the suggested topology[209]. Traditionally, statistical support has been provided through so-called "bootstrapping" – i.e. when positions in the original alignment are randomly resampled (permuted) multiple times (usually 100-1000 times) with replacement to produce a set of pseudo-replicate alignments[210]. A phylogeny is then reconstructed from each of these pseudo-replicates, and support for a cluster can then be assessed as the number of independent pseudo-replicates in which that cluster occurs – i.e. if a cluster occurs in 95 of 100 pseudo-phylogenies, the assigned bootstrap value would be 95%[192,211]. A limitation with the bootstrap approach is that it can be challenging to select a reasonable cut-off for significance – although bootstrap values >70% have been suggested to indicate strong support for a cluster as bootstrap values can be thought of as conservative measurements[212]. An alternative to bootstrapping is the Shimodaira-Hasegawa-like approximate Likelihood Ratio Test (aLRT-SH). This test is based on a likelihood ratio test to evaluate if a specific branch is significantly longer than zero or not, and aLRT-SH values ≥0.90 are commonly considered significant[212,213]. Moreover, the aLRT-SH is implemented within most fast ML tree estimation programs, such as PHYML and IQ-tree (both adopted in this thesis), resulting in fast computation of branch support[207,208].

Bayesian inference (like ML) also requires a user-specified nucleotide substitution model but performs a search for the tree that maximises the probability of seeing a tree given both the data and the nucleotide substitution model – i.e. P(Tree/Data). At the core of Bayesian inference is Bayes' theorem for calculating conditional probabilities. Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the independent events[214]. That is, the probability of the outcome of event A does not depend on the probability of the outcome of event B. According to the Bayes' theorem, the (posterior) probability P of a hypothesis (A|B) given the conditional event B is summarised in **Fig 17**.

$$\mathbf{P(A|B) = \frac{Pr(B|A) \times Pr(A)}{Pr(B)}}$$

Where:
P(A|B) – the probability of event A occurring, given event B has occurred
P(B|A) – the probability of event B occurring, given event A has occurred
P(A) – the probability of event A
P(B) – the probability of event B

Figure 17. The Bayes Theorem. Bayes' mathematical formula is used to describe the probability of an event based on prior knowledge of the conditions that might be relevant to the event, where A and B are independent events (i.e., the probability of the outcome of event A does not depend on the probability of the outcome of event B).

Here is an example to contextualize Bayes' theorem. As a medical statistician, you need to calculate the probability of a patient having lung cancer if they are a smoker (it is generally said that "smoking" is linked to the development of lung cancer). In this example, **A** represents the event "patient has lung cancer", and based on your past data, you know that 10% of patients visiting your clinic have lung cancer. Therefore, **P(A) = 0.10**. On the other hand, **B** represents the event that "the patient is a smoker", and from recorded patient demographic data, 5% of the clinic's patients are smokers. Therefore **P(B) = 0.05.** You might also have prior information that 7% of the patients diagnosed with lung cancer are smokers. Thus, the probability that a patient is a smoker given that they have lung cancer is 7%, and this is your **B|A**. Based on Bayes' theorem, **P(A|B) = (0.07×0.1)/0.05 = 0.14** meaning that, if the patient is a smoker, their chances of having lung cancer is 0.14 (14%). This is a larger proportion than the 10% suggested by past data – although it is still unlikely that any patient has lung cancer (i.e. there are uncertainties).

In phylogenetics, Bayesian inference is computed using the Markov Chain Monte Carlo (MCMC) methods proposed by Metropolis and Hastings[215,216]. Based on the MCMC approach, the process of a heuristic tree search involves a random walk over the space of all possible tree combinations. The "heuristic tree search" process is comparable to the "hill-climbing process in a forested terrain" intending to find the highest peak and involves several steps. One random tree *t1* is selected as the current tree and compared to a second tree *t2*; if the likelihood *L1* of *t1* is less than the likelihood *L2* of *t2*, then *t2* replaces *t1* as the current tree (or solution) and the process proceeds one step up-hill. However, if *LI* is more than *L2*, *t1* is retained as the current solution; *t1* is saved (sampled) and the whole process is reiterated several times (usually millions of times depending on the chain length as specified by the user). The likelihood values characteristically increase rapidly during the initial stage of the tree search because the starting point is far away from the highest peak (i.e. regions in parameter space with high posterior probability). Trees logged in the initial stage of the Markov chain (typically referred to as the burn-in stage) are discarded as they are biased by the starting point. However, as the chain moves towards and around the highest likelihood, the likelihood values tend to reach a plateau stage where the chain "converges" – i.e. stays in the region with the highest

posterior probability. The frequency at which a tree is 're-visited' is proportional to its likelihood given the data (thus the solutions are biased based on their likelihood score). At the end of the process, the algorithm produces "a set of trees" that has been visited repeatedly. The sampling frequency is also user-specified. One can choose to sample/log every 1000th tree in a chain length of 10 million generations thus yielding a set of 10000 trees, all with a specific posterior distribution likelihood, and one typically chooses to discard the first 10% trees as chain burn-in.

An advantage of the Bayesian inference over the ML method is that many trees with a high likelihood are generated during the MCMC chain (the posterior likelihood). This means that, if the Markov chain runs for a sufficient number of generations, the amount of time it spends sampling a particular parameter value is proportional to the posterior probability of that value or interval[192]. For instance, if one clade is present in 85% of all sampled trees, then the likelihood that the clade is correct is 85% given the assumed nucleotide substitution model. An example of a program that implements the Bayesian phylogenetics is BEAST (Bayesian Evolutionary Analysis Sampling Trees) which was adopted in this thesis[217,218].

Molecular evolution and the molecular clock

The molecular clock hypothesis postulates that the rate of molecular evolution is constant over time or among species – where mutations accumulate at a uniform rate after species divergence, keeping time like a timepiece[219]. To infer divergence dates, it may seem fitting to assume a constant rate of evolution throughout the tree[220]. However, unless when analysing sequences from the same species (i.e. population data) or very closely related species, the concept of a perfectly constant rate of evolution (i.e. the 'strict' clock) has been challenged by results from datasets showing considerable departure from clock-like evolution[221-223]. Such rate variation among lineages can bias the estimated divergence dates[224] as well as phylogenetic inference[206].

Therefore, models that "relax" the assumptions of the molecular clock by allowing evolutionary rates to vary over time and across lineages have been developed[225-228]. The approach allows for co-estimation of both the phylogeny and the divergence dates under a relaxed molecular clock[229,230]. However, performance is subject to the size of the dataset under investigation (analysing thousands of sequences using the MCMC procedure is extremely computationally intensive and MCMC parameters often fail to converge)[231].

Population dynamics and the coalescent

Evolution is a gradual process that defines the characteristics of a population over time. Thus, molecular sequences, whether sampled simultaneously (homochronous) or serially through time (heterochronous), can be used to reconstruct the demographic history of natural populations[232]. Though observing evolutionary

variations in samples obtained from a population, we can retrospectively address biological questions regarding the historical demography of the population as well as the evolutionary dynamics that yielded the observed patterns in the population[192]. The coalescent population genetic model was developed to infer historical dynamics from contemporary character states in a population, and typically refers to a stochastic process that allows the estimation of historical states of a population from a genealogy of random samples from the population[233-237]. The theory assumes that all individuals in a population descend from one common ancestor and that genetic material is transmitted from ancestors to descendants along the branches of the phylogenetic tree (**Fig 18**). As we move back in time from the present, one can follow the number of lineages in the genealogy in each generation – the number of lineages decreases with every coalescent event (i.e. when two lineages share one common ancestor) but increases with every sampling event.

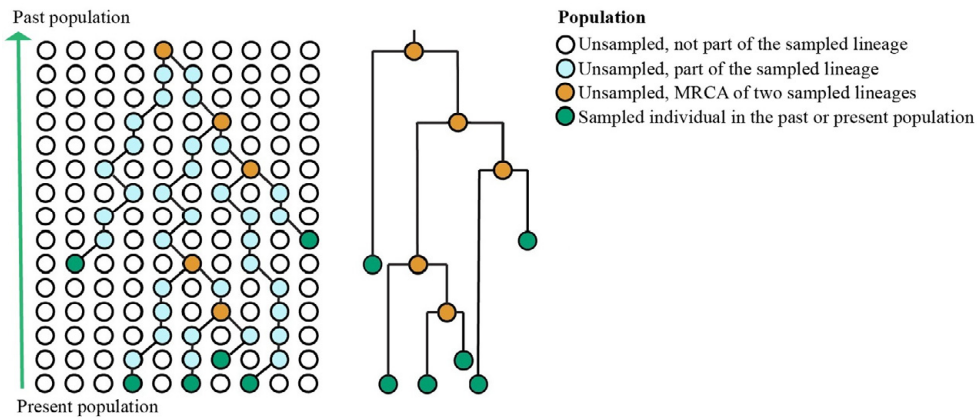


Figure 18. Schematic representation of the relationship between the demographic history and the genealogy of individuals sampled assuming a constant population size. The phylogeny includes six individuals (green dots) that have been sampled over 15 generations. As we move backwards in time, the number of lineages per generation decreases where two individuals have a shared common ancestor (a coalescent event, orange dots), and increases when sampled individuals are encountered (a sampling event). The figure was simplified based on Drummond *et al.*, (2003)[232].

The probability of a coalescence event at a given time is inversely proportional to the population size at the denoted time (i.e. the more the lineages, the faster the rate of convergence). The pattern of observed coalescence and sampling events can therefore be useful in estimating the demographic history of a study population[218,232]. When the molecular clock (strict or relaxed) is assumed, the coalescent theory permits dating the time to the most recent common ancestor (tMRCA) of a sub-population as well as estimating parameters such as nucleotide substitution rate (μ, site^{-1} year^{-1}), growth rate (r), and effective population size (N_e) based on the tree topology.

The effective population size can be defined as the size of an idealized Wright-Fisher population (with discrete, nonoverlapping generations)[238] which would have the same carrying capacity for genetic variation as the given (census) population and thus loses or gains genetic diversity at the same rate as the census population size[192]. To reduce variability in the inferred estimates of population size and substitution rate, it is advisable to use heterochronous sequences (with broad temporal sampling). The selected demographic model influences the estimated demographic pattern, and a large variety of existing models has therefore been proposed, such as the constant size, exponential growth, logistic growth, expansion growth, and Bayesian skygrid or the Skyride model[218,237,239,240].

Furthermore, combining Bayesian inference with the coalescence theory and individual-specific demographic traits (such as geographic or transmission risk group) allows investigations into spatiotemporal evolution of (mostly fast-evolving) organisms/viruses based on genetic sequences[6,31,32,240-242].

Materials and methods

Systematic literature review
For the literature review (**paper VI**), an exhaustive search of the PubMed database (https://pubmed.ncbi.nlm.nih.gov/) was carried out by analysing peer-reviewed research articles on HIV-1 phylodynamics in sSA published in English 1995-2021. Review articles, book chapters, editorials and articles published in other languages were excluded from the search.

The MeSH terms (HIV-1) AND (Africa) were used to select HIV-1 articles from African countries. The keywords "phylogenetic analysis" OR "phylodynamics" OR "evolution" OR "phylogeny" OR "molecular epidemiology" OR "transmission" were used to widen the scope and to ensure that all relevant research articles were included. Filters on the year of publication, language and article type were applied to refine the search.

Two investigators carried out the selection process independently. The articles were manually screened, first by title, then by abstract to assess relevance based on our eligibility criteria (i.e. description of HIV-1 mixing within and between geographic regions and risk groups). Any discordance between the two independent reviewers on the eligibility of articles was resolved through discussions for a consensus.

Shortlisted articles were imported into EndNote X8 (Clarivate, Philadelphia, Pennsylvania, USA) for further management, and to compile the information presented in this review.

Study population and partial HIV-1 *pol* sequence datasets
The HIV-1 *pol* sequences analysed in this doctoral dissertation (**papers I-V)** were either newly generated or published sequences retrieved from the Los Alamos HIV sequence database[243].

New HIV-1 *pol* sequences from Kenya (**papers I-IV**) were generated from blood plasma obtained through studies conducted through the MSM Health Research Consortium – a multi-site collaboration between researchers affiliated with KEMRI-Wellcome Trust (KWTRP) in Coastal Kenya, Nyanza Reproductive Health Society (NRHS) in Western Kenya, Kenya AIDS Vaccine Initiative's Institute of Clinical Research (KAVI-ICR), and Sex Workers Outreach Program (SWOP) clinics in Nairobi.

Table 3 shows a summary of the source of new HIV-1 sequences from Kenya. These included samples from Coast – derived from participants in a prospective observational cohort (2006-2019)[244], samples from Nairobi from a respondent-

driven sample survey (TRANSFORM, 2017)[245], an unpublished study at the Kenya AIDS Vaccine Initiative's Institute of Clinical Research, and samples from Nyanza derived from the Anza Mapema cohort (2015-2017)[246]. Additional nationwide HIV-1 *pol* sequences (2008-2018) were obtained from the national HIV-1 reference laboratory at the Kenya Medical Research Institute (KEMRI) – Centre for Global Health Research.

Site	**Risk group**				
	HET	**MSM**	**FSW**	**PWID**	**Total**
KWTRP	48	21	107	0	176
NHRS	0	57	14	0	71
KAVI	30	0	7	0	37
SWOP	0	50	0	0	50
TRANSFORM	0	85	0	0	85
KEMRI-CGHR	336	0	0	0	336
Total	414	213	128	0	755

Table 3. The source of new HIV-1 sequences in Kenya. Abbreviations: HET, heterosexual adults; MSM, men who have sex with men; FSW, female sex workers; PWID, people who inject drugs. Site abbreviations: KWTRP, Kenya Medical Research Institute (KEMRI)-Wellcome Trust (Coastal Kenya); NHRS, Nyanza Reproductive Health Society (in Western Kenya); KAVI-ICR, Kenya AIDS Vaccine Initiative's Institute of Clinical Research (in Nairobi, Central Kenya); SWOP, Sex Workers Outreach Program clinics in Nairobi, TRANSFORM, a cohort of transfeminine people and cisgender men who have sex with men in Nairobi; KEMRI-CGHR, Kenya Medical Research Institute (KEMRI) – Centre for Global Health Research (Western Kenya).

New sequences were supplemented with published Kenyan HIV-1 *pol* sequences (HXB2 positions 2000-3600) retrieved from the Los Alamos HIV-1 sequence database (1986-2019)[243]. In cases where more than one sequence per individual was available, the oldest sequence was retained. All sequences were annotated with sampling date, sampling location, treatment status, age, sex, and risk group – MSM (men who have sex with men); PWID (people who inject drugs); FSW (female sex workers); and HET (presumed heterosexuals including men and women for whom risk assessment was not available). Missing information for published sequences was retrieved from relevant studies or obtained through communication with study authors.

New HIV-1 sequences from Pakistan (**paper V**) were from an individually matched case-control study[247] that recruited cases, defined as children aged 0–15 years registered for HIV-1 care at the Paediatric Treatment Center, Shaikh Zayed Children's Hospital that was established in response to the outbreak. Overall, we analysed 532 HIV-1 partial *pol* sequences, including outbreak sequences (N=344) and previously published sequences (N=188) representing Pakistani PWID, heterosexuals, sex workers, and other individuals with unknown transmission risks.

Partial HIV-1 *pol* sequencing

The HIV-1 *pol* gene was selected for sequencing based on prior knowledge that it was the most abundant HIV-1 gene among all published Kenyan sequences in the Los Alamos HIV sequence database (and would thus increase sample size in our inference). Also, this region would enable the characterisation of HIV-1 drug resistance.

For **papers I-IV**, HIV-1 RNA was extracted from patient plasma samples using the RNeasy Lipid Tissue Mini Kit (QIAGEN) with modifications from the manufacturer's protocol[248]. Briefly, 100µl patient blood plasma was lysed in 1000µl Qiazol Reagent. An in-column DNase I treatment step was included to ensure the purity of the extracted RNA – before elution in 40µl polymerase chain reaction (PCR)-clean water.

Reverse transcription and amplification of partial HIV-1 *pol* gene were performed using the One-Step Superscript III RT/Platinum Taq High Fidelity Enzyme Mix (ThermoFisher Scientific™) with the *pol*-specific primer pair JA269 and JA272[249]. First-round PCR products were amplified in a nested PCR with DreamTaq Green DNA Polymerase (ThermoFisher Scientific™) using *pol*-specific primers JA271 and JA270[249]. For **paper V**, proviral DNA was extracted from blood samples using Qiagen's QIAamp DNA blood mini kit according to the manufacturer's protocol, followed by HIV-1 *pol* gene nested PCR amplification using a separate set of *pol*-specific primers[250]. PCR products were sequenced in both directions with the nested PCR primers using the BigDye terminator kit v1.1 (Applied Biosystems). New HIV-1 *pol* sequences (approximately 1020 nucleotides (nt), HXB2 (K03455) positions 2267-3287 for Kenyan sequences and approximately 1200 nt, HXB2 positions 2078-3320 for Pakistani sequences) were determined on an ABI PRISM 3130×1 Genetic Analyzer (Applied Biosystems).

HIV-1 subtype analysis

HIV-1 *pol* sequences were aligned with the HIV-1 Group M (subtypes A-K + Recombinants) subtype reference dataset (available at the Los Alamos HIV database, http://www.hiv.lanl.gov) using Clustal X2 (v2.1) and the MAFFT algorithm in Geneious Prime 2019[251,252]. The resulting alignments were used to construct ML phylogenetic trees in PhyML using the general time-reversible substitution model with a gamma-distributed rate variation and proportion of invariant sites (GTR+Γ4+I)[207]. Branch support was assessed using the Shimodaira-Hasegawa-like approximate Likelihood Ratio Test (aLRT-SH) in PhyML, with aLRT-SH ≥0.90 considered as significant[212,213]. The Subtype/CRF-resolved phylogenies were visualized using FigTree v1.4.4 (https://github.com/rambaut/figtree/releases). Unique recombinant forms (URFs) were initially detected using the REGA HIV-1 subtyping tool (version 3) and further characterised by boot-scan analysis in SimPlot [253,254].

Cluster analysis

Based on the subtyping results, sequences were grouped into the main subtypes that were observed in the datasets. For each subtype-specific dataset, a search for related sequences was done separately using the NCBI GenBank BLAST tool, with results limited to a threshold of 10 similar hits per patient sequence, as previously described[6,212,255,256]. Duplicate sequences were removed based on the sequence identifiers and accession numbers. Redundant sequences were manually removed, and every single hit was further explored to identify and exclude previously published volunteer sequences. All sequences were aligned by subtype and subtype-specific phylogenies were reconstructed in PhyML[257].

Country-specific clusters (i.e. Kenyan or Pakistani clusters) were defined as any cluster with aLRT-SH ≥0.90 and comprising ≥80% country-specific sequences, regardless of the genetic distances within the cluster (exemplified for Kenya clusters in **Fig. 19**)[6,212,255,256]. In **paper IV**, potentially active transmission clusters were identified using aLRT-SH ≥0.90, and a genetic distance ≥0.015 substitutions/site in Cluster Picker[212,258].

Identified clusters were classified into dyads (2 sequences), networks (3-14 sequences), or large clusters (>14 sequences)[6].

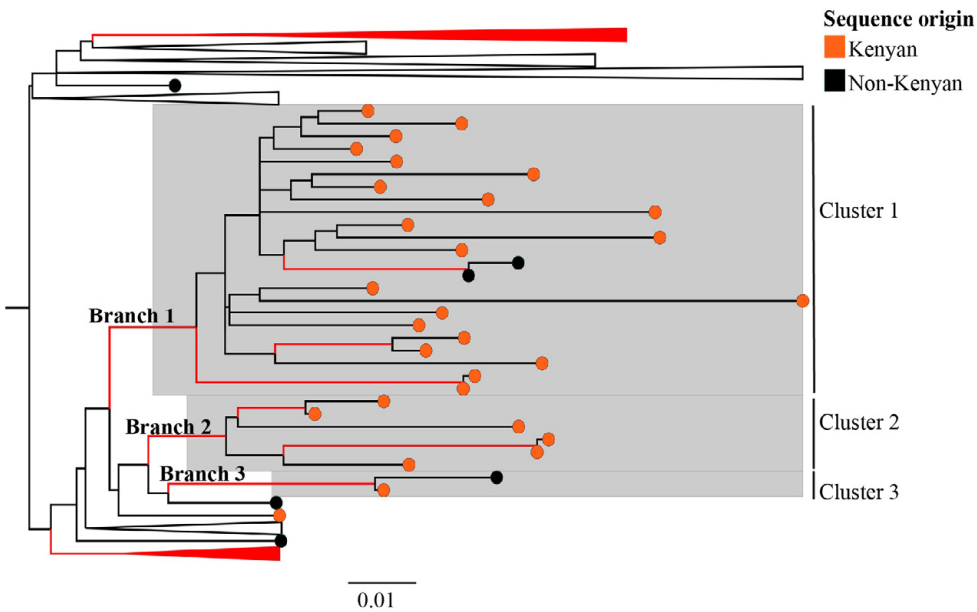


Figure 19. Phylogenetic identification of Kenyan HIV-1 subtype G clusters. A phylogeny of HIV-1 subtype G sequences from Kenya (shown as orange dots), and similar non-Kenyan GenBank reference sequences (black dots). Three monophyletic clusters are highlighted grey – but only cluster 1 and cluster 2 are considered Kenyan clusters as they meet both criteria (i.e. SH-aLRT branch support ≥0.90 and comprising ≥80% Kenyan HIV-1 sequences). Cluster 3 is not Kenyan as per the definition (although it has SH-aLRT branch support ≥0.90, it comprises <80% Kenyan sequences).

Bayesian phylodynamic inference
HIV-1 evolutionary origins and past population dynamics were determined using a Bayesian approach – for sequences with information on sampling dates. The temporal signal of each dataset was initially assessed in TempEst (v1.5.3)[259]. Bayesian inferences were done in BEAST 1.10.4 using the Bayesian Skygrid model with an uncorrelated lognormal relaxed clock and inferred under the GTR+Γ4+I substitution model[217,218,260,261]. Where appropriate, and to enhance precision in estimating evolutionary parameters within and between clusters from different risk groups, a previously described hierarchical phylogenetic model (HPM) was used[262]. BEAST runs of 100 million - 500 million generations were computed, sampling every 10000th-50000th iteration, and discarding the first 10% as burn-in. Convergence was determined in Tracer v.1.7.0 and defined as effective sample sizes (ESS) ≥100[217].

Bayesian phylogeographic inference
A discrete phylogeographic inference using an empirical tree distribution was computed where the expected number of HIV-1 migrations for every pathway were inferred on a branch-by-branch basis (**papers I and III)**, as previously described[31,231]. The geographic area of sampling and/or risk group were used as discrete states. The asymmetric continuous-time Markov chain (CTMC) model was preferentially used as it relaxes the assumption of constant diffusion rates through time to realistically model phylogeographic processes[231,263].

A robust counting approach implemented in BEAST was used to estimate the forward and reverse HIV-1 movement events (Markov jumps) between locations and risk group states along the branches of dated phylogenetic trees[264]. Well-supported movements and Bayes factors (BF) assessing statistical support were summarized using SPREAD v1.0.7, (BF≥3 was considered significant)[231]. Maximum clade credibility (MCC) trees annotated with key demographic/epidemiological data were summarized in Tree-Annotator v1.10.4 (BEAST suite) and visualized in FigTree (v1.4.4).

A phylogeographic analysis is sensitive to sampling size as small sample sizes might not be informative enough to infer migration profiles, and large sample sizes may be too computationally intensive to analyse resulting in evolutionary parameters that fail to converge)[31,231,265]. Therefore, several approaches were used to limit sampling bias arising from the disproportionate allocation of sequences from some discrete states.

In **paper I**, in the first senario, sequences in the sub-subtype A1 dataset (sampled during 2004-2019) were sub-sampled proportional to the HIV-1 prevalence per geographic province. This procedure was independently replicated 30 times – resulting in 30 datasets each having 892 sequences of which 35% were from

Nyanza, 17% Rift Valley, 13% Nairobi, and 7% Coast. A similar approach was taken with risk group as a discrete state – resulting in thirty datasets each having 802 sequences of which 64% were from HET, 14% FSW, 15% MSM, and 4% PWID. Cluster analysis (as described above) was performed independently for each dataset. Clusters having >14 sequences were identified – and discrete state phylogeographic analysis with Markov jumps inferences were then performed independently for each of the identified clusters.

In the second sensitivity analysis, HIV-1 A1 sequences collected during recent years (i.e. 2010-2019) were sub-sampled proportionally as was done in the first scenario – resulting in five independent datasets with location-annotation (each having 144 sequences – 35% from Nyanza, 17% Rift Valley, 13% Nairobi, and 7% Coast), and five independent datasets with risk group annotation (each having 97 sequences – 64% HET, 14% FSW, 15% MSM, and 4% PWID). However, unlike in the cluster-wise approach, the complete sub-sampled datasets were used directly to estimate virus migration between states.

In the third sensitivity analysis, HIV-1 A1 sequences collected during 2010-2019 were sub-sampled uniformly into five datasets with equal number of sequences per discrete state. The location-annotated dataset had 100 sequences (25 sequences from each province), while the dataset annotated for risk group had 108 sequences (27 sequences for each risk group).

Additionally, in **paper III**, the sequences were first grouped by subtype (A1, C and D), and the phylogeographic inference was then performed by subtype using all available sequences. Second, to reduce sampling bias arising from the disproportional allocation of sequences per location, sequences in the sub-subtype A1-specific dataset (the largest of the three subtypes) were randomised and sub-sampled into a dataset with an equal number of sequences per province, followed by phylogeographic inference.

Characterisation of HIV-1 drug resistance

For **papers II and V**, HIV-1 sequences were submitted to the Stanford HIV drug resistance database using the calibrated population resistance tool to screen for drug resistance-associated mutations (http://cpr.stanford.edu/cpr.cgi). Drug resistance mutations were identified based on the WHO list for surveillance of genotypic drug resistance mutations[266,267]. In this thesis, HIVDR mutations detected among ART-naïve individuals were classified as pre-treatment drug resistance whilst those among individuals on treatment were classified as acquired drug resistance. In addition, phylogenetic cluster analysis was repeated, as described above, to assess for clustering amongst isolates identified with surveillance drug resistance mutations.

Statistical analysis
Frequencies and percentages were used to describe the distribution of sequences within the study population. Changes in the proportion of variables over time were assessed using the *nptrend* non-parametric test for trends (a Stata plugin)[268]. Multivariable logistic regression models were used to assess associations between variables, and $p<0.05$ was defined as statistically significant. A Kruskal-Wallis H test and a *post hoc* Dunn's test with Bonferroni correction for multiple comparisons were conducted to determine differences in HIV-1 evolutionary rate, cluster growth rates, and time to the most recent common ancestor (tMRCA) estimates among clusters from multiple risk groups. Statistics and summary plots were done using Stata 15 (College Station, Texas, USA) and RStudio (version 1.2.5001) with the packages: *yarrr, circlize* and *ggplot2*[269-271].

Ethical considerations
Molecular epidemiology studies involve linking virus sequences to patient socio-demographic data. This raises the issue of possible disclosure of confidential and private data, particularly among vulnerable populations. Individuals anonymity was ensured by delinking sequence identifiers with patient identifiers.

For **papers I-IV**, plasma samples used to generate new HIV-1 sequences were obtained from ongoing or concluded studies approved by the Kenya Medical Research Institute (KEMRI) Scientific and Ethics Review Unit (SERU 3747, 3280 and 3520, and SSC 894). Since published sequences were obtained from an open-access public domain, informed consent was not retrospectively obtained. Instead, we sought approval through a study protocol that was reviewed by KEMRI/SERU (SERU 3547).

For **paper V**, the study protocol was reviewed and approved by Aga Khan University Ethics Review Committee (ERC #2019-1536-4200). Written informed consent to participate in this study was provided by the participant's legal guardian/next of kin.

Data availability
Newly generated nucleotide sequences were deposited in GenBank under the following accession numbers: OM109695-OM110282, MT084914-MT085076. (**papers I-IV**), and MN698251, MN698252, MN698253, MN698255, MN698256, MN698257, MN698258, MN698259, MN698260, MN698261, MN698262, MN698263, MN698264, MN752136, MN752137, and MT748850-MT749178 (**paper V**).

Main findings and discussion

Molecular description of the HIV-1 epidemic in Kenya

Main findings

- HIV-1 subtype A (sub-subtype A1) was the dominant subtype among risk groups and geographic locations in Kenya and had increasing proportions between 2004-2019 compared to other identified strains (**papers I-IV**).
- HIV-1 transmission between risk groups was rare, most of the HIV-1 transmission occurred within-risk groups (**papers I and IV**).
- Albeit infrequent, viruses flow mostly from HET to key populations than vice-versa (**paper I**).
- The sub-epidemic among PWID was separate from all other risk groups (**papers I and IV**).
- There was extensive geographic HIV-1 mixing in Kenya and significant transmission from high-to-low HIV-1 prevalence regions (**papers I and IV**).
- In the MSM HIV-1 epidemic, there was more virus flow from Coastal Kenya to other provinces than vice-versa (**paper III**).
- Pre-treatment HIVDR increased from 6.9% 1986-2005 to 24.2% 2016-2020. No integrase-inhibitors drug resistance was detected in Kenya (**paper II**).

HIV-1 subtype A (sub-subtype A1) dominated the epidemic, with increasing proportions (2004-2019)

In **paper I**, we analysed 4058 sequences, of which the majority (N=3401, 83.8%) were HET followed by MSM (N=372, 9.2%), FSW (N=227, 5.6%), and PWID (N=58, 1.4%). Overall, these numbers represent an estimated sampling density of 0.3% of the HIV-1 epidemic in Kenya, and specific sampling densities of 10.8% for MSM, 1.7% for PWID, 0.6% for FSW, and 0.3% for HET. Sequences were available from seven of eight former administrative provinces in Kenya: Nairobi (N=1440, 35.5% of the sequences in this study); Coast (N=1061, 26.2%); Nyanza (N=665, 16.4%); Rift Valley (N=508, 12.5%); Western (N=158, 3.8%); Central (N=44, 1.1%); Eastern (N=6, 0.2%); and 176 (4.3%) sequences with missing data on sampling location.

Irrespective of geographic sampling province or risk group, HIV-1 sub-subtype A1 was the most dominant strain (**Fig. 20**). Temporal trend analysis (2004-2019) in subtype distribution revealed that whilst the proportion in sub-subtype A1 infections increased from 59.7% to 78.3%, (linear-by-linear trend test, $p<0.001$), there was no significant change in subtype C (linear-by-linear trend test, $p=0.30$) or subtype D

(linear-by-linear trend test, p=0.59), whereas subtype G decreased from 1.2% to 0.0%, (linear-by-linear trend test, p=0.013), and CRFs decreased from 2.7% to 0.0% (linear-by-linear trend test, p=0.001, **Fig. 21**).

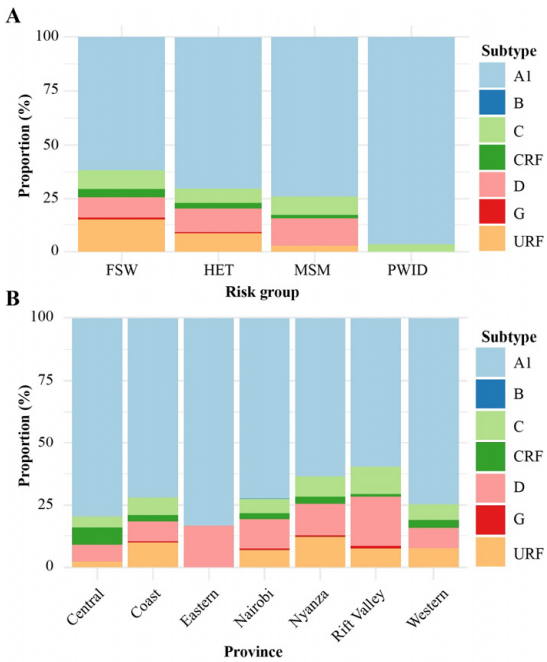


Figure 20. Distribution of HIV-1 subtypes by risk group and geographic provinces in Kenya. HIV-1 subtype A (A1) was the most dominant subtype among (A) different risk groups, and (B) geographic provinces in Kenya. Abbreviations: CRF, circulating recombinant form; URF, unique recombinant form; HET, heterosexual; MSM, men who have sex with men; FSW, female sex work; PWID, people who inject drugs.

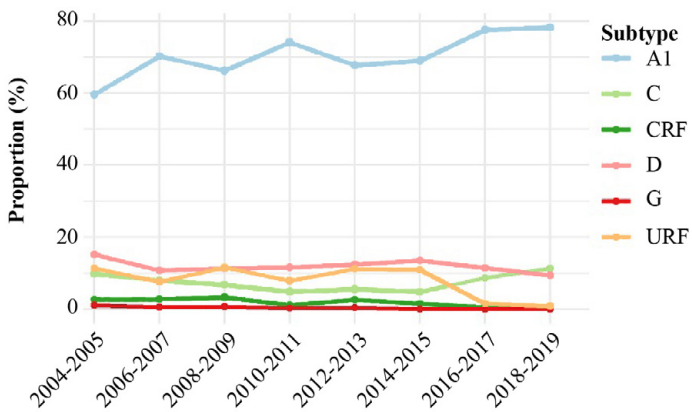


Figure 21. Temporal changes (2004-2019) in the overall proportion of HIV-1 subtypes and recombinants over two-year intervals in Kenya. The proportion of sub-subtype A1 increased significantly over the study period (2004-2019, 59.7% to 78.3%, linear-by-linear trend test, p<0.001).

HIV-1 transmission was compartmentalised by risk groups

We investigated HIV-1 clustering by risk groups: MSM, PWID, FSW, and HET. By integrating HIV-1 phylogenetic and patient epidemiological data, we showed that on a nationwide scale, HIV-1 transmission in Kenya was largely compartmentalized by risk groups. This result was based on the identification of 409 statistically supported phylogenetic clusters (**Table 4**) – where a majority (88.5%) represented within-risk-group clustering (**paper I**). Furthermore, we found that 11.5% of the clusters represented HIV-1 mixing between risk groups – in detail, mixing between MSM/HET (3.7% of all clusters), FSW/HET (3.7%), MSM/FSW/HET (2.2%), MSM/FSW (1.5%), MSM/PWID/FSW/HET (0.2%), and PWID/HET (0.2%). Furthermore, this translated to 7.6% HIV-1 mixing between MSM and HET in Kenya (**paper I**). Sequences from HET females in clusters dominated by MSM sequences provided evidence for heterosexual linkages in these clusters.

These findings at the countrywide scale were also consistent with our phylogenetic estimates specific for the epidemic in Coastal Kenya (**paper IV**) which demonstrated frequent (85%) within-risk group clustering, and minimal (15%) HIV-1 mixing between MSM and the heterosexual population[183]. A previous analysis by Bezemer *et al*. (albeit with a small sample size i.e. N=674) had also concluded that there was infrequent mixing between MSM and HET in Kenya[272].

Combined, finding from our study and the previous analysis by Bezemer and colleagues demonstrate that HIV-1 transmission involved predominantly within-risk group transmission chains.

	Dyads[a]	**Networks[b]**	**Large clusters[c]**	**Total (N,%)**
Subtype				
A (A1)	182 (59%)	105 (34%)	19 (6%)	306 (75%)
C	16 (64%)	8 (32%)	1 (4%)	25 (6%)
D	51 (65%)	27 (35%)	0 (0%)	78 (19%)
Risk category				
HET	204 (65%)	101 (32%)	11 (3%)	316 (77%)
Mixed[*]	24 (51%)	16 (34%)	7 (15%)	47 (11%)
MSM	13 (35%)	23 (62%)	1 (3%)	37 (9%)
FSW	7 (100.0%)	0 (0%)	0 (0%)	7 (2%)
PWID	1 (50%)	0 (0%)	1 (50%)	2 (<1%)
Total	249 (61%)	140 (34%)	20 (5%)	409

Table 4. The distribution of Kenyan HIV-1 clusters (N=409) by HIV-1 subtype and transmission route. Abbreviations: HET, heterosexual transmission; Mixed, clusters with sequences from different risk groups; MSM, men who have sex with men; FSW, female sex work; PWID, people who inject drugs. [*]Risk groups in mixed clusters (N, proportion in all clusters): mixing between MSM/HET (3.7% of all clusters), FSW/HET (3.7%), MSM/FSW/HET (2.2%), MSM/FSW (1.5%), MSM/PWID/FSW/HET (0.2%), and PWID/HET (0.2%). [a]Dyads: clusters of 2 sequences; [b]Networks: clusters of 3-14 sequences; [c]Large clusters: clusters of >14 sequences.

Interestingly, the majority of the MSM in mixed clusters also reported having female sexual partners (i.e. bisexual) – indicating that this group, in addition to female sex workers, could have been an important transmission link to the HET epidemic (**papers I and IV**). Also, the majority (73%) of HIV-1 sequences from transgender people clustered with HIV-1 sequences from MSM – although there was considerable (28.6%) HIV-1 mixing between transgender people and HET, suggesting that transgender people in Kenya could also have linked HIV-1 transmission between HET and MSM. HIV-1 among transgender people (and MSM) has been exceptionally understudied in the African setting. The reason for this has been that these populations are particularly hard to reach in most African countries, which has made it exceedingly difficult to obtain samples from these populations. The HIV-1 sub-epidemics among key populations, and how it influences the mixed epidemic in Africa, warrant further research.

In **paper I**, we phylodynamically quantified the number of virus jumps between different risk groups. We found that when HIV-1 jumped between risk groups it was more often from heterosexual to key populations than vice-versa. In detail, we found significantly more HIV-1 flow from heterosexual to key populations than vice versa (82.9% vs. 12.8%, Student's T-test, $p<0.001$) – these results were confirmed by multiple sensitivity analyses testing the robustness of our data.

It should be emphasised that the detected virus jumps represented rare events because overall transmission between risk groups was itself rare in the Kenyan epidemic. Put otherwise, transitions/jumps between populations identified in the phylogeographic analyses may not be equated with transmission events because phylogeographic modelling only counts jumps between populations. Hence, most transmission events that occur within risk groups were not counted as jumps (as also supported by the clustering analysis).

From **paper I and paper IV,** it is, therefore, possible that HIV-1 key populations may not have had disproportionately transmitted HIV-1 to heterosexuals in the general epidemic, as had been hypothesised previously[13]. Indeed, it has been well established that the vast majority of HIV-1 transmission in Kenya could be attributed to risky heterosexual behaviours[17,273]. However, this needs to be further assessed in future studies based on larger datasets representative of all key populations and the general population.

We found that the HIV-1 sub-epidemic among PWID may have been separate from all other risk groups (**paper 1, and IV**). Amongst PWID (all derived from Coastal Kenya), only two clusters were identified (one dyad and one large cluster, both PWID exclusive), with the large cluster comprising 80% of all PWID sequences in the dataset (N=41). This suggested that the majority of the HIV-1 PWID epidemic

was most likely introduced from one single source followed by a long-term gradual spread within the PWID population – a pattern that distinguished PWID transmission from that of other risk groups in Kenya.

On the other hand, the MSM HIV-1 sub-epidemic showed numerous phylogenetically linked (MSM-exclusive) HIV-1 clusters, consistent with multiple introductions and ongoing infections among MSM within close networks in Kenya (**papers I, III and IV**)[6,274,275]. Half of the clusters comprised sequences collected from MSM from different geographic regions – indicating geographically extensive MSM HIV-1 linkages. High rates of clustering involving HIV-1 in MSM could be linked to an increased risk of infection among MSM within close networks, both in our setting and in other higher-income settings[6,272,274,276,277]. We also estimated that a high proportion (65%) of HIV-1 transmissions occurred between 2000 and 2014 and that several clusters extended over multiple years, suggesting onward HIV-1 transmission among MSM within geographically diverse HIV-1 networks (**paper III**). MSM HIV-1 sequences in this study were not closely related to reference sequences from the global epidemic, implying that the HIV-1 epidemic among MSM in Kenya had been sustained locally.

Overall, our results are of relevance for the roll-out of interventions designed to target key populations. For example, although such targeted interventions may reduce HIV-1 incidence and transmission among the targeted key populations, it is possible that that under similar conditions as in the mixed Kenyan HIV-1 epidemic, targeting key populations may only have a limited effect on HIV-1 incidence in the much larger heterosexual population and other intervention strategies may, therefore, be more cost-effective.

Evidence of geographic HIV-1 mixing in Kenya

Of the 409 HIV-1 clusters identified (**paper I**), the majority (60.6%) were province-exclusive irrespective of transmission risk group, including clusters from Nairobi (26.2% of all clusters), Coast (14.2%), Nyanza (12.5%), Rift Valley (5.6%), Western (1.5%), and Central provinces (0.7%). The remaining clusters (39.4%) were mixed between different geographic provinces.

Pairs of geographic provinces located next to each other were involved in an extensive cyclic HIV-1 exchange – and West-to-East migration accounted for the majority (76.1%) of all within-country jumps (compared to East-to-West migration which accounted for only 23.9% jumps between provinces). These phylodynamic estimates reveal a pattern of HIV-1 transmission from higher-to-lower HIV-1 prevalence regions than vice versa (76.1% vs. 23.9%, Student's T-test, p=0.001; with both uniform and proportional sub-sampling, as well as when restricting the temporal focus to recent years, i.e. 2010-2019). Irrespective of transmission risk, the

largest number of people with HIV-1, and approximately 40% of all newly diagnosed HIV-1 infections in Kenya have occurred in Western Kenya[11].

Interestingly, inter-provincial transmission dynamics in the MSM HIV-1 epidemic (**paper III**) revealed a high proportion of HIV-1 export from Coast province to other provinces (Nairobi and Nyanza) – implying that the Coast province could have been a major geographic source of HIV-1 transmission amongst Kenyan MSM. Of all provinces in Kenya, the Coast province has had the highest prevalence of HIV-1 among MSM[278]. In addition, MSM in Coastal Kenya have been known to be highly mobile, and some reported engaging in sex work in different locations across the country[277]. This might explain the high rates of HIV-1 export from Coast observed in this study.

Taken together results from **papers I and III** demonstrate that transmission dynamics involving HIV-1 key populations in Africa may vary compared to dynamics in the heterosexual epidemic. There is a need to increase HIV-1 research involving different risk populations in Africa.

The effective population size had stabilised at a high level

Phylodynamic analysis investigating the evolutionary dynamics of HIV-1 in various risk groups revealed that the number of effective infections[279] had stabilised at a high level (**Fig. 22**). The estimated trends in past effective population sizes also mirrored overall temporal trends in the Kenyan HIV-1 epidemic (as shown in **Fig. 3**). The epidemic grew exponentially during the mid-1980s to late-1990s but had stabilised during recent years, perhaps owing to the national roll-out of ART in 2004 and increased linkage to HIV-1 care programs which may have reduced the number of new infections.

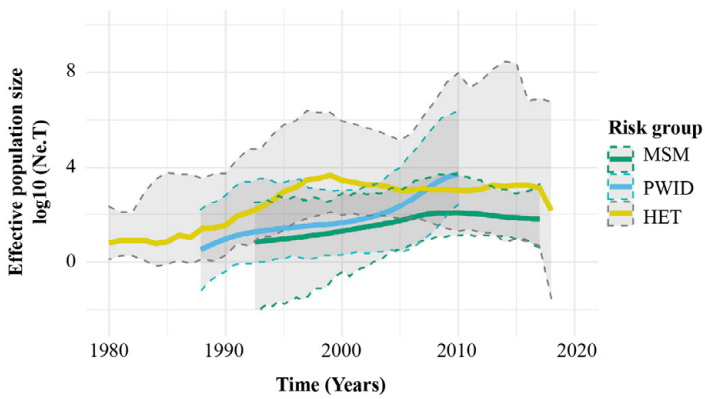


Figure 22. **Population dynamics in the HIV-1 epidemic among HET and mixed-risk group clusters.** Bayesian Skygrid plots showing historical population dynamics in different risk groups in Kenya. Median estimates of the effective population size over time are shown as continuous lines (Yellow, HET; Blue, PWID; and Green, MSM). The shaded area represents the 95% higher posterior density (HPD) intervals of the inferred effective population size.

Likewise, a closer look at the MSM HIV-1 epidemic (**paper III**) revealed a declining effective population size for the dominant strain (HIV-1 A1, 2017-2019, albeit with a broad confidence interval, **Fig. 23**). This could probably reflect benefits from earlier ART initiation[280], risk reduction counselling[281,282], or early recognition of acute HIV-1 infections[283-285]. Some studies have also shown some uptake of pre-exposure prophylaxis among MSM in recent years[286-289]. Although it has been difficult to model the effectiveness of these recent interventions on the epidemic size, future studies may be able to do so using MSM HIV-1 sequences that are currently being sampled in various ongoing cohorts.

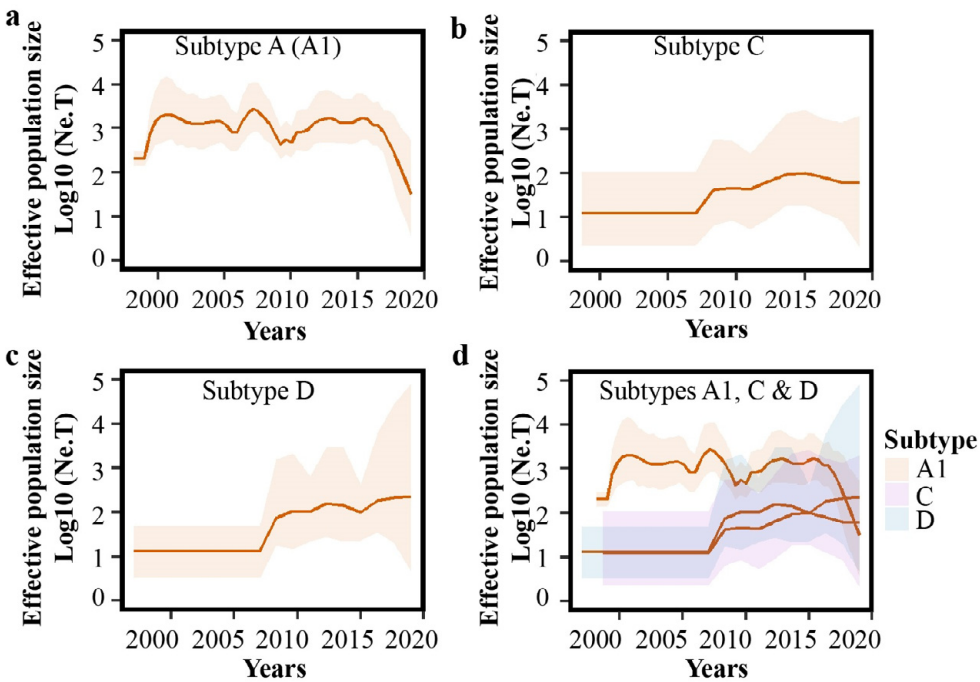


Figure 23. Population dynamics of HIV-1 sub-subtype A1, subtype D and subtype C lineages among MSM in Kenya. Bayesian Skygrid plots showing population dynamics of the (a) HIV-1 sub-subtype A1, (b) HIV-1 subtype C and (c) HIV-1 subtype D lineages in Kenyan MSM. Median estimates of the number of MSM contributing to new infections are shown as continuous lines (Red for sub-subtype A1, Brown for subtype C, and Blue for subtype D). The shaded area represents the 95% higher posterior density (HPD) intervals of the inferred effective population size for each lineage. HIV-1 A1 had declining dynamics approaching 2020.

Overall, in addition to existing broad epidemic control strategies in the general epidemic, increasing access to treatment – as well as de-stigmatisation and diversification of providers may further reduce HIV-1 incidence among key populations in Kenya[245].

High levels of pre-treatment and acquired drug resistance in different risk groups

The overall prevalence of pre-treatment HIVDR in Kenya was high (>15%) – exceeding WHO's 10% threshold for changing first-line NNRTI-based ART to INSTI-based ART[135]. Compared to HET, pre-treatment HIVDR was higher among PWID (adjusted odds ratio, aOR, 3.5, 95% confidence interval, CI: 1.7-5.4, p<0.001), and children (aOR, 4.3, 95% CI: 2.7-7.0, p<0.001, **paper II**).

The high pre-treatment HIVDR observed mostly reflected high levels of NNRTI drug resistance in all risk groups. There was no significant difference in pre-treatment HIVDR between HET and MSM (aOR, 1.0, 95% CI: 0.7-1.5, p=0.921) and between HET and FSW (aOR, 1.1, 95% CI: 0.6-1.7, p=0.842).

Notably, pre-treatment NRTI drug resistance was only moderate (<10%) and there was no pre-treatment HIVDR to INSTIs, a key component of globally recommended regimens for HIV-1 treatment.

Interestingly, whilst overall (and NNRTI) pre-treatment HIVDR among FSW and MSM increased during 2015-2020, there was a decreasing trend among HET during 2016-2020 which coincided with the nationwide transition from NNRTI to INSTI-based ART regimens[290]. Thus, the pre-treatment HIVDR identified against older drug classes (NNRTIs) may not be clinically relevant, and current ART guidelines combining DTG with two NRTIs may remain effective in controlling the nationwide epidemic[290].

In the recent past, ART coverage in Kenya has been much lower among key populations compared to HET in the general population[10]. As of 2020, ART coverage was 73% in FSW, 68% in PWID, and 63% in MSM, compared to 86% in the general (largely heterosexual) epidemic[10]. Low ART coverage (and consequently lower virus suppression rates) among key populations could have increased the risk of transmission of resistant strains within these groups. This may explain the higher proportion of pre-treatment HIVDR among MSM and PWID compared to lower risk HET with better access to HIV-1 treatment services.

The drug resistance levels estimated among key populations in Kenya were higher than estimates from higher-income settings with equitable access to ART and better patient monitoring[291]. Our findings may reflect the status of HIVDR in other global regions with a similar HIV epidemic as Kenya. Further research may be necessary to monitor trends in HIV-1 drug resistance trends among different populations in sSA.

Conclusions from a review of studies on HIV-1 phylogenetic linkages between populations in sSA

In summary, in **paper VI**[9], the HIV-1 sub-epidemics in several African countries appeared to be localized in specific communities (with limited HIV-1 flow between neighboring communities) – although some HIV-1 geographic hotspots have been identified. Furthermore, human migration linked to economic activities (including mining and fishing) may have contributed to increased HIV-1 transmission.

Few studies had investigated HIV-1 clustering between different risk groups in sSA[9]. Overall, in addition to the expected HIV-1 links between HET and FSW owing to sex work, several studies had observed low rates of HIV-1 clustering between HET females and MSM – proof of HIV-1 mixing between MSM and HET. However, HIV-1 mixing appeared to be at relatively low rates across the region (although this had been difficult to quantify empirically because of the shortage of HIV-1 sequence data from MSM, FSW, and PWID). Further research to reveal the factors driving the HIV-1 epidemic in sSA is needed.

Molecular characterization of a paediatric HIV-1 outbreak in Larkana, Pakistan

Main findings

- HIV-1 CRF02_AG and sub-subtype A1 were the dominant HIV-1 variants in the epidemic (**paper V)**.
- Outbreak sequences exhibited no phylogenetic mixing with sequences from other HIV-1 infected key populations in Pakistan (**paper V**).
- Multiple clusters were indicative of a multi-source, and not a single-source outbreak (**paper V**).

Previously, the HIV-1 epidemic in Pakistan has concentrated among PWID, MSM, transgender people, and sex workers[292]. However, in April 2019, an extensive HIV-1 outbreak involving more than 1000 children occurred in the Larkana District, Pakistan[293]. Perinatal transmissions were ruled out because the majority of the children were from HIV-1 seronegative mothers[294]. Clinicians and researchers suspected that HIV-1 transmission in this outbreak was linked to poor infection prevention control practices including reuse of needles and inadequate blood screening. Initial reports alluded to a single-source outbreak[247]. In **paper V**, we investigated HIV-1 clustering patterns using HIV-1 sequences from the outbreak and HIV-1 key populations (PWID, FSW, and MSM)[250].

HIV-1 sequences from the outbreak belonged to different HIV-1 strains (mostly the recombinant CRF02_AG and sub-subtype A1). Four phylogenetically linked clusters within the outbreak were identified. The HIV-1 outbreak sequences exhibited no phylogenetic mixing with sequences from other key populations of

Pakistan. The presence of multiple clusters of different subtypes was evidence of several introductions of HIV-1 into the children population in Larkana, contrasting speculations of a single source outbreak from a single health practitioner[247]. The median tMRCAs of the Larkana outbreak sequences were estimated to 2016 for both the CRF02_AG and the subtype A1 clusters. This suggested longstanding transmissions going back several years before the reported outbreak in 2019.

The multiple introductions were likely a consequence of ongoing transmission within key populations in Larkana, and possibly, the Larkana strain may have been introduced into the general population through poor infection prevention control practices in healthcare settings[295-297]. Overall, the study (**paper V**) highlights the need to scale up HIV-1 prevention programmes among different population groups, to improve blood safety and infection prevention control, and to eliminate structural and social barriers to health service delivery to different HIV key populations in Pakistan, and other LMICs.

Limitations, potential solutions, and future perspectives

Determining the driving factors in the global HIV-1 epidemic may be important to guide targeted HIV-1 prevention[298]. Phylogenetic methods could help in characterizing such driving factors but rely on the availability of densely sampled virus sequences obtained from well-characterised cohorts. Unfortunately, HIV-1 key populations are particularly hard to reach in most African countries, which makes it exceedingly difficult to obtain virus samples from these populations. However, where some HIV-1 data are available from key populations (as in this dissertation), it is possible to discern patterns in HIV-1 spread between various risk groups, even in a mixed epidemic.

Phylogenetic clustering represents indirect evidence of epidemiologic linkages and might not fully represent the true transmission networks – especially in datasets with low sequence coverage in the studied epidemic. Studies with well characterised patient demographics and dense sampling among infected individuals have provided useful information for HIV-1 prevention (especially in Europe and Northern America settings)[4,5,274,299-303]. However, low sampling density (and the shortage of HIV-1 sequences from key populations) is a constant limitation to phylogenetic studies in sSA[9]. A low sampling density generally results in missing links and smaller HIV-1 clusters, and may therefore limit the reliability of phylogenetic inference[304].

In sSA, few phylogenetic studies have investigated HIV-1 clustering between HIV-1 key populations and HET (summarized in the review **paper VI)**[9]. In addition to the expected HIV-1 links between HET and FSW owing to sex work, several other studies have reported low rates of HIV-1 clustering between HET females and MSM – concrete proof of HIV-1 mixing between MSM and HET. Likewise, findings from studies in this dissertation (**paper I and IV**) support the hypothesis of limited HIV-1 mixing between HET and key populations (MSM and PWID) in Kenya. This fi ding likely applies to other countries with a similar epidemic in sSA.

However, future studies in sSA need to achieve larger and proportional sample coverage across all risk groups and geographic locations. Concurrent with increasing sampling coverage, emphasis should be made to capture patient demographic information (and data on sampling dates and location) during sampling. We observed that many publicly available HIV-1 sequences from Africa lack accompanying patient demographic data. Despite this shortcoming, such published sequences still represent an incredible source of HIV-1 sequence data (although findings based on such sequences may require careful interpretation). For

instance, in the case of published sequences lacking risk data, based on phylogenetic clustering, the probable risk group for nodes within a cluster with inadequate annotation may be deduced from association with nodes with a known risk group. This approach was recently used to identify potential non-disclosed MSM (self-reported HET men who clustered only with men) in the United Kingdom[301]. Yet, given that low HIV-1 sampling in sSA may persist in the unforeseen future, statistical and or phylogenetic models that control for missed sampling may need to be developed.

While phylogenetic models may reveal and quantify the movement of viruses between locations and risk groups, they are limited in the in-depth determination of how and where virus transmission have occurred without additional information, e.g. on human movement. Residents in a community may get infected while living or travelling outside their homes and such external introductions could be further disentangled by combining movement and migration data with virus data. Although these mobility methodologies have been developed and used to quantify the impact of human mobility on malaria transmission in different African countries, their application in deciphering HIV-1 transmission is limited[305-307]. Therefore, there is a need to incorporate mobility networks into the phylogenetic spatiotemporal models to quantify the HIV-1 movement patterns and links between locations and communities more precisely.

Overall, some consortia (such as the MSM Health Research Consortium in Kenya and PANGEA-HIV: phylogenetics for generalized epidemics in Africa) have been established to sample various populations across Africa. There are also plans to expand sampling – especially among MSM and other HIV key populations in various countries in sSA. It is possible that a more homogenous and dense sampling from the participating countries may be achieved in the near future to improve and strengthen the limitations of phylodynamic methods in characterizing the mixed HIV-1 epidemic in sSA[8].

Acknowledgements

I acknowledge with appreciation all the help I have received in the years during which studies in this thesis were conceived and realised.

To my PhD advisors: Thank you, Joakim Esbjörnsson – you have been incredibly supportive all along, and I am glad to have been your PhD student. I learned a lot from you as a person, a supervisor, and a career mentor. I am extremely grateful for the opportunity to belong and to grow scientifically in your research group. Thank you Eduard Sanders for taking a chance on me – and for your hard work to advance research on key populations in Africa. Your resilience is admirable and has rubbed off on me! Thank you, Amin Hassan, for your patience and cool demeanour – and for your very well-informed guidance in designing the projects within and beyond this dissertation. Thank you Patrik Medstrand for being the nicest group leader and for accepting to co-supervise my PhD research. Thanks for providing comments which greatly improved studies in this thesis – as well as other studies in the pipeline. To Kamini Gounder, thank you for believing in my PhD projects – and for being part of some projects which we could not accomplish due to global restrictions following the COVID-19 Pandemic. I am sure we will build on these studies more in the coming year!

Thank you Marianne Jansson for your kindness and for providing opportunities to gain teaching experience – this was a lot of fun! To my colleagues and friends from the Lund University Virus Center (Sara, Malin N, Angelica, Nordine Zsófia, Sviataslau, Jamirah, Emil, Dawit, Oskar, Fregenet, Christian, Uroosa, Sugata, Malin H, Isak, Annie, Felicia, Lykke, Claus, Corrado, and Luis) and at the IAVI CRC, KEMRI-Wellcome Trust Research Program in Kilifi, Kenya (Elise van der Elst, Eunice Nduati, Lucie Ikumi, Keith Kipoto, Makobu Kimani, Elizabeth Wahome, Ian Oyaro, Sharon, Kelly Ramko, Caroline Ngetsa, James Chemweno, Karanja, Clara Agutu, Maartje Dijkstra, Shaun Palmer, and Sein Yiakon) – thank you for the many thoughtful discussions and a lot of fun events! It was a great pleasure to work with all of you and I hope you will remember these experiences as fondly as I do.

Special thanks to SANTHE for funding my PhD studies and for contributing to research capacity building among African scientists. Professor Thumbi Ndung'u, you have set the bar high, and I am glad that the next generation of African scientists has you to emulate. A big thanks also to the SANTHE fraternity (Denis, Kim, Sipho, Victoria and David Lavu, and all my fellow students and colleagues affiliated with SANTHE). Special thanks to IAVI, Swedish Research Council, and STINT for offering additional funds to support studies in this dissertation. I would also like to thank all the people involved in student capacity building including the IDeAL team (Sam Kinyanjui, Dorcas Mbuvi, Rita Baya, Florence Kirimi, and Charles Kamau)

at KEMRI-WTRP, and the doctoral administration at Lund University, without whom this thesis could not have been completed.

Philippe Lemey, and Guy Baele, Katholieke Universiteit Leuven – thank you for a nice exchange of ideas on Bayesian analyses. Special thanks to all co-authors in the studies that are part of this thesis – I appreciate your great feedback and brilliant suggestions. To my study participants – I am incredibly proud of you. Thank you for your contribution to science. I am grateful to members of the MSM Health Research Consortium (MHRC) and the staff in the HIV/STI project at the KEMRI-WTRP for their commitment to serving HIV-1 key populations. I am also grateful to all the people I have interacted with at conferences, seminars, workshops and elsewhere – who have provided useful feedback, and an avenue to discuss phylogenetic inference.

Finally, I will always be grateful to my family, and friends, for their constant support and encouragement along the way.

References

1 Joint United Nations Programme on HIV/AIDS (UNAIDS). *Seizing the moment, tackling entrenched inequalities to end epidemics, global AIDS update*, <https://aids2020.unaids.org/report/> (2020).

2 Joint United Nations Programme on HIV/AIDS. *Global data on HIV epidemiology and response*, <https://aidsinfo.unaids.org/> (2021).

3 Bain, L. E., Nkoke, C. & Noubiap, J. J. N. UNAIDS 90–90–90 targets to end the AIDS epidemic by 2020 are not realistic: comment on "Can the UNAIDS 90–90–90 target be achieved? A systematic analysis of national HIV treatment cascades". *BMJ global health* **2**, e000227 (2017).

4 Poon, A. F. *et al.* Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *The lancet HIV* **3**, e231-e238 (2016).

5 Ratmann, O. *et al.* Sources of HIV infection among men having sex with men and implications for prevention. *Science translational medicine* **8**, 320ra322-320ra322 (2016).

6 Esbjörnsson, J. *et al.* HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. *Virus evolution* **2**, vew010 (2016).

7 Joint United Nations Programme on HIV/AIDS (UNAIDS). *UNAIDS DATA 2018*, <https://www.unaids.org/sites/default/files/media_asset/unaids-data-2018_en.pdf> (2018).

8 Abeler-Dörner, L., Grabowski, M. K., Rambaut, A., Pillay, D. & Fraser, C. PANGEA-HIV 2: phylogenetics and networks for generalised epidemics in Africa. *Current Opinion in HIV and AIDS* **14**, 173 (2019).

9 Nduva, G. M., Nazziwa, J., Hassan, A. S., Sanders, E. J. & Esbjörnsson, J. The Role of Phylogenetics in Discerning HIV-1 Mixing among Vulnerable Populations and Geographic Regions in Sub-Saharan Africa: A Systematic Review. *Viruses* **13**, 1174 (2021).

10 (UNAIDS), J. U. N. P. o. H. A. *UNAIDS DATA 2021*, <https://www.unaids.org/sites/default/files/media_asset/JC3032_AIDS_Data_book_2021_En.pdf > (2021).

11 National AIDS and STI Control Programme (NASCOP). *Preliminary KENPHIA 2018 Report*, <https://www.nascop.or.ke/kenphia-report> (2020).

12 Kenya National AIDS Control Council. *Kenya AIDS Strategic Framework 2014/2015–2018/2019*, <http://nacc.or.ke/wp-content/uploads/2015/09/KASF_Final.pdf> (2019).

13 National AIDS and STI Control Programme. *Kenya HIV County Profiles 2016.*, <http://nacc.or.ke/wp-content/uploads/2016/12/Kenya-HIV-County-Profiles-2016.pdf> (2017).

14 Kenya National Bureau of Statistics. *2019 Kenya population and housing census Volume 1: Population by county and sub-county*, <https://www.knbs.or.ke/?wpdmpro=2019-kenya-population-and-housing-census-volume-i-population-by-county-and-sub-county> (2019).

15 Kenya National AIDS control council (NACC). *Kenya HIV estimates report 2018*, <https://nacc.or.ke/wp-content/uploads/2018/11/HIV-estimates-report-Kenya-20182.pdf> (2018).

16 National AIDS and STI Control Programme (NASCOP). *Key Population Mapping and Size Estimation in Selected Counties in Kenya: Phase 1*, <https://hivpreventioncoalition.unaids.org/wp-content/uploads/2020/02/KPSE-Phase1-Final-Report.pdf> (2019).

17 Kenya National AIDS Control Council. *Kenya HIV Prevention Response and Modes of Transmission Analysis.*, <https://icop.or.ke/wp-content/uploads/2016/09/KenyaMOT-2009.pdf> (2009).

18 Smith, A. D., Tapsoba, P., Peshu, N., Sanders, E. J. & Jaffe, H. W. Men who have sex with men and HIV/AIDS in sub-Saharan Africa. *The Lancet* **374**, 416-422 (2009).

19 Dwyer-Lindgren, L. *et al.* Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature* **570**, 189-193 (2019).

20 Control, C. f. D. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men-New York City and California. *mmwr* **30**, 305-308 (1981).

21 Greene, W. C. A history of AIDS: looking back to see ahead. *European journal of immunology* **37 Suppl 1**, S94-102, doi:10.1002/eji.200737441 (2007).

22 Sharp, P. M. & Hahn, B. H. Origins of HIV and the AIDS pandemic. *Cold Spring Harbor perspectives in medicine* **1**, a006841 (2011).

23 Barré-Sinoussi, F. *et al.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**, 868-871 (1983).

24 Gallo, R. C. *et al.* Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *science* **224**, 500-503 (1984).

25 Popovic, M., Sarngadharan, M. G., Read, E. & Gallo, R. C. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science* **224**, 497-500 (1984).

26 Clavel, F. *et al.* Isolation of a new human retrovirus from West African patients with AIDS. *Science* **233**, 343-346 (1986).

27 Letvin, N. L. *et al.* Acquired immunodeficiency syndrome in a colony of macaque monkeys. *Proceedings of the National Academy of Sciences* **80**, 2718-2722 (1983).

28 Daniel, M. *et al.* Isolation of T-cell tropic HTLV-III-like retrovirus from macaques. *Science* **228**, 1201-1204 (1985).

29 Sharp, P. M., Robertson, D. L., Gao, F. & Hahn, B. H. Origins and diversity of human immunodeficiency viruses. *Aids* **8**, S27-S42 (1994).

30 Hahn, B. H., Shaw, G. M., De, K. M. & Sharp, P. M. AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607-614 (2000).

31 Faria, N. R. *et al.* The early spread and epidemic ignition of HIV-1 in human populations. *science* **346**, 56-61 (2014).

32 Lemey, P. *et al.* Tracing the origin and history of the HIV-2 epidemic. *Proceedings of the National Academy of Sciences* **100**, 6588-6592 (2003).

33 Faria, N. R. *et al.* Phylogeographical footprint of colonial history in the global dispersal of human immunodeficiency virus type 2 group A. *The Journal of general virology* **93**, 889 (2012).

34 van der Loeff, M. F. S. *et al.* Sixteen years of HIV surveillance in a West African research clinic reveals divergent epidemic trends of HIV-1 and HIV-2. *International journal of epidemiology* **35**, 1322-1328 (2006).

35 Hamel, D. J. *et al.* Twenty years of prospective molecular epidemiology in Senegal: changes in HIV diversity. *AIDS research and human retroviruses* **23**, 1189-1196 (2007).

36 Månsson, F. *et al.* Prevalence and incidence of HIV-1 and HIV-2 before, during and after a civil war in an occupational cohort in Guinea-Bissau, West Africa. *Aids* **23**, 1575-1582 (2009).

37 Biague, A. *et al.* High sexual risk taking and diverging trends of HIV-1 and HIV-2 in the military of Guinea Bissau. *The Journal of Infection in Developing Countries* **4**, 301-308 (2010).

38 Månsson, F. *et al.* Trends of HIV-1 and HIV-2 prevalence among pregnant women in Guinea-Bissau, West Africa: possible effect of the civil war 1998–1999. *Sexually transmitted infections* **83**, 463-467 (2007).

39 Olesen, J. S. *et al.* HIV-2 continues to decrease, whereas HIV-1 is stabilizing in Guinea-Bissau. *Aids* **32**, 1193-1198 (2018).

40 Narayan, O. & Clements, J. E. Biology and pathogenesis of lentiviruses. *Journal of General Virology* **70**, 1617-1639 (1989).

41 Berger, E. A. *et al.* A new classification for HIV-1. *Nature* **391**, 240-240 (1998).

42 Ruelas, D. & Greene, W. An integrated overview of HIV-1 latency. *Cell* **155**, 519-529 (2013).

43 Ndung'u, T., McCune, J. M. & Deeks, S. G. Why and where an HIV cure is needed and how it might be achieved. *Nature* **576**, 397-405 (2019).

44 Robertson, D. L. *et al.* HIV-1 nomenclature proposal. *Science* **288**, 55-55 (2000).

45 Hemelaar, J. *et al.* Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis. *The Lancet infectious diseases* **19**, 143-155 (2019).

46 Bbosa, N., Kaleebu, P. & Ssemwanga, D. HIV subtype diversity worldwide. *Current Opinion in HIV and AIDS* **14**, 153-160, doi:10.1097/coh.0000000000000534 (2019).

47 LANL HIV-1 Database. *HIV Circulating Recombinant Forms (CRFs)*, <https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html> (2021).

48 Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711-716 (2009).

49 Blood, G. A. C. Human immunodeficiency virus (HIV). *Transfusion Medicine and Hemotherapy* **43**, 203 (2016).

50 Uchtenhagen, H. *et al.* Crystal structure of the HIV-2 neutralizing Fab fragment 7C8 with high specificity to the V3 region of gp125. *PLoS One* **6**, e18767 (2011).

51 Rizzuto, C. D. *et al.* A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science* **280**, 1949-1953 (1998).

52 Scarlatti, G. *et al.* In vivo evolution of HIV-1 co-receptor usage and sensitivity to chemokine-mediated suppression. *Nature medicine* **3**, 1259-1265 (1997).

53 Karlsson, I. *et al.* HIV biological variability unveiled: frequent isolations and chimeric receptors reveal unprecedented variation of coreceptor use. *Aids* **17**, 2561-2569 (2003).

54 Doranz, B. J. *et al.* A dual-tropic primary HIV-1 isolate that uses fusin and the β-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. *Cell* **85**, 1149-1158 (1996).

55 Choe, H. *et al.* The β-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. *Cell* **85**, 1135-1148 (1996).

56 Rucker, J. *et al.* Utilization of chemokine receptors, orphan receptors, and herpesvirus-encoded receptors by diverse human and simian immunodeficiency viruses. *Journal of Virology* **71**, 8999-9007 (1997).

57 Deng, H., Unutmaz, D., KewalRamani, V. N. & Littman, D. R. Expression cloning of new receptors used by simian and human immunodeficiency viruses. *Nature* **388**, 296-300 (1997).

58 Mild, M. & Avhandling, A. *Intrapatient evolution of HIV-1 in the context of coreceptor usage*. (Citeseer, 2007).

59 Weissenhorn, W., Dessen, A., Harrison, S., Skehel, J. & Wiley, D. Atomic structure of the ectodomain from HIV-1 gp41. *Nature* **387**, 426-430 (1997).

60 Miyauchi, K., Kim, Y., Latinovic, O., Morozov, V. & Melikyan, G. B. HIV enters cells via endocytosis and dynamin-dependent fusion with endosomes. *Cell* **137**, 433-444 (2009).

61 Suzuki, Y. & Craigie, R. The road to chromatin—nuclear entry of retroviruses. *Nature Reviews Microbiology* **5**, 187-196 (2007).

62 Pathak, V. K. & Hu, W.-S. in *Seminars in virology.* 141-150 (Elsevier).

63 Farnet, C. M. & Haseltine, W. A. Determination of viral proteins present in the human immunodeficiency virus type 1 preintegration complex. *Journal of virology* **65**, 1910-1915 (1991).

64 Henderson, L., Sowder, R., Copeland, T., Benveniste, R. & Oroszlan, S. Isolation and characterization of a novel protein (X-ORF product) from SIV and HIV-2. *Science* **241**, 199-201 (1988).

65 Bednarik, D. P. & Folks, T. M. Mechanisms of HIV-1 latency. *AIDS (London, England)* **6**, 3-16 (1992).

66 Guerrero, S. *et al.* HIV-1 replication and the cellular eukaryotic translation apparatus. *Viruses* **7**, 199-218 (2015).

67 Cann, A. J. & Karn, J. Molecular biology of HIV: new insights into the virus life-cycle. *Aids* **3**, S19-34 (1989).

68 Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582-1586 (1996).

69 Song, H. *et al.* Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nature communications* **9**, 1-15 (2018).

70 Esbjörnsson, J. *HIV-1 evolution, disease progression and molecular epidemiology of HIV-1 single and HIV-1 and HIV-2 dual-infected individuals in Guinea-Bissau*. (Lund University, 2010).

71 Shaw, G. M. & Hunter, E. HIV transmission. *Cold Spring Harbor perspectives in medicine* **2**, a006965 (2012).

72 Leynaert, B., Downs, A. M., de Vincenzi, I. & HIV, E. S. G. o. H. T. o. Heterosexual transmission of human immunodeficiency virus: variability of infectivity throughout the course of infection. *American journal of epidemiology* **148**, 88-96 (1998).

73 DeGruttola, V., Seage III, G. R., MAYER, K. H. & Horsburgh Jr, C. R. Infectiousness of HIV between male homosexual partners. *Journal of clinical epidemiology* **42**, 849-856 (1989).

74 Hudgens, M. G. *et al.* Subtype-specific transmission probabilities for human immunodeficiency virus type 1 among injecting drug users in Bangkok, Thailand. *American journal of Epidemiology* **155**, 159-168 (2002).

75 Baggaley, R. F., Boily, M.-C., White, R. G. & Alary, M. Risk of HIV-1 transmission for parenteral exposure and blood transfusion: a systematic review and meta-analysis. *Aids* **20**, 805-812 (2006).

76 Patel, P. *et al.* Estimating per-act HIV transmission risk: a systematic review. *AIDS (London, England)* **28**, 1509 (2014).

77 Quinn, T. C. *et al.* Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. *The New England journal of medicine* **342**, 921-929, doi:10.1056/nejm200003303421303 (2000).

78 Fideli, U. S. *et al.* Virologic and immunologic determinants of heterosexual transmission of human immunodeficiency virus type 1 in Africa. *AIDS Res Hum Retroviruses* **17**, 901-910, doi:10.1089/088922201750290023 (2001).

79 Miller, W. C., Rosenberg, N. E., Rutstein, S. E. & Powers, K. A. The role of acute and early HIV infection in the sexual transmission of HIV. *Current Opinion in HIV and AIDS* **5**, 277 (2010).

80 Eisinger, R. W., Dieffenbach, C. W. & Fauci, A. S. HIV viral load and transmissibility of HIV infection: undetectable equals untransmittable. *Jama* **321**, 451-452 (2019).

81 Auvert, B. *et al.* Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. *PLoS medicine* **2**, e298 (2005).

82 Bailey, R. C. *et al.* Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *The lancet* **369**, 643-656 (2007).

83 Coates, T. J., Richter, L. & Caceres, C. Behavioural strategies to reduce HIV transmission: how to make them work better. *The Lancet* **372**, 669-684 (2008).

84 Curran, K. *et al.* HIV-1 prevention for HIV-1 serodiscordant couples. *Current HIV/AIDS Reports* **9**, 160-170 (2012).

85 Baeten, J. M. *et al.* Use of a vaginal ring containing dapivirine for HIV-1 prevention in women. *New England Journal of Medicine* **375**, 2121-2132 (2016).

86 Rhodes, T. *et al.* Is the promise of methadone Kenya's solution to managing HIV and addiction? A mixed-method mathematical modelling and qualitative study. *BMJ open* **5**, e007198 (2015).

87 Calabrese, S. K. & Mayer, K. H. Providers should discuss U= U with all patients living with HIV. *The lancet HIV* **6**, e211-e213 (2019).

88 Smith, P. *et al.* Undetectable= untransmittable (U= U) messaging increases uptake of HIV testing among men: results from a pilot cluster randomized trial. *AIDS and Behavior* **25**, 3128-3136 (2021).

89 McMichael, A. J., Borrow, P., Tomaras, G. D., Goonetilleke, N. & Haynes, B. F. The immune response during acute HIV-1 infection: clues for vaccine development. *Nature Reviews Immunology* **10**, 11-23 (2010).

90 Margolick, J. B. *et al.* Failure of T-cell homeostasis preceding AIDS in HIV-1 infection. *Nature medicine* **1**, 674-680 (1995).

91 Tindall, B. & Cooper, D. A. Primary HIV infection: host responses and intervention strategies. *Aids* **5**, 1-14 (1991).

92 Epstein, F., Pantaleo, G., Graziosi, C. & Fauci, A. The immunopathogenesis of human immunodeficiency virus infection. *N. Engl. J. Med* **328**, 327-335 (1993).

93 Fraser, C., Hollingsworth, T. D., Chapman, R., de Wolf, F. & Hanage, W. P. Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis. *Proceedings of the National Academy of Sciences* **104**, 17441-17446, doi:10.1073/pnas.0708559104 (2007).

94 Lavreys, L. *et al.* Higher set point plasma viral load and more-severe acute HIV type 1 (HIV-1) illness predict mortality among high-risk HIV-1–infected African women. *Clinical Infectious Diseases* **42**, 1333-1339 (2006).

95 Sterling, T. R. *et al.* Initial plasma HIV-1 RNA levels and progression to AIDS in women and men. *The New England journal of medicine* **344**, 720-725, doi:10.1056/nejm200103083441003 (2001).

96 Kiwanuka, N. *et al.* 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**, 707-713, doi:10.1086/527416 (2008).

97 Wymant, C. *et al.* A highly virulent variant of HIV-1 circulating in the Netherlands. *Science* **375**, 540-545 (2022).

98 McPhee, E. *et al.* The interaction of HIV set point viral load and subtype on disease progression. *AIDS research and human retroviruses* **35**, 49-51 (2019).

99 Pantaleo, G. & Fauci, A. S. New concepts in the immunopathogenesis of HIV infection. *Annual review of immunology* **13**, 487-512 (1995).

100 Arrildt, K. T., Joseph, S. B. & Swanstrom, R. The HIV-1 env protein: a coat of many colors. *Current HIV/AIDS reports* **9**, 52-63, doi:10.1007/s11904-011-0107-3 (2012).

101 Penn, M. L., Grivel, J.-C., Schramm, B., Goldsmith, M. A. & Margolis, L. 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 (1999).

102 Bruchfeld, J., Correia-Neves, M. & Källenius, G. Tuberculosis and HIV coinfection. *Cold Spring Harbor perspectives in medicine* **5**, a017871 (2015).

103 Huang, L. *et al.* HIV-associated Pneumocystis pneumonia. *Proceedings of the American Thoracic Society* **8**, 294-300 (2011).

104 Perry, C. M. Maraviroc. *Drugs* **70**, 1189-1213 (2010).

105 Lalezari, J. P. *et al.* Enfuvirtide, an HIV-1 fusion inhibitor, for drug-resistant HIV infection in North and South America. *New England Journal of Medicine* **348**, 2175-2185 (2003).

106 Hervey, P. S. & Perry, C. M. Abacavir. *Drugs* **60**, 447-479 (2000).

107 Bang, L. M. & Scott, L. J. Emtricitabine. *Drugs* **63**, 2413-2424 (2003).

108 Jarvis, B. & Faulds, D. Lamivudine. *Drugs* **58**, 101-141 (1999).

109 Gallant, J. E. & Deresinski, S. Tenofovir disoproxil fumarate. *Clinical Infectious Diseases* **37**, 944-950 (2003).

110 Langtry, H. D. & Campoli-Richards, D. M. Zidovudine. *Drugs* **37**, 408-450 (1989).

111 Feng, M. *et al.* Doravirine suppresses common nonnucleoside reverse transcriptase inhibitor-associated mutants at clinically relevant concentrations. *Antimicrobial Agents and Chemotherapy* **60**, 2241-2247 (2016).

112 Adkins, J. C. & Noble, S. Efavirenz. *Drugs* **56**, 1055-1064 (1998).

113 Schöller-Gyüre, M., Kakuda, T. N., Raoof, A., De Smedt, G. & Hoetelmans, R. M. Clinical pharmacokinetics and pharmacodynamics of etravirine. *Clinical pharmacokinetics* **48**, 561-574 (2009).

114 Mirochnick, M., Clarke, D. F. & Dorenbaum, A. Nevirapine. *Clinical pharmacokinetics* **39**, 281-293 (2000).

115 Sharma, M. & Saravolatz, L. D. Rilpivirine: a new non-nucleoside reverse transcriptase inhibitor. *Journal of Antimicrobial Chemotherapy* **68**, 250-256 (2013).

116 Cvetkovic, R. S. & Goa, K. L. Lopinavir/ritonavir. *Drugs* **63**, 769-802 (2003).

117 Plosker, G. L. & Noble, S. Indinavir. *Drugs* **58**, 1165-1203 (1999).

118 Walmsley, S. *et al.* Lopinavir–ritonavir versus nelfinavir for the initial treatment of HIV infection. *New England Journal of Medicine* **346**, 2039-2046 (2002).

119 Harrison, T. S. & Scott, L. J. Atazanavir. *Drugs* **65**, 2309-2336 (2005).

120 Shimura, K. *et al.* Broad antiretroviral activity and resistance profile of the novel human immunodeficiency virus integrase inhibitor elvitegravir (JTK-303/GS-9137). *Journal of virology* **82**, 764-774 (2008).

121 Katlama, C. & Murphy, R. Dolutegravir for the treatment of HIV. *Expert opinion on investigational drugs* **21**, 523-530 (2012).

122 Tsiang, M. *et al.* Antiviral activity of bictegravir (GS-9883), a novel potent HIV-1 integrase strand transfer inhibitor with an improved resistance profile. *Antimicrobial agents and chemotherapy* **60**, 7086-7097 (2016).

123 Buzón, M. J. *et al.* HIV-1 replication and immune dynamics are affected by raltegravir intensification of HAART-suppressed subjects. *Nature medicine* **16**, 460-465 (2010).

124 Link, J. O. *et al.* Clinical targeting of HIV capsid protein with a long-acting small molecule. *Nature* **584**, 614-618 (2020).

125 Singh, K. *et al.* GS-CA compounds: first-in-class HIV-1 capsid inhibitors covering multiple grounds. *Frontiers in microbiology* **10**, 1227 (2019).

126 Fischl, M. A. *et al.* The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New England Journal of Medicine* **317**, 185-191 (1987).

127 Vella, S., Schwartländer, B., Sow, S. P., Eholie, S. P. & Murphy, R. L. The history of antiretroviral therapy and of its implementation in resource-limited areas of the world. *Aids* **26**, 1231-1241 (2012).

128 Palella Jr, F. J. *et al.* Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine* **338**, 853-860 (1998).

129 Gulick, R. M. *et al.* Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. *New England Journal of Medicine* **337**, 734-739 (1997).

130 Hammer, S. M. *et al.* A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine* **337**, 725-733 (1997).

131 Brady, M. T. *et al.* Declines in mortality rates and changes in causes of death in HIV-1-infected children during the HAART era. *Journal of acquired immune deficiency syndromes (1999)* **53**, 86 (2010).

132 Margolis, A. M., Heverling, H., Pham, P. A. & Stolbach, A. A review of the toxicity of HIV medications. *Journal of Medical Toxicology* **10**, 26-39 (2014).

133 Melikian, G. L. *et al.* Non-nucleoside reverse transcriptase inhibitor (NNRTI) cross-resistance: implications for preclinical evaluation of novel NNRTIs and clinical genotypic resistance testing. *Journal of Antimicrobial Chemotherapy* **69**, 12-20 (2014).

134 Villa, G. *et al.* Drug resistance outcomes of long-term ART with tenofovir disoproxil fumarate in the absence of virological monitoring. *Journal of Antimicrobial Chemotherapy* **73**, 3148-3157 (2018).

135 Organization, W. H. Guidelines on the public health response to pretreatment HIV drug resistance: July 2017. (2017).

136 Osiyemi, O. *et al.* Efficacy and Safety of Switching to Dolutegravir/Lamivudine (DTG/3TC) Versus Continuing a Tenofovir Alafenamide–Based 3-or 4-Drug Regimen for Maintenance of Virologic Suppression in Adults Living With HIV-1: Results Through Week 144 From the Phase 3, Non-inferiority TANGO Randomized Trial. *Clinical Infectious Diseases* (2022).

137 Llibre, J. M. *et al.* Efficacy, safety, and tolerability of dolutegravir-rilpivirine for the maintenance of virological suppression in adults with HIV-1: phase 3, randomised, non-inferiority SWORD-1 and SWORD-2 studies. *The Lancet* **391**, 839-849 (2018).

138 Grulich, A. E. *et al.* Long-term protection from HIV infection with oral HIV pre-exposure prophylaxis in gay and bisexual men: findings from the expanded and extended EPIC-NSW prospective implementation study. *The Lancet HIV* **8**, e486-e494 (2021).

139 Bachmann, N. *et al.* Importance of routine viral load monitoring: higher levels of resistance at ART failure in Uganda and Lesotho compared with Switzerland. *Journal of Antimicrobial Chemotherapy* **74**, 468-472 (2019).

140 Gupta, R. K. *et al.* Virological monitoring and resistance to first-line highly active antiretroviral therapy in adults infected with HIV-1 treated under WHO guidelines: a systematic review and meta-analysis. *The Lancet infectious diseases* **9**, 409-417 (2009).

141 Cohen, M. S. *et al.* Prevention of HIV-1 infection with early antiretroviral therapy. *New England journal of medicine* **365**, 493-505 (2011).

142 Tanser, F., Bärnighausen, T., Grapsa, E., Zaidi, J. & Newell, M.-L. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science* **339**, 966-971 (2013).

143 Clutter, D. S., Jordan, M. R., Bertagnolio, S. & Shafer, R. W. HIV-1 drug resistance and resistance testing. *Infection, Genetics and Evolution* **46**, 292-307 (2016).

144 DeGruttola, V. *et al.* The relation between baseline HIV drug resistance and response to antiretroviral therapy: re-analysis of retrospective and prospective studies using a standardized data analysis plan. *Antiviral therapy* **5**, 41-48 (2000).

145 Beyrer, C. & Pozniak, A. HIV drug resistance—an emerging threat to epidemic control. *New England Journal of Medicine* **377**, 1605-1607 (2017).

146 Scherrer, A. U. *et al.* Emergence of acquired HIV-1 drug resistance almost stopped in Switzerland: a 15-year prospective cohort analysis. *Clinical infectious diseases* **62**, 1310-1317 (2016).

147 Coffin, J. M. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* **267**, 483-489, doi:10.1126/science.7824947 (1995).

148 Dauber, D. S. *et al.* Altered substrate specificity of drug-resistant human immunodeficiency virus type 1 protease. *J Virol* **76**, 1359-1368, doi:10.1128/jvi.76.3.1359-1368.2002 (2002).

149 Malim, M. H. & Emerman, M. HIV-1 sequence variation: drift, shift, and attenuation. *Cell* **104**, 469-472, doi:10.1016/s0092-8674(01)00234-3 (2001).

150 Clavel, F. & Hance, A. J. HIV drug resistance. *New England Journal of Medicine* **350**, 1023-1035 (2004).

151 Kuritzkes, D. R. Preventing and managing antiretroviral drug resistance. *AIDS patient care and STDs* **18**, 259-273 (2004).

152 World Health Organization. *HIV Drug Resistance Report 2019. Geneva, Switzerland*, <https://www.who.int/hiv/pub/drugresistance/hivdr-report-2019/en/> (2019).

153 Tang, M. W. & Shafer, R. W. HIV-1 Antiretroviral Resistance. *Drugs* **72**, e1-e25, doi:10.2165/11633630-000000000-00000 (2012).

154 Jain, V. *et al.* Differential persistence of transmitted HIV-1 drug resistance mutation classes. *The Journal of infectious diseases* **203**, 1174-1181, doi:10.1093/infdis/jiq167 (2011).

155 Yang, W. L. *et al.* Assessing the Paradox Between Transmitted and Acquired HIV Type 1 Drug Resistance Mutations in the Swiss HIV Cohort Study From 1998 to 2012. *The Journal of infectious diseases* **212**, 28-38, doi:10.1093/infdis/jiv012 (2015).

156 Santos, A. F. & Soares, M. A. HIV genetic diversity and drug resistance. *Viruses* **2**, 503-531 (2010).

157 Häggblom, A., Svedhem, V., Singh, K., Sönnerborg, A. & Neogi, U. Virological failure in patients with HIV-1 subtype C receiving antiretroviral therapy: an analysis of a prospective national cohort in Sweden. *The lancet HIV* **3**, e166-e174 (2016).

158 Pingen, M. *et al.* Diminished transmission of drug resistant HIV-1 variants with reduced replication capacity in a human transmission model. *Retrovirology* **11**, 113, doi:10.1186/s12977-014-0113-9 (2014).

159 Chesney, M. A., Morin, M. & Sherr, L. Adherence to HIV combination therapy. *Social science & medicine* **50**, 1599-1605 (2000).

160 Boerma, R. *et al.* Alarming increase in pretreatment HIV drug resistance in children living in sub-Saharan Africa: a systematic review and meta-analysis. *Journal of Antimicrobial Chemotherapy* **72**, 365-371 (2017).

161 Gupta, R. K. *et al.* HIV-1 drug resistance before initiation or re-initiation of first-line antiretroviral therapy in low-income and middle-income countries: a systematic review and meta-regression analysis. *The Lancet infectious diseases* **18**, 346-355 (2018).

162 Hamers, R. L., Sigaloff, K. C. E., Kityo, C., Mugyenyi, P. & de Wit, T. F. R. Emerging HIV-1 drug resistance after roll-out of antiretroviral therapy in sub-Saharan Africa. *Current Opinion in HIV and AIDS* **8**, 19-26, doi:10.1097/COH.0b013e32835b7f94 (2013).

163 Rhee, S. Y. *et al.* HIV-1 transmitted drug resistance surveillance: shifting trends in study design and prevalence estimates. *Journal of the International AIDS Society* **23**, e25611 (2020).

164 Mansky, L. M. & Temin, H. M. 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**, 5087-5094 (1995).

165 Zhuang, J. *et al.* Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *Journal of virology* **76**, 11273-11282 (2002).

166 Keele, B. F. *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* **105**, 7552-7557, doi:10.1073/pnas.0802203105 (2008).

167 Frost, S. D. W. *et al.* Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection. *Proc Natl Acad Sci U S A* **102**, 18514-18519, doi:10.1073/pnas.0504658102 (2005).

168 Park, S. Y., Love, T. M., Perelson, A. S., Mack, W. J. & Lee, H. Y. Molecular clock of HIV-1 envelope genes under early immune selection. *Retrovirology* **13**, 38, doi:10.1186/s12977-016-0269-6 (2016).

169 Holmes, E. C., Zhang, L. Q., Simmonds, P., Ludlam, C. A. & Brown, A. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proceedings of the national Academy of Sciences* **89**, 4835-4839 (1992).

170 Leitner, T., Kumar, S. & Albert, J. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J Virol* **71**, 4761-4770 (1997).

171 Shankarappa, R. *et al.* Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection. *Journal of Virology* **73**, 10489-10502 (1999).

172 Lythgoe, K. A., Gardner, A., Pybus, O. G. & Grove, J. Short-sighted virus evolution and a germline hypothesis for chronic viral infections. *Trends in microbiology* **25**, 336-348 (2017).

173 Karlsson, A. C., Gaines, H., Sällberg, M., Lindbäck, S. & Sönnerborg, A. Reappearance of founder virus sequence in human immunodeficiency virus type 1-infected patients. *Journal of virology* **73**, 6191-6196 (1999).

174 Fraser, C. *et al.* Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. *Science* **343** (2014).

175 Rouzine, I. M., Weinberger, A. D. & Weinberger, L. S. An evolutionary role for HIV latency in enhancing viral transmission. *Cell* **160**, 1002-1012 (2015).

176 Lythgoe, K. A. & Fraser, C. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proceedings of the Royal Society B: Biological Sciences* **279**, 3367-3375 (2012).

177 Ngandu, N. K. *et al.* Brief report: selection of HIV-1 variants with higher transmission potential by 1% tenofovir gel microbicide. *Journal of acquired immune deficiency syndromes (1999)* **76**, 43 (2017).

178 Theys, K. *et al.* The impact of HIV-1 within-host evolution on transmission dynamics. *Current opinion in virology* **28**, 92-101 (2018).

179 Lemey, P., Rambaut, A. & Pybus, O. G. HIV evolutionary dynamics within and among hosts. *Aids Rev* **8**, 125-140 (2006).

180 Carlson, J. M. *et al.* Impact of pre-adapted HIV transmission. *Nature medicine* **22**, 606-613 (2016).

181 Tully, D. C. *et al.* Differences in the selection bottleneck between modes of sexual transmission influence the genetic composition of the HIV-1 founder virus. *PLoS pathogens* **12**, e1005619 (2016).

182 Carlson, J. M. *et al.* Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science* **345**, 1254031 (2014).

183 Nduva, G. M. *et al.* HIV-1 transmission patterns within and between risk groups in coastal Kenya. *Scientific reports* **10**, 1-10 (2020).

184 Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *science* **303**, 327-332 (2004).

185 Via, S. Natural selection in action during speciation. *Proceedings of the National Academy of Sciences of the United States of America* **106 Suppl 1**, 9939-9946, doi:10.1073/pnas.0901397106 (2009).

186 Arnold, M. L. *Natural hybridization and evolution*. (Oxford University Press on Demand, 1997).

187 Benton, M. J. Diversification and extinction in the history of life. *Science* **268**, 52-58, doi:10.1126/science.7701342 (1995).

188 Linnaeus, C. v. Systema naturae, vol. 1. *Systema naturae, Vol. 1* (1758).

189 Drummond, A., Pybus, O. G. & Rambaut, A. Inference of viral evolutionary rates from molecular sequences. *Adv Parasitol* **54**, 331-358 (2003).

190 Posada, D. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular biology and evolution* **19**, 708-717 (2002).

191 Grassly, N. C., Harvey, P. H. & Holmes, E. C. Population dynamics of HIV-1 inferred from gene sequences. *Genetics* **151**, 427-438 (1999).

192 Lemey, P., Salemi, M. & Vandamme, A.-M. *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. (Cambridge University Press, 2009).

193 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *bioinformatics* **23**, 2947-2948 (2007).

194 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).

195 Katoh, K., Misawa, K., Kuma, K. i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30**, 3059-3066 (2002).

196 Arenas, M. Trends in substitution models of molecular evolution. *Frontiers in genetics* **6**, 319 (2015).

197 Jukes, T. H. & Cantor, C. R. Evolution of protein molecules. *Mammalian protein metabolism* **3**, 21-132 (1969).

198 Lefort, V., Longueville, J.-E. & Gascuel, O. SMS: smart model selection in PhyML. *Molecular biology and evolution* **34**, 2422-2424 (2017).

199 Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution* **17**, 368-376 (1981).

200 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* **16**, 111-120 (1980).

201 Hasegawa, M., Kishino, H. & Yano, T.-a. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution* **22**, 160-174 (1985).

202 Posada, D. jModelTest: phylogenetic model averaging. *Molecular biology and evolution* **25**, 1253-1256 (2008).

203 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4**, 406-425 (1987).

204 Sokal, R. R. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.* **38**, 1409-1438 (1958).

205 Page, R. D. & Holmes, E. C. *Molecular evolution: a phylogenetic approach*. (John Wiley & Sons, 2009).

206 Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic zoology* **27**, 401-410 (1978).

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

208 Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**, 268-274 (2015).

209 Bunupuradah, T. *et al.* Low-dose versus standard-dose ritonavir-boosted atazanavir in virologically suppressed Thai adults with HIV (LASA): a randomised, open-label, non-inferiority trial. *The lancet HIV* **3**, e343-e350 (2016).

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

211 Levin, S. A. *Encyclopedia of biodiversity*. (Elsevier Inc., 2013).

212 Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J. & Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data. *AIDS (London, England)* **31**, 1211 (2017).

213 Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* **55**, 539-552, doi:10.1080/10635150600755453 (2006).

214 Bayes, T. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, 370-418 (1763).

215 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics* **21**, 1087-1092 (1953).

216 Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. (1970).

217 Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus evolution* **4**, vey016 (2018).

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

219 Dos Reis, M., Donoghue, P. C. & Yang, Z. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics* **17**, 71-80 (2016).

220 Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *Journal of theoretical biology* **8**, 357-366 (1965).

221 Britten, R. J. Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**, 1393-1398 (1986).

222 Ayala, F. J. Vagaries of the molecular clock. *Proceedings of the National Academy of Sciences* **94**, 7776-7783 (1997).

223 HASEGAWA, M. & KISHINO, H. Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. *The Japanese Journal of Genetics* **64**, 243-258 (1989).

224 Yoder, A. D. & Yang, Z. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* **17**, 1081-1090 (2000).

225 Thorne, J. L., Kishino, H. & Painter, I. S. Estimating the rate of evolution of the rate of molecular evolution. *Molecular biology and evolution* **15**, 1647-1657 (1998).

226 Huelsenbeck, J. P., Larget, B. & Swofford, D. A compound Poisson process for relaxing the molecular clock. *Genetics* **154**, 1879-1892 (2000).

227 Aris-Brosou, S. & Yang, Z. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Systematic Biology* **51**, 703-714 (2002).

228 Rambaut, A. & Bromham, L. Estimating divergence dates from molecular sequences. *Molecular biology and evolution* **15**, 442-448 (1998).

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

230 Cranston, K. & Rannala, B. Closing the gap between rocks and clocks. *Heredity* **94**, 461-462 (2005).

231 Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS computational biology* **5** (2009).

232 Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R. & Rodrigo, A. G. Measurably evolving populations. *Trends in ecology & evolution* **18**, 481-488 (2003).

233 Kingman, J. F. C. The coalescent. *Stochastic processes and their applications* **13**, 235-248 (1982).

234 Kingman, J. F. On the genealogy of large populations. *Journal of applied probability* **19**, 27-43 (1982).

235 Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology* **23**, 183-201 (1983).

236 Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).

237 Griffiths, R. C. & Tavare, S. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **344**, 403-410 (1994).

238 Ewens, W. J. *Mathematical population genetics: theoretical introduction*. Vol. 1 (Springer, 2004).

239 Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular biology and evolution* **30**, 713-724 (2013).

240 Pybus, O., Drummond, A., Nakano, T., Robertson, B. & Rambaut, A. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Molecular biology and evolution* **20**, 381-387 (2003).

241 Worobey, M. *et al.* The emergence of sars-cov-2 in europe and north america. *Science* **370**, 564-570 (2020).

242 Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nature microbiology* **4**, 10-19 (2019).

243 Los Alamos National Laboratory. *HIV-1 database at the Los Alamos National Laboratory*, <http://www.hiv.lanl.gov/> (2019).

244 Sanders, E. J. *et al.* High HIV-1 incidence, correlates of HIV-1 acquisition, and high viral loads following seroconversion among MSM. *Aids* **27**, 437-446, doi:10.1097/QAD.0b013e32835b0f81 (2013).

245 Smith, A. D. *et al.* HIV burden and correlates of infection among transfeminine people and cisgender men who have sex with men in Nairobi, Kenya: an observational study. *The lancet. HIV*, doi:10.1016/s2352-3018(20)30310-6 (2021).

246 Kunzweiler, C. P. *et al.* Depressive Symptoms, Alcohol and Drug Use, and Physical and Sexual Abuse Among Men Who Have Sex with Men in Kisumu, Kenya: The Anza Mapema Study. *AIDS and behavior* **22**, 1517-1529, doi:10.1007/s10461-017-1941-0 (2018).

247 Siddiqui, A. R. *et al.* Investigation of an extensive outbreak of HIV infection among children in Sindh, Pakistan: protocol for a matched case–control study. *BMJ open* **10**, e036723 (2020).

248 Esbjörnsson, J. *et al.* Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease-indication of an evolving epidemic in West Africa. *Retrovirology* **7**, 23 (2010).

249 Hedskog, C. *et al.* Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PloS one* **5**, e11345 (2010).

250 Abidi, S. H. *et al.* Phylogenetic and Drug-Resistance Analysis of HIV-1 Sequences From an Extensive Paediatric HIV-1 Outbreak in Larkana, Pakistan. *Frontiers in Microbiology* **12**, doi:10.3389/fmicb.2021.658186 (2021).

251 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *bioinformatics* **23**, 2947-2948 (2007).

252 Los Alamos National Library. *HIV-1 database at the Los Alamos National Library*, <http://www.hiv.lanl.gov/> (2019).

253 Lole, K. S. *et al.* Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *Journal of virology* **73**, 152-160 (1999).

254 Pineda-Peña, A.-C. *et al.* 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**, 337-348 (2013).

255 Nazziwa, J. *et al.* Characterisation of HIV-1 Molecular Epidemiology in Nigeria: Origin, Diversity, Demography and Geographic Spread. *Scientific reports* **10** (2020).

256 Nduva, G. M. *et al.* HIV-1 Transmission Patterns Within and Between Risk Groups in Coastal Kenya. *Scientific reports* **10**, 6775, doi:10.1038/s41598-020-63731-z (2020).

257 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307-321 (2010).

258 Ragonnet-Cronin, M. *et al.* Automated analysis of phylogenetic clusters. *BMC bioinformatics* **14**, 1-10 (2013).

259 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution* **2**, vew007 (2016).

260 Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular biology and evolution* **29**, 2157-2167 (2012).

261 Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol* **30**, 713-724, doi:10.1093/molbev/mss265 (2013).

262 Suchard, M. A., Kitchen, C. M., Sinsheimer, J. S. & Weiss, R. E. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic biology* **52**, 649-664 (2003).

263 Edwards, C. J. *et al.* Ancient hybridization and an Irish origin for the modern polar bear matriline. *Current Biology* **21**, 1251-1258 (2011).

264 Minin, V. N. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of mathematical biology* **56**, 391-412 (2008).

265 Bbosa, N. *et al.* Phylogeography of HIV-1 suggests that Ugandan fishing communities are a sink for, not a source of, virus from general populations. *Scientific reports* **9**, 1-8 (2019).

266 Stanford University. *HIVdb version 9.0*, <Available at: https://hivdb.stanford.edu/hivdb/> (2021).

267 Bennett, D. E. *et al.* Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PloS one* **4**, e4724 (2009).

268 Cuzick, J. A Wilcoxon-type test for trend. *Statistics in medicine* **4**, 87-90 (1985).

269 Wickham, H. *ggplot2: elegant graphics for data analysis*. (springer, 2016).

270 Phillips, N. D. Yarrr! The pirate's guide to R. *APS Observer* **30** (2017).

271 Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).

272 Bezemer, D. *et al.* HIV Type 1 transmission networks among men having sex with men and heterosexuals in Kenya. *AIDS research and human retroviruses* **30**, 118-126 (2014).

273 Gouws, E. & Cuchi, P. Focusing the HIV response through estimating the major modes of HIV transmission: a multi-country analysis. *Sexually transmitted infections* **88**, i76-i85 (2012).

274 Sallam, M. *et al.* Molecular epidemiology of HIV-1 in Iceland: Early introductions, transmission dynamics and recent outbreaks among injection drug users. *Infection, Genetics and Evolution* **49**, 157-163 (2017).

275 Skar, H. *et al.* Dynamics of two separate but linked HIV-1 CRF01_AE outbreaks among injection drug users in Stockholm, Sweden, and Helsinki, Finland. *Journal of virology* **85**, 510-518 (2011).

276 Hassan, A. S. *et al.* HIV-1 subtype diversity, transmission networks and transmitted drug resistance amongst acute and early infected MSM populations from Coastal Kenya. *PLoS One* **13**, e0206177, doi:10.1371/journal.pone.0206177 (2018).

277 Geibel, S. *et al.* Factors associated with self-reported unprotected anal sex among male sex workers in Mombasa, Kenya. *Sexually transmitted diseases* **35**, 746-752 (2008).

278 Kenya National AIDS Control Council. *Kenya HIV Prevention Response and Modes of Transmission Analysis.*, <http://siteresources.worldbank.org/INTHIVAIDS/Resources/375798-1103037153392/KenyaMOT22March09Final.pdf> (2009).

279 Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J. & Frost, S. D. Phylodynamics of infectious disease epidemics. *Genetics* **183**, 1421-1430 (2009).

280 Ministry of Health, N. A. S. C. P. *Guidelines on Use of Antiretroviral Drugs for Treating and Preventing HIV Infections in Kenya 2016*, <https://www.prepwatch.org/wp-content/uploads/2016/08/Guidelines-on-ARV-for-Treating-Preventing-HIV-Infections-in-Kenya.pdf> (2016).

281 Graham, S. M. *et al.* A randomized controlled trial of the Shikamana intervention to promote antiretroviral therapy adherence among gay, bisexual, and other men who have sex with men in Kenya: feasibility, acceptability, safety and initial effect size. *AIDS and behavior*, 1-14 (2020).

282 Möller, L. M. *et al.* Changes in sexual risk behavior among MSM participating in a research cohort in coastal Kenya. *AIDS (London, England)* **29**, S211 (2015).

283 Mugo, P. M. *et al.* Effect of text message, phone call, and in-person appointment reminders on uptake of repeat HIV testing among outpatients screened for acute HIV infection in Kenya: a randomized controlled trial. *PLoS One* **11**, e0153612 (2016).

284 Sanders, E. J. *et al.* Acute HIV-1 infection is as common as malaria in young febrile adults seeking care in coastal Kenya. *AIDS (London, England)* **28**, 1357 (2014).

285 Sanders, E. J. *et al.* Targeted screening of at-risk adults for acute HIV-1 infection in sub-Saharan Africa. *AIDS (London, England)* **29**, S221 (2015).

286 Kimani, M. *et al.* Pr EP interest and HIV-1 incidence among MSM and transgender women in coastal Kenya. *Journal of the International AIDS Society* **22**, e25323 (2019).

287 Wahome, E. *et al.* An empiric risk score to guide PrEP targeting among MSM in coastal Kenya. *AIDS and behavior* **22**, 35-44 (2018).

288 Graham, S. M. *et al.* Development and pilot testing of an intervention to promote care engagement and adherence among HIV-positive Kenyan MSM. *AIDS (London, England)* **29**, S241 (2015).

289 van der Elst, E. M. *et al.* Strengthening healthcare providers' skills to improve HIV services for MSM in Kenya. *AIDS (London, England)* **29**, S237 (2015).

290 National AIDS & STI Control Program, M. o. H. K. *Guidelines on Use of Antiretroviral Drugs for Treating and Preventing HIV Infection in Kenya – 2018 Edition*, <https://www.nascop.or.ke/new-guidelines> (2018).

291 Macdonald, V. *et al.* Prevalence of pretreatment HIV drug resistance in key populations: a systematic review and meta-analysis. *Journal of the International AIDS Society* **23**, e25656 (2020).

292 Raees, M. A., Abidi, S. H., Ali, W., Khanani, M. R. & Ali, S. HIV among women and children in Pakistan. *Trends in microbiology* **21**, 213-214 (2013).

293 Mir, F. *et al.* Factors associated with HIV infection among children in Larkana District, Pakistan: a matched case-control study. *The Lancet HIV* **8**, e342-e352 (2021).

294 Mir, F. *et al.* HIV infection predominantly affecting children in Sindh, Pakistan, 2019: a cross-sectional study of an outbreak. *The Lancet Infectious Diseases* **20**, 362-370 (2020).

295 Altaf, A., Pasha, S., Vermund, S. H. & Shah, S. A. A second major HIV outbreak in Larkana, Pakistan. *JPMA. The Journal of the Pakistan Medical Association* **66**, 1510-1511 (2016).

296 Siddiqui, A. R. *et al.* Investigation of an extensive outbreak of HIV infection among children in Sindh, Pakistan: protocol for a matched case-control study. *BMJ Open* **10**, e036723, doi:10.1136/bmjopen-2019-036723 (2020).

297 Mir, F. *et al.* HIV infection predominantly affecting children in Sindh, Pakistan, 2019: a cross-sectional study of an outbreak. *The Lancet. Infectious diseases* **20**, 362-370, doi:10.1016/s1473-3099(19)30743-1 (2020).

298 Anderson, S.-J. *et al.* Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study. *The Lancet* **384**, 249-256 (2014).

299 Vasylyeva, T. I. *et al.* Molecular epidemiology reveals the role of war in the spread of HIV in Ukraine. *Proceedings of the National Academy of Sciences* **115**, 1051-1056, doi:10.1073/pnas.1701447115 (2018).

300 Volz, E. M. *et al.* HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS Med* **10**, e1001568; discussion e1001568, doi:10.1371/journal.pmed.1001568 (2013).

301 Ragonnet-Cronin, M. *et al.* Non-disclosed men who have sex with men in UK HIV transmission networks: phylogenetic analysis of surveillance data. *The Lancet HIV* **5**, e309-e316, doi:https://doi.org/10.1016/S2352-3018(18)30062-6 (2018).

302 Fisher, M. *et al.* Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. *Aids* **24**, 1739-1747 (2010).

303 Kouyos, R. D. *et al.* Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *The Journal of infectious diseases* **201**, 1488-1497 (2010).

304 Novitsky, V., Moyo, S., Lei, Q., DeGruttola, V. & Essex, M. Impact of sampling density on the extent of HIV clustering. *AIDS research and human retroviruses* **30**, 1226-1235 (2014).

305 Ihantamalala, F. A. *et al.* Estimating sources and sinks of malaria parasites in Madagascar. *Nature Communications* **9**, doi:10.1038/s41467-018-06290-2 (2018).

306 Wesolowski, A. *et al.* Quantifying the Impact of Human Mobility on Malaria. *Science* **338**, 267-270, doi:10.1126/science.1223467 (2012).

307 Okano, J. T., Sharp, K., Valdano, E., Palk, L. & Blower, S. HIV transmission and source-sink dynamics in sub-Saharan Africa. *The lancet. HIV* **7**, e209-e214, doi:10.1016/s2352-3018(19)30407-2 (2020).

Quantifying rates of HIV-1 flow between risk groups and geographic locations in Kenya: a country-wide phylogenetic study

George M. Nduva[1,2*], Frederick Otieno[3], Joshua Kimani[4,5], Elizabeth Wahome[2], Lyle R. McKinnon[4,5,6], Francois Cholette[5,7], Maxwell Majiwa[8], Moses Masika[9], Gaudensia Mutua[9], Omu Anzala[9], Susan M. Graham[2,10], Larry Gelmon[4,5], Matt A. Price[11,12], Adrian D. Smith[15], Robert C. Bailey[3,13], Guy Baele[14], Philippe Lemey[14], Amin S. Hassan[1,2#], Eduard J. Sanders[2,15#], and Joakim Esbjörnsson[1,15#]

[#]Equal contribution as senior authors

[1]Lund University, Lund, Sweden, [2]Kenya Medical Research Institute-Wellcome Trust Research Programme, Kilifi, Kenya, [3]Nyanza Reproductive Health Society, Kisumu, Kenya, [4]University of Nairobi, Nairobi, Kenya, [5]University of Manitoba, Winnipeg, Canada, [6]Centre for the AIDS Programme of Research in South Africa (CAPRISA), South Africa, [7]National Microbiology Laboratory at the JC Wilt Infectious Diseases Research Centre, Public Health Agency of Canada, Winnipeg, Canada, [8]Kenya Medical Research Institute / Center for Global Health Research, Kisumu, Kenya, [9]KAVI Institute of Clinical Research, University of Nairobi, Nairobi, Kenya, [10]University of Washington, Seattle, USA, [11]IAVI, San Francisco, USA, [12]University of California, San Francisco, USA, [13]University of Illinois at Chicago, USA, [14]KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory of Clinical and Evolutionary Virology, Leuven, Belgium, [15]The University of Oxford, Oxford, United Kingdom.

[*]Corresponding Author:
George M. Nduva
Faculty of Medicine / Department of Translational Medicine
Postal address: Systems Virology, BMC B13, 221 84 Lund, Sweden
Email: george.makau_nduva@med.lu.se

Joakim Esbjörnsson
Faculty of Medicine / Department of Translational Medicine
Postal address: Systems Virology, BMC B13, 221 84 Lund, Sweden
Email: joakim.esbjornsson@med.lu.se

Short title: HIV-1 Molecular Epidemiology in Kenya

Keywords: HIV-1; key populations; molecular epidemiology; transmission

ABSTRACT
In Kenya, HIV-1 key populations including men having sex with men (MSM), people who inject drugs (PWID) and female sex workers (FSW) are thought to significantly contribute to HIV-1 transmission in the wider, mostly heterosexual (HET) HIV-1 transmission network. However, clear data on HIV-1 transmission dynamics within and between these groups are limited. We aimed to empirically quantify rates of HIV-1 flow between key populations and the HET population, as well as between different geographic regions to determine HIV-1 "hotspots" and their contribution to HIV-1 transmission in Kenya. We used maximum-likelihood phylogenetic and Bayesian inference to analyse 4058 HIV-1 *pol* sequences (representing 0.3% of the epidemic in Kenya) sampled 1986–2019 from individuals of different risk groups and regions in Kenya. We found 89% within-risk group transmission and 11% mixing between risk groups, cyclic HIV-1 exchange between adjoining geographic provinces and strong evidence of HIV-1 dissemination from (i) West-to-East (i.e. higher-to-lower HIV-1 prevalence regions), and (ii) heterosexual-to-key populations. Low HIV-1 prevalence regions and key populations are sinks rather than major sources of HIV-1 transmission in Kenya. Targeting key populations in Kenya needs to occur concurrently with strengthening interventions in the general epidemic.

INTRODUCTION

The world is off-track on the United Nations Programme on HIV and AIDS (UNAIDS) objective to reduce the global HIV-1 incidence rate, with an estimated 1.7 million new HIV-1 infections in 2019[1]. To fast-track reduction in global HIV-1 incidence whilst also achieving efficiency gains, UNAIDS directs national governments to invest strategically in HIV-1 programmes. This includes directing treatment and prevention to HIV-1 key populations (defined as UNAIDS as gay men and other men who have sex with men [MSM], people who inject drugs [PWID], sex workers [FSW], transgender people, and sex partners of key populations)[2]. An approach to inform decision-making is to identify populations populations that contribute with a disproportionate number of infections in local epidemic and to eliminate structural and social barriers to health service delivery among key populations[3,4].

In North America and European settings, the HIV-1 epidemic mainly affects HIV-1 key populations, and the availability of large numbers of HIV-1 genetic sequences and associated patient risk group information have allowed extensive characterisation of HIV-1 networks[5-7]. In contrast, in sub-Saharan Africa (accounting for 65% of all new HIV-1 infections globally), the HIV-1 epidemic mainly affects the heterosexual population (HET). However, pockets of concentrated sub-epidemics involving high-risk groups have also been described[8-10]. Additionally, there is evidence of overlapping sexual networks and phylogenetic linkages between HIV-1 key populations and HET[10]. However, the scarcity of HIV-1 sequences from key populations has limited phylogenetic assessment of HIV-1 transmsission within and between key populations and lower-risk populations in sub-Saharan Africa.

Kenya has the fifth-largest number of people with HIV-1 in the world, and the early HIV-1 epidemic in the country was defined exclusively as heterosexual and involving FSW and long-distance truck drivers[11,12]. As a consequence, governmental HIV-1 surveillance did not focus on other marginalised key populations such as MSM and PWID[4,13,14]. The Kenyan Ministry of Health has reported high HIV-1 prevalence among key populations (29.3% among FSW, 18.2% among MSM and 18.2% among PWID, compared to 4.5% in the general epidemic)[15,16]. As a consequence, directed programmes for key populations have been initiated based on the assumption that they contribute with a disproportionate number of infections to the larger HIV-1 transmission network in the nationwide epidemic[16,17]. However, phylogenetic studies in Coastal Kenya have suggested that most HIV-1 transmissions occur within risk groups (with only 15% of the identified clusters reflecting mixing between MSM, FSW, and HET in Coastal Kenya)[18,19]. Moreover, to the best of our knowledge, no study has empirically assessed the rates of HIV-1 flow between key populations and the heterosexual population in Kenya. Also, spatial mapping of the Kenyan epidemic has revealed extensive geographic heterogeneity with HIV-1 prevalence ranging from less than 1% in the North Eastern province to more than 20% around the shores of Lake Victoria in the Western regions of the country [11]. Such spatial differences in HIV-1 distribution likely influence HIV-1 diffusion dynamics [20,21], but HIV-1 transmission rates between different geographic areas in Kenya are still unknown.

Phylodynamic analysis has been widely used to determine HIV-1 networks, reconstruct virus historical spatial dissemination, as well as assessing rates of virus flow between populations with varying HIV-1 prevalence[7,18,19,22-29]. However, due to the scarcity of HIV-1 sequences from key populations, phylogeographic assessment of HIV-1 transmission rates between populations are rare in sub-Saharan Africa[28]. Here, we combined HIV-1 phylogenetic and epidemiological data to reconstruct HIV-1 networks and to empirically quantify rates of HIV-1 flow between risk groups and geographic regions to identify and determine the contribution of HIV-1 "hotspots" in sustaining HIV-1 transmission in Kenya. We hypothesised that virus flow would be predominantly from high prevalence "hotspots" to the lower prevalence populations.

METHODS

Study population and sequence dataset

New HIV-1 *pol* sequences were generated from blood plasma obtained through studies conducted through the MSM Health Research Consortium – a multi-site collaboration between researchers affiliated with KEMRI-Wellcome Trust (KWTRP) in Coastal Kenya, Nyanza Reproductive Health Society (NRHS) in Western Kenya, Kenya AIDS Vaccine Initiative's Institute of Clinical Research (KAVI-ICR), and Sex Workers Outreach Program (SWOP) clinics in Nairobi. These included samples from Coast derived from participants in a prospective observational cohort (2006-2019)[30], samples from Nairobi from a respondent-driven sample survey (TRANSFORM, 2017)[31], and samples from Nyanza derived from the Anza Mapema cohort (2015-2017)[32]. Additional nationwide HIV-1 *pol* sequences (2008-2018) were obtained from the national HIV-1 reference laboratory at the Kenya Medical Research Institute (KEMRI) – Centre for Global Health Research.

In addition, all published Kenyan HIV-1 *pol* sequences (1986-2019, corresponding to HXB2 positions 2000-3600) available in the Los Alamos HIV-1 sequence database were retrieved March 19th 2020[33]. In cases where more than one sequence per individual was available, the oldest sequence was retained. Newly generated and publicly available sequences were annotated with sampling date, sampling location (province), treatment status, age, sex, and risk group (MSM [men who reported having sex with men]; PWID [men and women who inject drugs]; FSW [female sex workers]; and HET [presumed heterosexuals including men and women for whom risk assessment was not available]). Missing information for published sequences was retrieved from relevant studies or obtained through communication with study authors [19,24,34-42].

RNA extraction, DNA amplification, and partial HIV-1 *pol* sequencing

HIV-1 RNA was extracted from blood plasma samples using the RNeasy Lipid Tissue Mini Kit (QIAGEN) with modifications from the manufacturer's standard protocol[43]. Briefly, 100 µl patient blood plasma was lysed in 1000 µl Qiazol Reagent. Reverse transcription and amplification of partial HIV-1 *pol* gene were performed using the One-Step Superscript III RT/Platinum Taq High Fidelity Enzyme Mix (ThermoFisher Scientific™) with the *pol*-specific primer pair JA269 and JA272[44]. First-round PCR products were amplified in a nested PCR with DreamTaq Green DNA Polymerase (ThermoFisher Scientific™) using *pol*-specific primers JA271 and JA270[44]. PCR products were sequenced in both directions with the nested PCR primers using the BigDye terminator kit v1.1 (Applied Biosystems). New HIV-1 *pol* sequences (approximately 1020 nucleotides [nt], HXB2 [K03455] positions 2267-3287) were determined on an ABI PRISM 3130×1 Genetic Analyzer (Applied Biosystems).

Population estimates and sampling density

Sampling density (the proportion of genotyped HIV-1 sequences in the estimated number of HIV-infected individuals per geographic location and risk group) was computed based on national HIV-1 prevalence estimates [11,16,17,45-47].

Subtype analysis

All Kenyan HIV-1 *pol* sequences were combined and aligned with the Los Alamos HIV-1 Group M (subtypes A-K + Recombinants) subtype reference dataset (http://www.hiv.lanl.gov) using the MAFFT algorithm in Geneious Prime 2019 [48]. The HIV-1 subtype/circulating recombinant form (CRF) for each sequence was determined by maximum-likelihood (ML) phylogenetic analysis in PhyML using the general time-reversible substitution model with a gamma-distributed rate variation and proportion of invariant sites (GTR+Γ4+I) [49]. Branch support was determined by the approximate likelihood ratio test with the Shimodaira-Hasegawa-like procedure (SH-aLRT) in PhyML, and SH-aLRT support values ≥0.90 were considered significant [49]. The Subtype/CRF-resolved phylogeny was visualized using FigTree v1.4.4 (https://github.com/rambaut/figtree/releases). Unique recombinant forms (URFs) were characterised by boot-scan analysis in SimPlot [50-52].

Cluster analysis
Sequences were grouped into subtype-specific datasets and the most similar non-Kenyan sequences for each available Kenyan sequence were determined by a BLAST, as previously described[7,18,26]. Redundant sequences or clonal sequences from the same individual were removed from the dataset. All sequences were aligned by subtype and subtype-specific, and alignments were manually edited to exclude codon positions associated with drug resistance. Maximum-likelihood phylogenies were reconstructed in PhyML[49]. For each subtype, monophyletic clades with aLRT-SH support ≥0.9 and which were dominated (≥80%) by Kenyan sequences (compared to reference sequences) were defined as Kenyan HIV-1[7,18,26,53]. Identified clusters were classified into dyads (2 sequences), networks (3-14 sequences), or large clusters (>14 sequences)[7].

Bayesian phylodynamic inference
HIV-1 evolutionary origins and past population dynamics were determined using subsets of the main subtypes as well as for the large clusters identified in the cluster analysis. Only sequences with information on sampling dates were included in this analysis. The temporal signal was assessed in TempEst (v1.5.3) [54]. Bayesian inferences were done in BEAST 1.10.4 using the Bayesian Skygrid model with an uncorrelated lognormal relaxed clock and inferred under the GTR + Γ4 + I substitution model[55-58]. To enhance precision in estimating evolutionary parameters within and between clusters from different risk groups, a previously described hierarchical phylogenetic model (HPM) was specified on evolutionary parameters[59]. Each MCMC chain was run for 300 million states, sampling every 30,000th iteration and discarding the first 10% as burn-in. Convergence was determined in Tracer v.1.7.0 and defined as effective sample sizes (ESS) ≥200[55] – and where this was not achieved, the burnin was adjusted or the analysis re-run with a longer chain[60].

Bayesian phylogeographic inference
We computed a discrete phylogeographic inference using an empirical tree distribution – where the expected number of HIV-1 migrations for every pathway were inferred on a branch-by-branch basis as previously described[20,61]. Sampling province and risk group were used as independent discrete states. The asymmetric continuous-time Markov chain (CTMC) model was preferentially used as it relaxes the assumption of constant diffusion rates through time to realistically model phylogeographic processes[61,62]. A robust counting approach implemented in BEAST was used to estimate the forward and reverse HIV-1 movement events (Markov jumps) between locations and risk group states along the branches of time dated phylogenetic trees [63]. Well-supported movements and Bayes factors (BF) assessing statistical support were summarized using SPREAD v1.0.7, (BF≥3 was considered significant) [61]. Maximum clade credibility (MCC) trees annotated with key demographic and epidemiological data were summarized in Tree-Annotator v1.10.4 (BEAST suite) and visualized in Figtree (v1.4.4).

Sensitivity analysis
In Kenya, the vast majority (35%) of people with HIV-1 are in Nyanza province, followed by Rift Valley (17%), Nairobi (13%), Western (9%), Central (9%), Eastern (9%), Coast (7%), North Eastern (<1%) – and modes of transmission estimates have shown that 64% of infections result from heterosexual contact among casual or married couples, female sex work (14%), men having sex with men (15%) and injection drug use (4%)[11,16,46].

Phylogeographic analysis is sensitive to sampling size (on one hand, a small sample size might not be informative enough to infer migration profiles and on the other hand, analyzing thousands of sequences using the MCMC procedure is extremely computationally intensive and MCMC parameters often fail to converge)[20,28,61]. In addition, skewed sampling may further bias inference due to over-sampling some traits compared to others. It is therefore essential that the sampling strategy ensures a sufficiently representative number of samples from each discrete trait to avoid over-scoring transitions or counts in the empirical tree distribution. This necessitates down-sampling over-sampled traits to reduce bias, and excluding under-represented traits from the analysis[60,64,65]. In our dataset, Western, Central, Eastern and North-Eastern provinces were underrepresented and hence excluded, and temporal focus was limited to sequences collected after 2004. Focus was on transitions between four locations (Nyanza, Rift Valley,

Nairobi, and Coast), and between risk groups (MSM, PWID, FSW, and HET), and several approaches were used to limit sampling bias arising from the disproportional allocation of sequences per discrete state (described in detail below). HIV-1 sequences were first annotated with the year of sampling (2004-2019) and a discrete trait (risk group or location). In-house Perl scripts were used to randomize and select a set of sequences with uniform or proportional probability whilst also ensuring temporal sampling fidelity[60].

In detail, in the first scenario, location-annotated HIV-1 sequences were sub-sampled proportional to the HIV-1 prevalence per geographic province. This procedure was independently replicated 30 times – resulting in 30 datasets each having 892 sequences of which 35% were from Nyanza, 17% Rift Valley, 13% Nairobi, and 7% Coast. A similar approach was taken with risk group as a discrete state – resulting in thirty datasets each having 802 sequences of which 64% were from HET, 14% FSW, 15% MSM, and 4% PWID. Cluster analysis (as described above) was performed independently for each dataset. Clusters having >14 sequences were identified – and discrete state phylogeographic analysis with Markov jumps inferences were then performed independently for each of the identified clusters.
Next, we further explored whether the population dynamics seen in recent years (i.e. 2010-2019) were different from those observed in the complete dataset (i.e. 2004-2019). In the second sensitivity analysis, HIV-1 A1 sequences collected during 2010-2019 were sub-sampled proportionally as was done in the first scenario – resulting in five independent datasets with location-annotation (each having 144 sequences – 35% from Nyanza, 17% Rift Valley, 13% Nairobi, and 7% Coast), and five independent datasets with risk group annotation (each having 97 sequences – 64% HET, 14% FSW, 15% MSM, and 4% PWID). However, unlike in the cluster-wise approach, the complete sub-sampled datasets were used directly to estimate virus migration between states. In the third sensitivity analysis, HIV-1 A1 sequences collected during 2010-2019 were sub-sampled uniformly into five datasets with equal number of sequences per discrete state. The location-annotated dataset had 100 sequences (25 sequences from each province), while the dataset annotated for risk group had 108 sequences (27 sequences for each risk group).

Statistical analysis

Changes in the proportion of HIV-1 subtypes and recombinants over time were assessed using the *nptrend* non-parametric test for trends [66]. Frequencies and percentages were used to describe the distribution of sequences within the study population. A logistic regression model was used to assess associations between individual sequence characteristics (e.g. subtype/CRF, location of sampling, risk group, and year [range] of sampling) and phylogenetic clustering. Variables with $p<0.1$ in exploratory bivariable analyses were included in the multivariable model, in which $p<0.05$ was defined as statistically significant. A Kruskal-Wallis H test and a post hoc Dunn's test with Bonferroni correction for multiple comparisons were conducted to determine differences in HIV-1 evolutionary rate, cluster growth rates, and time to the most recent common ancestor (tMRCA) estimates among clusters from multiple risk groups. Statistics and summary plots were done using Stata 15 (StataCorp LLC, College Station, Texas, USA) and RStudio (version 1.2.5001) with the packages: *yarrr, circlize* and *ggplot2* [67-69].

Ethical considerations

All research was performed following relevant guidelines/regulations. For the newly generated sequences, informed consent for use of plasma samples was obtained from all participants from respective studies. Since published sequences were obtained from an open-access public domain, informed consent was not retrospectively obtained. Instead, we sought approval through a study protocol that was reviewed by the Kenya Medical Research Institute (KEMRI) Scientific and Ethics Review Unit (SERU 3547).

Data availability

Newly generated nucleotide sequences were deposited in GenBank under the following accession numbers: MT084914-MT085076, and OM109695-OM110282.

RESULTS

Study population and sequence dataset

We analysed 4058 HIV-1 *pol* sequences collected 1986-2019, of which 3303 (81.4%) were previously published and 755 (18.6%) newly generated for this study (Table 1, Supplementary Figure S1, and Supplementary Table S1). Most sequences were from HET (N=3401, 83.8%), followed by MSM (N=372, 9.2%), FSW (N=227, 5.6%), and PWID (N=58, 1.4%). Overall, these numbers represent an estimated sampling density of 0.3% of the HIV-1 epidemic in Kenya, and specific sampling densities of 10.8% for MSM, 1.7% for PWID, 0.6% for FSW, and 0.3% for HET (Supplementary Table S2). Sequences were available from seven (of eight) former administrative provinces in Kenya: Nairobi (N=1440, 35.5% of the sequences in this study); Coast (N=1061, 26.2%); Nyanza (N=665, 16.4%); Rift Valley (N=508, 12.5%); Western (N=158, 3.8%); Central (N=44, 1.1%); Eastern (N=6, 0.2%); and 176 (4.3%) sequences with missing data on sampling location (Table 1, and Figure 1). All PWID sequences were derived from the Coast province. Sampling year and place were missing for 176 (4.3%) of the newly generated HET sequences. These sequences were included in the assessment of subtype diversity in Kenya but excluded from the Bayesian phylodynamic analysis (which necessitates information on sampling date). In our dataset, 14 MSM identified as transgender persons. Subsequent sub-analyses were made to tease out clustering patterns specific for transgender persons relative to other risk groups.

HIV-1 sub-subtype A1 and subtype D dominated the epidemic in Kenya

Among the combined new and published Kenyan sequences (N=4058, Supplementary Table S3), HIV-1 sub-subtype A1 was most common (70.5%) followed by subtype D (11.4%, Supplementary Figure S2). Sub-subtype A1 was also the most common HIV-1 strain in all provinces and amongst all risk groups (Supplementary Table S4, and Supplementary Table S5, respectively). Temporal trend analysis in subtype distribution was restricted to the period after 2004 that comprised 92.0% of the sequences (Supplementary Figure S2). Sub-subtype A1 infections increased from 59.7% to 78.3%, 2004-2019 ($p<0.001$). No significant change was seen for subtype C ($p=0.30$) or subtype D ($p=0.59$), whereas subtype G decreased from 1.2% to 0.0%, 2004-2019 ($p=0.013$). Overall, CRFs decreased from 2.7% to 0.0%, 2004-2019 ($p=0.005$), whereas URFs decreased from 11% to 0.9%, 2004-2019 ($p=0.001$). Bayesian inference also revealed that the effective population size estimates for HIV-1 sub-subtype A1 were consistently higher than those for HIV-1 subtypes C and D throughout the study period (Figure 2).

HIV-1 geographic mixing within and between provinces in Kenyan

Overall, 1832 (45%) of Kenyan sequences were found in 409 clusters including sub-subtype A1 (N=306, 74.8%), subtype C (N=25, 6.1%), and subtype D (N=78, 19.1%) clusters (Table 2, Supplementary Table S6, Supplementary Figure S3, and Supplementary Figure S4).

Overall, 1485 (51.9%) of sub-subtype A1 sequences, 137 (48.1%) subtype C, and 210 (45.6%) subtype D formed clusters. The remaining 1375 (48.1%) sub-subtype A1, 148 (51.9%) subtype C, and 251 (54.5%) sequences were singletons (Supplementary Table S6). Majority (N=248, 60.6%) were province-exclusive, including clusters from Nairobi (N=107, 26.2%), Coast (N=58, 14.2%), Nyanza (N=51, 12.5%), Rift Valley (N=23, 5.6%), Western (N=6, 1.5%), and Central (N=3, 0.7%). The remaining clusters (N=161, 39.4%) were mixed between different geographic provinces (Supplementary Figure S5a).

Within-risk group clustering dominated among Kenyan HIV-1 clusters

Majority (N=362, 88.5%) of the clusters represented within-risk group HIV-1 transmission including HET (N=316; 72.1%), MSM (N=37, 9.1%), FSW (N=7, 1.7%) and PWID (N=2, 0.5%). Further and amongst PWID, only two clusters were identified (one dyad and one large cluster, both PWID exclusive), with the large cluster comprising 80% of all PWID sequences in the dataset (N=41). The remaining clusters (N=47, 11.5%) involved mixed linkages between different risk groups including MSM/HET (N=15, 3.7% of all clusters), FSW/HET (N=15, 3.7%), MSM/FSW/HET (N=9, 2.2%), MSM/FSW (N=6, 1.5%), MSM/PWID/FSW/HET (N=1, 0.2%), and PWID/HET (N=1, 0.2%) mixed clusters (Table 2, Supplementary Figure S5b). A sub-analysis of clustering patterns involving transgender people showed that nine of 14 (64.3%) clustered with MSM, four clustered with HET (28.6%), and one did not cluster with any other sequences in the dataset (7.1%). Compared to HET,

MSM and PWID sequences were more likely to cluster (adjusted odds ratio [aOR] 4.4, 95% confidence interval [CI] 3.2-6.0, p<0.001; and aOR 3.4, CI 1.8-6.5, p<0.001, respectively, Table 3).

The effective population size has stabilised over time amongst all risk groups

The correlation between divergence from root and time of sampling was low in our dataset (i.e. R^2 = 0.139, 0.136, and 0.121 for the sub-subtype A1, subtype C, and subtype D datasets, respectively, Supplementary Figure S6). Thus normal priors were specified for the time of the most recent common ancestor (tMRCA) of sub-subtype A1, subtype C and subtype D, based on previous estimations[20,22]. The inference of HIV-1 dynamics in the Kenyan epidemic was based on a Bayesian phylodynamic analysis of the large Kenyan HIV-1 clusters (19 sub-subtype A1 and one subtype C cluster (Supplementary Table S7). All sub-subtype A1 HET clusters exhibited similar dynamics (Supplementary Figure S7) and were merged in one plot to assess overall dynamics among HET (Figure 3a). The number of effective infections (proportional to the transmission rate over the prevalence) for HET increased over time from 1987 to the mid-2000s, after which infections stabilised. The number of Kenyan PWID contributing to new HIV-1 infections over time increased gradually from 1987 to 2010, the latest sampling date for PWID (Figure 3c), whereas the MSM-exclusive cluster showed stable dynamics with no periods of exponential growth between 1991 and 2019, the latest sampling date for MSM (Figure 3d). The only large subtype C cluster that was found was a HET cluster – this cluster showed similar dynamics as the sub-subtype A1 HET clusters, with increasing effective population size from 1983 to the early 2000s followed by a stabilisation (Figure 3b).

Evolutionary parameters were similar among clusters of different risk groups

Subtype C had the earliest tMRCA (1977, 95% higher posterior density [HPD] interval: 1968-1985) of all clusters. The median tMRCA estimates of sub-subtype A1 clusters indicated multiple introductions into Kenya over 42 years (1978-2019), with most clusters introduced between the late 1980s and early 1990s. The earliest tMRCA for a Kenyan HET cluster was estimated to 1978 (95% HPD interval: 1971-1990); MSM to 1991 (HPD interval: 1974-2004); and PWID to 1987 (HPD interval: 1985-1990). The median HIV-1 evolutionary rates ranged from 1.01×10^{-3} to 1.3×10^{-3} substitutions site^{-1} year^{-1} (s/s/y) for subtype A1 in HET clusters and 1.28×10^{-3} to 1.34×10^{-3} s/s/y for mixed-risk group clusters. The median HIV-1 evolutionary rate for the only large MSM cluster was 9.80×10^{-4} s/s/y, and 1.06×10^{-3} s/s/y for the only large PWID cluster. Pairwise comparison of median evolutionary rates (with Bonferroni correction for multiple comparisons) showed no difference in evolutionary rates between HET and MSM (p=0.169), HET and PWID (p=1.00), and MSM and PWID (p=0.297). No statistical differences were found between tMRCA estimates or cluster growths between clusters of different risk groups, respectively (p=0.822, and p=0.321, Table 4, Figure 4).

Evidence of West-to-East HIV-1 migration, and transmission from HET to key populations

Phylogeographic analysis was based on HIV-1 sub-subtype A1 – the strain with the highest number of sequences in our study, and the most dominant strain circulating strain in Kenya. In all sensitivity analyses, Western, Central and Eastern provinces were excluded as they had the smallest number of sequences in the study, and sequences from transgender people and MSM were analysed together as one risk group. The Markov jumps estimates from the cluster-wise phylogeographic inference indicated that the majority (62.6%) of HIV-1 jumps occurred within Kenyan borders whilst the remaining involved HIV-1 export (24.1%) from Kenya to other countries, and HIV-1 import (13.2%) to Kenya (Table 5). The proportion of West-to-East jumps over time was significantly higher than that of East-to-West jumps (p=0.001, Figure 5a, and 4b). West-to-East migration accounted for the majority (76.1%) of all within-country jumps – including jumps from Nyanza to Nairobi (10.3%), Rift Valley to Nairobi (9.8%), Nyanza to Rift Valley (9.2%), Nyanza to Coast (6.3%), Rift Valley to Coast (6.3%), and Nairobi to Coast (5.7%). East-to-West migration accounted for only 23.9% within-country jumps and comprised jumps from Rift Valley to Nyanza (7.5%), Nairobi to Nyanza (4.6%), and Nairobi to Rift Valley (2.9%, Figure 5b). Pairs of geographic provinces located next to each other were involved in an extensive cyclic HIV-1 exchange – including transmission from Nyanza to Rift Valley (9.2% forward jumps versus 7.5% reverse jumps) and Rift Valley to Nairobi (9.8% vs 2.9%). Although Coast province received a significant proportion of translocated HIV-1 lineages (18.3% of all HIV-1 jumps), no within-country HIV-1 jumps were observed as originating from Coast province. Uniform and proportional sub-

sampling of the sequences collected 2010-2019 indicated more West-to-East virus flow than vice-versa ($p<0.001$ for all comparisons, Table 6, Supplementary Figure S8a and Supplementary Figure S8b).

The cluster-wise phylogeographic inference showed that 82.9% of virus jumps between risk groups were from HET (involving HET-to-FSW [34.0%], HET-to-MSM [31.9%], and HET-to-PWID [17.0%]). Only 12.8% virus jumps were from key populations (involving MSM-to-HET [6.4%] and PWID-to-HET [6.4%], Figure 5d). The remaining were MSM-to-FSW virus jumps (4.3%, Table 5). Also, the proportion of virus jumps from HET to key populations over time was significantly higher compared with virus jumps from key populations to HET ($p<0.001$, Figure 5c). The earliest estimated Markov jump event from HET-to-FSW occurred in 1981, followed by HET-to-MSM (1986), and HET-to-PWID (1990, Figure 5d). Virus jumps among HET were common as early as during the 1980s while virus jumps among MSM (i.e. MSM-to-MSM) and among PWID (i.e. PWID-to-PWID) increased during the 1990s and 2000s, respectively (Figure 5d). Uniform and proportional sub-sampling of the sequences collected 2010-2019 indicated more HIV-1 jumps from HET to key populations than vice-versa ($p<0.001$ for all comparisons, Table 6, Supplementary Figure S8c and Supplementary Figure S8d)

DISCUSSION

We show that HIV-1 transmission in Kenya was largely compartmentalized by risk groups. This result is based on the identification of 409 statistically supported phylogenetic clusters – where a majority (88.5%) represents within-risk group clustering. Furthermore, we found that 11.5% of the clusters represented HIV-1 mixing between risk groups – including approximately 7.6% HIV-1 mixing between MSM and HET in Kenya. These findings are consistent with previous phylogenetic data in Coastal Kenya demonstrating minimal HIV-1 mixing between key populations and the heterosexual population[18,19]. We have previously estimated frequent (85%) within-risk group clustering, and minimal (15%) HIV-1 mixing between MSM and the HET in Coastal Kenya[18]. Likewise, Bezemer and colleagues – albeit with a small sample size and sequences only from Nairobi and Coast province only found one HIV-1 MSM/HET link, indicating infrequent HIV-1 mixing between MSM and HET[19]. The phylogeographic inference indicated a higher proportion of HIV-1 jumps from HET to MSM, FSW and PWID. However, the detected virus jumps represent rare events as overall transmission between risk groups is itself rare in the Kenyan epidemic (as shown in the cluster analysis). Yet, these findings indicate that contrary to concerns by the Ministry of Health in Kenya[16], HIV-1 key populations may not disproportionately transmit HIV-1 to heterosexuals in the general epidemic. Also, it is well established that the vast majority of HIV-1 transmission in Kenya could be attributed to risky heterosexual behaviours[15,70].

Overall, our study highlights important dynamics in HIV-1 spread in the context of a mixed HIV-1 epidemic and support the hypothesis of frequent within-risk group transmission and limited between-risk group transmission[18,19]. This hypothesis is further strengthened by findings from a review of 35 studies assessing HIV-1 mixing between HIV-1 populations in sub-Saharan Africa highlighting the predominance of within-risk group transmission chains in most countries[10]. To reduce population-level HIV-1 incidence in sub-Saharan Africa, HIV-1 control programs may require both broad-reaching interventions aimed at the general epidemic, as well as strengthening micro-strategies that address disparities among population categories (including scale-up of ART, HIV-1 testing and other prevention programs directed towards key populations such as MSM, PWID and FSW who are most-at-risk of infection)[2,31,71-74].

In this study, HIV-1 transmission in Kenya involved predominantly West-to-East dissemination, notably from high HIV-1 prevalence regions (including the former Nyanza province in Western Kenya) to comparatively lower HIV-1 prevalence regions (including former Coastal province). Irrespective of transmission risk, the largest number of people with HIV-1, and approximately 40% of all newly diagnosed HIV-1 infections have been suggested to occur in Western Kenya[11]. It is therefore plausible that the observed HIV-1 dissemination pattern reflects considerable HIV-1 transmission from high-to-low HIV-1 prevalence regions, a finding that likely applies to other sub-Saharan African countries with substantial within-country variation in the prevalence of HIV-1. However, our findings contrast data from Uganda showing significant virus flow from low-to-high HIV-1 prevalence populations along the Lake Victoria[21,27,28]. In the current study, we did not have data on fishing folk and we did not assess transmission between fishing folk and inland communities. Yet, it is possible that some undisclosed fishing-folk were grouped with HET (unless where the risk group was known) and classified as belonging to the Nyanza province. The gradient in HIV-1 prevalence in Kenya decreases Eastwards, and we observe an overall higher proportion of HIV-1 migration from provinces in the West (Nyanza and Rift valley) towards provinces in the East (such as the Coast province). Mathematical modelling and empirical evidence have shown that directed approaches may reduce HIV-1 incidence across sub-Saharan Africa[76-79]. Optimizing existing prevention strategies in geographic HIV-1 hotspots[75] in sub-Saharan Africa (such as Western Kenya) may therefore result in declining population-level HIV-1 incidence[3,80].

Our study represents one of the largest national-level analyses of HIV-1 *pol* diversity that has been done in Africa. However, we were still limited by a low sampling density and data on how the study participants in the published studies were identified for sequencing. Low sampling likely resulted in missing links in identified Kenyan clusters and low probability of detecting some rare subtypes circulating in Kenya[81]. Moreover, PWID and their partners, as well as the clients of sex workers, were

less likely to get into treatment studies and were therefore underrepresented in this study. It is therefore likely that the rates of HIV-1 transmissions from FSW, MSM and PWID to the HET population were underestimated owing to those missing links. Despite the lower sampling density of HET compared to MSM, PWID, and FSW sequences in the full dataset, our sensitivity analyses controlling for sampling bias indicated more virus jumps from HET to key populations. The observed links would likely not be fewer if additional HET samples were included to match the higher sampling density among MSM, PWID and FSW. Also, excluding some geographic locations from our sensitivity analysis due to few numbers of sequences from these provinces in our dataset may have resulted in missing transmission chains and links which may have implications in the dynamics of geographic HIV-1 spread in Kenya[53,81]. Nonetheless, the excluded provinces have HIV-1 prevalence rates lower than the national average and based on findings from this analysis, it is unlikely that they would be major sources of HIV-1 in Kenya. Lastly, we assessed HIV-1 flow between populations, not between individuals, and these population-level inferences may not be extrapolated to individual transmissions. Also, virus jumps between risk populations in the phylogeographic analyses may not be equated with transmission events because the discrete phylogeographic modelling used in this analysis only accounts for between-risk group jump, and not within-risk group jumps. Other similar studies from developed settings with concentrated epidemics and dense sampling among infected individuals (as well as readily available patient demographic data) have provided information useful in HIV-1 prevention[5,6,23,82-86]. To minimise phylogenetic uncertainties arising from low sample coverage, future studies in sub-Saharan Africa should aim to achieve higher sampling densities and aim to include sequences collected in years that are more recent to determine more active Kenyan clusters.

In conclusion, we have estimated the rates of transmission between the general heterosexual population and HIV-1 key populations, and between geographic regions with varying HIV-1 prevalence in Kenya. We showed that high HIV-1 prevalence regions may be important sources of HIV-1 to lower-prevalence regions, and that the Kenyan HIV-1 epidemic is largely compartmentalized by risk groups and that the contribution of key populations to the wider heterosexual transmission network may be significantly lower than vice versa. In the mixed Kenyan HIV-1 epidemic, targeting HIV-1 key populations needs to occur concurrently with strengthening broad interventions in the general population. These findings could pave the way towards strengthening HIV-1 control in Kenya and other countries in sub-Saharan Africa.

ADDITIONAL INFORMATION

Acknowledgements

We thank the staff affiliated with the MSM Health Research Consortium (MHRC) and IAVI for supporting studies involving key populations in Kenya. This manuscript was submitted for publication with permission from the Director of the Kenya Medical Research Institute (KEMRI).

Author contributions

A.S.H., E.J.S., and J.E conceptualized and designed the study. A.S.H., E.J.S. and J.E provided funding for the study. E.J.S., S.M.G., J.K, L.M, F.C, G.M, M.M, O.A, L.G, A.S, and R.B provided samples from which new sequences used in the study were generated. G.N.M performed lab work, inferential analyses and produced all figures and tables. G.B and PL assisted with phylogeographic analysis and E.W helped with data analysis. G.N.M and J.E. wrote the manuscript and all the authors reviewed, edited, and approved the manuscript for submission.

Competing Interests

The authors declare no competing interests.

Funding information

This work was supported through the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant #DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant #107752/Z/15/Z] and the UK government. This work was also supported in part by funding from the Swedish Research Council (grant #2016–01417) and the Swedish Society for Medical Research (grant #SA-2016). The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS). PL acknowledges support by the Research Foundation - Flanders ('Fonds voor Wetenschappelijk Onderzoek - Vlaanderen', G066215N, G0D5117N and G0B9317N). GB acknowledges support from the Interne Fondsen KU Leuven / Internal Funds KU Leuven under grant agreement C14/18/094, and the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek - Vlaanderen', G0E1420N). IAVI's support is made possible by the generous support of the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) through the United States Agency for International Development (USAID). The full list of IAVI donors is available at www.iavi.org. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, IAVI, PEPFAR, USAID or the United States Government, Swedish Research Council, or the UK government.

References

1 Joint United Nations Programme on HIV/AIDS (UNAIDS). *UNAIDS Global AIDS Report*, <https://www.unaids.org/sites/default/files/media_asset/2020_global-aids-report_en.pdf> (2020).

2 Kelly, S. L. *et al.* The global Optima HIV allocative efficiency model: targeting resources in efforts to end AIDS. *The Lancet HIV* **5**, e190-e198 (2018).

3 Anderson, S.-J. *et al.* Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study. *The Lancet* **384**, 249-256 (2014).

4 Smith, A. D., Tapsoba, P., Peshu, N., Sanders, E. J. & Jaffe, H. W. Men who have sex with men and HIV/AIDS in sub-Saharan Africa. *Lancet (London, England)* **374**, 416-422, doi:10.1016/s0140-6736(09)61118-1 (2009).

5 Poon, A. F. *et al.* Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *The lancet HIV* **3**, e231-e238 (2016).

6 Ratmann, O. *et al.* Sources of HIV infection among men having sex with men and implications for prevention. *Science translational medicine* **8**, 320ra322-320ra322 (2016).

7 Esbjörnsson, J. *et al.* HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. *Virus evolution* **2**, vew010 (2016).

8 Joint United Nations Programme on HIV/AIDS (UNAIDS). *UNAIDS DATA 2018*, <https://www.unaids.org/sites/default/files/media_asset/unaids-data-2018_en.pdf> (2018).

9 Abeler-Dörner, L., Grabowski, M. K., Rambaut, A., Pillay, D. & Fraser, C. PANGEA-HIV 2: phylogenetics and networks for generalised epidemics in Africa. *Current Opinion in HIV and AIDS* **14**, 173 (2019).

10 Nduva, G. M., Nazziwa, J., Hassan, A. S., Sanders, E. J. & Esbjörnsson, J. The Role of Phylogenetics in Discerning HIV-1 Mixing among Vulnerable Populations and Geographic Regions in Sub-Saharan Africa: A Systematic Review. *Viruses* **13**, 1174 (2021).

11 National AIDS and STI Control Programme (NASCOP). *Preliminary KENPHIA 2018 Report*, <https://www.nascop.or.ke/kenphia-report> (2020).

12 Kreiss, J. K. *et al.* AIDS virus infection in Nairobi prostitutes. Spread of the epidemic to East Africa. *N Engl J Med* **314**, 414-418, doi:10.1056/nejm198602133140704 (1986).

13 Makofane, K., van der Elst, E. M., Walimbwa, J., Nemande, S. & Baral, S. D. From general to specific: moving past the general population in the HIV response across sub-Saharan Africa. *Journal of the International AIDS Society* **23**, e25605 (2020).

14 Sanders, E. J. *et al.* HIV-1 infection in high risk men who have sex with men in Mombasa, Kenya. *Aids* **21**, 2513-2520 (2007).

15 Kenya National AIDS Control Council. *Kenya HIV Prevention Response and Modes of Transmission Analysis.*, <https://icop.or.ke/wp-content/uploads/2016/09/KenyaMOT-2009.pdf> (2009).

16 National AIDS and STI Control Programme. *Kenya HIV County Profiles 2016.*, <http://nacc.or.ke/wp-content/uploads/2016/12/Kenya-HIV-County-Profiles-2016.pdf> (2017).

17 Kenya National AIDS Control Council. *Kenya AIDS Strategic Framework 2014/2015–2018/2019*, <http://nacc.or.ke/wp-content/uploads/2015/09/KASF_Final.pdf> (2019).

18 Nduva, G. M. *et al.* HIV-1 Transmission Patterns Within and Between Risk Groups in Coastal Kenya. *Sci Rep* **10**, 6775, doi:10.1038/s41598-020-63731-z (2020).

19 Bezemer, D. *et al.* HIV Type 1 transmission networks among men having sex with men and heterosexuals in Kenya. *AIDS research and human retroviruses* **30**, 118-126 (2014).

20 Faria, N. R. *et al.* The early spread and epidemic ignition of HIV-1 in human populations. *science* **346**, 56-61 (2014).

21 Grabowski, M. K. *et al.* Migration, hotspots, and dispersal of HIV infection in Rakai, Uganda. *Nature communications* **11**, 1-12 (2020).

22 Faria, N. R. *et al.* Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa. *PLoS pathogens* **15**, e1007976-e1007976 (2019).

23 Sallam, M. *et al.* Molecular epidemiology of HIV-1 in Iceland: Early introductions, transmission dynamics and recent outbreaks among injection drug users. *Infection, Genetics and Evolution* **49**, 157-163 (2017).
24 Hassan, A. S. *et al.* HIV-1 subtype diversity, transmission networks and transmitted drug resistance amongst acute and early infected MSM populations from Coastal Kenya. *PloS one* **13**, e0206177 (2018).
25 Esbjörnsson, J., Mild, M., Månsson, F., Norrgren, H. & Medstrand, P. HIV-1 molecular epidemiology in Guinea-Bissau, West Africa: origin, demography and migrations. *PloS one* **6**, e17025 (2011).
26 Nazziwa, J. *et al.* Characterisation of HIV-1 Molecular Epidemiology in Nigeria: Origin, Diversity, Demography and Geographic Spread. *Sci Rep* **10** (2020).
27 Ratmann, O. *et al.* Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda. *The Lancet HIV* **7**, e173-e183 (2020).
28 Bbosa, N. *et al.* Phylogeography of HIV-1 suggests that Ugandan fishing communities are a sink for, not a source of, virus from general populations. *Scientific reports* **9**, 1-8 (2019).
29 De Oliveira, T. *et al.* Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *The lancet HIV* **4**, e41-e50 (2017).
30 Sanders, E. J. *et al.* High HIV-1 incidence, correlates of HIV-1 acquisition, and high viral loads following seroconversion among MSM. *Aids* **27**, 437-446, doi:10.1097/QAD.0b013e32835b0f81 (2013).
31 Smith, A. D. *et al.* HIV burden and correlates of infection among transfeminine people and cisgender men who have sex with men in Nairobi, Kenya: an observational study. *The lancet. HIV*, doi:10.1016/s2352-3018(20)30310-6 (2021).
32 Kunzweiler, C. P. *et al.* Depressive Symptoms, Alcohol and Drug Use, and Physical and Sexual Abuse Among Men Who Have Sex with Men in Kisumu, Kenya: The Anza Mapema Study. *AIDS Behav* **22**, 1517-1529, doi:10.1007/s10461-017-1941-0 (2018).
33 Los Alamos National Laboratory. *HIV-1 database at the Los Alamos National Laboratory*, <http://www.hiv.lanl.gov/> (2019).
34 Hassan, A. S. *et al.* Low prevalence of transmitted HIV type 1 drug resistance among antiretroviral-naive adults in a rural HIV clinic in Kenya. *AIDS Res Hum Retroviruses* **29**, 129-135, doi:10.1089/aid.2012.0167 (2013).
35 Hué, S. *et al.* HIV type 1 in a rural coastal town in Kenya shows multiple introductions with many subtypes and much recombination. *AIDS research and human retroviruses* **28**, 220-224 (2012).
36 Zeh, C. *et al.* Molecular Epidemiology and Transmission Dynamics of Recent and Long-Term HIV-1 Infections in Rural Western Kenya. *PLoS One* **11**, e0147436, doi:10.1371/journal.pone.0147436 (2016).
37 Tovanabutra, S. *et al.* Evaluation of HIV type 1 strains in men having sex with men and in female sex workers in Mombasa, Kenya. *AIDS research and human retroviruses* **26**, 123-131 (2010).
38 Sigaloff, K. C. *et al.* High prevalence of transmitted antiretroviral drug resistance among newly HIV type 1 diagnosed adults in Mombasa, Kenya. *AIDS research and human retroviruses* **28**, 1033-1037 (2012).
39 Onywera, H. *et al.* Surveillance of HIV-1 pol transmitted drug resistance in acutely and recently infected antiretroviral drug-naive persons in rural western Kenya. *PloS one* **12**, e0171124 (2017).
40 Gounder, K. *et al.* Complex Subtype Diversity of HIV-1 Among Drug Users in Major Kenyan Cities. *AIDS Res Hum Retroviruses* **33**, 500-510, doi:10.1089/aid.2016.0321 (2017).
41 Yang, C. *et al.* Genetic diversity and high proportion of intersubtype recombinants among HIV type 1-infected pregnant women in Kisumu, western Kenya. *AIDS Res Hum Retroviruses* **20**, 565-574, doi:10.1089/088922204323087822 (2004).
42 Hamers, R. L. *et al.* HIV-1 drug resistance in antiretroviral-naive individuals in sub-Saharan Africa after rollout of antiretroviral therapy: a multicentre observational study. *The Lancet. Infectious diseases* **11**, 750-759, doi:10.1016/s1473-3099(11)70149-9 (2011).

43 Esbjörnsson, J. *et al.* Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease-indication of an evolving epidemic in West Africa. *Retrovirology* **7**, 23 (2010).
44 Hedskog, C. *et al.* Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PloS one* **5**, e11345 (2010).
45 Kenya National Bureau of Statistics. *2019 Kenya population and housing census Volume 1: Population by county and sub-county*, <https://www.knbs.or.ke/?wpdmpro=2019-kenya-population-and-housing-census-volume-i-population-by-county-and-sub-county> (2019).
46 Kenya National AIDS control council (NACC). *Kenya HIV estimates report 2018*, <https://nacc.or.ke/wp-content/uploads/2018/11/HIV-estimates-report-Kenya-20182.pdf> (2018).
47 National AIDS and STI Control Programme (NASCOP). *Key Population Mapping and Size Estimation in Selected Counties in Kenya: Phase 1*, <https://hivpreventioncoalition.unaids.org/wp-content/uploads/2020/02/KPSE-Phase1-Final-Report.pdf> (2019).
48 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *bioinformatics* **23**, 2947-2948 (2007).
49 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307-321 (2010).
50 Lole, K. S. *et al.* Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *Journal of virology* **73**, 152-160 (1999).
51 Struck, D., Lawyer, G., Ternes, A.-M., Schmit, J.-C. & Bercoff, D. P. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic acids research* **42**, e144-e144 (2014).
52 Pineda-Peña, A.-C. *et al.* 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**, 337-348 (2013).
53 Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J. & Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data. *AIDS (London, England)* **31**, 1211 (2017).
54 Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution* **2**, vew007 (2016).
55 Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus evolution* **4**, vey016 (2018).
56 Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular biology and evolution* **29**, 2157-2167 (2012).
57 Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol* **30**, 713-724, doi:10.1093/molbev/mss265 (2013).
58 Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution* **22**, 1185-1192 (2005).
59 Suchard, M. A., Kitchen, C. M., Sinsheimer, J. S. & Weiss, R. E. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic biology* **52**, 649-664 (2003).
60 Hall, M. D., Woolhouse, M. E. & Rambaut, A. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods: A simulation study. *Virus evolution* **2** (2016).
61 Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS computational biology* **5** (2009).
62 Edwards, C. J. *et al.* Ancient hybridization and an Irish origin for the modern polar bear matriline. *Current Biology* **21**, 1251-1258 (2011).
63 Minin, V. N. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of mathematical biology* **56**, 391-412 (2008).
64 Volz, E. M. Complex population dynamics and the coalescent under neutrality. *Genetics* **190**, 187-201 (2012).

65 de Silva, E., Ferguson, N. M. & Fraser, C. Inferring pandemic growth rates from sequence data. *Journal of The Royal Society Interface* **9**, 1797-1808 (2012).
66 Cuzick, J. A Wilcoxon-type test for trend. *Statistics in medicine* **4**, 87-90 (1985).
67 Wickham, H. *ggplot2: elegant graphics for data analysis*. (springer, 2016).
68 Phillips, N. D. Yarrr! The pirate's guide to R. *APS Observer* **30** (2017).
69 Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).
70 Gouws, E. & Cuchi, P. Focusing the HIV response through estimating the major modes of HIV transmission: a multi-country analysis. *Sexually transmitted infections* **88**, i76-i85 (2012).
71 Cremin, I. *et al.* PrEP for key populations in combination HIV prevention in Nairobi: a mathematical modelling study. *The lancet. HIV* **4**, e214-e222, doi:10.1016/s2352-3018(17)30021-8 (2017).
72 Koss, C. A. *et al.* HIV incidence after pre-exposure prophylaxis initiation among women and men at elevated HIV risk: A population-based study in rural Kenya and Uganda. *PLoS medicine* **18**, e1003492 (2021).
73 Tago, A. *et al.* Declines in HIV prevalence in female sex workers accessing an HIV treatment and prevention programme in Nairobi, Kenya over a 10-year period. *AIDS* **35**, 317-324, doi:10.1097/qad.0000000000002747 (2021).
74 Tago, A. *et al.* Declines in HIV prevalence in female sex workers (FSWs) accessing an HIV treatment and prevention programme in Nairobi, Kenya over a 10-year period. *Aids*, doi:10.1097/qad.0000000000002747 (2020).
75 Dwyer-Lindgren, L. *et al.* Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature* **570**, 189-193 (2019).
76 Gerberry, D. J., Wagner, B. G., Garcia-Lerma, J. G., Heneine, W. & Blower, S. Using geospatial modelling to optimize the rollout of antiretroviral-based pre-exposure HIV interventions in Sub-Saharan Africa. *Nature communications* **5**, 1-15 (2014).
77 McGillen, J. B., Anderson, S.-J., Dybul, M. R. & Hallett, T. B. Optimum resource allocation to reduce HIV incidence across sub-Saharan Africa: a mathematical modelling study. *The lancet HIV* **3**, e441-e448 (2016).
78 Grabowski, M. K. *et al.* HIV prevention efforts and incidence of HIV in Uganda. *New England Journal of Medicine* **377**, 2154-2166 (2017).
79 Vandormael, A. *et al.* Declines in HIV incidence among men and women in a South African population-based cohort. *Nature communications* **10**, 1-10 (2019).
80 Bailey, R. C. *et al.* Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *The lancet* **369**, 643-656 (2007).
81 Novitsky, V., Moyo, S., Lei, Q., DeGruttola, V. & Essex, M. Impact of sampling density on the extent of HIV clustering. *AIDS research and human retroviruses* **30**, 1226-1235 (2014).
82 Vasylyeva, T. I. *et al.* Molecular epidemiology reveals the role of war in the spread of HIV in Ukraine. *Proceedings of the National Academy of Sciences* **115**, 1051-1056, doi:10.1073/pnas.1701447115 (2018).
83 Volz, E. M. *et al.* HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS Med* **10**, e1001568; discussion e1001568, doi:10.1371/journal.pmed.1001568 (2013).
84 Ragonnet-Cronin, M. *et al.* Non-disclosed men who have sex with men in UK HIV transmission networks: phylogenetic analysis of surveillance data. *The Lancet HIV* **5**, e309-e316, doi:https://doi.org/10.1016/S2352-3018(18)30062-6 (2018).
85 Fisher, M. *et al.* Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. *Aids* **24**, 1739-1747 (2010).
86 Kouyos, R. D. *et al.* Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *The Journal of infectious diseases* **201**, 1488-1497 (2010).

TABLES

Table 1. Demographics and distribution of newly generated and published Kenyan HIV-1 *pol* sequences by risk group.

Category		Risk group				
		HET	**MSM**	**FSW**	**PWID**	**Total**
Sequences	Published	2987 (87.8%)	159 (42.7%)	99 (43.6%)	58 (100.0%)	3303 (81.4%)
	New	414 (12.2%)	213 (57.3%)	128 (56.4%)	0 (0.0%)	755 (18.6%)
Province	Nairobi	1212 (35.6%)	137 (36.8%)	91 (40.1%)	0 (0.0%)	1440 (35.5%)
	Coast	704 (20.7%)	178 (47.9%)	121 (53.3%)	58 (100.0%)	1061 (26.2%)
	Nyanza	594 (17.5%)	57 (15.3%)	14 (6.2%)	0 (0.0%)	665 (16.4%)
	Rift Valley	507 (14.9%)	0 (0.0%)	1 (0.4%)	0 (0.0%)	508 (12.5%)
	Western	158 (4.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	158 (3.9%)
	Central	44 (1.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	44 (1.1%)
	Eastern	6 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	6 (0.2%)
	Missing*	176 (5.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	176 (4.3%)
Year (range)	2001-2010	2077 (64.4%)	118 (31.7%)	170 (74.9%)	58 (100.0%)	2423 (59.7%)
	2011-2019	1070 (33.2%)	254 (68.3%)	36 (15.9%)	0 (0.0%)	1360 (33.5%)
	1986-2000	78 (2.4%)	0 (0.0%)	21 (9.3%)	0 (0.0%)	99 (2.4%)
	Missing*	176 (5.2%%)	0 (0.0%)	1 (0.0%)	2 (0.0%)	176 (4.3%)
Total		**3401 (83.8%)**	**372 (9.2%)**	**227 (5.6%)**	**58 (1.4%)**	**4058 (100.0%)**

Abbreviations: MSM, men who have sex with men; PWID, people who inject drugs; FSW, female sex worker; HET, at-risk men and women who did not report sex work or male same-sex behaviour. *Sequences lacking information on year and geographic area of sampling. *Estimated number of people with HIV-1 as per geographic and transmission route category in Kenya[11,45-47]. **Number of people with HIV-1 included in the study based on the estimated number of people with HIV-1 in Kenya.

Table 2. Kenyan HIV-1 clusters (N=409) grouped into different subtypes and HIV-1 transmission routes.

	Dyads [a]	Networks [b]	Large clusters [c]	Total (N,%)
Subtype				
A (A1)	182 (59%)	105 (34%)	19 (6%)	306 (75%)
C	16 (64%)	8 (32%)	1 (4%)	25 (6%)
D	51 (65%)	27 (35%)	0 (0%)	78 (19%)
Risk category				
HET	204 (65%)	101 (32%)	11 (3%)	316 (77%)
Mixed*	24 (51%)	16 (34%)	7 (15%)	47 (11%)
MSM	13 (35%)	23 (62%)	1 (3%)	37 (9%)
FSW	7 (100.0%)	0 (0%)	0 (0%)	7 (2%)
PWID	1 (50%)	0 (0%)	1 (50%)	2 (<1%)
Total	249 (61%)	140 (34%)	20 (5%)	409

Abbreviations: HET, heterosexual transmission; Mixed; MSM, men who have sex with men; FSW, female sex work; PWID, people who inject drugs. *Risk groups in mixed clusters (N, proportion of mixed clusters) : MSM/HET (15, 32%), FSW/HET (15, 32%), MSM/FSW/HET (9, 19%), MSM/FSW (6, 13%), MSM/PWID/FSW/HET (1, 2%), and PWID/HET (1, 2%). [a] Dyads: clusters of 2 sequences [b] Networks: clusters of 3-14 sequences [c] Large clusters: clusters of >14 sequences.

Table 3. Factors associated with clustering among HIV-1 sequences from Kenya.

Characteristics		Bivariate Analysis*		Multivariate Analysis**	
		OR (95% CI)	p-value	aOR (95% CI)	p-value
Risk category	HET	Reference			
	MSM	3.8 (3-4.8)	<0.001	4.4 (3.2-6.0)	<0.001
	PWID	4.7 (2.5-8.8)	<0.001	3.4 (1.8-6.5)	<0.001
	FSW	0.6 (0.5-0.9)	0.003	1.2 (0.8-1.7)	0.391
Subtype	A1	Reference			
	C	0.9 (0.7-1.1)	0.215		
	D	0.8 (0.6-0.9)	0.011	0.68 (0.6-0.9)	<0.001
Year (range)	1986-2000	Reference		Reference	
	2001-2010	3.7 (2.2-6.2)	<0.001	3.9 (2.1-7.0)	<0.001
	2011-2019	5.1 (3.0-8.7)	<0.001	5.3 (2.9-9.9)	<0.001
Province	Central	Reference			
	Coast	1.3 (0.7-2.4)	0.383		
	Eastern	0.3 (0-3)	0.314		
	Nairobi	1.6 (0.9-2.9)	0.141		
	Nyanza	1.4 (0.7-2.6)	0.297		
	Rift Valley	0.8 (0.4-1.6)	0.576		
	Western	1 (0.5-1.9)	0.936		
	Unknown	1 (0.5-1.9)	0.945		
Sequence category	New	Reference			
	Published	1.2 (1.1-1.5)	0.007	0.6 (0.5-0.8)	<0.001

Abbreviations: MSM, men who have sex with men; PWID, people who inject drugs; FSW, female sex worker; HET, at-risk men and women who did not report sex work or male same-sex behaviour. *Only variables with a p<0.1 in the bivariate analysis were included in the multivariate model (thus subtype C and province were excluded from the multivariate analysis).

Table 4. Estimated dates of origin and evolutionary parameters of the large Kenyan HIV-1 clusters.

Cluster	tMRCA*	Evolutionary rate (E^{-3})	Growth rate (per year)
A1.1.MIX	1989 [1984, 1994]	1.32 [1.00, 1.66]	0.16 [0.11, 0.21]
A1.2.HET	1986 [1977, 1993]	1.05 [0.73, 1.39]	0.18 [0.12, 0.25]
A1.3.HET	1982 [1971, 1990]	1.05 [0.72, 1.39]	0.24 [0.13, 0.36]
A1.4.MIX	1989 [1983, 1996]	1.31 [0.97, 1.67]	0.28 [0.17, 0.41]
A1.6.PWID	1987 [1985, 1990]	1.06 [0.67, 1.52]	0.15 [0.07, 0.26]
A1.7.MIX	1988 [1977, 1997]	1.28 [0.93, 1.64]	0.21 [0.12, 0.30]
A1.8.MIX	1978 [1963, 1993]	1.32 [0.97, 1.69]	0.15 [0.09, 0.23]
A1.9.HET	1998 [1992, 2004]	1.09 [0.69, 1.71]	0.31 [0.15, 0.55]
A1.10.MIX	1993 [1984, 2000]	1.34 [0.99, 1.70]	0.07 [0.02, 0.12]
A1.11.HET	1998 [1993, 2001]	1.31 [0.91, 1.71]	0.07 [0.04, 0.12]
A1.12.HET	1991 [1983, 1999]	1.08 [0.73, 1.50]	0.19 [0.10, 0.33]
A1.13.HET	1987 [1977, 1995]	1.05 [0.72, 1.40]	0.22 [0.12, 0.36]
A1.14.HET	1991 [1981, 2001]	1.03 [0.69, 1.39]	0.21 [0.09, 0.37]
A1.15.MSM	1991 [1974, 2004]	0.98 [0.65, 1.29]	0.19 [0.09, 0.31]
A1.16.HET	1991 [1983, 1998]	1.06 [0.73, 1.47]	0.19 [0.09, 0.33]
A1.17.HET	1992 [1982, 2000]	1.07 [0.71, 1.54]	0.29 [0.15, 0.49]
A1.19.HET	1983 [1971, 1991]	1.01 [0.67, 1.35]	0.25 [0.17, 0.47]
C.1.HET	1977 [1968, 1985]	1.48 [1.09, 1.95]	0.07 [0.01, 0.14]

Abbreviations: HET, Heterosexual transmission; Mixed; MSM, men who have sex with men; FSW, female sex work; MTMC, perinatal transmission; PWID, people who inject drugs. Results are not shown for two clusters (A1.5.HET and A1.18.HET) whose parameters did not converge. *HPD: Higher posterior density interval. *TMRCA: time to the most recent common ancestor. Data are median and 95% higher posterior density intervals.

Table 5. Number of expected (Markov) jumps (BF≥3) inferred for HIV-1 migration between geographic locations and between risk groups based on the cluster-wise sub-sampling approach.

The direction of migration events (from-to)	Number of HIV-1 jumps (N, %)
Geographic	**174 (100.0%)**
Within-country	**109 (62.6%)**
Nyanza-Nairobi	18 (10.3%)
Rift Valley-Nairobi	17 (9.8%)
Nyanza-Rift Valley	16 (9.2%)
Rift Valley-Nyanza	13 (7.5%)
Nyanza-Coast	11 (6.3%)
Rift Valley-Coast	11 (6.3%)
Nairobi-Coast	10 (5.7%)
Nairobi-Nyanza	8 (4.6%)
Nairobi-Rift Valley	5 (2.9%)
Export from Kenya	**42 (24.1%)**
Nyanza-Ref	20 (11.5%)
Rift Valley-Ref	13 (7.5%)
Nairobi-Ref	6 (3.4%)
Coast-Ref	3 (1.7%)
Import into Kenya	**23 (13.2%)**
Ref-Coast	9 (5.2%)
Ref-Nyanza	5 (2.9%)
Ref-Rift Valley	5 (2.9%)
Ref-Nairobi	4 (2.3%)
Risk group	**47 (100.0%)**
HET-FSW	16 (34.0%)
HET-MSM	15 (31.9%)
HET-PWID	8 (17.0%)
PWID-HET	3 (6.4%)
MSM-HET	3 (6.4%)
MSM-FSW	2 (4.3%)

Abbreviations: Ref, reference HIV-1 *pol* sequences from the global epidemic that clustered closely with Kenyan sequences; HET, heterosexual transmission; Mixed; MSM, men who have sex with men; FSW, female sex work; MTMC, perinatal transmission; PWID, people who inject drugs.

Table 6. The number of HIV-1 jumps (2010-2019) based on proportional and uniform sub-sampling.

Jumps direction (from-to)	Number of Jumps (N)	
Jumps between locations	**Proportional sub-sampling**	**Uniform sub-sampling**
West to East	319 (88%)	213 (78%)
Nyanza-Rift Valley	129 (36%)	50 (18%)
Nyanza-Nairobi	113 (31%)	73 (27%)
Nyanza-Coast	50 (14%)	54 (20%)
Nairobi-Coast	8 (2%)	19 (7%)
Rift Valley-Nairobi	14 (4%)	8 (3%)
Rift Valley-Coast	5 (1%)	9 (3%)
East to west	43 (12%)	61 (22%)
Rift Valley-Nyanza	11 (3%)	6 (2%)
Nairobi-Rift Valley	9 (2%)	21 (8%)
Nairobi-Nyanza	9 (2%)	25 (9%)
Coast-Nyanza	7 (2%)	3 (1%)
Coast-Nairobi	4 (1%)	3 (1%)
Coast-Rift Valley	3 (1%)	3 (1%)
Jumps between risk groups		
HET to key populations	126 (94%)	126 (72%)
HET-FSW	64 (48%)	75 (43%)
HET-MSM	58 (43%)	46 (26%)
HET-PWID	4 (3%)	5 (3%)
Key populations to HET	3 (2%)	20 (11%)
FSW-HET	1 (1%)	15 (9%)
PWID-HET	1 (1%)	3 (2%)
MSM-HET	1 (1%)	2 (1%)
Key populations to others	5 (4%)	29 (17%)
FSW-MSM	2 (1%)	14 (8%)
FSW-PWID	1 (1%)	4 (2%)
MSM-FSW	2 (1%)	9 (5%)
MSM-PWID	0 (0%)	1 (1%)
PWID-FSW	0 (0%)	1 (1%)
PWID-MSM	0 (0%)	0 (0%)

Abbreviations: HET, heterosexual; MSM, men who have sex with men; FSW, female sex workers; PWID, people who inject drugs.

FIGURES

Figure 1. Map of Kenya highlighting geographic locations and sampling density.
Map of Kenya highlighting geographic locations (former administrative provinces), HIV-1 burden per province (proportion of people with HIV-1 as per province in Kenya [11,45-47]), and the sampling density (number of people with HIV-1 included in the study based on the estimated number of people with HIV-1 in Kenya).

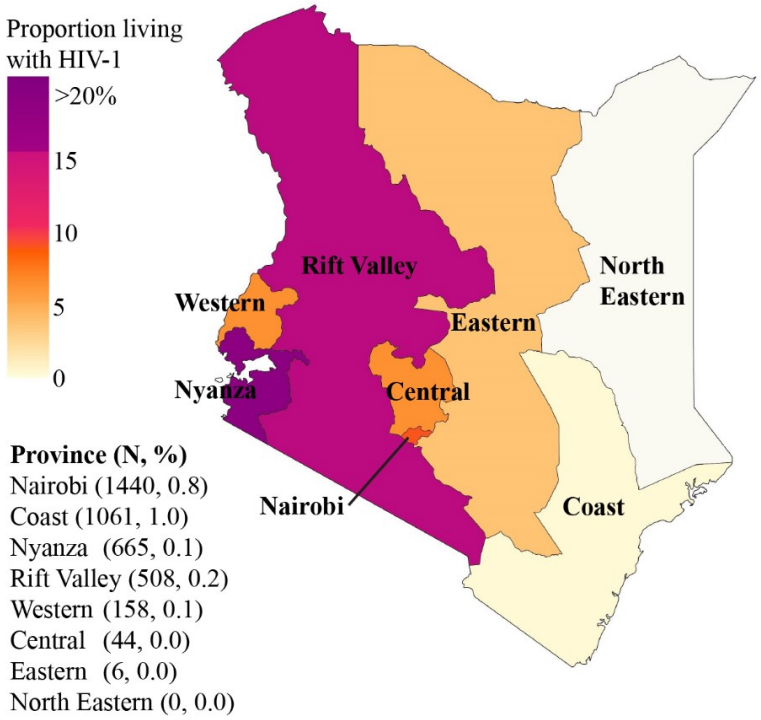


Figure 2. Population dynamics of HIV-1 sub-subtype A1, subtype D and subtype C lineages in Kenya.
Bayesian Skygrid plots showing effective population size of the (a) HIV-1 sub-subtype A1, (b) HIV-1 subtype C and (c) HIV-1 subtype D lineages in the Kenyan dataset. Median estimates of the effective population size overtime are shown as a continuous line in each plot (coloured Red for sub-subtype A1, Brown for subtype C, and Blue for subtype D). The shaded area represents the 95% higher posterior density intervals of the inferred effective population size for each lineage.

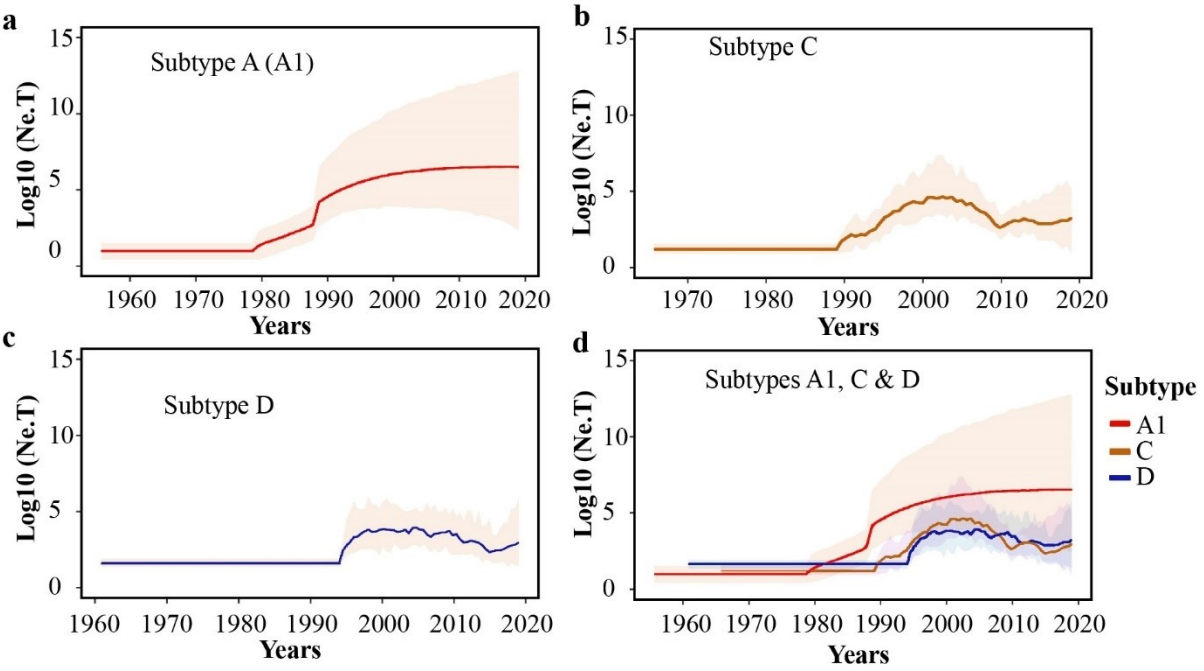


Figure 3. HIV-1 risk group-specific estimates in the effective population size through time in Kenya.
Bayesian Skygrid plots showing historical population dynamics of (**a**) the main HIV-1 sub-subtype A1 HET clusters, (**b**) the only large subtype C HET cluster, (**c**) the only large HIV-1 sub-subtype A1 PWID cluster and (**d**) the only large HIV-1 sub-subtype A1 MSM cluster in Kenya. Median estimates of the number of individuals contributing to new infections over time are shown as a continuous line coloured as per the dominant risk group per cluster (bluish-green: MSM; sky blue: PWID; and yellow: HET). The area shaded grey represents the 95% higher posterior density intervals of the inferred effective population size. Information on geographic representation per cluster is provided in the figure legends.

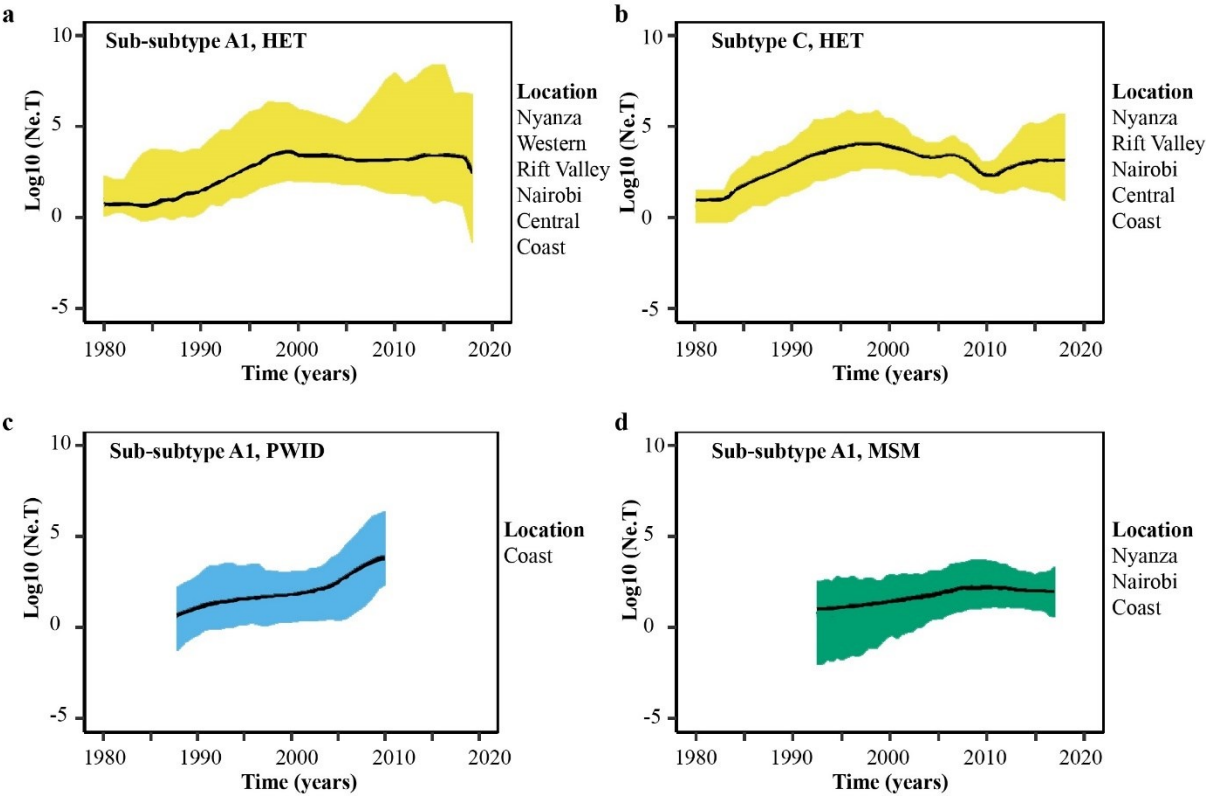


Figure 4. Date of origin, evolutionary rate, and growth rate among sub-subtype A1 and subtype C clusters of different risk groups. Time to the most recent common ancestor (**a**), evolutionary rate (**b**), and growth rate (**c**) estimates among seventeen sub-subtype A1 and one subtype C clusters. Median estimates and 95% higher posterior density interval are shown for the different categories per cluster, coloured by the dominant risk group per cluster. Results are not shown for two clusters (A1.5.HET and A1.18.HET) whose parameters did not converge.

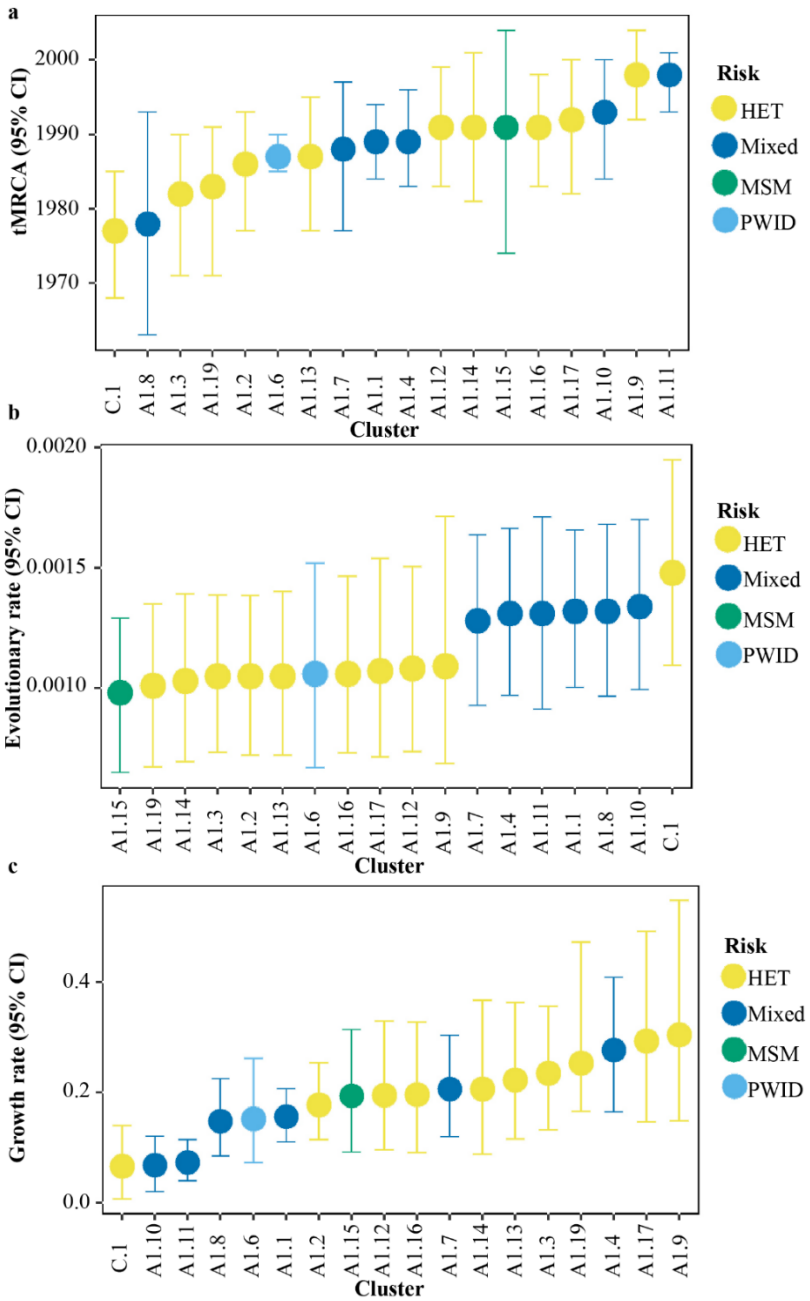
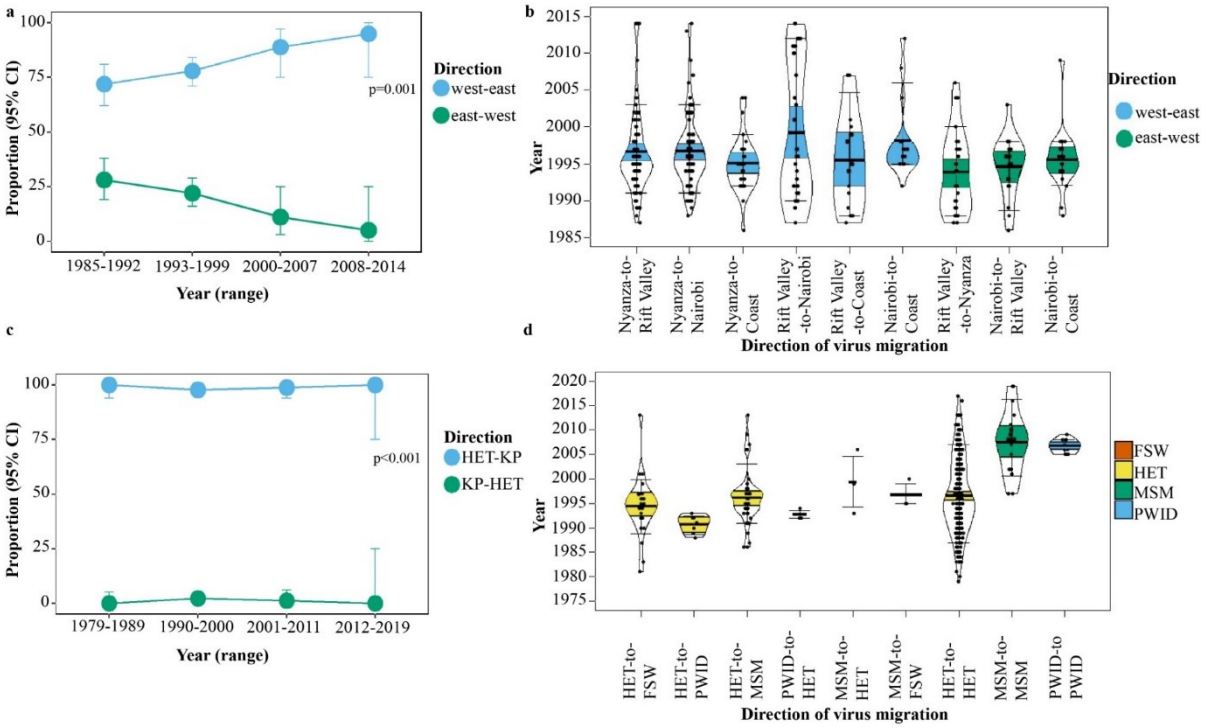


Figure 5. Proportion and dates of HIV-1 transitions between geographic provinces and risk groups.
Dates of HIV-1 transitions between geographic provinces and risk groups summarised from trait-annotated maximum clade credibility trees. Plots represent (**a**) proportion of West-to-East vs East-to-West geographic migration over time, (**b**) dates of HIV-1 dissemination between different geographic locations (where group median and interquartile range are coloured by the direction of transmission – coloured sky blue: West-to-East, and vermillion: East-to-West), (**c**) proportion of HIV-1 transmission from heterosexuals to key populations and vice-versa over time, and (**d**) dates of HIV-1 transmission within and between different risk groups (where group median and interquartile range are coloured by "source" risk group – coloured green: MSM; sky blue: PWID; vermillion: FSW; yellow: HET). Only transitions with a posterior probability higher than 0.90 are plotted. Dots in the pirate plots represent HIV-1 migration events.



SUPPLEMENTARY DATA

Files in this Data Supplement:

Supplementary tables

Table S1. The number and sources of plasma samples used to generate new HIV-1 pol sequences

Table S2. The number of Kenyan HIV-1 partial *pol* sequences (N=4058, 1986-2019) sequences analysed in this study compared to national estimates of the number of people living with HIV-1 in Kenya belonging to different risk groups and geographic regions.

Table S3. The number and temporal distribution of Kenyan HIV-1 partial *pol* sequences (N=4058, 1986-2019).

Table S4. Distribution of HIV-1 subtypes by geographic area.

Table S5. Distribution of HIV-1 subtypes by risk groups.

Table S6. Proportions of Kenyan sequences in clusters relative to Kenyan sequences that did not cluster and their distribution into risk groups and geographic provinces.

Table S7. Characteristics of large Kenyan clusters (N=20) used in the inference of past population dynamics.

Supplementary figures.

Figure S1. A summary scheme of sampling criteria in this study.

Figure S2. Distribution of HIV-1 sequences and subtypes (1986-2019).

Figure S3. Maximum-likelihood trees used to identify transmission clusters.

Figure S4. Size and subtype distribution of 409 Kenyan HIV-1 clusters identified in this study.

Figure S5. Graphical summary of the distribution of 409 Kenyan clusters by geographic locations and risk groups.

Figure S6. Root-to-tip regression analyses of phylogenetic temporal signal.

Figure S7. Population dynamics in the HET and mixed-risk group HIV-1 clusters.

Figure S8. Pirate plots quantifying direction and number of HIV-1 jumps between geographic locations and risk groups.

Tables

Table S1. A summary scheme of sampling criteria in this study.

Site	Risk group			
	HET	MSM	FSW	PWID
KWTRP	48	21	107	0
NHRS	0	57	14	0
KAVI-ICR	30	0	7	0
SWOP	0	50	0	0
TRANSFORM	0	85	0	0
KEMRI-CGHR	336	0	0	0
Total	414	213	128	0

Abbreviations: HET, heterosexual adults; MSM, men who have sex with men; FSW, female sex workers; PWID, people who inject drugs. Site abbreviations: KWTRP, Kenya Medical Research Institute (KEMRI) -Wellcome Trust (Coastal Kenya); NHRS, Nyanza Reproductive Health Society (in Western Kenya); KAVI-ICR, Kenya AIDS Vaccine Initiative's Institute of Clinical Research (in Nairobi, Central Kenya); SWOP, Sex Workers Outreach Program clinics in Nairobi, TRANSFORM, a cohort of transfeminine people and cisgender men who have sex with men in Nairobi; KEMRI-CGHR, Kenya Medical Research Institute (KEMRI) – Centre for Global Health Research (Western Kenya).

Table S2. The number of Kenyan HIV-1 partial *pol* sequences (N=4058, 1986-2019) analysed in this study compared to national estimates of the number of people with HIV-1 in Kenya belonging to different risk groups and geographic regions.

Characteristic		[a]Total population estimates	[b]PWHIV (N)	[c]Sample size (N)	[d]Sampling density (%)
Overall	Kenya	47,564,296	1,493,413 (100%)	4,058	0.3
Sampling location	Nyanza	7,163,260	526,972 (35%)	665	0.1
	Rift Valley	12,752,966	247,127 (17%)	508	0.2
	Nairobi	4,397,073	190,993 (13%)	1,440	0.8
	Western	4,128,162	141,561 (9%)	158	0.1
	Central	5,482,239	141,306 (9%)	44	0.0
	Eastern	6,821,049	132,232 (9%)	6	0.0
	Coast	4,329,474	108,994 (7%)	1,061	1.0
	North Eastern	2,490,073	4,196 (<1%)	0	0.0
Risk group	HET	26,642,987	1,341,164 (90%)	3,401	0.3
	FSW	133,675	40,103 (3%)	227	0.6
	MSM	19,175	3,452 (>1%)	372	10.8
	PWID	18,327	3,482 (>1%)	58	1.7
	Children (>15 years)	2,0750,132	105,213 (7%)	0	0.0

Abbreviations: HET, heterosexual adults; MSM, men who have sex with men; FSW, female sex workers; PWID, people who inject drugs; PWHIV, people with HIV. [a]Kenyan population estimates as of 2019 (Kenya National Bureau of Statistics, 2019). [b]The estimated number of PWHIV as per geographic area [computed from national population data (Kenya National Bureau of Statistics, 2019) and HIV-1 prevalence data per geographic region (National AIDS and STI Control Programme (NASCOP), 2020)], and risk groups [computed from key populations estimates (National AIDS and STI Control Programme (NASCOP), 2019) and HIV-1 prevalence per risk group (Kenya National AIDS control council (NACC), 2018)] in Kenya. [c]The number of people living with HIV-1 included in the study, and [d]the estimated proportion of people living with HIV-1 in Kenya included in the study.

Table S3. The number and temporal distribution of Kenyan HIV-1 partial *pol* sequences (N=4058, 1986-2019).

	Risk group				Province							
Sampling year	**HET**	**MSM**	**FSW**	**PWID**	**Nairobi**	**Coast**	**Nyanza**	**Rift Valley**	**Western**	**Central**	**Eastern**	**Total**
1986	0	0	2	0	2	0	0	0	0	0	0	**2**
1991	4	0	0	0	4	0	0	0	0	0	0	**4**
1993	1	0	0	0	1	0	0	0	0	0	0	**1**
1994	0	0	1	0	1	0	0	0	0	0	0	**1**
1996	4	0	12	0	0	12	4	0	0	0	0	**16**
1997	13	0	6	0	6	0	13	0	0	0	0	**19**
1998	7	0	0	0	0	0	7	0	0	0	0	**7**
1999	18	0	0	0	4	0	13	1	0	0	0	**18**
2000	31	0	0	0	7	8	8	8	0	0	0	**31**
2001	8	0	5	0	11	1	0	1	0	0	0	**13**
2002	4	0	7	0	11	0	0	0	0	0	0	**11**
2003	3	0	0	0	0	2	0	1	0	0	0	**3**
2004	199	0	0	0	5	0	171	23	0	0	0	**199**
2005	198	0	10	0	44	18	125	21	0	0	0	**208**
2006	380	28	20	0	154	57	16	201	0	0	0	**428**
2007	554	12	35	0	282	293	26	0	0	0	0	**601**
2008	236	19	75	0	97	230	3	0	0	0	0	**330**
2009	228	33	16	0	61	173	0	43	0	0	0	**277**
2010	267	26	2	58	127	118	107	1	0	0	0	**353**
2011	243	13	2	0	96	26	6	129	1	0	0	**258**
2012	284	6	4	0	144	6	5	0	139	0	0	**294**
2013	201	5	0	0	137	31	2	36	0	0	0	**206**
2014	94	6	2	0	36	11	55	0	0	0	0	**102**
2015	58	22	2	0	25	5	19	0	0	33	0	**82**
2016	44	68	22	0	37	40	50	2	3	1	1	**134**
2017	56	118	4	0	126	14	17	8	8	5	0	**178**
2018	90	2	0	0	22	2	18	33	7	5	5	**92**
2019	0	14	0	0	0	14	0	0	0	0	0	**14**
***Missing**	176	0	0	0	0	0	0	0	0	0	0	**176**
Total	3,401	372	227	58	1,440	1,061	665	508	158	44	6	**4,058**

Abbreviations: HET, heterosexual; MSM, men who have sex with men; FSW, female sex work; PWID, people who inject drugs.
*Missing: some (N=176, 4% of all sequences, all HET) of the newly generated sequences lacked data on the geographic area of sampling.

Table S4. Distribution of HIV-1 subtypes by geographic provinces in Kenya.

Subtype (N, %)	Central	Coast	Eastern	Nairobi	Nyanza	Rift Valley	Western	Unknown	Total
A1	35 (1.2%)	765 (26.8%)	5 (0.2%)	1044 (36.5%)	424 (14.8%)	304 (10.6%)	118 (4.1%)	165 (5.8%)	2860 (70.5%)
B	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
C	2 (0.7%)	75 (26.3%)	0 (0.0%)	84 (29.5%)	53 (18.6%)	55 (19.3%)	10 (3.5%)	6 (2.1%)	285 (7.0%)
CRF01_AE	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
CRF02_AG	0 (0.0%)	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
CRF10_CD	0 (0.0%)	8 (32%)	0 (0.0%)	7 (28%)	4 (16%)	2 (8%)	4 (16%)	0 (0.0%)	25 (0.6%)
CRF16_A2D	3 (7%)	9 (20.9%)	0 (0.0%)	19 (44.2%)	10 (23.3%)	2 (4.7%)	0 (0.0%)	0 (0.0%)	43 (1.1%)
CRF18_cpx	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
CRF21_A2D	0 (0.0%)	8 (38.1%)	0 (0.0%)	6 (28.6%)	4 (19.1%)	2 (9.5%)	1 (4.8%)	0 (0.0%)	21 (0.5%)
CRF43_02G	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
D	3 (0.7%)	86 (18.7%)	1 (0.2%)	169 (36.7%)	84 (18.2%)	100 (21.7%)	13 (2.8%)	5 (1.1%)	461 (11.4%)
G	0 (0.0%)	4 (20%)	0 (0.0%)	8 (40%)	3 (15%)	5 (25%)	0 (0.0%)	0 (0.0%)	20 (0.5%)
URF	1 (0.3%)	105 (31.1%)	0 (0.0%)	100 (29.6%)	82 (24.3%)	38 (11.2%)	12 (3.6%)	0 (0.0%)	338 (8.3%)
Total	44 (1.1%)	1061 (26.2%)	6 (0.2%)	1440 (35.5%)	665 (16.4%)	508 (12.5%)	158 (3.8%)	176 (4.3%)	4058 (100.0%)

Abbreviations: CRF, circulating recombinant form; URF, unique recombinant form; HET, heterosexual; MSM, men who have sex with men; FSW, female sex work; PWID, people who inject drugs. *Missing: some of the newly generated sequences (N=176, 4% of all sequences, all HET) had missing information on the geographic area of sampling.

Table S5. Distribution of HIV-1 subtypes by risk groups in Kenya.

Subtype	Risk group				
	HET	**MSM**	**FSW**	**PWID**	**Total**
A1	2388 (70.2%)	276 (74.2%)	140 (61.7%)	56 (96.6%)	2860 (70.5%)
B	1 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
C	232 (6.8%)	31 (8.3%)	20 (8.8%)	2 (3.4%)	285 (7.0%)
CRF01_AE	1 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
CRF02_AG	1 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
CRF10_CD	23 (0.7%)	1 (0.3%)	1 (0.4%)	0 (0.0%)	25 (0.6%)
CRF16_A2D	36 (1.1%)	4 (1.1%)	3 (1.3%)	0 (0.0%)	43 (1.1%)
CRF18_cpx	0 (0.0%)	0 (0.0%)	1 (0.4%)	0 (0.0%)	1 (0.0%)
CRF21_A2D	16 (0.5%)	1 (0.3%)	4 (1.8%)	0 (0.0%)	21 (0.5%)
CRF43_02G	1 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
D	391 (11.5%)	49 (13.2%)	21 (9.3%)	0 (0.0%)	461 (11.4%)
G	18 (0.5%)	0 (0.0%)	2 (0.9%)	0 (0.0%)	20 (0.5%)
URF	293 (8.6%)	10 (2.7%)	35 (15.4%)	0 (0.0%)	338 (8.3%)
Total	3401 (100.0%)	372 (100.0%)	227 (100.0%)	58 (100.0%)	4058 (100.0%)

Abbreviations: CRF, circulating recombinant form; URF, unique recombinant form; HET, heterosexual; MSM, men who have sex with men; FSW, female sex work; PWID, people who inject drugs. *Missing: some of the newly generated sequences (N=176, 4% of all sequences, all HET) had missing information on the geographic area of sampling.

Table S6. Proportions of Kenyan sequences in clusters relative to Kenyan sequences that did not cluster and their distribution into subtype, risk groups and geographic provinces.

	Clustered (N, %)	Did not cluster (N, %)	Total (N, %)
Subtype			
A1	1485 (51.9%)	1375 (48.1%)	2860 (100.0%)
C	137 (48.1%)	148 (51.9%)	285 (100.0%)
D	210 (45.6%)	251 (54.5%)	461 (100.0%)
Risk group			
HET	1441 (42.4%)	1960 (57.6%)	3401 (100.0%)
MSM	273 (73.4%)	99 (26.6%)	372 (100.0%)
FSW	73 (32.2%)	154 (67.8%)	227 (100.0%)
PWID	45 (77.6%)	13 (22.4%)	58 (100.0%)
Sampling location			
Nairobi	720 (50%)	720 (50%)	1440 (100.0%)
Coast	481 (45.3%)	580 (54.7%)	1061 (100.0%)
Nyanza	311 (46.8%)	354 (53.2%)	665 (100.0%)
Rift Valley	175 (34.5%)	333 (65.6%)	508 (100.0%)
Western	60 (38%)	98 (62%)	158 (100.0%)
Central	17 (38.6%)	27 (61.4%)	44 (100.0%)
Eastern	1 (16.7%)	5 (83.3%)	6 (100.0%)
*Missing	67 (38.1%)	109 (61.9%)	176 (100.0%)
Total	1832 (45.2%)	2226 (54.9%)	4058 (100.0%)

Abbreviations: HET, heterosexual; MSM, men who have sex with men; FSW, female sex work; PWID, people who inject drugs.
*Missing information on the geographic area of sampling

Table S7. Characteristics of large Kenyan clusters (N=20) used in the inference of past population dynamics.

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Subtype	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	C
Sampling location (%)																				
Central	1	0	0	0	0	0	14	0	4	4	0	0	0	0	0	0	6	0	0	4
Coast	31	29	26	47	3	100	24	30	4	29	4	14	29	33	29	15	63	60	13	28
Nairobi	31	44	15	32	60	0	38	48	38	39	96	36	29	28	52	30	19	20	53	30
Nyanza	25	15	41	18	0	0	14	0	21	18	0	23	17	11	14	30	13	7	0	19
Ref	0	0	0	0	0	0	0	0	0	0	0	0	17	6	5	5	0	13	20	5
Rift Valley	8	7	18	0	28	0	3	15	21	11	0	18	4	22	0	20	0	0	7	14
Western	3	4	0	3	10	0	7	6	13	0	0	9	4	0	0	0	0	0	7	0
Risk groups (%)																				
FSW	7	1	4	9	0	0	3	6	3	0	0	4	4	0	0	5	6	7	7	5
HET	81	95	96	83	100	0	92	84	93	42	97	92	96	94	0	95	94	80	93	87
MSM	12	4	0	9	0	0	6	11	3	58	0	4	0	6	100	0	0	7	0	3
PWID	0	0	0	0	0	100	0	0	0	0	4	0	0	0	0	0	0	6	0	5

Abbreviations: Ref, reference HIV-1 *pol* sequences from the global epidemic that clustered closely with Kenyan sequences; HET, heterosexual; MSM, men who have sex with men; FSW, female sex work; PWID, people who inject drugs.

Supplementary figures

Figure S1. Study scheme
A summary scheme of sampling criteria in this study.

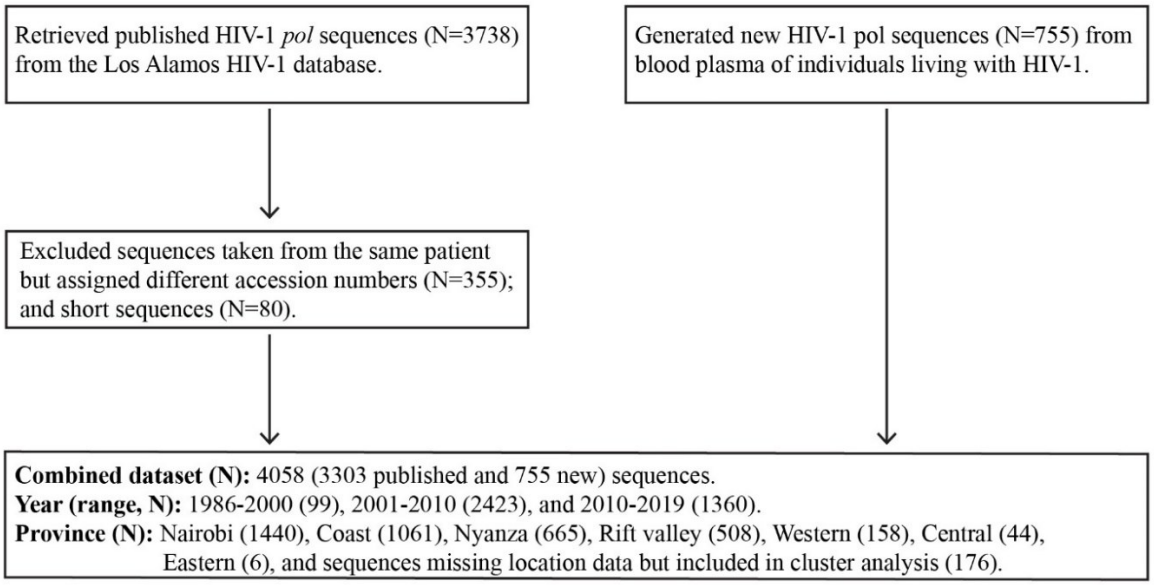
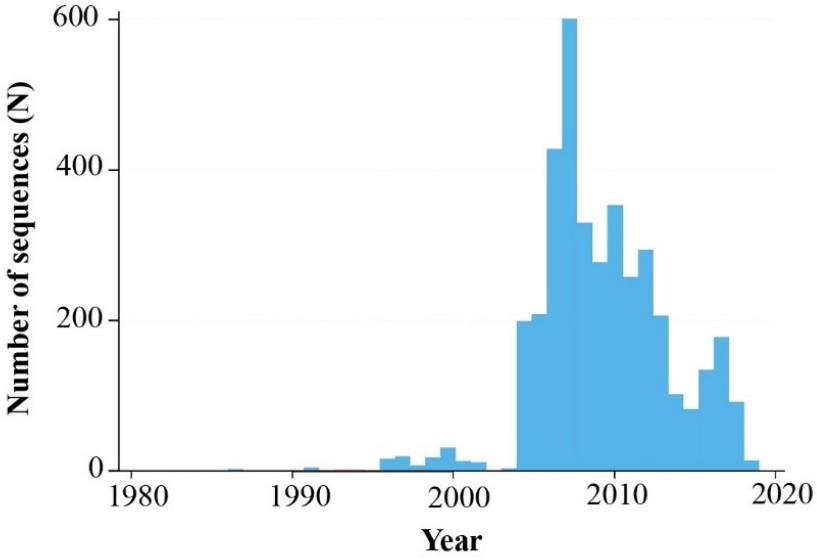
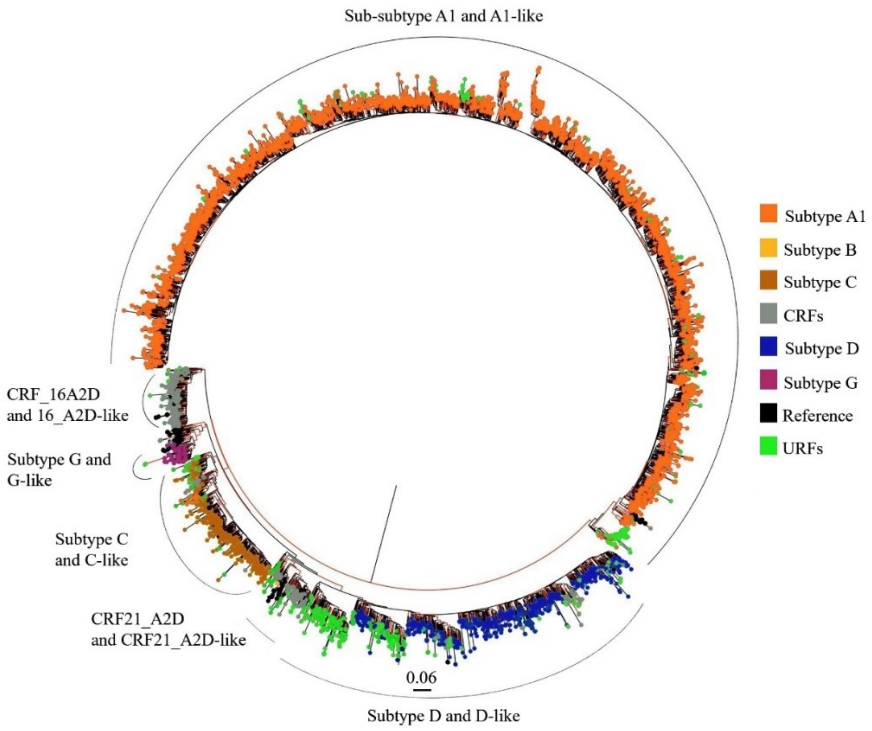


Figure S2. Distribution of HIV-1 sequences and subtypes in this study (1986-2019).
(**a**) Temporal distribution of the number of Kenyan HIV-1 sequences in this study. (**b**) ML phylogenetic reconstruction of HIV-1 group M genetic diversity based on genetic sequences (N=4058) from Kenya. Branch tips on the phylogenetic tree and proportion of HIV-1 lineages in different geographical locations are coloured according to subtypes (orange: sub-subtype A1; yellow: subtype B; brown: subtype C; blue: subtype D; maroon: subtype G; grey: circulating recombinant forms (CRFs); green: unique recombinant forms (URFs); black: HIV-1 group M reference sequences. (**c**) Temporal changes (2004-2019) in the overall proportion of HIV-1 subtypes and recombinants over two-years intervals in Kenya. A $p<0.05$ denotes a statistically significant increase or decrease in the proportion of respective circulating strains over time.

a


600
400
200
0
Number of sequences (N)
1980
1990
2000
2010
2020
Year

b


Sub-subtype A1 and A1-like
Subtype A1
Subtype B
Subtype C
CRFs
Subtype D
Subtype G
Reference
URFs
CRF_16A2D
and 16_A2D-like
Subtype G and
G-like
Subtype C
and C-like
CRF21_A2D
and CRF21_A2D-like
0.06
Subtype D and D-like

c

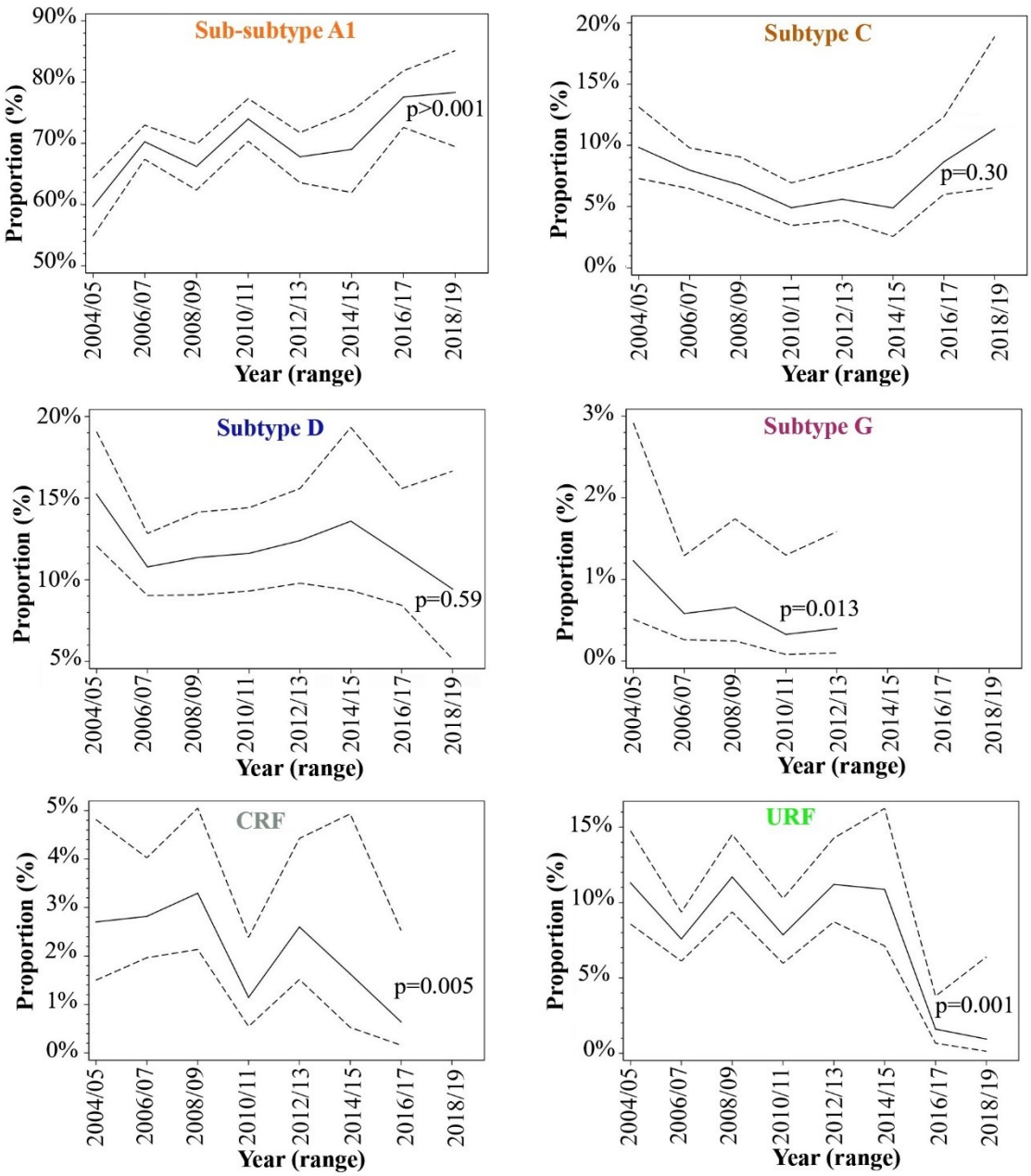
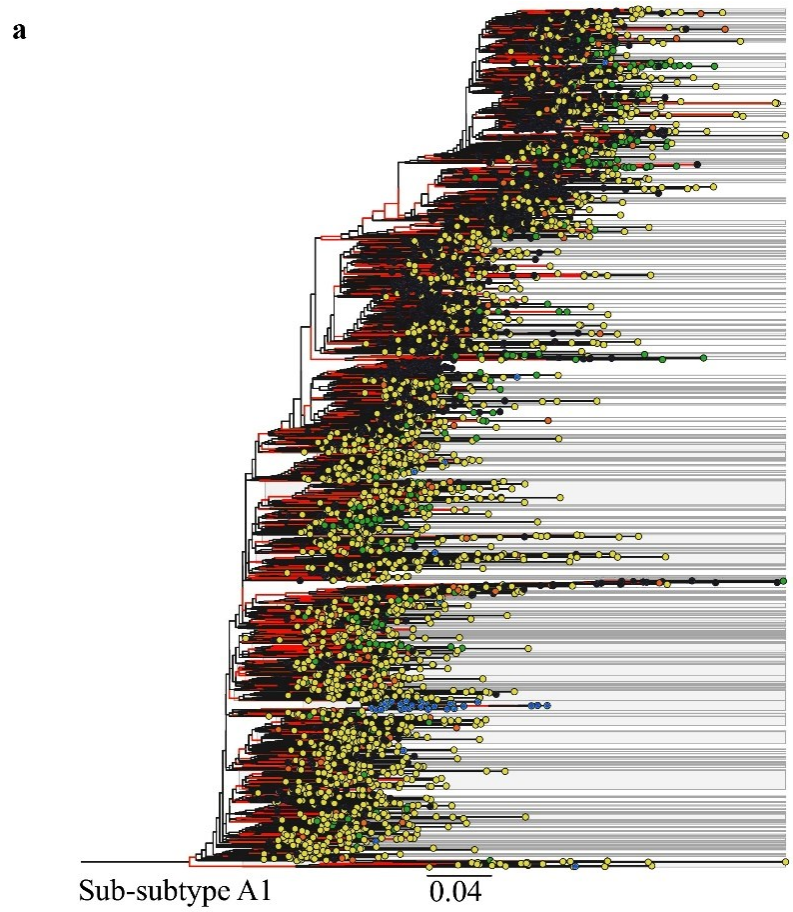
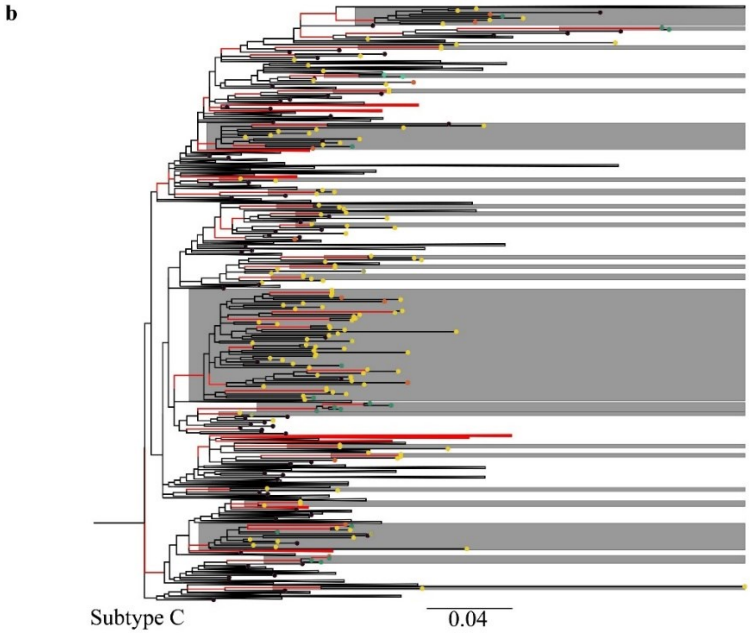

Sub-subtype A1
p>0.001
Subtype C
p=0.30
Subtype D
p=0.59
Subtype G
p=0.013
CRF
p=0.005
URF
p=0.001
Proportion (%)
Year (range)
2004/05
2006/07
2008/09
2010/11
2012/13
2014/15
2016/17
2018/19

Figure S3. Maximum-likelihood trees used to identify transmission clusters.
Maximum-likelihood trees used for identification of Kenya HIV-1 clusters. Trees represent (**a**) sub-subtype A1, (**b**) subtype C, and (**c**) subtype D transmission clusters, respectively. Each phylogeny is rooted at the midpoint. Monophyletic clusters with SH-aLRT support ≥0.9 and which have ≥80% sequences from Kenya are highlighted in grey. To enhance cluster visualization, some branches containing either reference sequences or Kenyan sequences that did not clusters have been collapsed (shown as black triangles). Branch tips within respective clusters are coloured as per cluster risk group (green: MSM; sky blue: PWID; vermillion: FSW; yellow: HET; and black: Reference sequences). Red bars in the respective trees represent statistically supported branches (i.e. branches with SH-aLRT support ≥0.9). Scale bars represent the genetic distance in substitutions per site in all phylogenies.





c

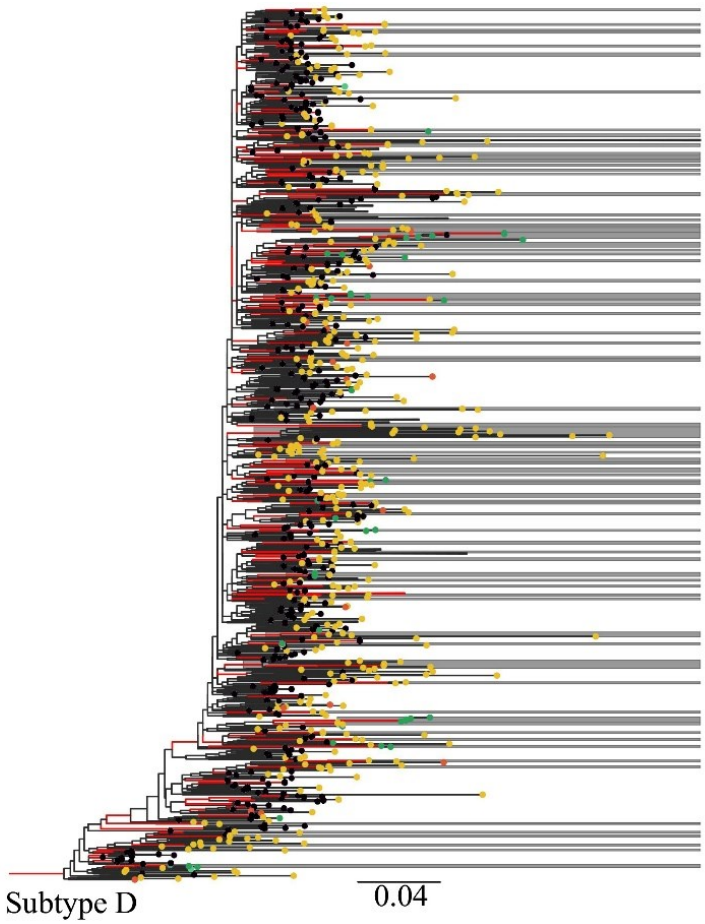


Figure S4. Distribution of clusters (N=409) of different subtypes by cluster size.
Size and subtype distribution of 409 Kenyan HIV-1 clusters identified in this study. The number of clusters per subtype are shown in the Y-axis (coloured by subtype: Red; sub-subtype A1, Brown; subtype C, and Deep Blue; subtype D clusters). The number of sequences per cluster is depicted in the X-axis.

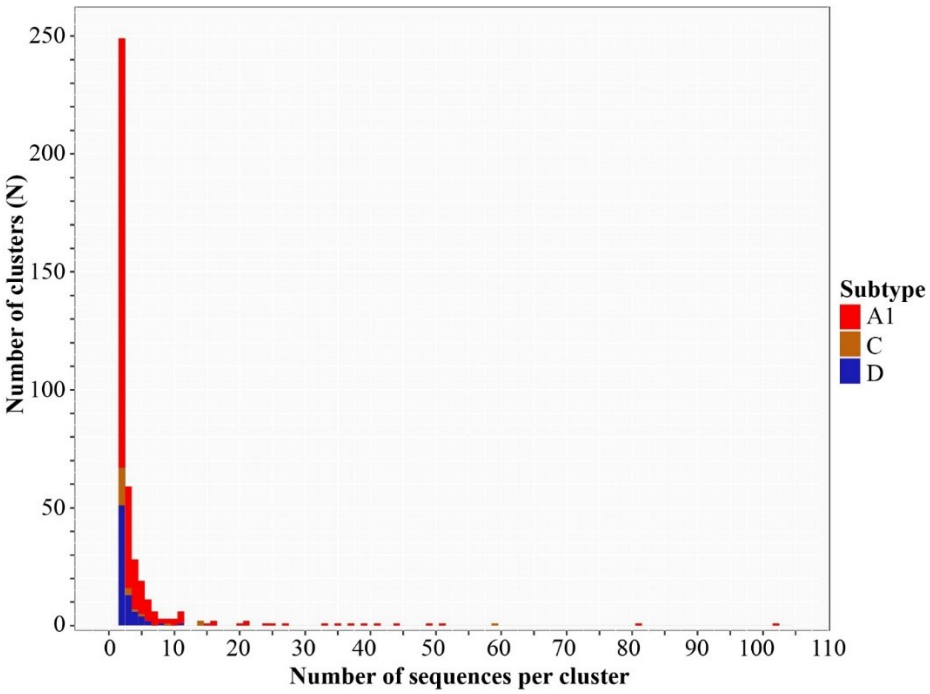
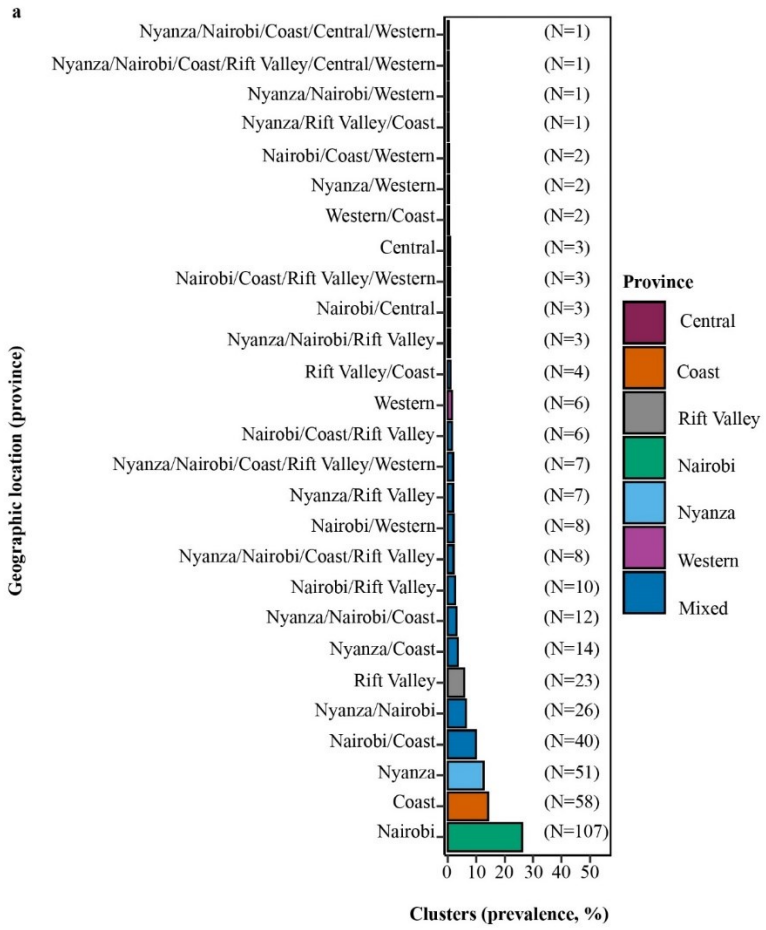


Figure S5. Summary of Kenyan clusters (N=409) by geographic and risk group.
Graphical summary of the distribution of 409 Kenyan clusters by (**a**) geographic locations (i.e. province) and (**b**) risk groups.



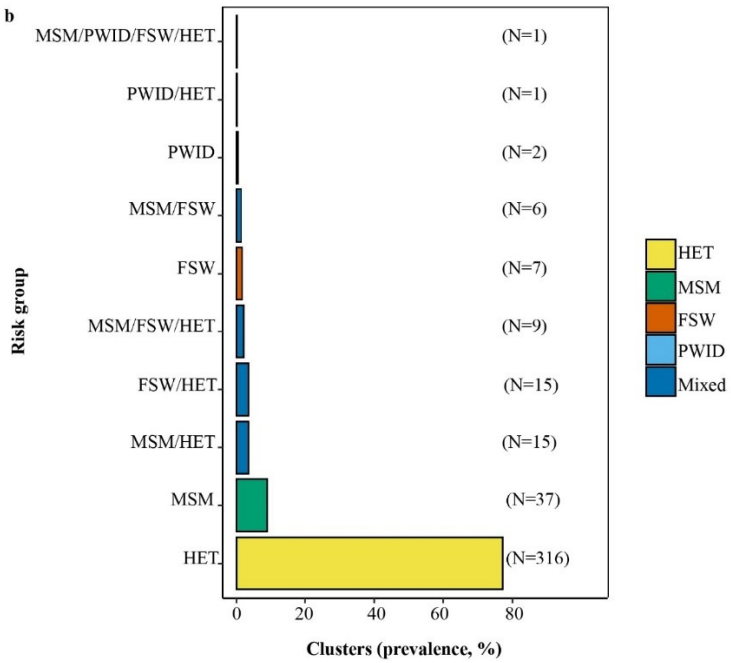


Figure S6. Root-to-tip regression analyses of phylogenetic temporal signal.
Root-to-tip regression analyses of phylogenetic temporal signal for sub-subtype A1, subtype C, and subtype D sequences from Kenya. Correlation and determination coefficient (R2) were estimated with TempEst.

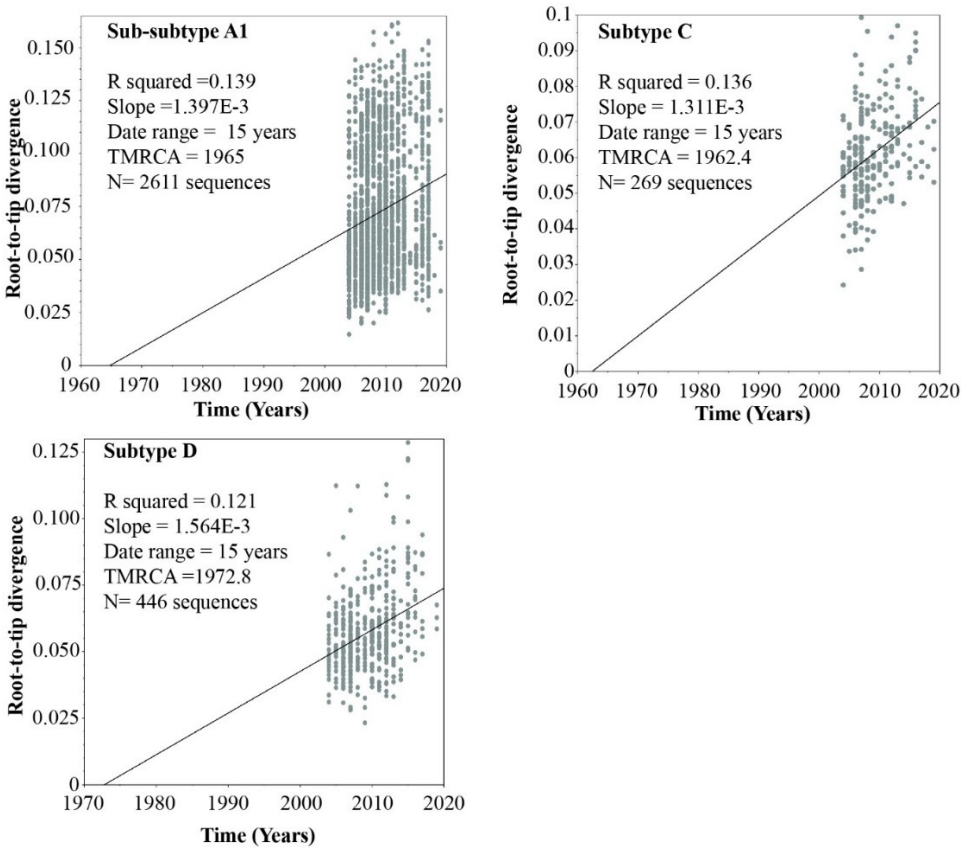


Figure S7. Population dynamics in the HIV-1 epidemic among HET and mixed-risk group clusters.
Bayesian Skygrid plots showing historical population dynamics of the main (**a**) HET and (**b**) mixed-risk group HIV-1 sub-subtype A1 clusters. Median estimates of the number of individuals contributing to new infections over time are shown as a continuous black line. The shaded area represents the 95% higher posterior density intervals of the inferred effective population size. Figure legends highlight information on dominating risk group per cluster, and the provinces of sampling.

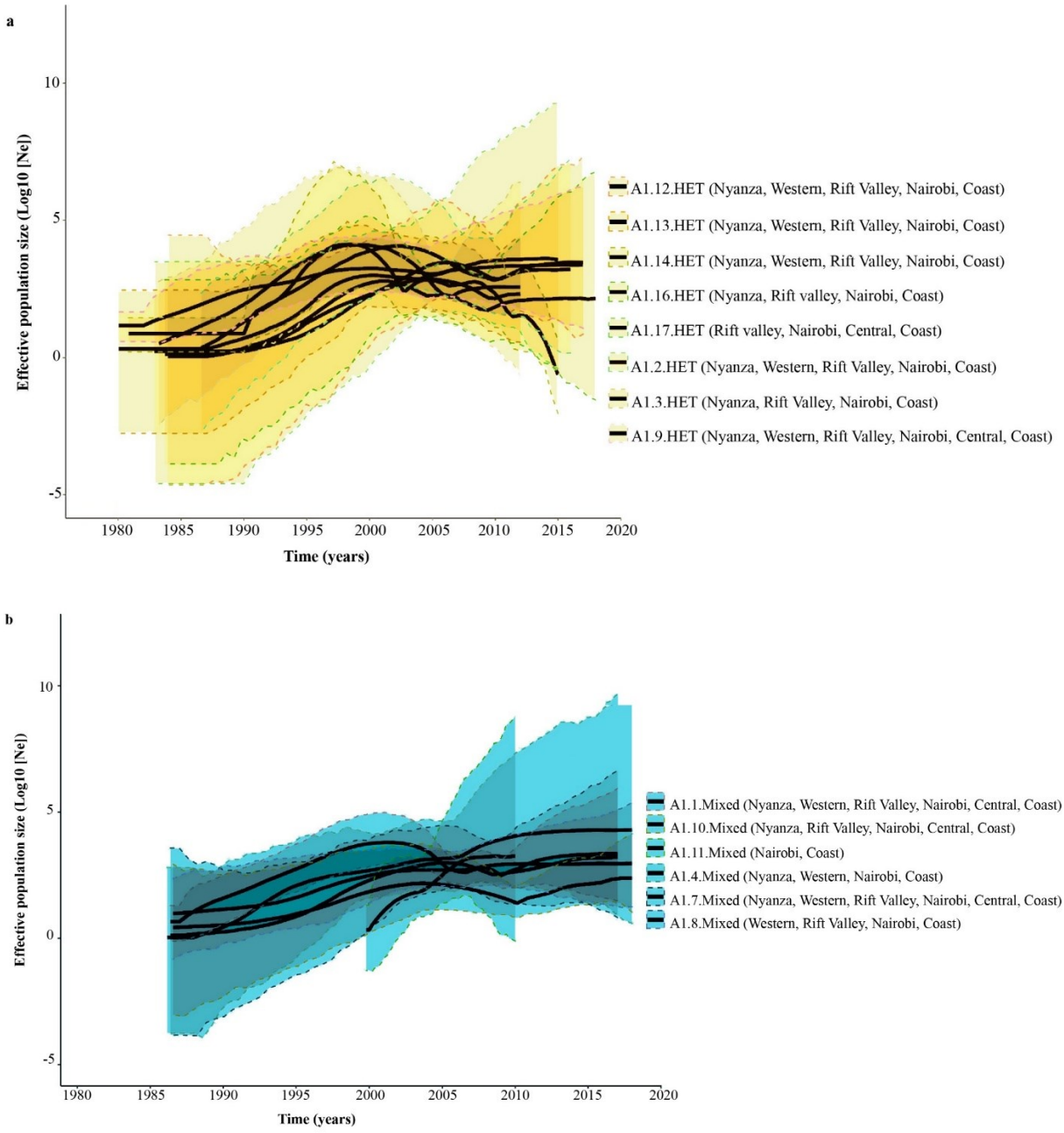
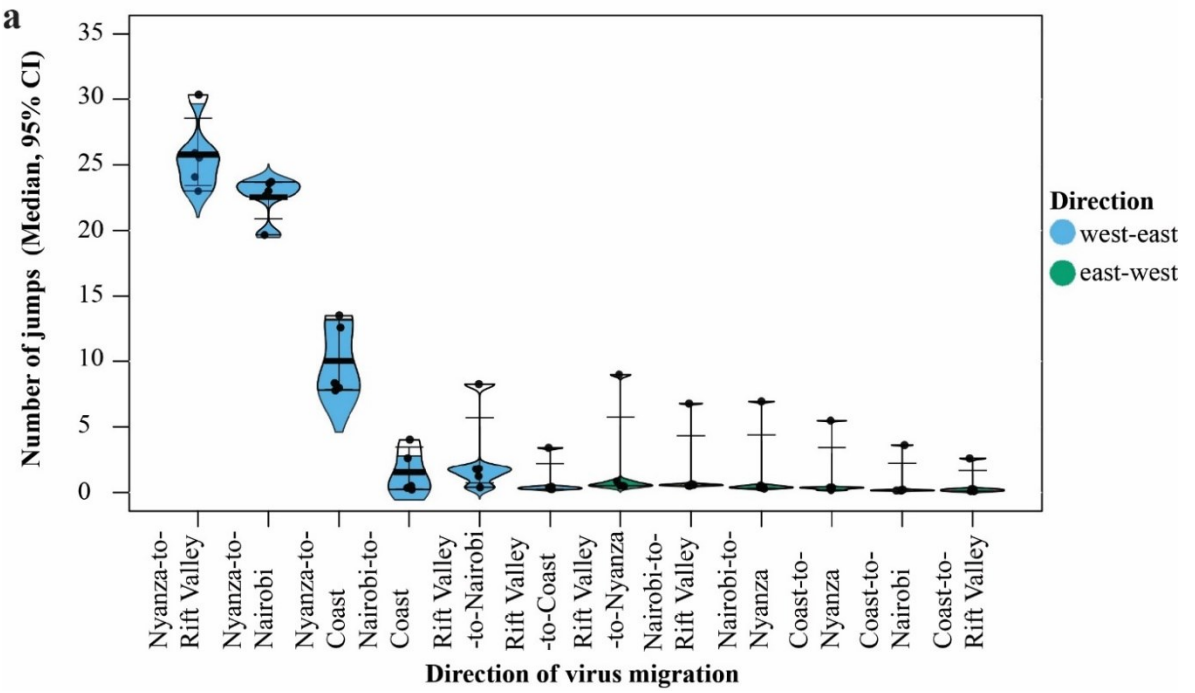
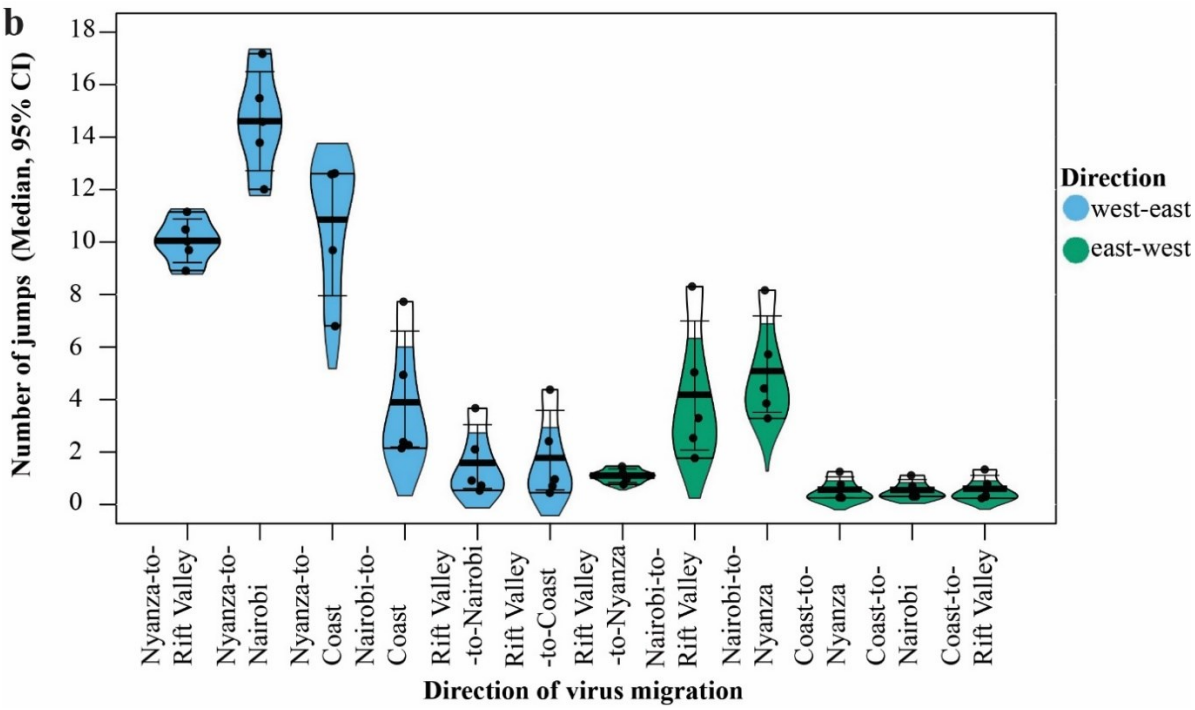
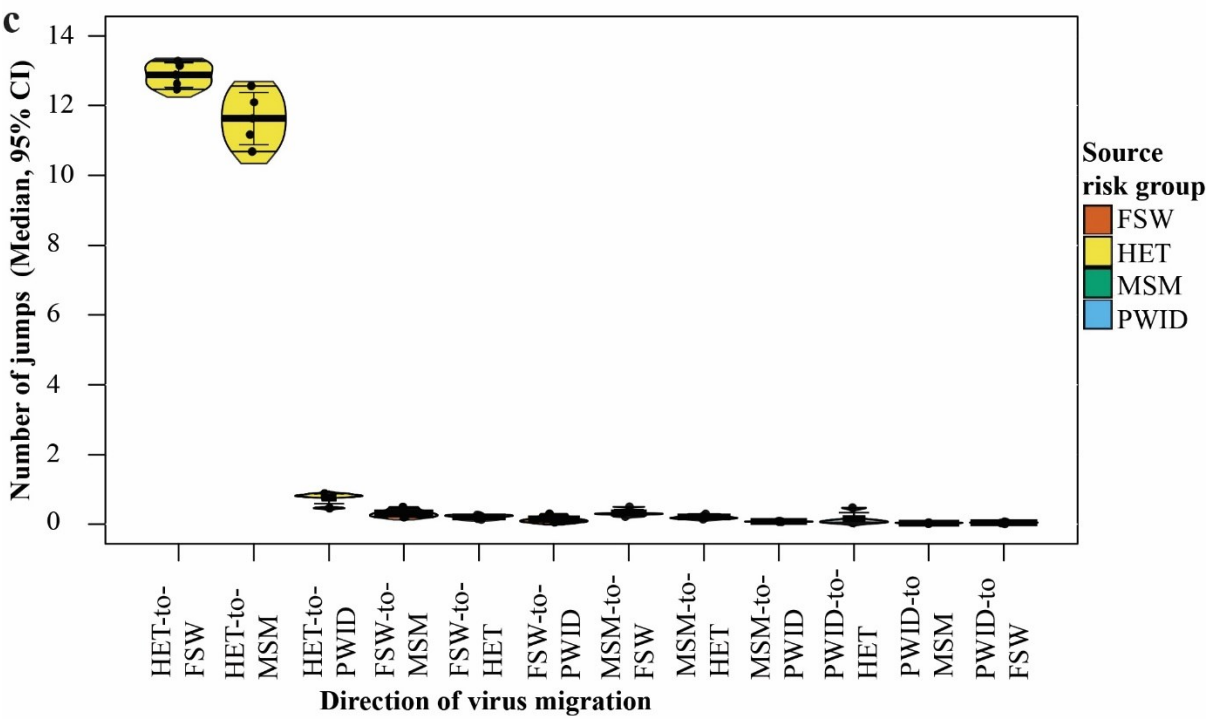


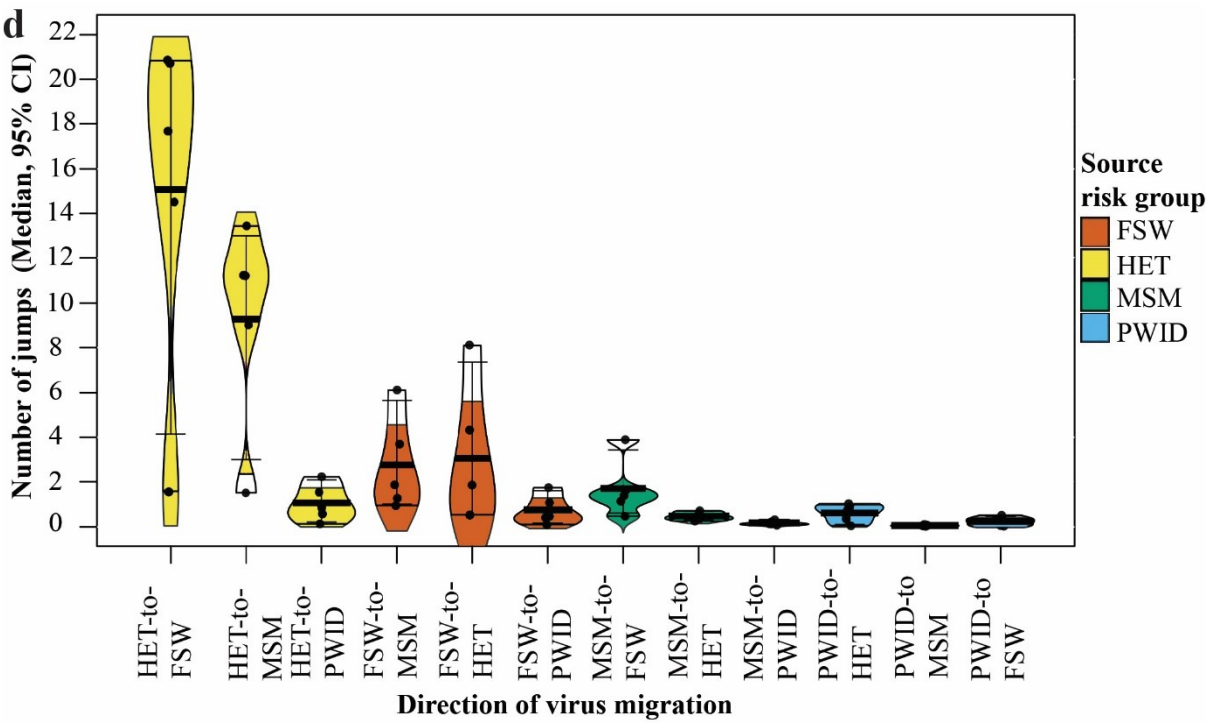
Figure S8. Number and direction of HIV-1 jumps between geographic locations and risk groups based on uniform and proportional sub-sampling of 1147 HIV-1 sub-subtype A1 sequences sampled 2010-2019 in Kenya.

Pirate plots quantifying direction and number of HIV-1 jumps between geographic locations and risk groups. Graphs represent (a) virus jumps between geographic provinces based on proportional sub-sampling (n=5 datasets; 70 sequences from Nyanza, 34 sequences from Rift Valley, 26 sequences from Nairobi, and 14 sequences from Coast); (b) virus jumps between geographic provinces based on uniform sub-sampling (n=5 datasets; 25 sequences from Nyanza, 25 sequences from Rift Valley, 25 sequences from Nairobi, and 25 sequences from Coast); (c) virus jumps between risk groups based on proportional sub-sampling (n=5 datasets; 64 sequences from HET, 14 sequences from FSW, 15 sequences from MSM, and 4 sequences from PWID); and (d) virus jumps between risk groups based on uniform sub-sampling (n=5 datasets; 27 sequences from HET, 27 sequences from FSW, 27 sequences from MSM, and 27 sequences from PWID). Black lines in the plots represent median jumps estimates (and 95% confidence intervals). The geographic plots are coloured as per the direction of transmission (blue: West-to-East; and green: East-to-West) whilst the risk group plots are coloured as per the source risk group (green: MSM; sky blue: PWID; vermillion: FSW; and yellow: HET). Overall, there was significantly more West-to-East virus flow than from East-to-West ($p<0.0001$; both uniform and proportional sub-sampling), and more virus flow from HET-to-key populations than vice-versa ($p<0.001$; both uniform and proportional sub-sampling).






c
Number of jumps (Median, 95% CI)
0
2
4
6
8
10
12
14
HET-to-FSW
HET-to-MSM
HET-to-PWID
FSW-to-MSM
FSW-to-HET
FSW-to-PWID
MSM-to-FSW
MSM-to-HET
MSM-to-PWID
PWID-to-HET
PWID-to MSM
PWID-to FSW
Direction of virus migration
Source risk group
FSW
HET
MSM
PWID


d
Number of jumps (Median, 95% CI)
0
2
4
6
8
10
12
14
16
18
20
22
HET-to-FSW
HET-to-MSM
HET-to-PWID
FSW-to-MSM
FSW-to-HET
FSW-to-PWID
MSM-to-FSW
MSM-to-HET
MSM-to-PWID
PWID-to-HET
PWID-to MSM
PWID-to FSW
Direction of virus migration
Source risk group
FSW
HET
MSM
PWID

SUPPLEMENTARY REFERENCES

Kenya National AIDS control council (NACC). (2018, 2019). Kenya HIV estimates report 2018. Retrieved from https://nacc.or.ke/wp-content/uploads/2018/11/HIV-estimates-report-Kenya-20182.pdf

Kenya National Bureau of Statistics. (2019). 2019 Kenya population and housing census Volume 1: Population by county and sub-county. Retrieved from https://www.knbs.or.ke/?wpdmpro=2019-kenya-population-and-housing-census-volume-i-population-by-county-and-sub-county

National AIDS and STI Control Programme (NASCOP). (2019). Key Population Mapping and Size Estimation in Selected Counties in Kenya: Phase 1. Retrieved from https://hivpreventioncoalition.unaids.org/wp-content/uploads/2020/02/KPSE-Phase1-Final-Report.pdf

National AIDS and STI Control Programme (NASCOP). (2020). Preliminary KENPHIA 2018 Report. Retrieved from https://www.nascop.or.ke/kenphia-report

Phylogeographic assessment reveals geographic sources of HIV-1 dissemination among men who have sex with men in Kenya

George M. Nduva[1,2*], Frederick Otieno[3], Joshua Kimani[4,5], Lyle R. McKinnon[4,5,6], Francois Cholette[5,7], Paul Sandstrom[7], Susan M. Graham[2,8], Matt A. Price[9,10], Adrian D. Smith[12], Robert C. Bailey[3,11], Amin S. Hassan[1,2#], Joakim Esbjörnsson[1,12#], and Eduard J. Sanders[2,12#]

[#]Equal contribution as senior authors

[1]Lund University, Lund, Sweden, [2]Kenya Medical Research Institute-Wellcome Trust Research Programme, Kilifi, Kenya, [3]Nyanza Reproductive Health Society, Kisumu, Kenya, [4]University of Nairobi, Nairobi, Kenya, [5]University of Manitoba, Winnipeg, Canada, [6]Centre for the AIDS Programme of Research in South Africa (CAPRISA), South Africa, [7]National Microbiology Laboratory at the JC Wilt Infectious Diseases Research Centre, Public Health Agency of Canada, Winnipeg, Canada, [8]University of Washington, Seattle, USA, [9]IAVI, San Francisco, USA, [10]University of California, San Francisco, USA, [11]University of Illinois at Chicago, USA, [12]The University of Oxford, Oxford, United Kingdom.

Corresponding Author:
George Nduva
Department of Translational Medicine, Lund University, 221 84 Lund, Sweden
Email: george.makau_nduva.7410@med.lu.se

Word count:
Short title: Molecular Epidemiology of HIV-1 in Kenyan MSM
Keywords: HIV-1, molecular epidemiology, phylogeographic, MSM, Kenya.

ABSTRACT
HIV-1 transmission dynamics involving men who have sex with men (MSM) in Africa are not well understood. We investigated the rates of HIV-1 transmission between MSM across three regions in Kenya: Coast, Nairobi and Nyanza. We analysed 372 HIV-1 partial *pol* sequences sampled during 2006-2019 from MSM in Coast (N=178, 47.9%), Nairobi (N=137, 36.8%), and Nyanza (N=57, 15.3%) provinces in Kenya. Maximum-Likelihood (ML) phylogenetics and Bayesian inference were used to determine HIV-1 clusters, evolutionary dynamics, and virus migration rates between geographic regions. HIV-1 sub-subtype A1 (72.0%) was most common followed by subtype D (11.0%), unique recombinant forms (8.9%), subtype C (5.9%), CRF 21A2D (0.8%), subtype G (0.8%), CRF 16A2D (0.3%), and subtype B (0.3%). Forty-six clusters (size range 2-20 sequences) were found – half (50.0%) of which had evidence of extensive HIV-1 mixing among different provinces. Data revealed an exponential increase in infections among MSM during the early-to-mid 2000s, and stable or decreasing transmission dynamics in recent years (2017-2019). Phylogeographic inference showed significant (Bayes Factor, BF>3) HIV-1 dissemination from Coast to Nairobi and Nyanza provinces, and from Nairobi to Nyanza province. Strengthening HIV-1 prevention programmes to MSM in geographic locations with higher HIV-1 prevalence among MSM (such as Coast and Nairobi) may reduce HIV-1 incidence among MSM in Kenya.

INTRODUCTION

In sub-Saharan Africa, the HIV-1 epidemic among men who have sex with men (MSM) has only recently received attention – and the role of MSM in HIV-1 transmission has been acknowledged[1-3]. In Kenya, the national HIV-1 prevalence is 4.9% in the adult population, but is three-fold higher in MSM than in heterosexual men[4,5]. HIV-1 prevalence among MSM in Kenya varies between regions – and ranges from 17.8% in Kisumu (Western Kenya)[6] to 24.5% in Coastal Kenya[7], and from 25.0% to 26.4% in Nairobi[8,9]. There is evidence of high mobility of MSM sex workers between regions which could link HIV-1 transmissions in different regions[10].The Ministry of Health in Kenya through the National AIDS Control Council (NACC) has made efforts to strengthen HIV healthcare services for MSM[11,12]. Yet, stigma against male-same-sex practices and policies criminalizing consensual same-sex sexual practices have obstructed progress[12-14]. In the past, geographic mobility has been shown to play an important role in HIV-1 dispersal[15,16]. Taken together, it is possible that spatial differences in HIV-1 distribution in Kenya combined with geographically mobile MSM sex workers could impact HIV-1 spread among MSM throughout the country[15,16]. However, clear data on HIV-1 transmission dynamics within and between MSM in different geographic regions are lacking in Kenya.

HIV-1 transmission dynamics can be assessed by linking socio-demographic, clinical, and behavioural data with HIV-1 sequence data through phylogenetics[17-26]. While limited HIV-1 sequences have been obtained from blood plasma from MSM living with HIV in Kenya, phylogenetic determination of patterns of HIV-1 transmission among Kenyan MSM suggest extensive MSM HIV-1 clustering (and infrequent HIV-1 mixing between MSM and presumed heterosexuals in the general population)[2,27-30]. In addition, a phylogenetic study in 2013 reported frequent HIV-1 gene flow between MSM in Coastal Kenya and Nairobi – albeit with small sample size and limited geographic coverage[29]. In the period 2005-2019, more MSM HIV-1 sequences have become available from diverse geographical locations in Kenya, allowing in-depth characterization of evolutionary dynamics in the MSM HIV-1 epidemic in Kenya. Here, we used HIV-1 *pol* data to phylodynamically infer HIV-1 transmission rates among MSM in three different geographic regions in Kenya.

MATERIALS AND METHODS

Study population

New sequences were generated from blood plasma obtained through studies conducted through the MSM Health Research Consortium – a multi-site collaboration between researchers affiliated with KEMRI-Wellcome Trust (KWTRP) in Coastal Kenya, Nyanza Reproductive Health Society (NRHS) in Nyanza, and Sex Workers Outreach Program (SWOP) clinics in Nairobi. These included samples from Coast derived from participants in a prospective observational cohort (2006-2019)[31], samples from Nairobi from a respondent-driven sample survey (TRANSFORM, 2017)[32], and samples from Nyanza derived from the Anza Mapema cohort (2015-2017)[33].

HIV-1 *pol* sequence dataset

The HIV-1 *pol* sequences were comprised of 1020 nucleotides, HXB2 [K03455] positions 2267-3287. HIV-1 RNA was purified from patient blood plasma using the RNeasy Lipid Tissue Mini Kit (QIAGEN) as previously described[34]. Reverse transcription and amplification of partial *pol* gene were performed using the One-Step Superscript III RT/Platinum Taq High Fidelity Enzyme Mix (ThermoFisher Scientific™) with the *pol*-specific primer pair JA269 and JA272[35]. First-round PCR products were amplified in a nested PCR with DreamTaq Green DNA Polymerase (ThermoFisher Scientific™) using *pol*-specific primers JA271 and JA270[35]. PCR products were sequenced in both directions with the nested PCR primers using the BigDye terminator kit v1.1 (Applied Biosystems) and the sequences were determined on an ABI PRISM 3130×1 Genetic Analyzer (Applied Biosystems).

Additional Kenyan HIV-1 *pol* sequences (referred to as published sequences, 2006-2019) were retrieved (October 11th 2021) from the Los Alamos HIV-1 sequence database[36]. The combined new and published sequences (referred to as Kenyan dataset) were annotated with information on sampling dates and geographical area of residence during sampling (i.e. province; Coast, Nairobi, Nyanza).

HIV-1 Subtyping

The Kenyan dataset was aligned with the HIV-1 Group M (subtypes A-K + Recombinants) subtype reference dataset (available at the Los Alamos HIV database, http://www.hiv.lanl.gov) using the MAFFT algorithm in Geneious Prime 2019[37,38]. The resulting alignment was used to construct a Maximum-Likelihood phylogenetic tree in PhyML using the general time-reversible substitution model with a gamma-distributed rate variation and proportion of invariant sites (GTR+Γ4+I)[39]. Branch support was assessed using the Shimodaira-Hasegawa like approximate Likelihood Ratio Test (aLRT-SH) in PhyML, with aLRT-SH ≥0.90 considered as significant[18,40]. Subtypes were assigned based on the Subtype/CRF-resolved phylogeny visualized using FigTree v1.4.4 (https://github.com/rambaut/figtree/releases). Subtype assignment was further verified using the REGA HIV-1 Subtyping Tool (v.3.0) and unique recombinant forms (URFs) were detected using the jumping profile Hidden Markov Model (jpHMM)[41,42].

HIV-1 Cluster analysis

Sequences were grouped into subtype-specific datasets and a search for related sequences was done for each subtype-specific (A1, C and D) dataset using the NCBI GenBank BLAST tool, limiting results to the 10 most similar hits per sequence, and retaining the oldest sequence per individual[23,24,43]. Kenyan sequences and reference sequences were combined and aligned using the MAFFT algorithm in Geneious Prime 2019[37]. Subtype-specific alignments were edited to exclude codon positions associated with drug resistance, and maximum-likelihood phylogenies were reconstructed in PhyML. For each subtype, monophyletic clades with aLRT-SH support ≥0.9 and which were dominated (≥80%) by Kenyan sequences (compared to reference sequences) were defined as Kenyan HIV-1 clusters[18]. Clusters were classified based on the number of sequences per cluster into dyads (2 sequences), networks (3-14 sequences) and large clusters (>14 sequences)[24,28,44].

Bayesian phylodynamic and discrete phylogeographic inference

To date clusters and to estimate the effective population size through time ($N_{e.T}$), Bayesian phylodynamic inference was performed in BEAST 1.10.4 using the Bayesian Skygrid model, an uncorrelated lognormal relaxed clock, and the general time-reversible substitution model with a

gamma-distributed rate variation and proportion of invariant sites (GTR+Γ4+I)[45-48]. Only sequences classified as pure A1, C, and D subtypes were analysed. BEAST runs were computed with a chain length of 100-300 million generations for each dataset, sampling every 10,000th – 30,000th iteration, and discarding the first 10% as burn-in. Convergence was determined in Tracer v.1.7.0 and defined as effective sample sizes (ESS) ≥100[45]. Maximum clade credibility (MCC) trees were summarised using Tree-Annotator v1.8.2 (BEAST suite).

To infer the direction of virus movements between geographic locations from HIV-1 sequence data, a discrete phylogeographic inference was computed, using specific locations as independent discrete states[15,49,50]. Several sensitivity analyses were performed to test the robustness of our data. First, the Kenyan dataset was grouped by subtype (A1, C and D), and the phylogeographic inference was performed using all the sequences per subtype. Secondly, to reduce sampling bias arising from the unproportionable allocation of sequences per location, sequences in the subtype A1-specific dataset (the largest of the three subtypes) were randomised and sub-sampled into a dataset with an equal number of sequences per province using in-house Perl scripts (available upon request). Lastly, subtype A1 sequences from Coast Province were sub-sampled uniformly and used to estimate virus migration between three geographically distinct regions in Coastal Kenya (i.e. Mombasa, South Coast, and North Coast).

In the phylogeographic inference, the asymmetric model was adopted (over the alternative symmetric model) as it relaxes the assumption of constant diffusion rates through time to realistically model the location-exchange processes[15,50]. In addition to estimating the direction of HIV-1 migration, the proportions of forward and reverse rates of migrations between geographic locations were quantified using a robust counting approach (Markov jumps) implemented in BEAST[51]. Maximum clade credibility (MCC) trees annotated with demographic and epidemiological data were summarized in Tree-Annotator v1.10.4 (BEAST suite) and visualized in Figtree (v1.4.4). Well-supported virus movements and Bayes factors (BF) assessing statistical support were summarized using SPREAD v1.0.7, and (BF≥3 was considered significant)[49].

Statistical analysis

Continuous data were presented using medians and interquartile ranges (IQR). Frequencies and percentages were used to describe categorical data. A multivariable logistic regression model was used to assess associations between individual sequence characteristics (e.g. subtype, location of sampling, year [range] of sampling, and source of sequence data – i.e. published or newly generated) and phylogenetic clustering. Statistics and summary plots were done using Stata 15 (StataCorp LLC, College Station, Texas, USA) and RStudio (version 1.2.5001) with the packages: *yarrr,* and *ggplot2*[52,53].

Nucleotide sequence accession numbers

Nucleotide sequences were deposited in GenBank under the following accession numbers: OM109723-OM109725, OM109756-OM109766, OM109772-OM109799, OM109814-OM109862, OM109879-OM109949, OM110011-OM110019, OM110126-OM110127, OM110136-OM110149, OM110169-OM110170, OM110171,OM110174, OM110178-OM110181, OM110193-OM110194, OM110212-OM110218, OM110229-OM110240, OM110245-OM110246, and OM110272-OM110282.

Ethical consideration

Plasma samples used to generate the new sequences were obtained from ongoing or concluded studies that were also approved by Kenya Medical Research Institute (KEMRI) Scientific and Ethics Review Unit (SERU 3747, 3280 and 3520, and SSC 894). Since published sequences were obtained from an open-access public domain, informed consent was not retrospectively obtained. Instead, we sought approval through a study protocol that was reviewed by KEMRI/SERU (SERU 3547).

RESULTS

Study Population, sequence dataset and subtype distribution

Among the 372 HIV-1 partial *pol* sequences analysed, 213 (57.3%) were generated in this study, and 159 (42.7%) were previously published. The majority (N=178, 47.9%) of the sequences were from the Coast province, 137 (36.8%) Nairobi province, and 57 (15.3%) Nyanza province. (Figure1, Table 1, Supplementary Tables S1, and S2, and Supplementary Figure S1). Sequences belonged to sub-subtype A1 (N=268, 72.0%), subtype D (N=41, 11.0%), subtype C (N=22, 5.9%), subtype G (N=3, 0.8%), CRF 21A2D (N=3, 0.8%), CRF 16A2D (N=1, 0.3%), and subtype B (N=1, 0.3%). Unique recombinant forms (URFs) identified included A1D (N=19, 5.1%), A1C (N=7, 1.9%), D01AE (N=5, 1.3%), A1B (N=1, 0.3%), DB (N=1, 0.3%, Figure 2).

MSM HIV-1 clusters

Clusters were determined from Maximum-likelihood (ML) phylogenies reconstructed for the most prevalent HIV-1 subtypes in the population (subtypes A (A1), C, and D – cumulatively comprising 89.0% of the sequences in the Kenyan dataset). Non-Kenyan HIV-1 reference sequences were obtained from GenBank based on similarity (where of 931 participant-unique sub-subtype A1 sequences remained after removal of redundancies; 488 for subtype C; and 350 for subtype D). Of 331 (A1, C and D) sequences in the cluster analysis, 229 sequences (61.2%) formed 46 statistically supported clusters (size range: 2-20 sequences). Dyad/pairs were most common (N=25, 54.4% of all clusters), followed by networks having 3-14 sequences (N=18, 39.1%), and large clusters having more than 14 sequences (N=3, 6.5%). The majority (N=34, 73.9%) were sub-subtype A1 clusters, followed by subtype D (N=8, 17.4%), and subtype C (N=4, 8.7%, Table 2, and Supplementary Figure S2)

Geographic stratification of clustering patterns

Stratification of clusters by geographic regions showed two distinct clustering patterns. First, some clusters (N=23, 50.0%) had sequences belonging exclusively to one specific province including Coast (N=14, 30.4%), Nairobi (N=6, 13.0%), and Nyanza (N=3, 6.5%) province-exclusive clusters. The remaining clusters (N=23, 50.0%) were mixed between different provinces where HIV-1 mixing between Coast and Nairobi was most common (N=13, 28.3% clusters), followed by mixing between Nyanza, Nairobi, and Coast (N=5, 10.9%), Nyanza and Nairobi (N=3, 6.5%), and Nyanza and Coast (N=2, 4.4%, Table 2, and Supplementary Figure S2). Sequences from Nairobi province were more likely to cluster compared to sequences from Coast province (adjusted odds ratio [aOR] 3.5, 95% confidence interval [CI] 1.2-10..4, P=0.022, Table 3).

Estimating effective population size through time and dating clusters

In-depth phylodynamic analysis indicated that the number of MSM contributing to new HIV-1 A1 infections over time increased exponentially during the early 2000s, followed by a period with some fluctuation (but largely steady) between 2000 and 2017, and mostly decreasing dynamics during recent years (2017-2019, Figure 3a). Likewise, for both subtype C and D lineages, the effective population size increased exponentially during 2007-2008 and has stabilized in recent years (2016-2019, Figure 3b, and 3c).

Estimating dates of origins of all clusters indicated that the majority (65%) of transmissions within clusters took place between 2000 and 2014. The oldest sub-subtype A1 cluster had 9 MSM from Nyanza, Nairobi, and Coast, and had originated during 1987, whilst the youngest cluster was dated to 2014 among MSM in Nyanza (Figure 4a, Supplementary Table S3, and Supplementary Figure S3). The largest A1 cluster (N=20, 2008-2017) had remained active over 20 years since the estimated time to the most recent common ancestor (tMRCA) in 1997 and was geographically spread out to Nyanza, Nairobi, and Coast provinces. The second-largest A1 cluster (N=19, 2008-2017) originated in 1996 and had sequences from Nyanza, Nairobi, and Coast provinces. The four subtype C clusters originated during 1988, 1998, 2009, and 2014, respectively, whilst the earliest subtype D cluster originated during 1976 and the youngest during 2014 (Figure 4b, Figure 4c, and Supplementary Table S3). Overall, there was evidence of onward HIV-1 transmission among MSM, within longstanding and geographically diverse HIV-1 networks.

HIV-1 migration between provinces in Kenya

Ancestral locations and rates in historical virus jumps were first estimated based on all subtype-specific sequences in the Kenyan dataset (i.e. 268 sub-subtype A1, 41 subtype D, and 22 subtype C sequences). Phylogeographic analysis indicated significant support (Bayes Factor, BF≥3) for virus migration from Coast to Nairobi (BF=3716; subtype A1, BF=268; subtype C; and BF=16; subtype D) and from Nairobi to Nyanza (BF=3716; subtype A1, BF=43; subtyped D, Supplementary Table S4). Exploring temporal trends in virus transitions between geographic provinces summarised from trait-annotated maximum clade credibility trees indicated that the proportion of virus export from Coast to Nairobi increased from 4.2% before 2000 to 14.2% during 2001-2010, and declined to 4.9% during 2011-2020. Likewise, virus export from Nairobi to Nyanza increased from 2.4% in 2000-2010 to 10.8% in 2011-2020, whilst reverse transitions were rare and occurred only from Nyanza to Nairobi (Supplementary Table S5, Suplementary Figure S4 and Suplementary Figure S5).

A sensitivity analysis with uniform sampling per province was performed to confirm the robustness of the initial phylogeographic inference. The uniformly sub-sampled dataset comprised 135 HIV-1 sub-subtype A1 sequences (45 sequences each from Nairobi, Mombasa, and Nyanza province). Based on this analysis, there was significant support for HIV-1 migration from Coast to Nairobi (BF=7766), Nairobi to Nyanza (BF=1293), and Coast to Nyanza (BF=336, Table 4). Furthermore, Markov jumps estimates with uniform sampling indicated that the majority (80.3%) of HIV-1 jumps between provinces occurred from Coast to other provinces including jumps from Coast to Nyanza (N=26, 42.6% of all virus jumps between provinces), and from Coast to Nairobi (N=23, 37.7%, Table 5, Figure 5). There was also some (N=10, 16.4%) virus exchange between Nairobi and Nyanza, such that virus jumps Nairobi to Nyanza (N=7, 11.5%) was two-fold higher than from Nyanza to Nairobi (N=3, 4.9%, Table 5).

DISCUSSION

We found high rates of HIV-1 geographic mixing and a high proportion of HIV-1 sequences exported from the Coast and Nairobi to Nyanza province – implying that the Coast and Nairobi provinces could be a major geographic sources of HIV-1 transmission amongst Kenyan MSM. Of all provinces in Kenya, the Coast and Nairobi provinces have the highest prevalence of HIV-1 among MSM[54]. In addition, MSM in Coastal Kenya are known to be highly mobile, and some engage in sex work in different locations across the country[10]. Taken together, our findings suggest that regions with the highest HIV-1 prevalence among MSM (such as Coast and Nairobi) may also have disseminated HIV-1 disproportionately to regions with lower HIV-1 prevalence among MSM (such as Nyanza province) in Kenya.

There are a few presumed mechanisms by which Coastal Kenya may serve as an important source of infections among MSM. One plausible explanation might be that as a very well recognised destination for domestic tourism and sex tourism, MSM (or non-disclosing HET) visit the area for sex tourism, effectively disseminating the virus upon returning from Coast. A second potential determinant could be connected to geographically mobile MSM sex workers – hypothetically, HIV-1 may first be acquired and/or amplified in the Coast, and then exported to other provinces. Thus the regional difference observed could potentially reflect amplification behaviour within Coastal Kenya – and onward spread to other provinces linked to an MSM migration gradient. Data on migration were not available during the current analysis but future studies may investigate this in detail. Future studies may also potentially investigate potential underlying demographic transitions – speculatively, young MSM sex workers may be drawn to the Coast province whilst older or socially privileged MSM or MSM sex workers may leave the region for other provinces. Overall, implementing HIV-1 prevention and care directed to MSM in Kenya (and considering areas with higher rates of HIV-1 dissemination such as Coast and Nairobi) might reduce ongoing HIV-1 transmission at a countrywide scale, as has been shown in other settings[55-58].

The majority (61.2%) of sequences analyzed in this study formed phylogenetically linked HIV-1 clusters, consistent with multiple introductions and ongoing infections among MSM within close networks in Kenya[23,24,59]. Half of the clusters comprised sequences collected from MSM from different geographic regions – indicating geographically extensive HIV-1 linkages. High rates of clustering involving HIV-1 in MSM have been reported both in our setting and other higher-income settings and could be linked to an increased risk of infection among MSM within close networks, involving geographically mobile individuals[10,23,24,27,29]. We estimated that a high proportion (65%) of HIV-1 transmissions occurred between 2000 and 2014 and that several clusters extended over multiple years, suggesting onward HIV-1 transmission among MSM within geographically diverse HIV-1 networks. HIV-1 sequences in this study were not closely related to reference sequences from the global epidemic, implying that the HIV-1 epidemic among MSM in Kenya is sustained locally.

In a broader context, several phylogenetic studies have revealed that the HIV-1 epidemic in Kenya is compartmentalized – where the majority of HIV-1 transmission occurs within risk groups[28-30]. Our recent work at a countrywide scale has demonstrated a minor (8%) proportion of HIV-1 MSM and heterosexual clustering[30]. Taken together, these studies indicate that ongoing transmission among MSM rarely impacts the general heterosexual HIV-1 epidemic in Kenya. MSM in Kenya have a high burden of HIV risk – to bring reduce overall HIV-1 incidence in Kenya, there is a need to implement directed HIV-1 prevention and treatment to MSM in Kenya.

The phylodynamic analysis investigating the evolutionary dynamics of the HIV-1 MSM sub-epidemic revealed an exponential increase in the number of infections during the early-to-mid 2000s (for HIV-1 A1, C and D lineages) – indicative of multiple HIV-1 outbreaks among Kenyan MSM[23,24,59]. Interestingly, the effective population size did not decrease following the nationwide introduction and scale-up of combination antiretroviral therapy (ART) in 2004. One potential reason for this is sub-optimal access to HIV-1 treatment and prevention services by MSM in Kenya due to fear of legal and social stigma and discrimination[4,60,61]. Nevertheless, the effective population size for the dominant strain (HIV-1 A1) showed fewer new infections in recent years (2017-2019) – possibly reflecting earlier ART

initiation due to changes in treatment recommendations[62] as well as some impact of risk reduction counselling, adherence support interventions[63,64], early recognition of acute HIV-1 infections, especially on the Kenyan Coast[65-67], and some uptake of pre-exposure prophylaxis targeting MSM in recent years[68-71]. Overall, increasing access to treatment – as well as destigmatisation and diversification of providers may further reduce HIV-1 incidence among MSM[32].

The major strength of our study is the use of HIV-1 sequences from well-characterized acute and early infected MSM cohorts sampled over 14 years in a sub-Saharan African setting. A limitation is that the study had a small sample size, which limited the identification of HIV-1 links in the entire MSM HIV-1 epidemic in Kenya. Incomplete sampling likely resulted in missing links and reduced clustering of HIV-1 sequences[72]. However, our sensitivity analyses before and after controlling for sampling bias indicated more jumps from Coastal Kenya to other provinces (and from Nairobi to Nyanza) than vice versa, indicating the robustness of the analysed HIV-1 sequence dataset. Another limitation is skewed spation-temporal sampling, and variations in sampling methods between studies which may have resulted in overrepresentation of some types of location-specific and/or subtype-specific clusters. Indeed, the HIV-1 C and HIV D lineages did not have a decreasing trend in recent years (2017-2019, compared to HIV-1 A1) – the reason for this could be realeted to skewed sampling over time in various geographic locations in this study. In addition, although the conflation of MSM and transgender people may have relevance for the distinction between sexual network types, we did not have data on gender identity – thus some transgender people may have been conflated for MSM.

In conclusion, we demonstrated extensive HIV-1 mixing among MSM in different regions in Kenya, where Coast and Nairobi provinces appears to have been a major source of virus dissemination. We hypothesise that MSM in these provinces may have disseminated HIV-1 disproportionately to MSM in other regions in the country. Increasing PrEP uptake and access to ART among MSM (and destigmatisation and diversification of providers) is necessary to reduce ongoing HIV-1 transmission among MSM in Kenya.

ADDITIONAL INFORMATION

Acknowledgements
We thank the staff affiliated with the MSM Health Research Consortium (MHRC) and IAVI for supporting studies involving key populations in Kenya. This manuscript was submitted for publication with permission from the Director of the Kenya Medical Research Institute (KEMRI).

Author contributions
G.N.M., A.S.H., E.J.S., and J.E conceptualized and designed the study. G.N.M, A.S.H., E.J.S. and J.E provided funding for the study. F.O., J.K., L.R.M., P.S., S.M.G., M.A.P., A.S., R.C.B, and E.J.S., provided samples from which new sequences used in the study were generated. G.N.M performed lab work, inferential analyses and produced all figures and tables. F.C helped with virus sequencing. G.N.M wrote the manuscript and all the authors reviewed, edited, and approved the manuscript for submission.

Competing Interests
The authors declare no competing interests.

Funding information
This work was supported through the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant # 107752/Z/15/Z] and the UK government. This work was also supported by funding from the Swedish Research Council (grant # 2016–01417) and the Swedish Society for Medical Research (grant # SA-2016). This work was funded in part by IAVI and made possible by the support of many donors, including the United States Agency for International Development (USAID). The full list of IAVI donors is available at http://www.iavi.org. The contents of this manuscript are the responsibility of the authors and do not necessarily reflect the views of USAID or the US Government, AAS, NEPAD Agency, Wellcome Trust, IAVI, Swedish Research Council, or the UK government.

Supplementary information
Accompanies this paper.

REFERENCES

1 Beyrer, C. *et al.* The expanding epidemics of HIV type 1 among men who have sex with men in low-and middle-income countries: diversity and consistency. *Epidemiologic reviews* **32**, 137-151 (2010).

2 Nduva, G. M., Nazziwa, J., Hassan, A. S., Sanders, E. J. & Esbjörnsson, J. The Role of Phylogenetics in Discerning HIV-1 Mixing among Vulnerable Populations and Geographic Regions in Sub-Saharan Africa: A Systematic Review. *Viruses* **13**, 1174 (2021).

3 Sanders, E. J., Jaffe, H., Musyoki, H., Muraguri, N. & Graham, S. M. Kenyan MSM: no longer a hidden population. *AIDS* **29**, S195-S199, doi:10.1097/qad.0000000000000928 (2015).

4 Kenya National AIDS Control Council. *Kenya AIDS Strategic Framework 2014/2015–2018/2019*, <http://nacc.or.ke/wp-content/uploads/2015/09/KASF_Final.pdf> (2019).

5 National AIDS and STI Control Programme (NASCOP). *Preliminary KENPHIA 2018 Report*, <https://www.nascop.or.ke/kenphia-report> (2020).

6 Kunzweiler, C. P. *et al.* Factors associated with prevalent HIV infection among Kenyan MSM: the Anza Mapema study. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **76**, 241-249 (2017).

7 Sanders, E. J. *et al.* HIV-1 infection in high risk men who have sex with men in Mombasa, Kenya. *Aids* **21**, 2513-2520 (2007).

8 Smith, A. D. *et al.* Disparities in HIV/STI burden and care coverage among men and transgender persons who have sex with men in Nairobi, Kenya: a cross-sectional study. *BMJ Open* **11**, e055783 (2021).

9 Smith, A. D. *et al.* HIV burden and correlates of infection among transfeminine people and cisgender men who have sex with men in Nairobi, Kenya: an observational study. *The Lancet HIV* **8**, e274-e283 (2021).

10 Geibel, S. *et al.* Factors associated with self-reported unprotected anal sex among male sex workers in Mombasa, Kenya. *Sexually transmitted diseases* **35**, 746-752 (2008).

11 Gruskin, S. & Tarantola, D. Universal access to HIV prevention, treatment and care: assessing the inclusion of human rights in international and national strategic plans. *AIDS (London, England)* **22**, S123 (2008).

12 van der Elst, E. M. *et al.* A more responsive, multi-pronged strategy is needed to strengthen HIV healthcare for men who have sex with men in a decentralized health system: qualitative insights of a case study in the Kenyan coast. *Journal of the International AIDS Society* **23**, e25597 (2020).

13 Cohen, M. S. *et al.* Antiretroviral treatment of HIV-1 prevents transmission of HIV-1: where do we go from here? *The Lancet* **382**, 1515-1524 (2013).

14 van der Elst, E. M. *et al.* Experiences of Kenyan healthcare workers providing services to men who have sex with men: qualitative findings from a sensitivity training programme. *Journal of the International AIDS Society* **16**, 18741 (2013).

15 Faria, N. R. *et al.* The early spread and epidemic ignition of HIV-1 in human populations. *science* **346**, 56-61 (2014).

16 Grabowski, M. K. *et al.* Migration, hotspots, and dispersal of HIV infection in Rakai, Uganda. *Nature communications* **11**, 1-12 (2020).

17 Pybus, O. G., Tatem, A. J. & Lemey, P. Virus evolution and transmission in an ever more connected world. *Proceedings of the Royal Society B: Biological Sciences* **282**, 20142878 (2015).

18 Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J. & Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data. *AIDS (London, England)* **31**, 1211 (2017).

19 Brenner, B. G. *et al.* High rates of forward transmission events after acute/early HIV-1 infection. *The Journal of infectious diseases* **195**, 951-959, doi:10.1086/512088 (2007).

20 Volz, E. M. *et al.* HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS medicine* **10**, e1001568; discussion e1001568, doi:10.1371/journal.pmed.1001568 (2013).

21 Ratmann, O. *et al.* Sources of HIV infection among men having sex with men and implications for prevention. *Science translational medicine* **8**, 320ra322-320ra322 (2016).

22 Poon, A. F. *et al.* Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *The lancet HIV* **3**, e231-e238 (2016).
23 Sallam, M. *et al.* Molecular epidemiology of HIV-1 in Iceland: Early introductions, transmission dynamics and recent outbreaks among injection drug users. *Infection, Genetics and Evolution* **49**, 157-163 (2017).
24 Esbjörnsson, J. *et al.* HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. *Virus evolution* **2**, vew010 (2016).
25 Bruhn, C. A. *et al.* The origin and emergence of an HIV-1 epidemic: from introduction to endemicity. *AIDS* **28**, 1031-1040, doi:10.1097/QAD.0000000000000198 (2014).
26 Frentz, D. *et al.* Patterns of transmitted HIV drug resistance in Europe vary by risk group. *PloS one* **9**, e94495 (2014).
27 Hassan, A. S. *et al.* HIV-1 subtype diversity, transmission networks and transmitted drug resistance amongst acute and early infected MSM populations from Coastal Kenya. *PLoS One* **13**, e0206177, doi:10.1371/journal.pone.0206177 (2018).
28 Nduva, G. M. *et al.* HIV-1 Transmission Patterns Within and Between Risk Groups in Coastal Kenya. *Sci Rep* **10**, 6775, doi:10.1038/s41598-020-63731-z (2020).
29 Bezemer, D. *et al.* HIV Type 1 transmission networks among men having sex with men and heterosexuals in Kenya. *AIDS research and human retroviruses* **30**, 118-126 (2014).
30 Nduva, G. M. *et al.* Quantifying rates of HIV-1 flow between risk groups and geographic locations in Kenya: a country-wide phylogenetic study. *(Submited)* (2021).
31 Sanders, E. J. *et al.* High HIV-1 incidence, correlates of HIV-1 acquisition, and high viral loads following seroconversion among MSM. *Aids* **27**, 437-446, doi:10.1097/QAD.0b013e32835b0f81 (2013).
32 Smith, A. D. *et al.* HIV burden and correlates of infection among transfeminine people and cisgender men who have sex with men in Nairobi, Kenya: an observational study. *Lancet HIV*, doi:10.1016/s2352-3018(20)30310-6 (2021).
33 Kunzweiler, C. P. *et al.* Depressive Symptoms, Alcohol and Drug Use, and Physical and Sexual Abuse Among Men Who Have Sex with Men in Kisumu, Kenya: The Anza Mapema Study. *AIDS and behavior* **22**, 1517-1529, doi:10.1007/s10461-017-1941-0 (2018).
34 Esbjörnsson, J. *et al.* Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease-indication of an evolving epidemic in West Africa. *Retrovirology* **7**, 23 (2010).
35 Hedskog, C. *et al.* Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PloS one* **5**, e11345 (2010).
36 Los Alamos National Laboratory. *HIV-1 database at the Los Alamos National Laboratory*, <http://www.hiv.lanl.gov/> (2019).
37 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *bioinformatics* **23**, 2947-2948 (2007).
38 Los Alamos National Library. *HIV-1 database at the Los Alamos National Library*, <http://www.hiv.lanl.gov/> (2019).
39 Guindon, S., Lethiec, F., Duroux, P. & Gascuel, O. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic acids research* **33**, W557-W559 (2005).
40 Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* **55**, 539-552, doi:10.1080/10635150600755453 (2006).
41 Schultz, A.-K. *et al.* jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic acids research* **37**, W647-W651 (2009).
42 Pineda-Peña, A.-C. *et al.* 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**, 337-348 (2013).
43 Kouyos, R. D. *et al.* Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *The Journal of infectious diseases* **201**, 1488-1497, doi:10.1086/651951 (2010).

44 Abidi, S. H. *et al.* Phylogenetic and drug-resistance analysis of HIV-1 sequences from an extensive paediatric HIV-1 outbreak in Larkana, Pakistan. *Frontiers in microbiology*, 2305 (2021).
45 Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4**, vey016 (2018).
46 Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular biology and evolution* **29**, 2157-2167 (2012).
47 Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol* **30**, 713-724, doi:10.1093/molbev/mss265 (2013).
48 Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution* **22**, 1185-1192 (2005).
49 Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS computational biology* **5** (2009).
50 Edwards, C. J. *et al.* Ancient hybridization and an Irish origin for the modern polar bear matriline. *Current Biology* **21**, 1251-1258 (2011).
51 Minin, V. N. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of mathematical biology* **56**, 391-412 (2008).
52 Wickham, H. *ggplot2: elegant graphics for data analysis*. (springer, 2016).
53 Phillips, N. D. Yarrr! The pirate's guide to R. *APS Observer* **30** (2017).
54 Kenya National AIDS Control Council. *Kenya HIV Prevention Response and Modes of Transmission Analysis.*, <http://siteresources.worldbank.org/INTHIVAIDS/Resources/375798-1103037153392/KenyaMOT22March09Final.pdf> (2009).
55 Gerberry, D. J., Wagner, B. G., Garcia-Lerma, J. G., Heneine, W. & Blower, S. Using geospatial modelling to optimize the rollout of antiretroviral-based pre-exposure HIV interventions in Sub-Saharan Africa. *Nature communications* **5**, 1-15 (2014).
56 McGillen, J. B., Anderson, S.-J., Dybul, M. R. & Hallett, T. B. Optimum resource allocation to reduce HIV incidence across sub-Saharan Africa: a mathematical modelling study. *The lancet HIV* **3**, e441-e448 (2016).
57 Anderson, S.-J. *et al.* Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study. *The Lancet* **384**, 249-256 (2014).
58 Bailey, R. C. *et al.* Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *The lancet* **369**, 643-656 (2007).
59 Skar, H. *et al.* Dynamics of two separate but linked HIV-1 CRF01_AE outbreaks among injection drug users in Stockholm, Sweden, and Helsinki, Finland. *Journal of virology* **85**, 510-518 (2011).
60 Micheni, M. *et al.* Risk of sexual, physical and verbal assaults on men who have sex with men and female sex workers in coastal Kenya. *AIDS (London, England)* **29**, S231 (2015).
61 Stannah, J. *et al.* HIV testing and engagement with the HIV treatment cascade among men who have sex with men in Africa: A systematic review and meta-analysis. *Lancet HIV* (2019).
62 Ministry of Health, N. A. S. C. P. *Guidelines on Use of Antiretroviral Drugs for Treating and Preventing HIV Infections in Kenya 2016*, <https://www.prepwatch.org/wp-content/uploads/2016/08/Guidelines-on-ARV-for-Treating-Preventing-HIV-Infections-in-Kenya.pdf> (2016).
63 Graham, S. M. *et al.* A randomized controlled trial of the Shikamana intervention to promote antiretroviral therapy adherence among gay, bisexual, and other men who have sex with men in Kenya: feasibility, acceptability, safety and initial effect size. *AIDS and behavior*, 1-14 (2020).
64 Möller, L. M. *et al.* Changes in sexual risk behavior among MSM participating in a research cohort in coastal Kenya. *AIDS (London, England)* **29**, S211 (2015).

65 Mugo, P. M. *et al.* Effect of text message, phone call, and in-person appointment reminders on uptake of repeat HIV testing among outpatients screened for acute HIV infection in Kenya: a randomized controlled trial. *PLoS One* **11**, e0153612 (2016).
66 Sanders, E. J. *et al.* Acute HIV-1 infection is as common as malaria in young febrile adults seeking care in coastal Kenya. *AIDS (London, England)* **28**, 1357 (2014).
67 Sanders, E. J. *et al.* Targeted screening of at-risk adults for acute HIV-1 infection in sub-Saharan Africa. *AIDS (London, England)* **29**, S221 (2015).
68 Kimani, M. *et al.* Pr EP interest and HIV-1 incidence among MSM and transgender women in coastal Kenya. *Journal of the International AIDS Society* **22**, e25323 (2019).
69 Wahome, E. *et al.* An empiric risk score to guide PrEP targeting among MSM in coastal Kenya. *AIDS and behavior* **22**, 35-44 (2018).
70 Graham, S. M. *et al.* Development and pilot testing of an intervention to promote care engagement and adherence among HIV-positive Kenyan MSM. *AIDS (London, England)* **29**, S241 (2015).
71 van der Elst, E. M. *et al.* Strengthening healthcare providers' skills to improve HIV services for MSM in Kenya. *AIDS (London, England)* **29**, S237 (2015).
72 Novitsky, V., Moyo, S., Lei, Q., DeGruttola, V. & Essex, M. Impact of sampling density on the extent of HIV clustering. *AIDS research and human retroviruses* **30**, 1226-1235 (2014).
73 National AIDS and STI Control Programme (NASCOP). *Key Population Mapping and Size Estimation in Selected Counties in Kenya: Phase 1*, <https://hivpreventioncoalition.unaids.org/wp-content/uploads/2020/02/KPSE-Phase1-Final-Report.pdf> (2019).

TABLES

Table 1. Distribution of HIV-1 *pol* sequences (N=372) from Kenyan MSM, overall and by geographic location.

Category	**Number of sequences (N, %)**			
Geographic region	**Coast**	**Nairobi**	**Nyanza**	**Total**
Year (range)				
2006–2010	117 (65.7%)	1 (0.7%)	0 (0.0%)	118 (31.7%)
2011–2015	32 (18.0%)	1 (0.7%)	19 (33.3%)	52 (14.0%)
2016–2019	29 (16.3%)	135 (98.5%)	38 (66.7%)	202 (54.3%)
Sequences				
New	21 (11.8%)	135 (98.5%)	57 (100%)	213 (57.3%)
Published	157 (88.2%)	2 (1.5%)	0 (0.0%)	159 (42.7%)
Subtype				
A1	121 (68%)	102 (74.5%)	45 (79%)	268 (72%)
D	22 (12.4%)	13 (9.5%)	6 (10.5%)	41 (11%)
URF	16 (9%)	14 (10.2%)	3 (5.3%)	33 (8.9%)
C	14 (7.9%)	5 (3.7%)	3 (5.3%)	22 (5.9%)
21A2D	0 (0%)	3 (2.2%)	0 (0%)	3 (0.8%)
G	3 (1.7%)	0 (0%)	0 (0%)	3 (0.8%)
16A2D	1 (0.6%)	0 (0%)	0 (0%)	1 (0.3%)
B	1 (0.6%)	0 (0%)	0 (0%)	1 (0.3%)
Total	178 (47.9%)	137 (36.8%)	57 (15.3%)	372 (100%)

Abbreviations: MSM, men who have sex with men. URF; unique recombinant form, CRF; circulating recombinant form.

Table 2. The number of Kenyan MSM HIV-1 clusters by cluster size and geographic region.

	Dyads (2 sequences)	**Networks (3-14)**	**Large clusters (≥14)**	**Total clusters**
Subtype				
A1	12 (66.7%)	19 (76.0%)	3 (100%)	34 (73.9%)
C	2 (11.1%)	2 (8.0%)	0 (0.0%)	4 (8.7%)
D	4 (22.2%)	4 (16.0%)	0 (0.0%)	8 (17.4%)
Geographic region				
Coast	6 (24.0%)	8 (44.4%)	0 (0.0%)	14 (30.4%)
Coast/Nairobi	11 (44.0%)	2 (11.1%)	0 (0.0%)	13 (28.3%)
Nairobi	2 (8.0%)	4 (22.2%)	0 (0.0%)	6 (13.0%)
Nyanza/Nairobi/Coast	2 (8.0%)	0 (0.0%)	3 (100%)	5 (10.9%)
Nyanza	0 (0.0%)	3 (16.67%)	0 (0.0%)	3 (6.5%)
Nyanza/Nairobi	3 (12.0%)	0 (0.0%)	0 (0.0%)	3 (6.5%)
Nyanza/Coast	1 (4.0%)	1 (5.56%)	0 (0.0%)	2 (4.4%)
Total	25 (54.4%)	18 (39.1%)	3 (6.5%)	46 (100%)

Abbreviations: MSM, men who have sex with men. Clusters were classified based on the number of sequences per cluster into dyads (2 sequences), networks (3-14 sequences) and large clusters (>14 sequences)

Table 3. Factors associated with HIV-1 clustering among MSM with HIV-1 in Kenya.

Characteristics		**Multivariate Analysis**
		*aOR, 95% CI), p-value
Year (range)	2006-2010	Reference
	2011-2015	1.0 (0.4-2.2), 0.937
	2016-2020	1.1 (0.3-3.4), 0.932
Subtype	A1	Reference
	C	0.6 (0.2-1.5), 0.258
	D	1.0 (0.5-2.0), 0.884
Province	Coast	Reference
	Nairobi	3.5 (1.2-10.4), 0.022
	Nyanza	1.8 (0.5-5.9), 0.34
Sequence	Published	Reference
	Newly generated	2.5 (1.7-4.0), <0.001

Abbreviations: MSM, men who have sex with men; PWID; *aOR, adjusted odds ratio.

Table 4. HIV-1 migration rates (Bayes factor, BF≥3) between geographic locations in Kenya.

The direction of migration events (from, to)	Bayes Factor (BF)	Posterior probability
Migration between provinces		
Coast-to-Nairobi	7766	1
Nairobi-to-Nyanza	1293	1
Coast-to-Nyanza	336	1
Nyanza-to-Nairobi	3	0.7
Nyanza-to-Coast	3	0.7

Table 5. The number of expected (Markov) jumps inferred for HIV-1 A1 migration between geographic locations.

The direction of migration events (from, to)	Number of HIV-1 jumps (N, %)
Between provinces	**61 (100%)**
Coast-Nyanza	26 (42.6%)
Coast-Nairobi	23 (37.7%)
Nairobi-Nyanza	7 (11.5%)
Nyanza-Nairobi	3 (4.9%)
Nairobi-Coast	1 (1.6%)
Nyanza-Coast	1 (1.6%)

FIGURES

Figure 1. Map of Kenya showing the distribution of sequences in this study.
A map of Kenya showing the number of HIV-1 sequences from MSM analysed in this study, and distribution by different geographic regions. The map is coloured based on the estimated number of MSM as mapped at the county level during the 2018 key population size estimates national survey[73].

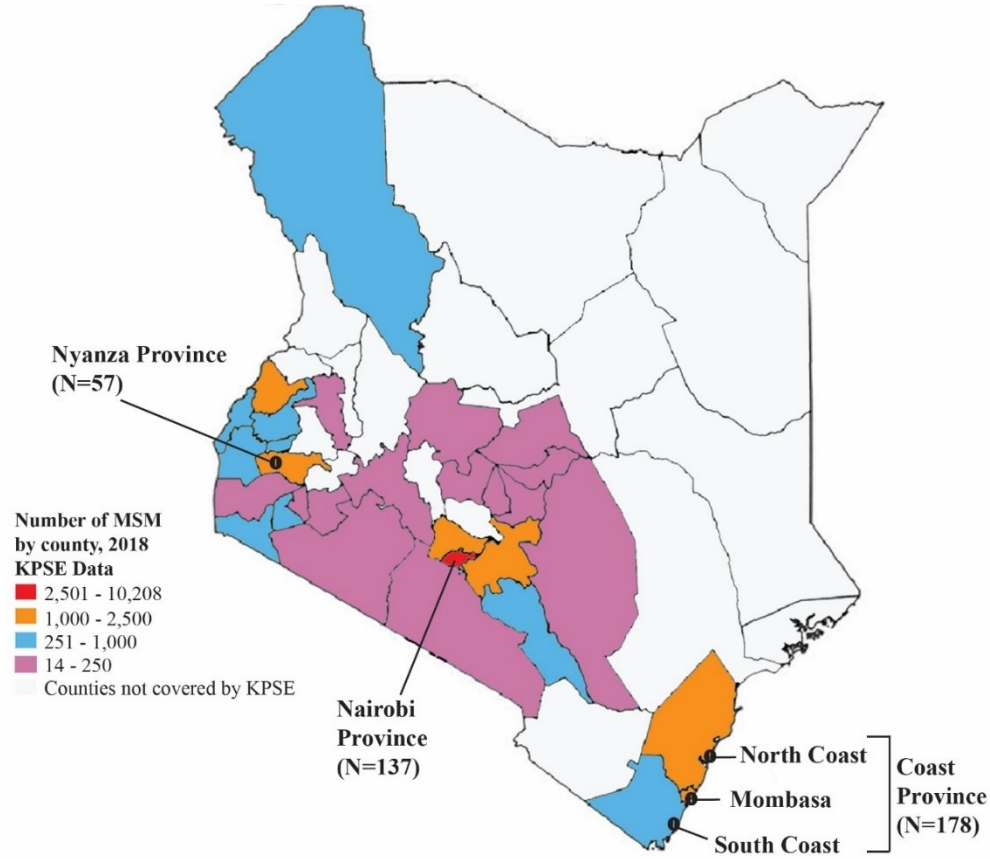


Figure 2. HIV-1 genotypes among 372 MSM sequences from Kenya.
Maximum-likelihood phylogenetic tree of 372 HIV-1 *pol* sequences from MSM living with HIV-1 in Kenya (and 194 HIV-1 Group M subtype reference sequences from the Los Alamos HIV database). Branch tips colours correspond to the respective HIV-1 subtype, sub-subtype or recombinant form as shown in the legend. Key nodes with aLRT-SH support ≥0.9 are highlighted with an asterisk. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.

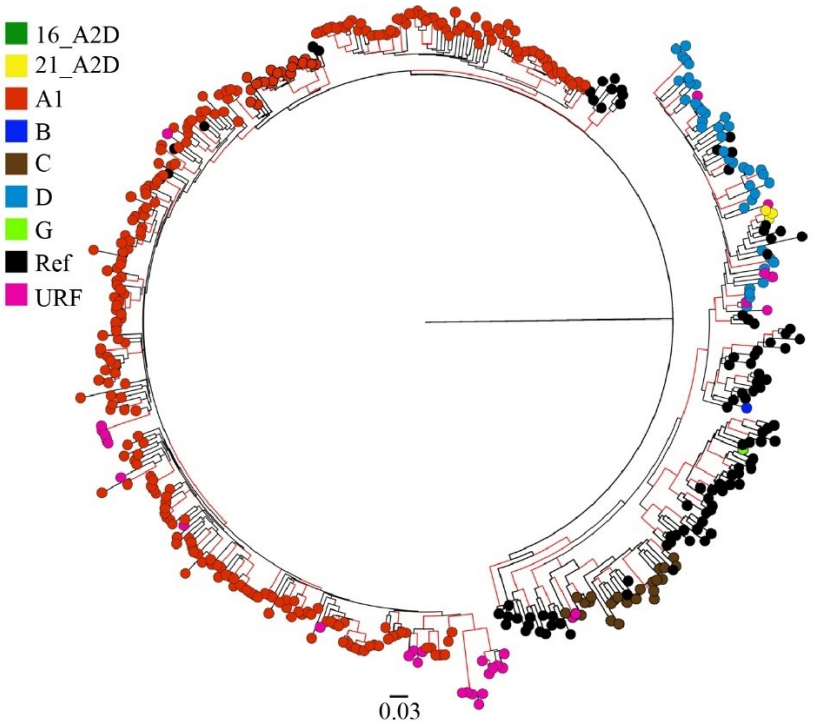


Figure 3. Population dynamics of HIV-1 sub-subtype A1, subtype D and subtype C lineages among MSM in Kenya.
Bayesian Skygrid plots showing population dynamics of the (a) HIV-1 sub-subtype A1, (b) HIV-1 subtype C and (c) HIV-1 subtype D lineages in Kenyan MSM. Median estimates of the number of MSM contributing to new infections are shown as a continuous line in each plot (coloured Red for sub-subtype A1, Brown for subtype C, and Blue for subtype D). The shaded area represents the 95% higher posterior density intervals of the inferred effective population size for each lineage.

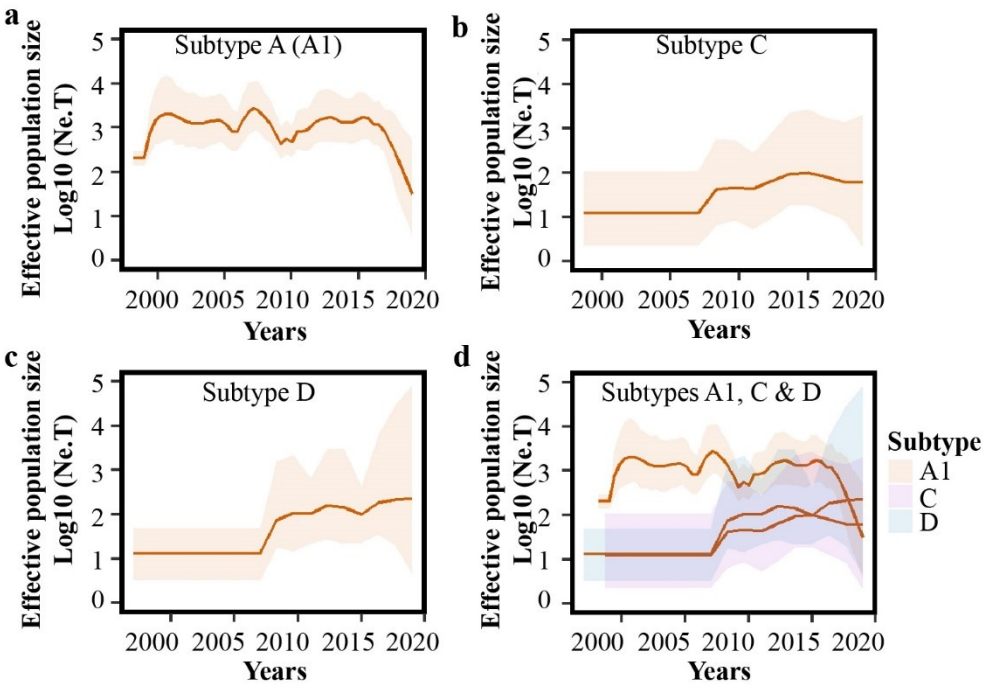


Figure 4. Characteristics and posterior distribution of time to most recent common ancestors estimated for all Kenya clusters.
Bayesian tMRCA estimates for (a) HIV-1 sub-subtype A1, (b) HIV-1 subtype C and (c) HIV-1 subtype D lineages in Kenyan MSM HIV-1 clusters. Dots represent the estimated tMRCA and are coloured as per the provinces represented by sequences in each cluster as shown in the legend. Black error bars represent sampling time (with lower interval representing the oldest sampling time per cluster and upper interval representing the most recent sampling time per cluster).

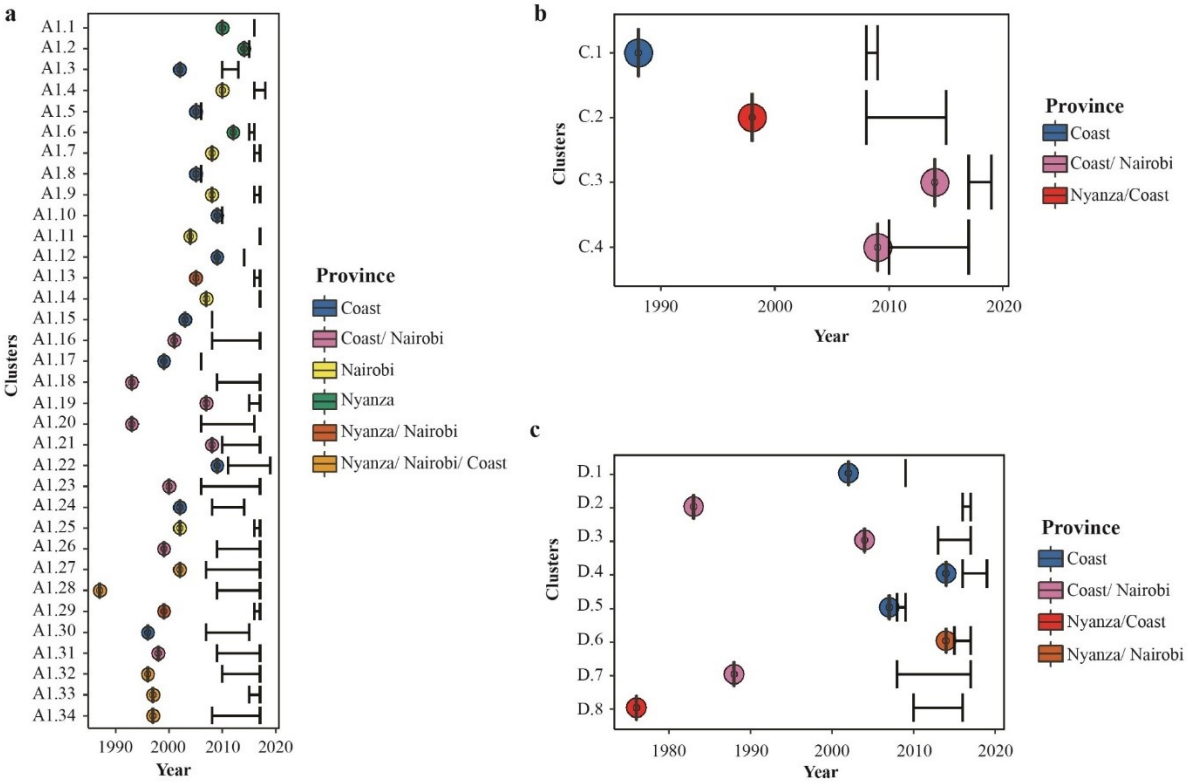
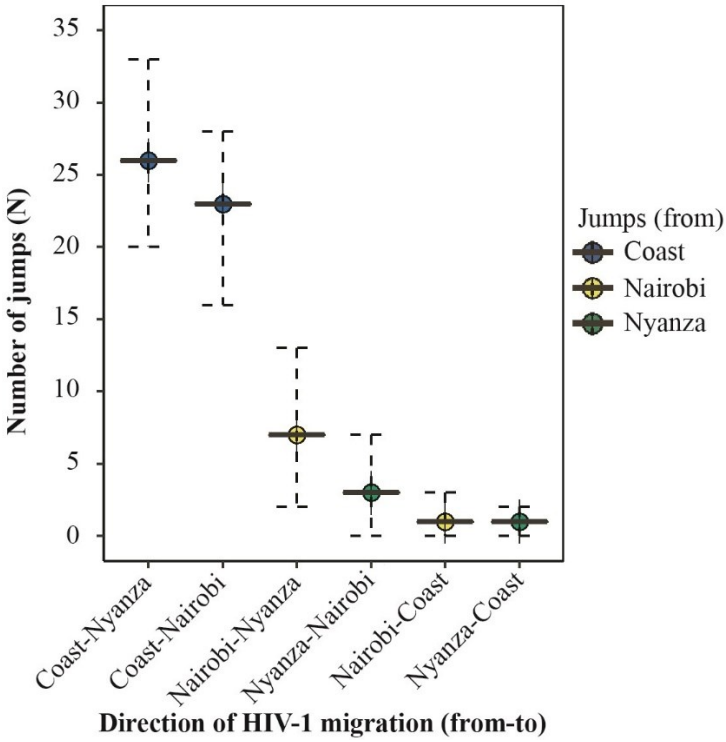


Figure 5. Summary of the expected number of HIV-1 migration between geographic regions in Kenya
Summary of the median number (and 95% HPD interval) of Markov jumps inferred with a uniform sampling of geographic regions. Plots represent (**a**) HIV-1 exchange between provinces. Plots are coloured by the "source" location as shown in the legend. Only statistically significant transitions (Bayes Factor (BF) ≥3) are plotted.



SUPPLEMENTARY DATA

Files in this Data Supplement:

Table S1. Distribution of HIV-1 subtypes by year (range) of sampling.
Table S2. Distribution of HIV-1 subtypes by year (range) and geographic province of sampling.
Table S3. Characteristics and posterior distribution of time to most recent common ancestors estimated for all Kenya clusters.
Table S4. Phylogeographic inference of HIV-1 migration rates (Bayes factor, BF≥3) between geographic locations in the full Kenyan dataset.

Legends for supplementary figures.

Figure S1. The frequency of HIV-1 subtypes per province and by year (range) of sampling.
Figure S2. The maximum-likelihood tree used to identify HIV-1 clusters.
Figure S3. The maximum clade credibility trees used to date clusters.
Figure S4. The maximum clade credibility tree summary of the Bayesian inference.
Figure S5. The proportion and dates of HIV-1 transitions between geographic provinces and risk groups.

TABLES

Table S1. Overall distribution of HIV-1 subtypes by year (range) of sampling.

HIV-1 Subtype	Years (Range)			
	2006-2010	**2011-2015**	**2016-2019**	**Total**
A1	84 (31.3%)	38 (14.2%)	146 (54.5%)	268 (72.0%)
D	13 (31.7%)	8 (19.5%)	20 (48.8%)	41 (11.0%)
URF	9 (27.3%)	2 (6.1%)	22 (66.7%)	33 (8.9%)
C	7 (31.8%)	4 (18.2%)	11 (50.0%)	22 (5.9%)
21_A2D	0 (0.0%)	0 (0.0%)	3 (100.0%)	3 (0.8%)
G	3 (100%)	0 (0.0%)	0 (0.0%)	3 (0.8%)
16_A2D	1 (100%)	0 (0.0%)	0 (0.0%)	1 (0.3%)
B	1 (100%)	0 (0.0%)	0 (0.0%)	1 (0.3%)
Total	118 (31.7%)	52 (14.0%)	202 (54.3%)	372 (100.0%)

Table S2. Distribution of HIV-1 subtypes by year (range) of sampling and geographic area of sampling.

Province	Subtype	Year range (N, %)			Total (N, %)
		2006-2010	**2011-2015**	**2016-2019**	
Coast	A1	84 (71.8%)	26 (81.3%)	11 (37.9%)	121 (100.0%)
	D	13 (11.1%)	4 (12.5%)	5 (17.2%)	22 (100.0%)
	URF	8 (6.8%)	1 (3.1%)	7 (24.1%)	16 (100.0%)
	C	7 (6%)	1 (3.1%)	6 (20.7%)	14 (100.0%)
	G	3 (2.6%)	0 (0.0%)	0 (0.0%)	3 (100.0%)
	16A2D	1 (0.9%)	0 (0.0%)	0 (0.0%)	1 (100.0%)
	B	1 (0.9%)	0 (0.0%)	0 (0.0%)	1 (100.0%)
Sub-total		117 (100.0%)	32 (100.0%)	29 (100.0%)	178 (100.0%)
Nairobi	A1	0 (0.0%)	1 (100%)	101 (74.8%)	102 (100.0%)
	URF	1 (100%)	0 (0.0%)	13 (9.6%)	14 (100.0%)
	D	0 (0.0%)	0 (0.0%)	13 (9.6%)	13 (100.0%)
	C	0 (0.0%)	0 (0.0%)	5 (3.7%)	5 (100.0%)
	21A2D	0 (0.0%)	0 (0.0%)	3 (2.2%)	3 (100.0%)
Sub-total		1 (100%)	1 (100%)	135 (100.0%)	137 (100.0%)
Nyanza	A1	0 (0.0%)	11 (57.9%)	34 (89.5%)	45 (100.0%)
	D	0 (0.0%)	4 (21.1%)	2 (5.3%)	6 (100.0%)
	C	0 (0.0%)	3 (15.8%)	0 (0.0%)	3 (100.0%)
	URF	0 (0.0%)	1 (5.3%)	2 (5.3%)	3 (100.0%)
Sub-total		0 (100.0%)	19 (100.0%)	38 (100.0%)	57 (100.0%)
Total		118 (31.7%)	52 (14.0%)	202 (54.3%)	372 (100.0%)

Table S3. Characteristics and posterior distribution of time to most recent common ancestors estimated for Kenyan HIV-1 clusters.

Cluster name[1]	Tips (N)[2]	Province	Year(s) of diagnosis[3]	tMRCA[4]
A1.28	9	Nyanza/Nairobi/Coast	2009-2017	1987
A1.18	4	Coast/Nairobi	2009-2017	1993
A1.20	4	Coast/Nairobi	2006-2016	1993
A1.30	11	Coast	2007-2015	1996
A1.32	16	Nyanza/Nairobi/Coast	2010-2017	1996
A1.33	19	Nyanza/Nairobi/Coast	2015-2017	1997
A1.34	20	Nyanza/Nairobi/Coast	2008-2017	1997
A1.31	13	Coast/Nairobi	2009-2017	1998
A1.17	4	Coast	2006	1999
A1.26	7	Coast/Nairobi	2009-2017	1999
A1.29	9	Nyanza/Nairobi	2016-2017	1999
A1.23	5	Coast/Nairobi	2006-2017	2000
A1.16	4	Coast/Nairobi	2008-2017	2001
A1.3	2	Coast	2010-2013	2002
A1.24	5	Coast	2008-2014	2002
A1.25	6	Nairobi	2016-2017	2002
A1.27	7	Nyanza/Nairobi/Coast	2007-2017	2002
A1.15	3	Coast	2008	2003
A1.11	2	Nairobi	2017	2004
A1.5	2	Coast	2006	2005
A1.8	2	Coast	2006	2005
A1.13	3	Nyanza/Nairobi	2016-2017	2005
A1.14	3	Nairobi	2017	2007
A1.19	4	Coast/Nairobi	2015-2017	2007
A1.7	2	Nairobi	2016-2017	2008
A1.9	2	Nairobi	2016-2017	2008
A1.21	4	Coast/Nairobi	2010-2017	2008
A1.10	2	Coast	2010	2009
A1.12	2	Coast	2014	2009
A1.22	5	Coast	2011-2019	2009
A1.1	2	Nyanza	2016	2010
A1.4	2	Nairobi	2016-2018	2010
A1.6	2	Nyanza	2015-2016	2012
A1.2	2	Nyanza	2015	2014
C.1	2	Coast	2008-2009	1988
C.2	2	Nyanza/Coast	2008-2015	1998
C.4	5	Coast/Nairobi	2010-2017	2009
C.3	3	Coast/Nairobi	2017-2019	2014
D.8	6	Nyanza/Coast	2010-2016	1976
D.2	2	Coast/Nairobi	2016-2017	1983
D.7	6	Coast/Nairobi	2008-2017	1988
D.1	2	Coast	2009	2002
D.3	2	Coast/Nairobi	2013-2017	2004
D.5	3	Coast	2008-2009	2007
D.4	2	Coast	2016-2019	2014
D.6	5	Nyanza/Nairobi	2015-2017	2014

[1]Clusters are named according to subtype/CRF, and risk group dominating the cluster.
[2]Number of sequences per cluster.
[3]The respective earliest and most recent date (year) of sampling of sequences in the cluster
[4]Estimated tMRCA: Median time to the most recent common ancestor of the cluster.

Table S4. Phylogeographic inference of HIV-1 migration rates (Bayes factor, BF≥3) between geographic locations in the full Kenyan dataset.

Bayes factor (BF) support and posterior probability inferred for HIV-1 transmission between geographic locations in the full Kenyan sub-subtype A1, subtype C and subtype D datasets. Only significant transitions (BF≥3) are shown.

	The direction of migration events (from-to)	**Bayes Factor (BF)**	**Posterior Probability**
Migration between provinces			
HIV-1 A1	Coast-to-Nairobi	3716	1
	Nairobi-to-Nyanza	3716	1
	Nyanza-to-Coast	4	0.8
HIV-1 C	Coast-to-Nairobi	268	1
	Coast-to-Nyanza	8	0.9
	Nyanza-to-Coast	3	0.7
	Nairobi-to-Coast	3	0.7
HIV-1 D	Nairobi-to-Nyanza	43	1
	Coast-to-Nairobi	16	0.9
	Nyanza-to-Coast	4	0.8

Table S5. Temporal proportion in transitions within and between geographic provinces in the HIV-1 A1 dataset.

Jumps (from – to)*	**Year (Range)**			**Total**
	1990-2000	**2001-2010**	**2011-2020**	
Within-provinces				
Coast-Coast	22 (91.7%)	117 (69.2%)	10 (9.8%)	149 (50.5%)
Nairobi-Nairobi	0 (0%)	14 (8.3%)	56 (54.9%)	70 (23.7%)
Nyanza-Nyanza	0 (0%)	4 (2%)	19 (19%)	23 (8%)
Between provinces				
Coast-Nairobi	1 (4.2%)	24 (14.2%)	5 (4.9%)	30 (10.2%)
Coast-Nyanza	1 (4.2%)	6 (3.6%)	0 (0%)	7 (2.4%)
Nairobi-Nyanza	0 (0%)	4 (2.4%)	11 (10.8%)	15 (5.1%)
Nyanza-Nairobi	0 (0%)	0 (0%)	1 (1%)	1 (0.3%)
Total	24 (100%)	169 (100%)	102 (100%)	295 (100%)

*Transitions between geographic provinces were summarised from the HIV-A1 trait-annotated maximum clade credibility tree which had denser sampling (number of sequences) and temporal coverage compared to other subtypes.

FIGURES

Figure S1. The frequency of HIV-1 subtypes per province and by year (range) of sampling.
The proportion of HIV-1 subtypes per province distributed into three time periods (i.e 2006-2010, 2011-2015, and 2016-2019).

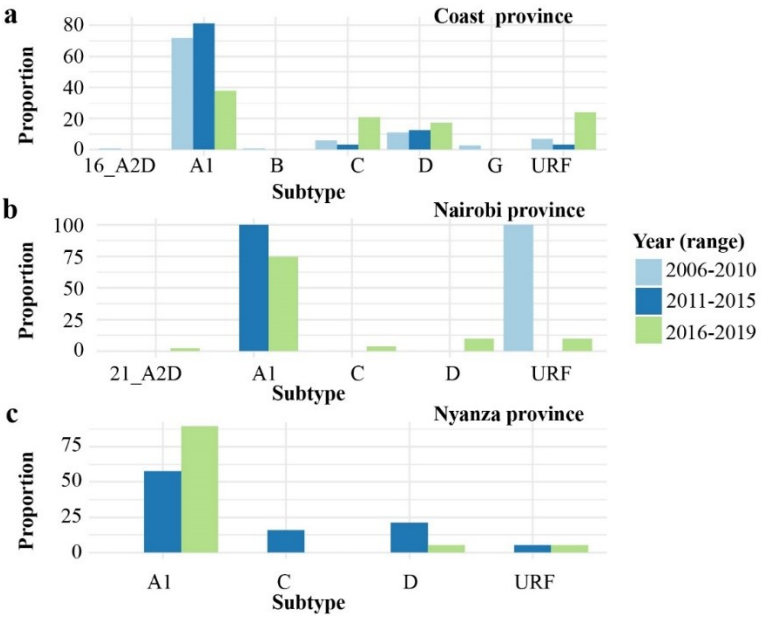


Figure S2. Maximum-likelihood trees used to identify HIV-1 clusters.
Maximum-likelihood trees used for the identification of MSM HIV-1 clusters. Trees represent A: Sub-subtype A1; B: Subtype C; and C: Subtype D HIV-1 clusters, respectively. Each phylogeny is rooted at the midpoint. Monophyletic clusters with aLRT-SH support ≥0.9 and which have ≥80% sequences from coastal Kenya are highlighted in grey. To enhance cluster visualization, some branches containing either reference sequences or Kenyan sequences that are not part of clusters have been collapsed (shown as black or red triangles, with the recent end of the triangle indicating the latest sampling date. Branch tips within respective clusters are coloured as per geographic province cluster (Orange: Coast; Green: Nairobi; Sky blue: Nyanza; and Black: Reference sequences). Scale bars represent a genetic distance of 0.01 in all phylogenies.

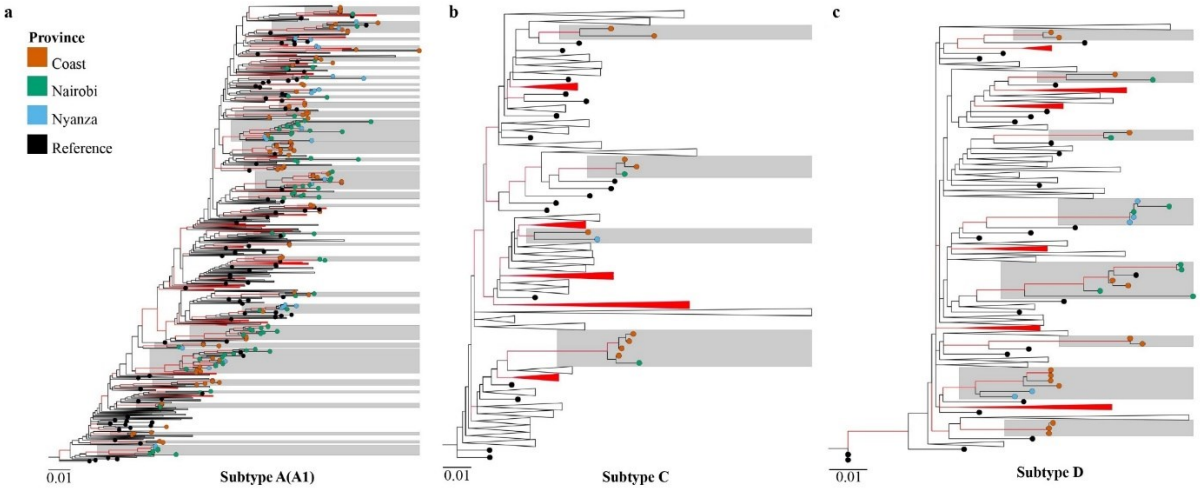


Figure S3. Maximum clade credibility trees used to date clusters.
Maximum clade credibility (MCC) trees used to determine the time to the most recent common ancestor of the Kenyan HIV-1 clusters. Trees represent A: Sub-subtype A1; B: Subtype C; and C: Subtype D, respectively. To enhance cluster visualization, some branches containing either reference sequences or coastal Kenya sequences that are not part of clusters have been collapsed (shown as non-coloured, black, or red triangles, with the recent end of the triangle indicating the latest sampling date. Branch tips are colour-coded as per geographic province cluster (Orange: Coast; Green: Nairobi; and Sky blue: Nyanza).

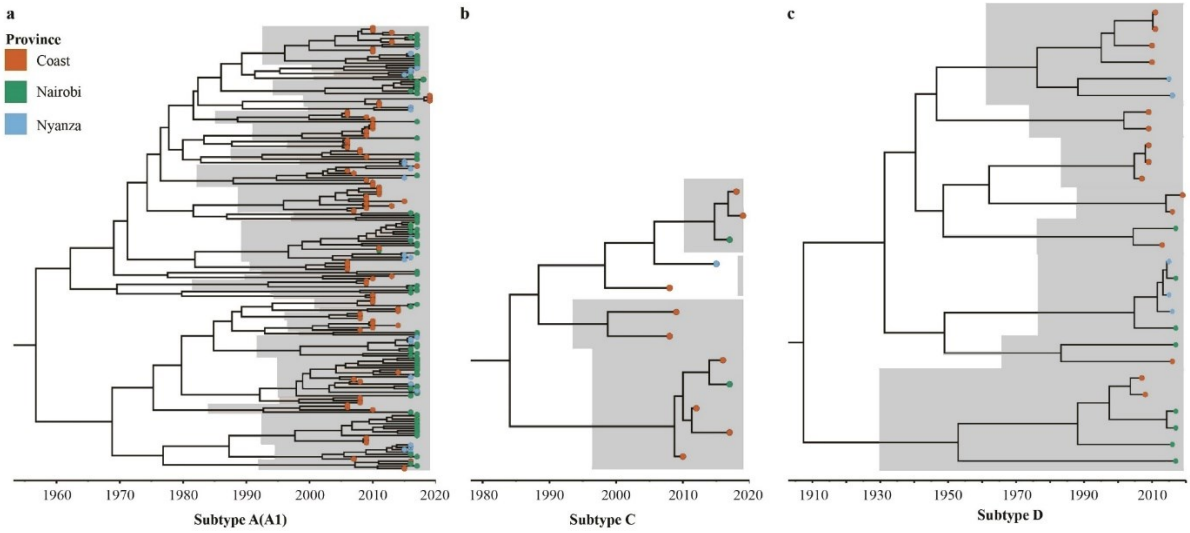


Figure S4. The maximum clade credibility tree summary of the Bayesian inference.
Maximum clade credibility trees revealing phylogeographical estimates of HIV-1 spread in three Kenyan provinces. Trees represent A: Sub-subtype A1; B: Subtype C; and C: Subtype D, respectively. Branch colours correspond to the province of origin as shown in the legend: Orange: Coast; Green: Nairobi; Sky blue: Nyanza.

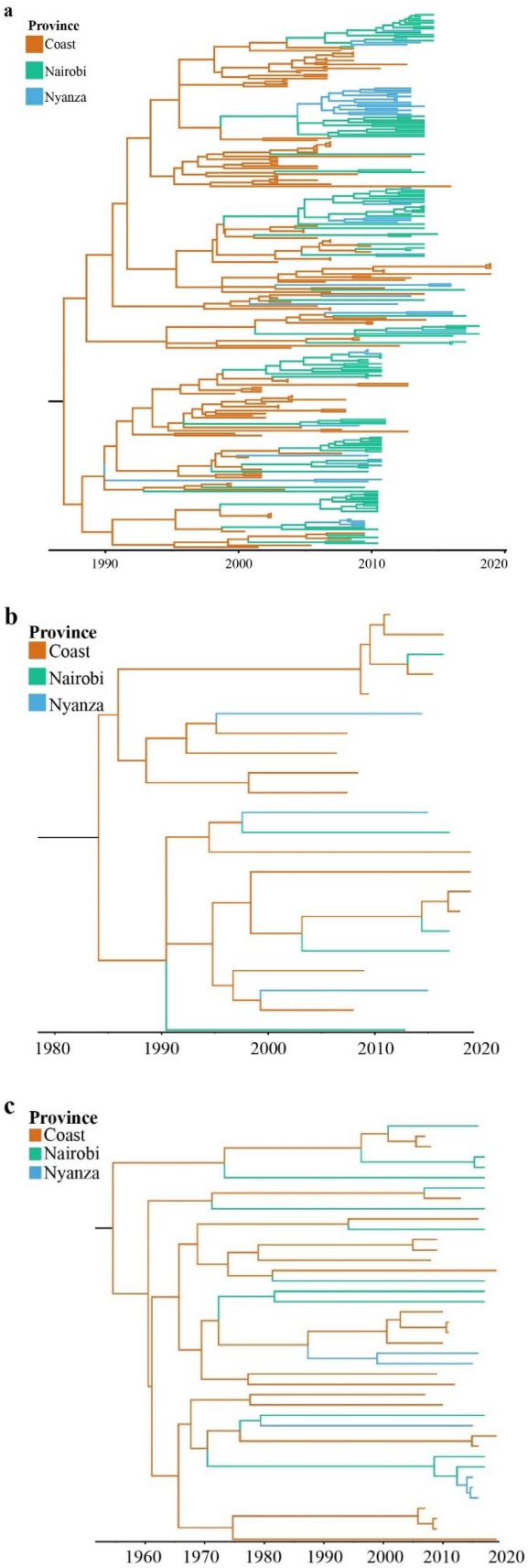
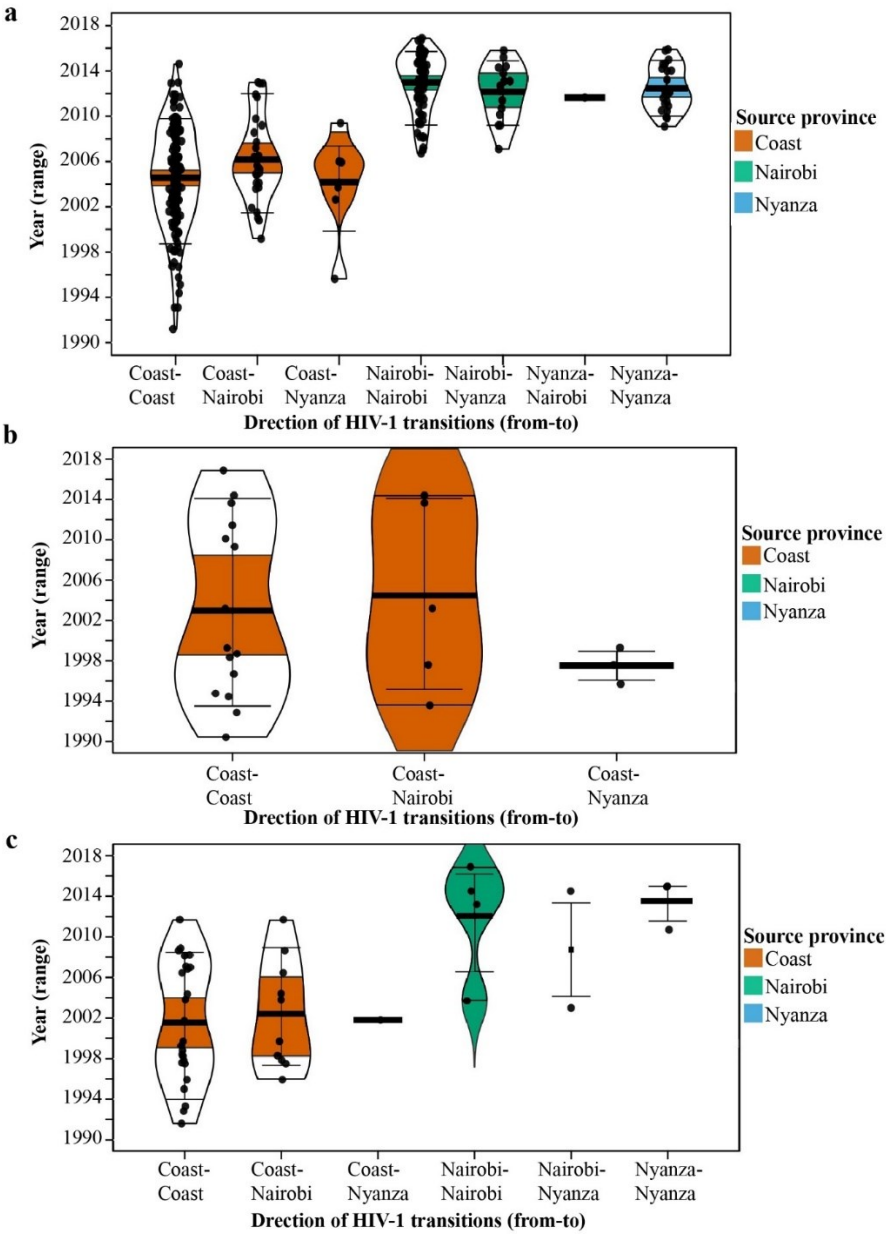


Figure S5. The estimated proportion and dates of HIV-1 transitions between geographic provinces and risk groups.
Pirate plots summarising the dates (year) and the frequency of HIV-1 transitions between geographic provinces summarised from trait-annotated maximum clade credibility trees. Plots represent (a) sub-Subtype A1, (b) subtype C, and (c) subtype-D transitions – where group median and interquartile range are coloured by the source of transition (Orange; transitions from Coast, Green; transitions from Nairobi, and Sky-Blue; transitions from Nyanza). Only transitions with a posterior probability higher than 0.90 are plotted. Dots in the pirate plots represent HIV-1 migration events.



SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN

HIV-1 Transmission Patterns Within and Between Risk Groups in Coastal Kenya

George M. Nduva[1,2], Amin S. Hassan[1,2], Jamirah Nazziwa[1], Susan M. Graham[2,3], Joakim Esbjörnsson[1,4,5] ✉ & Eduard J. Sanders[2,4,5]

HIV-1 transmission patterns within and between populations at different risk of HIV-1 acquisition in Kenya are not well understood. We investigated HIV-1 transmission networks in men who have sex with men (MSM), injecting drug users (IDU), female sex workers (FSW) and heterosexuals (HET) in coastal Kenya. We used maximum-likelihood and Bayesian phylogenetics to analyse new (N = 163) and previously published (N = 495) HIV-1 polymerase sequences collected during 2005–2019. Of the 658 sequences, 131 (20%) were from MSM, 58 (9%) IDU, 109 (17%) FSW, and 360 (55%) HET. Overall, 206 (31%) sequences formed 61 clusters. Most clusters (85%) consisted of sequences from the same risk group, suggesting frequent within-group transmission. The remaining clusters were mixed between HET/MSM (7%), HET/FSW (5%), and MSM/FSW (3%) sequences. One large IDU-exclusive cluster was found, indicating an independent sub-epidemic among this group. Phylodynamic analysis of this cluster revealed a steady increase in HIV-1 infections among IDU since the estimated origin of the cluster in 1987. Our results suggest mixing between high-risk groups and heterosexual populations and could be relevant for the development of targeted HIV-1 prevention programmes in coastal Kenya.

Approximately 5.6% in the population in Kenya are infected by HIV-1, with a more than three-fold higher HIV-1 prevalence among so-called high-risk groups – including men who have sex with men (MSM), injecting drug users (IDU) and female sex workers[1–4]. Modelling data on modes of HIV-1 transmission in Kenya have shown that at least one-third of all new infections occur among high-risk groups[5]. However, little is known about local HIV-1 networks and transmission within and between high-risk groups and the heterosexual (HET) population in African settings[6]. Molecular epidemiology studies in coastal Kenya have described a dynamic HIV-1 epidemic characterised by subtypes A, C, D, and different circulating recombinant forms (CRFs)[6–15]. These studies have indicated high proportions of recombinants within HET, but this was not evident among MSM in a recent study[8], alluding to separate epidemics. One study in coastal Kenya observed similar HIV-1 recombination patterns among HIV-1 strains in MSM and FSW, suggesting reinfections within mixed networks[13].

HIV-1 transmission dynamics can be assessed by linking socio-demographic, clinical and behavioural data with HIV-1 sequence data by phylogenetics[16,17]. With few exceptions, most phylogenetic studies of the HIV-1 epidemic in sub-Saharan Africa have focused on understanding HIV-1 subtype diversity and prevalence of antiretroviral resistance mutations[18–24]. Phylogenetic studies highlighting the dynamics of HIV-1 transmission and contribution of high-risk groups to onward viral transmission are common in North America and Europe, where largescale HIV-1 sequence data are available[25–32]. Due to the limited availability of HIV-1 sequences from sub-Saharan Africa, only a few phylogenetic studies have assessed the dynamics of the HIV-1 epidemic in the region[33–35]. Transmission networks studies in Kenya have demonstrated clustering of MSM sequences with evidence of transmission between different geographical regions, and limited mixing between MSM and HET[6,8,13]. We have also demonstrated extensive clustering of HIV-1 *pol* sequences from MSM who have sex with only men and MSM who have sex with both men and women in coastal Kenya[8]. Given that many MSM on the coast of Kenya have sex with both men and women, there is a possibility of HIV-1 transmission linkages between MSM and the local HET community[4]. The primary objective of the current study was to investigate transmission networks within and between MSM, IDU, FSW, and HET in coastal Kenya using both newly generated

[1]Lund University, Lund, Sweden. [2]KEMRI/Wellcome Trust Research Programme, Kilifi, Kenya. [3]University of Washington, Seattle, WA, USA. [4]The University of Oxford, Oxford, United Kingdom. [5]These authors contributed equally: Joakim Esbjörnsson and Eduard J. Sanders. ✉e-mail: Joakim.esbjornsson@med.lu.se

Risk group		MSM (N = 131, 20%)	IDU (N = 58, 9%)	FSW (N = 109, 17%)	HET (N = 360, 55%)	Total (N = 658, 100%)
Sequences	New	9 (7%)	0 (0%)	102 (94%)	52 (14%)	163 (25%)
	Published	122 (93%)	58 (100%)	7 (6%)	308 (86%)	495 (75%)
Subtype	A	92 (70%)	51 (88%)	71 (65%)	217 (60%)	431 (66%)
	C	9 (7%)	2 (3%)	9 (8%)	26 (7%)	46 (7%)
	D	13 (10%)	5 (9%)	12 (11%)	39 (11%)	69 (10%)
	Others*	17 (13%)	0 (0%)	17 (16%)	78 (22%)	112 (17%)
Year (range)	2005–2007	27 (21%)	0 (0%)	54 (50%)	24 (7%)	105 (16%)
	2008–2010	60 (46%)	58 (100%)	41 (38%)	302 (84%)	461 (70%)
	2011–2013	18 (14%)	0 (0%)	0 (0%)	0 (0%)	18 (3%)
	2014–2016	14 (11%)	0 (0%)	10 (9%)	32 (9%)	56 (8%)
	2017–2019	12 (9%)	0 (0%)	4 (4%)	2 (1%)	18 (3%)
Area	Mombasa county	74 (57%)	58 (100%)	7 (6%)	71 (20%)	210 (32%)
	Kilifi county	57 (44%)	0 (0%)	102 (94%)	289 (80%)	448 (68%)

Table 1. Demographics and distribution of newly generated and published coastal Kenya HIV-1 *pol* sequences by risk-group. Abbreviations: MSM, men who have sex with men; IDU, injecting drug user; FSW, female sex worker; HET, at-risk men and women who did not report sex work or male same-sex behaviour. *Subtype/recombinant (N, %): B (1, 0.2%), G (2, 0.5%), A1D (42, 6.4%), A1C (18, 2.7%), A2D (13, 2.0%), 16_A2D (10, 1.5%), A2_16A2D (6, 0.9%), CD (6, 0.9%), A1A2 (3, 0.5%), A1A2D (3, 0.5%), A1_16A2D (1, 0.2%), A1A2_16A2D (1, 0.2%), A1A2C (1, 0.2%), A1A2CD (1, 0.2%), CA1D (1, 0.2%), DG (1, 0.2%).

and previously published HIV-1 *pol* sequences. A secondary objective was to determine HIV-1 genetic diversity among different risk groups in coastal Kenya.

Results

Study Population, sequence dataset and sampling density. The analysed 658 coastal Kenyan HIV-1 partial *pol* sequences included both newly generated (N = 163) and previously published sequences (N = 495). Sequences were collected during 2005–2019 in the Mombasa (N = 210, 32%) and Kilifi (N = 448, 68%) counties in coastal Kenya (Table 1). The risk groups included MSM (N = 131, 20%), IDU (N = 58, 9%), FSW (N = 109, 17%), and HET (N = 360, 55%). Based on size estimation of risk groups and the number of infected populations infected with HIV-1 in Mombasa and Kilifi counties[36], our study was more powered to pick out MSM (sampling density, 51%) and IDU (sampling density, 12%) clusters compared with FSW (sampling density 3%) and HET (sampling density, 1%) clusters (Supplementary Table S1).

HIV-1 subtypes A, C, and D dominated the epidemic in coastal Kenya. Phylogenetic analysis was used to determine the subtype distribution in the full coastal Kenya sequence dataset (N = 658). In total, 431 subtype A (66%), 46 subtype C (7%), 69 subtype D (10%), 2 subtype G (>1%), and 110 CRF and unique recombinant form (URFs, 17%) sequences were identified (Table 1 and Supplementary Fig. S1). In addition, all subtype A sequences belonged to sub-subtype A1. Detailed recombination analyses of newly generated sequences demonstrated extensive recombination between subtypes, sub-subtypes, and recombinant forms (Supplementary Table S2).

Identification of coastal Kenya-specific HIV-1 transmission clusters. Maximum-likelihood (ML) phylogenies were reconstructed independently for the most prevalent HIV-1 subtypes in the population (subtypes A, C, and D). Reference sequences were obtained from GenBank based on similarity (whereof 731 participant-unique sub-subtype A1 sequences remained after removal of redundancies; 256 for subtype C; and 92 for subtype D).

Transmission networks were classified based on the number of sequences per cluster into dyads (2 sequences), networks (3–14 sequences), and large clusters (>14). Of the 658 coastal Kenyan sequences, 206 sequences (31%) formed 61 statistically supported clusters (size range: 2–41 sequences). These included 39 dyads (64% of all clusters), 21 networks (34%), and one large cluster (2%) (Table 2 and Supplementary Table S3). Most of the clusters were found among the subtype A sequences (N = 50, 82%), followed by the subtype D sequences (N = 7, 11%), and the subtype C sequences (N = 4, 7%) (Supplementary Fig. S2).

Risk-group specific clustering patterns. Stratification by risk group showed two distinct clustering patterns (Fig. 1). The first pattern represented exclusive within-risk group clustering, where sequences in a cluster belonged exclusively to one specific risk group. Compared to HET sequences, MSM and IDU sequences were more likely to cluster (adjusted odds ratio [aOR] 25.9, 95% confidence interval [CI] 10–63.9, $P < 0.001$; and aOR 31.5, CI 12.2–81.6, $P < 0.001$, respectively, Table 3). Of the 61 clusters observed, 85% were risk group exclusive clusters. These included 24 MSM clusters (11 dyads and 13 networks), four IDU clusters (three dyads and one large cluster), six FSW clusters (six dyads), and 18 HET clusters (13 dyads and five networks). The majority of the MSM sequences (N = 84, 64%) formed small independent clusters ranging in size from two to nine sequences per cluster. Likewise, the majority of IDU sequences (N = 47, 82%) formed clusters. Interestingly, most of the

Risk group	Dyads (2 sequences, N = 39, 64%)	Networks (3–14, N = 21, 34%)	Large clusters (≥14, N = 1, 2%)	Total clusters (N = 61, 100%)
MSM	11	13	0	24 (39%)
IDU	3	0	1	4 (7%)
FSW	6	0	0	6 (10%)
HET	13	5	0	18 (30%)
HET/FSW	2	1	0	3 (5%)
MSM/HET	2	2	0	4 (7%)
MSM/FSW	2	0	0	2 (3%)

Table 2. The number of coastal Kenyan transmission clusters by cluster size and risk group. Abbreviations: MSM, men who have sex with men; IDU, injecting drug user; FSW, female sex worker; HET, at-risk men and women who did not report sex work or male same-sex behaviour.

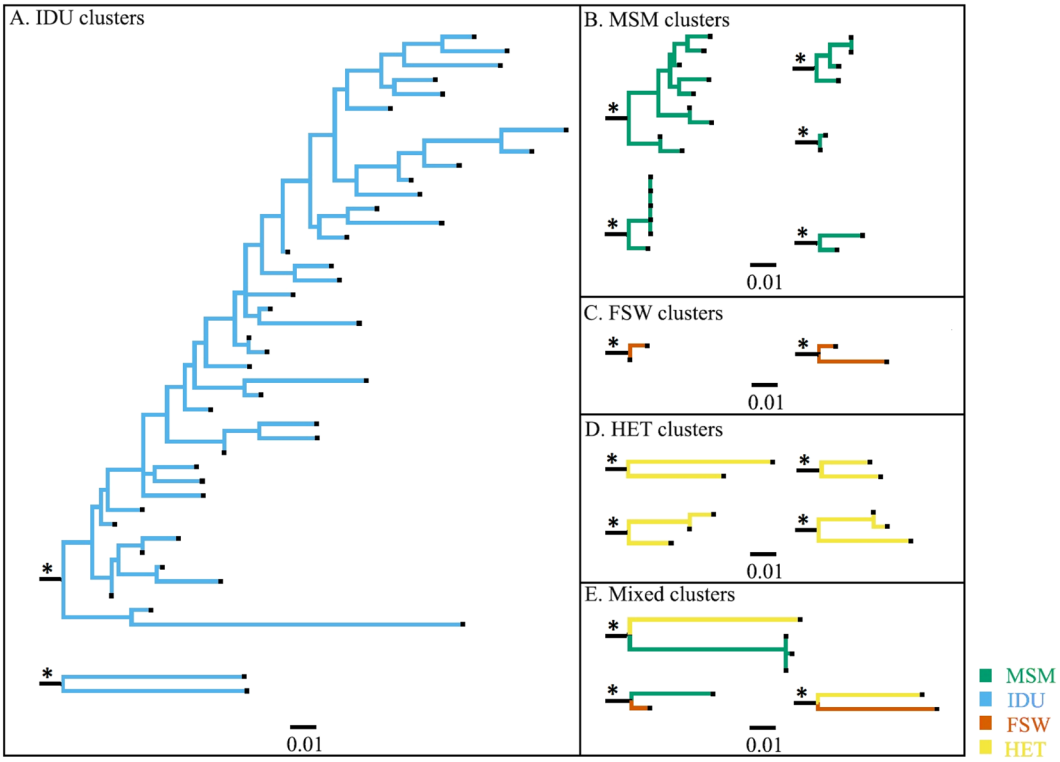


Figure 1. Clustering patterns of different risk groups in coastal Kenya. Representative clusters selected to highlight typical clustering patterns of the different risk groups. The branches are coloured according to the different risk groups (Bluish- green: MSM; Sky blue: IDU; Vermillion: FSW; Yellow: HET, and Black: reference sequences). As an overview, MSM formed several small clusters ranging in size from two to nine sequences per cluster (**A**). Most IDU sequences (N = 41) formed one single large cluster (**B**). In contrast, FSW and HET clusters were small (mostly dyads containing two sequences), although most FSW and HET sequences existed as single sequences or clustered with reference sequences (**C**). Asterisks have been used to highlight branches leading to significantly supported clusters (aLRT-SH branch support of ≥0.9).

clustering IDU sequences were of sub-subtype A1 and formed one single large cluster (N = 41, 80%). In contrast, only a small proportion of the FSW (N = 22, 20%) and HET (N = 67, 18%) sequences formed risk group-exclusive clusters.

In addition to risk group-exclusive clustering, 15% of all clusters were mixed between risk-groups (Table 2): Two (3%) MSM/FSW dyads, four (7%) MSM/HET mixed clusters (two dyads and two networks), and three (5%) FSW/HET clusters (two dyads and one network). Of relevance, both MSM/HET networks (one sub-subtype A1 and one subtype D) had sequences from four MSM and one HET female. Moreover, of the eleven MSM sequences that were found in mixed clusters, eight (73%) reported sex work in the three months preceding sample collection, and 10 (91%) reported bisexual behaviour. The FSW/HET network consisted of two FSW sequences and one HET male sequence.

Characteristics		Bivariable Analysis*		Multivariable Analysis**	
		OR (95% CI)	P-value	aOR (95% CI)	P-value
Risk group	HET	Reference		Reference	
	MSM	13.8 (8.6–22.2)	<0.001	25.8 (10–63.9)	<0.001
	IDU	29.1 (13.8–61.1)	<0.001	31.5 (12.2–81.6)	<0.001
	FSW	1.1 (0.6–2)	0.71	1.2 (0.5–2.9)	0.656
Subtype	A1	Reference			
	C	0.4 (0.2–0.8)	0.01	0.4 (0. 1–0.9)	0.025
	D	0.5 (0.3–0.8)	0.008	0.4 (0.2–0.8)	0.014
Year (range)	2005–2007	Reference			
	2008–2010	0.9 (0.6–1.4)	0.55		
	2011–2019	1.1 (0.6–1.9)	0.83		
Area (county)	Kilifi	Reference			
	Mombasa	3.9 (2.7–5.5)	<0.001	0.9 (0.5–1.6)	0.675
Sequence category	Published	Reference			
	New	0.3 (0.2–0.5)	<0.001	0.8 (0.4–1.8)	0.555

Table 3. Factors associated with clustering among 546 subtypes A1, C, and D HIV-1 sequences from MSM, IDU, FSW, and HET individuals from coastal Kenya. Abbreviations: MSM, men who have sex with men; IDU, injecting drug user; FSW, female sex worker; HET, at-risk men and women who did not report sex work or male same-sex behaviour. *Variables at a P-value of <0.1 in the bivariable analysis were included in the multivariable model. Circulating and unique recombinant forms were excluded from the multivariable analysis.

Most sequences from coastal Kenya clustered exclusively with sequences of Kenyan origin. Only three clusters with sequences from coastal Kenya had links to published sequences from outside coastal Kenya. One sequence in an MSM cluster was from an MSM from Nairobi. One sequence in another MSM cluster was from a German MSM, and the last sequence was found in a mixed MSM/HET cluster and was from a Canadian individual of unknown gender (Supplementary Table S3).

Genetic diversity between clusters of different risk groups. To further dissect differences in clustering patterns between risk groups, we determined the average genetic diversity in the identified clusters. A previously described ML bootstrap approach was employed to determine the genetic diversity (based on 1000 ML bootstrap trees)[37]. The median genetic diversity was 0.009 substitutions/site (s/s, IQR: 0.005–0.017 s/s) for MSM clusters, 0.03 s/s (IQR: 0.02–0.055 s/s) for IDU clusters, 0.008 s/s (IQR: 1×10^{-8}–0.018 s/s) for FSW clusters, 0.015 s/s (IQR: 0.006–0.023 s/s) for HET clusters, and 0.018 s/s (IQR: 0.013–0.024 s/s) for mixed clusters. A Kruskal-Wallis H test showed that the distribution of genetic diversity differed across the five groups ($\chi^2 = 11.074$, four degrees of freedom, P-value = 0.026). A Dunn's post hoc test for multiple comparisons using rank sums showed a significant difference in diversity between FSW and IDU (mean rank difference = 33.08, adjusted P-value = 0.039, Fig. 2, Supplementary Table S4).

Analysis of active transmission clusters. Among the 61 clusters defined by risk group, we identified 34 potentially active clusters as (determined by low genetic distance <1.5%), suggesting ongoing transmission at the time of sample collection. Stratification of the potentially active clusters by risk-group showed: Eight MSM dyads and five MSM networks; four IDU dyads; five FSW dyads; and seven HET dyads and two HET networks. Potentially active clusters with sequences from different risk groups included one FSW/HET network and two MSM/HET networks (Supplementary Table S5).

Estimation of time to the most recent common ancestor (tMRCA) and evolutionary rates. To gain insight into the evolutionary dynamics of the identified transmission clusters, we determined the evolutionary rate and the tMRCA of the clusters by Bayesian phylogenetic analysis. The median tMRCA of the 61 coastal Kenya clusters indicated that HIV-1 has been introduced in coastal Kenya several times over a period of 27 years (1985–2012, Supplementary Fig. S3 and Table S3). Only one cluster was large enough to allow for in-depth phylodynamic analysis. This cluster comprised 41 IDU sequences and the tMRCA for this cluster was determined to be 1987 (95% higher posterior density [HPD] interval: 1985–1990) (Fig. 3) with a median evolutionary rate of 6.4×10^{-3} substitutions/site/year (HPD interval: $3.9 \times 10^{-3} - 1.1 \times 10^{-2}$). The Skygrid analysis indicated that the number of IDU contributing to new HIV-1 infections over time increased gradually from 1987 to 2010.

Discussion

In this study, we found several HIV-1 links between MSM/HET, HET/FSW, and MSM/FSW, indicating mixing between these risk groups in coastal Kenya. Sequences from HET females in clusters dominated by MSM sequences provided evidence for heterosexual linkages in these clusters. The majority of the MSM in mixed clusters also reported having female sexual partners, indicating that this group, in addition to female sex workers, could be an important transmission link to the HET epidemic[4].

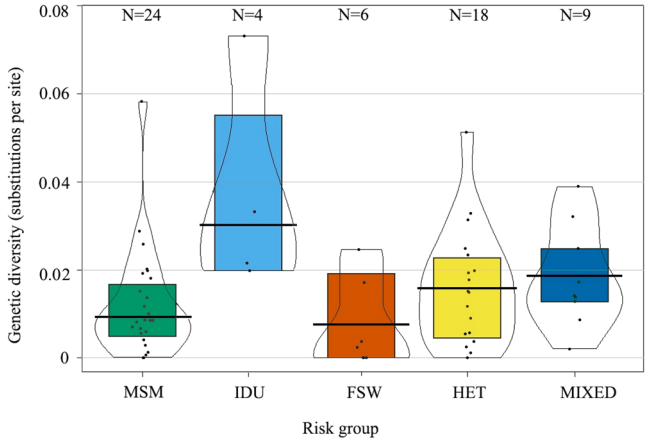


Figure 2. Genetic diversity of different risk group-specific clusters in coastal Kenya. A pirate plot[63] illustrating the differences in genetic diversity between MSM, IDU, FSW, HET and Mixed clusters. Black dots represent the median estimates of the genetic diversity per cluster. The group median and the interquartile range diversity estimates are indicated in box plots coloured by risk group (Bluish-green: MSM; Sky blue: IDU; Vermillion: FSW; Yellow: HET; Deep blue: Mixed risk groups).

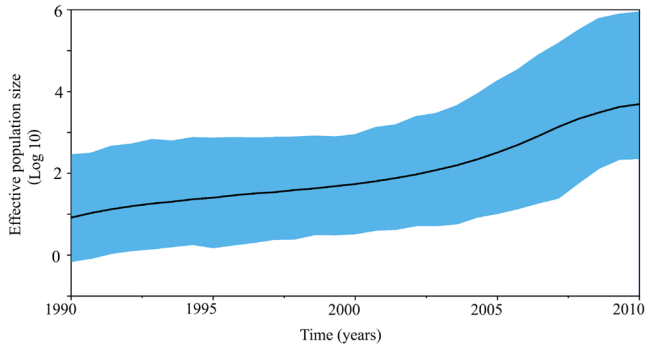


Figure 3. Population dynamics of the HIV-1 sub-epidemic among injecting drug users in coastal Kenya. A Bayesian Skygrid plot showing population dynamics of the HIV-1 sub-subtype A1 injecting drug users' sub-epidemic in coastal Kenya. Since the IDU *pol* sequence alignment did not contain temporal information (all sequences were sampled in 2010), the node height for this cluster was calibrated using information from the tMRCA posterior distribution obtained from dating the origin of subtype A1 Kenyan clusters[64]. Median estimates of the number of injecting drug users contributing to new infections are shown as a continuous black line. The shaded area represents the 95% higher posterior density intervals of the inferred effective population size.

Transmission linkages between MSM and HET in coastal Kenya might be expected to some extent, given that sexual interaction between MSM and other risk groups in the community is common[2–4,6,13]. In the only previous study of phylogenetic HIV-1 transmission linkages between MSM and HET in coastal Kenya, Bezemer and colleagues only found one single transmission pair of an MSM and a known female partner. Hence, extensive mixing between the MSM and HET epidemics was not found in that study[6]. In comparison, our analysis included a higher sampling density and availability of risk-group annotated sequences obtained in more recent years. This likely explains the significantly higher number of mixed clusters between MSM and HET sequences in the current study. In a broader context, our study complements existing research on the role of mixed networks in HIV-1 transmission – both globally and in sub-Saharan Africa[26,30,31,38–40].

Although we found several instances of mixed clusters, the majority of the coastal Kenya clusters represented within-risk-group HIV-1 transmission. MSM-exclusive and IDU-exclusive clusters were more common than FSW and HET clusters. High rates of clustering among MSM and IDU have been described before, both in our setting and elsewhere, and have been linked to an elevated risk of infection among MSM and IDU within close networks[6,8,17,29,30,41]. Whereas the MSM sequences were found in several smaller clusters, the vast majority of the IDU sequences formed one large cluster. This suggests that the majority of the HIV-1 IDU epidemic in Coastal Kenya was introduced from one single source followed by a long-term gradual spread within the IDU population – a pattern that distinguishes IDU transmission from that of other risk groups in coastal Kenya. In contrast to previous studies where IDU sequences clustered with very low genetic diversity[29,30,42], IDU clusters in our analysis had the highest genetic diversity compared to the other analysed risk groups. This indicates that the elapsed time

between HIV-1 infection and sample collection may have been longer among IDUs, and/or that the proportion of collected IDU sequences in relation to the true number of IDUs was lower compared to other risk groups. The underlying reasons for this are unknown and warrant further investigation. However, with or without missing links, such clustering pattern is indicative of long-standing HIV-1 transmission linked with intravenous drug use in coastal Kenya[17].

This is the first study in Africa to investigate the evolutionary dynamics of an HIV-1 sub-epidemic among IDU. The phylodynamic analysis indicated a steady increase in the epidemic among IDU in coastal Kenya from 1987 to 2010, with no evidence of a rapid exponential increase in the number infections that is typical of HIV-1 outbreaks among IDUs[29,30,42]. Still, the gradual increase in infections among IDU is compatible with a known period of increased injection of heroin in the region[43]. Interestingly, and given a general absence of epidemiological surveillance data, new infections among IDU did not seem to decrease with the national rollout of combination antiretroviral therapy (ART) in 2004. This could be a consequence of the unfavourable climate of stigma, discrimination and hostile legislation associated with IDU and most-at-risk-populations in Kenya, which impedes these populations from accessing medical services including ART[1,44]. Opioid substitution therapy for IDU, Pre-Exposure Prophylaxis (PrEP) targeting all high-risk groups, and initiation of ART immediately upon diagnosis have all been introduced and scaled up after 2010, when the IDU samples used for this study were collected[45,46]. As new sequence data are made available, future studies may shed light on the effectiveness of these strategies.

This study had some limitations: First, the identified transmission chains are likely to suffer from missing links due to low sampling density. A low sampling density generally results in reduced clustering of HIV-1 sequences[47]. Because majority of the studies in the coastal Kenya setting have mostly focused on recruiting MSM participants, FSW and HET in our analysis had low sampling densities and inevitably, several transmission chains may have been missed. Furthermore, it is important to acknowledge that the IDU sequences were from one study, using samples from one setting (Mombasa), and collected over a period of less than one year (2010). It is therefore likely that our findings are not representative of the entire IDU epidemic in coastal Kenya. Larger studies of HIV-1 transmission in the IDU population in coastal Kenya are therefore warranted; still we found clear branching patterns indicative of long-standing HIV-1 transmission associated with intravenous drug use. Second, skewed sampling between risk groups may result in overrepresentation of some types of risk group-specific and mixed clusters. Third, given that annotating sequences from sub-Saharan Africa with information about transmission risk factor has become common only in recent years, some published sequences used in this analysis lacked risk data and were assigned HET (by far the most dominant route of HIV-1 transmission in coastal Kenya). However, the risk group for nodes within a cluster that had inadequate annotation can often be deduced from association with nodes with a known risk group[48]. Since none of the presumed heterosexual sequences in this study was found in mixed clusters, it is unlikely that this potential caveat had any effect on our conclusions. Finally, few HIV-1 *pol* sequences were available after the year 2010 for all risk groups. This limited our analysis to the representation of some risk groups by the year and area of sampling, further restraining characterisation of recent clusters and ongoing transmission chains.

In conclusion, we demonstrated that there is HIV-1 mixing between high-risk groups and heterosexual populations in coastal Kenya, with frequent within-risk-group transmission. We highlight that high-risk groups could contribute to the epidemic either through seeding and maintaining new infections within their own risk group or through linking infections across different risk groups. It is possible that HIV-1 prevention programmes targeting FSW, MSM and IDU populations could reduce overall HIV-1 transmission in coastal Kenya. As more sequences become available from wider geographic regions, further and larger studies with uniform sampling densities across different risk groups will be necessary to estimate the impact of mixing between risk groups and the general population on HIV-1 spread and to determine the source populations that could most effectively be targeted to mitigate new infections in sub-Saharan Africa.

Methods

Study population and sequence dataset. All published HIV-1 *pol* sequences available in the HIV database at the Los Alamos National Laboratory (LANL) originating from coastal Kenya and collected 2005–2019 retrieved[49]. Sequences sampled from the same individuals were excluded from the data set, retaining only the oldest sequence per participant. Risk group information was obtained from LANL and any missing data were obtained by communication with study authors or inferred from reviewing literature from the respective studies[6,8–15]. In addition, new HIV-1 *pol* sequences were generated from samples collected 2005–2019 participants in an acute HIV-1 infection cohort and from a prospective observational study following high-risk volunteers in an HIV-1 vaccine feasibility cohort described elsewhere[3]. All new sequences were collected from treatment-naïve men and women aged 18 years and over. HIV-1 diagnosis for samples collected before 2016 was done using two rapid antibody tests in parallel (Determine, Abbott Laboratories; Unigold, Trinity Biotech), with conflicting results resolved by an enzyme-linked immunosorbent assay (ELISA, Genetic System HIV-1/2 plus O EIA; Bio-Rad). HIV-1 diagnosis for samples collected after 2016 was made using the GeneXpert HIV-1 Qual (Cepheid, Sunnyvale, CA, USA).

Sequences were annotated by date of sample collection, geographical area (Mombasa or Kilifi county), and by risk group. Sequences were classified into: MSM (men who reported having sex with men); IDU (individuals who reported injecting drugs with a needle and syringe); FSW (females who reported ever receiving money, gifts, or favours in exchange for sex); and HET (all other individuals [both male and female] who did not report engaging in sex work, male same-sex behaviour or injection drug use).

RNA extraction, amplification of HIV-1 pol region and sequencing. HIV-1 RNA was extracted from blood plasma samples using the RNeasy Lipid Tissue Mini Kit (QIAGEN) with modifications from the manufacturer's standard protocol[50]. Briefly, 100 µl of blood plasma was efficiently lysed in 1000 µl Qiazol Lysis

Reagent (Qiagen). DNA was removed by treating the column with RNase-free DNase 1 (Qiagen) prior to RNA elution in 40 µl nuclease-free water. Reverse transcription and amplification of partial *pol* gene were performed using the One-Step Superscript III RT/Platinum Taq High Fidelity Enzyme Mix (ThermoFisher Scientific™) with the *pol*-specific primer pair JA269 and JA272[51]. First-round PCR products were amplified in a nested PCR with DreamTaq Green DNA Polymerase (ThermoFisher Scientific™) using *pol*-specific primers JA271 and JA270. PCR products were sequenced in both directions with the nested PCR primers using the BigDye terminator kit v1.1 (Applied Biosystems) and the sequences were determined on an ABI PRISM 3130×1 Genetic Analyzer (Applied Biosystems).

Sampling density. County-exclusive estimates for HIV-1 prevalence for high-risk groups in Kenya were not available when this analysis was done. Hence, the national HIV-1 prevalence estimate for each risk group and the estimates of people infected with HIV-1 in Mombasa and Kilifi counties were used to estimate the sampling density of our study (defined as the proportion of genotyped viral sequences in the estimated number of HIV-infected individuals in a risk group)[36].

HIV-1 Subtyping. Newly generated consensus and published *pol* sequences from coastal Kenya were combined, codon-aligned using ClustalX 2.0.11, and manually adjusted in Geneious Prime 2019 (Version 2019 2.1) (https://www.geneious.com)[52]. The combined sequences were then aligned with the Group M (subtypes A-K + Recombinants) HIV-1 subtype reference dataset downloaded from Los Alamos HIV Database (http://www.hiv.lanl.gov/)[49].

To infer genetic relatedness, phylogenetic reconstruction was done in PhyML using the general time-reversible substitution model with a gamma-distributed rate variation and proportion of invariant sites (GTR + Γ4 + I)[53]. Branch support was estimated using the Shimodaira-Hasegawa approximate likelihood ratio test (aLRT-SH) as implemented in PhyML[54]. An aLRT-SH ≥0.9 was considered statistically significant[17]. The subtype-resolved maximum-likelihood phylogenetic tree was visualized in FigTree (v1.4.3). Unique recombinant forms (URFs) were further resolved by Bootscan analyses using SimPlot and breakpoints identified using a sliding window size of 300 bp and a step size of 20 bp[55].

Transmission Cluster analysis. To detect local transmission clusters, newly generated and published HIV-1 *pol* sequences were combined into one coastal Kenya sequence dataset. In addition, the ten most similar GenBank reference sequences for each coastal Kenya sequence were obtained through BLAST[29,30,56]. The unique coastal Kenya sequences and the reference sequences were aligned and analysed to determine HIV-1 transmission clusters. Subtype-specific maximum-likelihood phylogenies were reconstructed in PhyML. For each subtype, transmission clusters were manually determined by inspecting the ML tree from root to terminal tips to ensure sequences clustered with high branch support. Monophyletic clades with aLRT-SH support ≥0.9 and which were dominated (≥80%) by sequences from coastal Kenya (compared to reference sequences) were defined as coastal Kenya transmission clusters[17]. To determine active transmission clusters, sequences were further explored using a genetic distance threshold (≤1.5%) and aLRT-SH branch support of ≥0.9 in Cluster Picker[57]. Transmission networks were classified based on the number of sequences per cluster into dyads (2 sequences), networks (3–14 sequences) and large clusters (>14 sequences)[30].

Diversity analysis. A previously described ML bootstrap approach was employed to determine the genetic diversity in the identified clusters[37]. Briefly, all sequences in coastal Kenya transmission clusters were used to construct 1000 bootstrap ML phylogenies in Garli 2.0[58]. Diversity estimates were determined in Perl (version 5.30.0) using in-house Perl and Bioperl scripts[59]. The diversity for each pre-defined cluster was estimated by averaging the pairwise tree-distances between the cluster-specific sequences in each bootstrap tree, resulting in 1000 diversity estimates per cluster. Next, the medians of these 1000 diversity estimates were determined for each cluster and then used in the analysis as previously described[37]. The scripts used in this analysis is available from the authors upon request.

Bayesian phylogenetic analysis. To estimate the dates of origin (time to most recent common ancestor; tMRCA) of the coastal Kenyan clusters, maximum clade credibility trees were generated using a Bayesian Markov Chain Monte Carlo (MCMC) approach as implemented in BEASTv1.8.2[60,61]. Based on marginal likelihood estimators for model selection and testing in BEAST, the Bayesian Skygrid model with an uncorrelated lognormal relaxed clock and inferred under the GTR + Γ4 + I substitution model was adopted as the best fit model for subsequent inferences. BEAST runs of 100–300 million generations were performed, logging samples after every 10000–30000 steps with the initial 10–30% discarded as burn-in. The convergence of MCMC parameter estimates was accessed based on effective sample size estimates (ESS > 200) using Tracer v1.6[62]. Trees were summarized in Tree-Annotator v1.8.2 (BEAST suite) and maximum clade credibility (MCC) trees were visualized in Figtree. We also aimed to dissect the demographic history of the only large coastal Kenya cluster identified. Since the sequences in this cluster did not contain temporal information (all IDU sequences were sampled in 2010), the node height posterior distribution (tMRCA) for the IDU cluster from the transmission clusters analysis described above was used as a tree-root height calibration prior (fixed the tree root height to 1985), guiding a skygrid analysis to estimate the effective population size (N_e) of the large IDU cluster over time.

Statistical analysis. Frequencies and percentages were used to describe the distribution of sequences within the study population by risk group, HIV-1 subtype, calendar year of sampling and county of sampling. A logistic regression model was used to assess factors associated with clustering. Variables with P-values <0.1 in the

bivariable analysis were included in the multivariable model. A P-value of <0.05 was defined as statistically significant. A Kruskal-Wallis H test and a post hoc Dunn's test (with Bonferroni correction for multiple comparisons) were conducted to determine differences in genetic diversity estimates among clusters from multiple risk groups. Statistics were done using Stata version 15 and RStudio (version 1.2.5001), and the packages: *DescTools* (version 0.99.29, https://cran.r-project.org/package=DescTools) and *yarrr* (version 0.1.6)[63]. The full R code is available on request from the authors.

Nucleotide sequence accession numbers. Nucleotide sequences were deposited in GenBank under the following accession numbers: MT084914 - MT085076.

Ethical consideration. All research was performed in accordance with relevant guidelines/regulations. Informed consent was obtained from all participants who provided blood plasma samples from which new HIV-1 sequences were generated (informed consent was not required for subjects whose sequences were obtained from LANL). Plasma samples used to generate the new sequences were obtained from on-going or concluded studies that were also approved by KEMRI/SERU (SERU 3747, 3280 and 3520, and SSC 894). The current study is part of a parent protocol that was reviewed and approved by the Kenya Medical Research Institute (KEMRI) Scientific and Ethics Review Unit (SERU 3547).

Received: 13 November 2019; Accepted: 30 March 2020;
Published online: 21 April 2020

References

1. Kenya National AIDS Control Council. Kenya AIDS Strategic Framework 2014/2015–2018/2019, http://nacc.or.ke/wp-content/uploads/2015/09/KASF_Final.pdf (2019).
2. Sanders, E. J., Jaffe, H., Musyoki, H., Muraguri, N. & Graham, S. M. Kenyan MSM: no longer a hidden population. *AIDS* **29**, S195–S199, https://doi.org/10.1097/qad.0000000000000928 (2015).
3. Sanders, E. J. *et al.* High HIV-1 incidence, correlates of HIV-1 acquisition, and high viral loads following seroconversion among MSM. *Aids* **27**, 437–446, https://doi.org/10.1097/QAD.0b013e32835b0f81 (2013).
4. Smith, A. D. *et al.* Heterosexual behaviours among men who sell sex to men in coastal Kenya. *AIDS* **29**(Suppl 3), S201–210, https://doi.org/10.1097/QAD.0000000000000889 (2015).
5. Kenya National AIDS Control Council. Kenya HIV Prevention Response and Modes of Transmission Analysis., http://siteresources.worldbank.org/INTHIVAIDS/Resources/375798-1103037153392/KenyaMOT22March09Final.pdf (2009).
6. Bezemer, D. *et al.* HIV Type 1 transmission networks among men having sex with men and heterosexuals in Kenya. *AIDS Res. Hum. retroviruses* **30**, 118–126 (2014).
7. Gounder, K. *et al.* Complex Subtype Diversity of HIV-1 Among Drug Users in Major Kenyan Cities. *AIDS Res. Hum. Retroviruses* **33**, 500–510, https://doi.org/10.1089/aid.2016.0321 (2017).
8. Hassan, A. S. *et al.* HIV-1 subtype diversity, transmission networks and transmitted drug resistance amongst acute and early infected MSM populations from Coastal Kenya. *PLoS One* **13**, e0206177, https://doi.org/10.1371/journal.pone.0206177 (2018).
9. Hué, S. *et al.* HIV type 1 in a rural coastal town in Kenya shows multiple introductions with many subtypes and much recombination. *AIDS Res. Hum. retroviruses* **28**, 220–224 (2012).
10. Osman, S. *et al.* Diversity of HIV type 1 and drug resistance mutations among injecting drug users in Kenya. *AIDS Res. Hum. Retroviruses* **29**, 187–190, https://doi.org/10.1089/AID.2012.0182 (2013).
11. Price, M. A. *et al.* Transmitted HIV type 1 drug resistance among individuals with recent HIV infection in East and Southern Africa. *AIDS Res. Hum. Retroviruses* **27**, 5–12, https://doi.org/10.1089/aid.2010.0030 (2011).
12. Sigaloff, K. C. *et al.* High prevalence of transmitted antiretroviral drug resistance among newly HIV type 1 diagnosed adults in Mombasa, Kenya. *AIDS Res. Hum. retroviruses* **28**, 1033–1037 (2012).
13. Tovanabutra, S. *et al.* Evaluation of HIV type 1 strains in men having sex with men and in female sex workers in Mombasa, Kenya. *AIDS Res. Hum. Retroviruses* **26**, 123–131, https://doi.org/10.1089/aid.2009.0115 (2010).
14. Khamadi, S. A. *et al.* Genetic diversity of HIV type 1 along the coastal strip of Kenya. *AIDS Res. Hum. Retroviruses* **25**, 919–923, https://doi.org/10.1089/aid.2009.0005 (2009).
15. Hassan, A. S. *et al.* Low prevalence of transmitted HIV type 1 drug resistance among antiretroviral-naive adults in a rural HIV clinic in Kenya. *AIDS Res. Hum. Retroviruses* **29**, 129–135, https://doi.org/10.1089/aid.2012.0167 (2013).
16. Pybus, O. G., Tatem, A. J. & Lemey, P. Virus evolution and transmission in an ever more connected world. *Proc. R. Soc. B: Biol. Sci.* **282**, 20142878 (2015).
17. Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J. & Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data. *AIDS* **31**, 1211 (2017).
18. Onywera, H. *et al.* Surveillance of HIV-1 pol transmitted drug resistance in acutely and recently infected antiretroviral drug-naive persons in rural western Kenya. *PLoS one* **12**, e0171124 (2017).
19. Chimukangara, B. *et al.* Moderate-to-high levels of pretreatment HIV drug resistance in KwaZulu-Natal Province, South Africa. *AIDS Res. Hum. retroviruses* **35**, 129–138 (2019).
20. Rodgers, M. A. *et al.* Sensitive next-generation sequencing method reveals deep genetic diversity of HIV-1 in the Democratic Republic of the Congo. *J. virology* **91**, e01841–01816 (2017).
21. Tongo, M. *et al.* Unravelling the complicated evolutionary and dissemination history of HIV-1M subtype A lineages. *Virus Evolution* **4**, vey003 (2018).
22. Faria, N. R. *et al.* Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. *Infection, Genet. Evolution* **12**, 453–460 (2012).
23. Faria, N. R. *et al.* The early spread and epidemic ignition of HIV-1 in human populations. *science* **346**, 56–61 (2014).
24. Esbjörnsson, J., Mild, M., Månsson, F., Norrgren, H. & Medstrand, P. HIV-1 molecular epidemiology in Guinea-Bissau, West Africa: origin, demography and migrations. *PLoS one* **6**, e17025 (2011).
25. Brenner, B. G. *et al.* High rates of forward transmission events after acute/early HIV-1 infection. *J. Infect. Dis.* **195**, 951–959, https://doi.org/10.1086/512088 (2007).
26. Volz, E. M. et al. HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. PLoS Med 10, e1001568; discussion e1001568, https://doi.org/10.1371/journal.pmed.1001568 (2013).
27. Ratmann, O. *et al.* Sources of HIV infection among men having sex with men and implications for prevention. *Sci. Transl. Med.* **8**, 320ra322–320ra322 (2016).
28. Poon, A. F. *et al.* Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *lancet HIV.* **3**, e231–e238 (2016).

29. Sallam, M. *et al*. Molecular epidemiology of HIV-1 in Iceland: Early introductions, transmission dynamics and recent outbreaks among injection drug users. *Infection, Genet. Evolution* **49**, 157–163 (2017).
30. Esbjörnsson, J. *et al*. HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. *Virus evolution* **2**, vew010 (2016).
31. Bruhn, C. A. *et al*. The origin and emergence of an HIV-1 epidemic: from introduction to endemicity. *AIDS* **28**, 1031–1040, https://doi.org/10.1097/QAD.0000000000000198 (2014).
32. Frentz, D. *et al*. Patterns of transmitted HIV drug resistance in Europe vary by risk group. *PLoS one* **9**, e94495 (2014).
33. Grabowski, M. K. *et al*. The Role of Viral Introductions in Sustaining Community-Based HIV Epidemics in Rural Uganda: Evidence from Spatial Clustering, Phylogenetics, and Egocentric Transmission Models. *PLOS Med.* **11**, e1001610, https://doi.org/10.1371/journal.pmed.1001610 (2014).
34. De Oliveira, T. *et al*. Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *lancet HIV.* **4**, e41–e50 (2017).
35. Bbosa, N. *et al*. Phylogeography of HIV-1 suggests that Ugandan fishing communities are a sink for, not a source of, virus from general populations. *Sci. Rep.* **9**, 1051 (2019).
36. National AIDS and STI Control Programme. Kenya HIV County Profiles 2016., http://nacc.or.ke/wp-content/uploads/2016/12/Kenya-HIV-County-Profiles-2016.pdf (2017).
37. Esbjörnsson, J. *et al*. Inhibition of HIV-1 Disease Progression by Contemporaneous HIV-2 Infection. *N. Engl. J. Med.* **367**, 224–232, https://doi.org/10.1056/NEJMoa1113244 (2012).
38. Middelkoop, K. *et al*. Epidemiology of HIV-1 subtypes among men who have sex with men in Cape Town, South Africa. *JAIDS J. Acquired Immune Deficiency Syndromes* **65**, 473–480 (2014).
39. Konou, A. A. *et al*. Genetic diversity and transmission networks of HIV-1 strains among men having sex with men (MSM) in Lomé, Togo. *Infection, Genet. Evolution* **46**, 279–285 (2016).
40. Côté, A.-M. *et al*. Transactional sex is the driving force in the dynamics of HIV in Accra, Ghana. *Aids* **18**, 917–925 (2004).
41. Parczewski, M. *et al*. Expanding HIV-1 subtype B transmission networks among men who have sex with men in Poland. *PLoS One* **12**, e0172473 (2017).
42. Skar, H. *et al*. Dynamics of two separate but linked HIV-1 CRF01_AE outbreaks among injection drug users in Stockholm, Sweden, and Helsinki, Finland. *J. virology* **85**, 510–518 (2011).
43. Beckerleg, S., Telfer, M. & Hundt, G. L. The rise of injecting drug use in East Africa: a case study from Kenya. *Harm Reduct. J.* **2**, 12, https://doi.org/10.1186/1477-7517-2-12 (2005).
44. Stannah, J. et al. HIV testing and engagement with the HIV treatment cascade among men who have sex with men in Africa: A systematic review and meta-analysis. Lancet HIV (2019).
45. Kenya National AIDS Control Council. Kenya AIDS Responce Progress Report 2018, https://www.lvcthealth.org/wp-content/uploads/2018/11/KARPR-Report_2018.pdf (2018).
46. Rhodes, T. *et al*. Is the promise of methadone Kenya's solution to managing HIV and addiction? A mixed-method mathematical modelling and qualitative study. *BMJ open.* **5**, e007198 (2015).
47. Novitsky, V., Moyo, S., Lei, Q., DeGruttola, V. & Essex, M. Impact of sampling density on the extent of HIV clustering. *AIDS Res. Hum. retroviruses* **30**, 1226–1235 (2014).
48. Novitsky, V. *et al*. Phylodynamic analysis of HIV sub-epidemics in Mochudi, Botswana. *Epidemics* **13**, 44–55 (2015).
49. Los Alamos National Library. HIV-1 database at the Los Alamos National Library, http://www.hiv.lanl.gov/ (2019).
50. Esbjörnsson, J. *et al*. Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease-indication of an evolving epidemic in West Africa. *Retrovirology* **7**, 23 (2010).
51. Hedskog, C. *et al*. Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PLoS one* **5**, e11345 (2010).
52. Larkin, M. A. *et al*. Clustal W and Clustal X version 2.0. *bioinformatics* **23**, 2947–2948 (2007).
53. Guindon, S., Lethiec, F., Duroux, P. & Gascuel, O. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic acids Res.* **33**, W557–W559 (2005).
54. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55**, 539–552, https://doi.org/10.1080/10635150600755453 (2006).
55. Lole, K. S. *et al*. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. virology* **73**, 152–160 (1999).
56. Kouyos, R. D. *et al*. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J. Infect. Dis.* **201**, 1488–1497, https://doi.org/10.1086/651951 (2010).
57. Ragonnet-Cronin, M. *et al*. Automated analysis of phylogenetic clusters. *BMC Bioinforma.* **14**, 317, https://doi.org/10.1186/1471-2105-14-317 (2013).
58. Zwickl, D. J. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion, (2006).
59. Stajich, J. E. *et al*. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**, 1611–1618 (2002).
60. Suchard, M. A. *et al*. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus. *Evolution* **4**, vey016 (2018).
61. Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
62. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
63. Phillips, N. D. Yarrr! The pirate's guide to R. APS Observer 30 (2017).
64. Faria, N. R. *et al*. Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa. *PLoS Pathog.* **15**, e1007976–e1007976 (2019).

Acknowledgements

We thank the International AIDS Vaccine Initiative (IAVI) for supporting the HIV at-risk cohort studies in Kilifi, Kenya. We are also grateful to the staff in the HIV/STI project at the Kenya Medical Research Institute (KEMRI) in Kilifi for their commitment to serving MSM. Finally, we thank Professor Philippe Lemey, Katholieke Universiteit Leuven, for providing useful input for dating and determining past population dynamics of the coastal Kenyan IDU clusters. This manuscript was submitted for publication with the permission from the Director of the Kenya Medical Research Institute (KEMRI). This work was supported through the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15–006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant # 107752/Z/15/Z] and the UK government. This work was also supported in part by funding from the Swedish Research Council (grant

2016–01417) and the Swedish Society for Medical Research (grant # SA-2016). The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, IAVI, Swedish Research Council, or the UK government. Open access funding provided by Lund University.

Author contributions

A.S.H., J.E., G.N. and E.J.S. conceptualized and designed the study. A.S.H., J.E. and E.J.S. provided funding for the study. E.J.S. and S.M.G. provided samples from which new sequences used in the study were generated. G.N.M. performed lab work, inferential analyses and produced all figures and tables. J.N. assisted with data analysis. A.S.H., J.E. and E.J.S. provided supervisory guidance. G.N.M. wrote the original draft manuscript and all the authors reviewed and edited the manuscript prior to submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-63731-z.

Correspondence and requests for materials should be addressed to J.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

frontiers in Microbiology

ORIGINAL RESEARCH
published: 17 August 2021
doi: 10.3389/fmicb.2021.658186

Phylogenetic and Drug-Resistance Analysis of HIV-1 Sequences From an Extensive Paediatric HIV-1 Outbreak in Larkana, Pakistan

Syed Hani Abidi[1†], George Makau Nduva[2,3†], Dilsha Siddiqui[1†], Wardah Rafaqat[4], Syed Faisal Mahmood[5], Amna Rehana Siddiqui[6], Apsara Ali Nathwani[7], Aneeta Hotwani[7], Sharaf Ali Shah[8], Sikander Memon[9], Saqib Ali Sheikh[9], Palwasha Khan[10], Joakim Esbjörnsson[2,11‡], Rashida Abbas Ferrand[7,10‡] and Fatima Mir[7‡]*

[1] Department of Biological and Biomedical Sciences, Aga Khan University, Karachi, Pakistan, [2] Department of Translational Medicine, Lund University, Lund, Sweden, [3] Kenya Medical Research Institute-Wellcome Trust Research Programme, Kilifi, Kenya, [4] Medical College, Aga Khan University, Karachi, Pakistan, [5] Department of Medicine, Aga Khan University, Karachi, Pakistan, [6] Department of Community Health Sciences, Aga Khan University, Karachi, Pakistan, [7] Department of Pediatrics and Child Health, Aga Khan University, Karachi, Pakistan, [8] Bridge Consultants Foundation, Karachi, Pakistan, [9] Sindh AIDS Control Program, Ministry of Health, Karachi, Pakistan, [10] Department of Clinical Research, London School of Hygiene & Tropical Medicine, London, United Kingdom, [11] The Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

Introduction: In April 2019, an HIV-1 outbreak among children occurred in Larkana, Pakistan, affecting more than a thousand children. It was assumed that the outbreak originated from a single source, namely a doctor at a private health facility. In this study, we performed subtype distribution, phylogenetic and drug-resistance analysis of HIV-1 sequences from 2019 outbreak in Larkana, Pakistan.

Methods: A total of 401 blood samples were collected between April–June 2019, from children infected with HIV-1 aged 0–15 years recruited into a case-control study to investigate the risk factors for HIV-1 transmission. Partial HIV-1 *pol* sequences were generated from 344 blood plasma samples to determine HIV-1 subtype and drug resistance mutations (DRM). Maximum-likelihood phylogenetics based on outbreak and reference sequences was used to identify transmission clusters and assess the relationship between outbreak and key population sequences between and within the determined clusters. Bayesian analysis was employed to identify the time to the most recent common recent ancestor (tMRCA) of the main Pakistani clusters.

Results: The HIV-1 circulating recombinant form (CRF) 02_AG and subtype A1 were most common among the outbreak sequences. Of the treatment-naïve participants, the two most common mutations were RT: E138A (8%) and RT: K219Q (8%). Four supported clusters within the outbreak were identified, and the median tMRCAs of the Larkana outbreak sequences were estimated to 2016 for both the CRF02_AG and the subtype A1 clusters. Furthermore, outbreak sequences exhibited no phylogenetic mixing with sequences from other high-risk groups of Pakistan.

OPEN ACCESS

Edited by:
Arshan Nasir,
Los Alamos National Laboratory (DOE), United States

Reviewed by:
Teiichiro Shiino,
National Center For Global Health and Medicine, Japan
Thomas Leitner,
Los Alamos National Laboratory (DOE), United States

***Correspondence:**
Syed Hani Abidi
m.haniabidi@gmail.com

† These authors have contributed equally to this work and share first authorship

‡ These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: 25 January 2021
Accepted: 21 July 2021
Published: 17 August 2021

Citation:
Abidi SH, Nduva GM, Siddiqui D, Rafaqat W, Mahmood SF, Siddiqui AR, Nathwani AA, Hotwani A, Shah SA, Memon S, Sheikh SA, Khan P, Esbjörnsson J, Ferrand RA and Mir F (2021) Phylogenetic and Drug-Resistance Analysis of HIV-1 Sequences From an Extensive Paediatric HIV-1 Outbreak in Larkana, Pakistan. Front. Microbiol. 12:658186. doi: 10.3389/fmicb.2021.658186

Conclusion: The presence of multiple clusters indicated a multi-source outbreak, rather than a single source outbreak from a single health practitioner as previously suggested. The multiple introductions were likely a consequence of ongoing transmission within the high-risk groups of Larkana, and it is possible that the so-called Larkana strain was introduced into the general population through poor infection prevention control practices in healthcare settings. The study highlights the need to scale up HIV-1 prevention programmes among key population groups and improving infection prevention control in Pakistan.

Keywords: HIV-1, outbreak investigation, phylogenetic analysis, drug resistance, paediatric [MeSH]

INTRODUCTION

The HIV-1 pandemic has been established for 40 years and has resulted in approximately 32.7 million deaths worldwide (de Mendoza, 2019). One of the characteristic features of HIV-1 is its high mutation rate and recombination rate within and between hosts, leading to the emergence of distinct subtypes and circulating recombinant forms (CRFs) (Taylor et al., 2008; van Zyl et al., 2018). The subtypes can also recombine to give rise to unique recombinant forms (URFs) and minor HIV-1 variants (Taylor et al., 2008). The consequent genetic diversity that characterises HIV-1 infection has implications for virological control and transmission. Mutations encoded by the virus can interfere with epitope processing and recognition, leading to immune evasion. In addition, mutations may also lead to resistance to anti-retroviral drugs (van Zyl et al., 2018). Furthermore, certain immune- or drug-escape mutations may facilitate rapid transmission (van Zyl et al., 2018).

The phylogenetic and phylodynamic analysis of sequences derived from people living with HIV-1 (PLWH), particularly those who are part of an outbreak, can help answer fundamental questions such as the directionality and pattern of transmission and in understanding the introduction of HIV-1 into different regions, as well as identify clusters of transmission (Kosakovsky Pond et al., 2018; Dawson et al., 2020; Romero-Severson et al., 2020).

Pakistan has experienced a growing HIV-1 epidemic that is concentrated among three key population groups namely persons who inject drugs (PWID), transgender sex workers [also known as *Hijra* sex workers (HSW)], and men who have sex with men (MSM) (Raees et al., 2013) in whom prevalence is around 38–40, 11, and 7.5%, respectively (Baqi et al., 1998; Waheed and Waheed, 2017; Hasan et al., 2018). Only 36,000 of an estimated 160,000 PLWH in Pakistan were aware of their HIV-1 positive status in 2020, and only 24,606 PLWH were receiving HIV-1 treatment of whom 7,693 were PWID (Esbjörnsson et al., 2016).

In April 2019, a cluster of fourteen HIV-1 diagnoses in children was reported in Ratodero, a town in Larkana district, Pakistan (Siddiqui et al., 2020). By December 2019, 1,167 children have been diagnosed with HIV-1 through a screening programme established in response to the outbreak (Brenner et al., 2007). Larkana has had three previous outbreaks of HIV, the first among PWID in 2003, the second in 2016 among 12 children in a paediatric hospital, and the third in 2016 among 56 individuals in a renal dialysis unit (Brenner et al., 2007; Altaf et al., 2016). These outbreaks were linked to poor infection prevention control practices including reuse of needles and inadequate blood screening.

For Pakistan, the outbreak in 2019 is unprecedented in terms of predominantly affecting children and its magnitude: prior to the outbreak, 1041 children had ever registered for HIV-1 care nationally over the past 13 years (Mir et al., 2020a). Early during the outbreak, media reports implicated a local doctor who had treated several of the infected children and who was later diagnosed with HIV, in spreading HIV-1 infection (Abi-Habib and Masood, 2019; Siddiqui et al., 2020).

In this study, we conducted a phylogenetic analysis to investigate the HIV-1 outbreak subtype and the pattern and source of transmission, and specifically whether this was a single-source outbreak. Furthermore, we also analysed the sequences for presence of drug resistance mutations (DRMs).

MATERIALS AND METHODS

Study Design and Setting

This study was embedded in an individually matched case-control study that recruited 401 cases defined as children aged 0–15 years who registered for HIV-1 care at the Paediatric Treatment Center at Shaikh Zayed Children's Hospital (Siddiqui et al., 2020). This centre was established by the Sindh AIDS Control Program in response to the outbreak. Prior to the outbreak, the nearest paediatric HIV-1 services were in the provincial capital, Karachi, situated more than 400 kilometres from Ratodero. Age-, sex-, and neighbourhood-matched HIV-uninfected controls were also recruited. An interviewer-administered questionnaire collected data on risk factors for HIV-1 infection. A blood sample was collected for Hepatitis B and C serology, and for HIV-1 phylogenetic studies (in cases only). Written informed consent was obtained from guardians and assent from participants. The study was approved by the Aga Khan University Ethical Review Committee (ERC# 2019-1536-4200). Prior to sample collection, written informed consent was obtained from the guardians, and if the child was able to understand the study procedures, a written assent was obtained. The study objectives were explained to the patient at the time of taking consent/assent and patients were informed that their identities will remain confidential. The

participants were also informed that they had the opportunity to withdraw from the study at any given time and that this would have no consequences on the treatment or the care that they would receive.

DNA Amplification and HIV-1 Genotyping

Proviral DNA was extracted from blood samples obtained from cases using Qiagen's QIAamp DNA blood mini kit according to the manufacturer's instructions and stored at -80°C. The *pol* gene was amplified from each extracted DNA sample using a two-step nested polymerase chain reaction (PCR) strategy. Two sets of outer primers were used: Forward (POLOF CAGCATGYCAGGGAGTRGGRGGACC, amino acid; 1832-1856, HXB2, IBF1 5′-AAATGATGACAGCATGTCAG GGAGT-3′. nt 1823-1847, HXB2) and Reverse (IBR1 5′-AACTT CTGTATATCATTGACAGTCCA-3′. nt 3303-3328, HXB2). The first-round product was used as a template for the second round with primer set, Forward (POLIF 5′-AGGCTAATTTT TTAGGGAARATYTGGCCTTCC-3′. nt 2078-2109, amino acid PR: 1–9; HXB2) and reverse (RTOUT3 5′-TATGTCATT GACAGTCCAGCT-3′. nt 3300–3320, amino acid RT: 251–257 HXB2) (Tariq et al., 2018). PCR Mastermix (ABM) Bestaq (2X) cat# G464 and Hotstart (2X) cat# G906 were used to prepare a 25 ul reaction mixture, and 0.8 pmol and 0.6 pmol primer used for the first and second round, respectively. Thermo cycle conditions were as follows: denaturation at 95°C for 5 min, followed by 40 cycles of denaturation at 95°C for 1 min, annealing at 50°C for IBF1/IBR1 and 55°C for POLOF/IBR1 sets (round 1), 60°C (round 2) for 20 s, extension at 72°C for 1 min with a final extension of at 72°C for 7 min. The protocol was run with positive and negative control to confirm results. The amplicons underwent sequencing using the Sanger sequencing platform (Macrogen, South Korea) and the sequences were deposited in the GenBank and assigned the accession numbers MN698251-MN698253, MN698255-MN698264, MN752136, MN752137, and MT748850-MT749178.

Subtype Analysis

HIV-1 *pol* sequence data used in the study comprised of either newly generated sequences (referred to as outbreak sequences) or Pakistani HIV-1 *pol* sequences retrieved from the Los Alamos HIV sequence database[1] (referred to as published Pakistani sequences representing Pakistani people who inject drugs (PWID), heterosexuals, sex workers, and other individuals with unknown transmission risk) (Kuiken et al., 2003). Outbreak sequences and published Pakistani HIV-1 *pol* sequences (referred to as the Pakistani dataset) were aligned with HIV-1 Group M (subtypes A-K + Recombinants) subtype reference sequences[1] using the MAFFT algorithm in Geneious Prime 2019 (Larkin et al., 2007). Subtyping of each sequence was determined by maximum-likelihood (ML) phylogenetic analysis in PhyML using the general time-reversible substitution model with a gamma-distributed rate variation and proportion of invariant sites (GTR + Γ4 + I) (Guindon et al., 2010). Branch support was estimated using the approximate likelihood ratio test with the Shimodaira-Hasegawa-like procedure (SH-aLRT) in PhyML, where SH-aLRT support values ≥0.90 were considered significant (Guindon et al., 2010). Phylogenies were visualised in FigTree v1.4.4.[2]

Cluster Analysis

To identify local transmission clusters, subtype-specific ML phylogenies were reconstructed for the main/predominant HIV-1 strains identified in the outbreak. The most similar non-Pakistani sequences for each sequence in the Pakistani dataset were retrieved from the NCBI GenBank and used as reference sequences as previously described (Esbjörnsson et al., 2016; Nazziwa et al., 2020; Nduva et al., 2020). The Pakistani dataset and GenBank reference sequences were aligned by subtype or CRF and subtype/CRF-specific phylogenies were reconstructed in PhyML (Guindon et al., 2010). Monophyletic clusters having aLRT-SH ≥0.90 and comprising ≥80% Pakistani sequences were defined as Pakistani-specific clusters (Esbjörnsson et al., 2016; Hassan et al., 2017; Nazziwa et al., 2020; Nduva et al., 2020). Clusters were classified into dyads (2 sequences), networks (3–14 sequences), or large clusters (>14 sequences) (Esbjörnsson et al., 2016).

Estimating Dates of the Most Recent Common Ancestor for Each Cluster

The dates of origin (time to the most recent common ancestor; tMRCA) of the large Pakistani-specific HIV-1 clusters were estimated using Bayesian Markov Chain Monte Carlo (MCMC) inference in BEAST (v1.10.4) (Gill et al., 2020). All Larkana outbreak sequences were sampled in 2019 and did not independently have a sufficient temporal signal for inference of the dates of origin. Hence, supplementary non-Pakistani reference sequences (seven CRF02_AG and six sub-subtype A1 sequences), and Pakistani sequences from previous outbreaks (33 CRF02_AG and 11 sub-subtype A1 sequences) sampled from different years were used to inform the temporal signal (assessed in TempEst v1.5.3) (Rambaut et al., 2016). Subtype-specific Bayesian inferences were done in BEAST 1.10.4 using the Bayesian Skygrid model with an uncorrelated lognormal relaxed clock and inferred under the GTR + Γ4 + I substitution model (Drummond et al., 2005; Baele et al., 2012; Gill et al., 2013; Suchard et al., 2018). BEAST runs of 500 million generations were performed, sampling every 50,000th iteration, and discarding the first 10% of samples as burn-in. Convergence was determined in Tracer v.1.7.0, defined as effective sample sizes (ESS) ≥200 (Suchard et al., 2018). Maximum clade credibility (MCC) trees were summarised in Tree-Annotator v1.10.4 and visualised in Figtree (v1.4.4).

Drug Resistance Mutation Analysis

Mutations in the HIV-1 *pol* gene (protease and reverse transcriptase region) associated with resistance against protease and reverse transcriptase inhibitors was determined using the Stanford HIV-1 drug resistance database (Rhee et al., 2003),

[1] http://www.hiv.lanl.gov

[2] https://github.com/rambaut/figtree/releases

and confirmed using the 2019 Update of the Drug Resistance Mutations in HIV-1 by the International AIDS Society–United States (Wensing et al., 2019). The DRMs were also classified as those conferring high, intermediate, low, and potential low-level resistance using the algorithm described in Stanford HIV-1 drug resistance database and IAS-United States report.

RESULTS

Study Population

Of the blood samples obtained from the 401 cases enrolled in the case-control study, 344 were successfully amplified and sequenced. The remaining 57 samples failed to amplify possibly due to low viral load secondary to receiving antiretroviral treatment or due to genomic diversity attributed to quasi-species in an individual (Debyser et al., 1998; Gupta et al., 2017). Out of 344 cases, socio-demographic information was available for 321 sequences, while information for 23 cases was missing. The median age of participants was three (IQR: 2–5) years and 65% were male (**Table 1**). The majority (84.4%) were taking ART at the time of sampling for a median period of 41 days (range: 22–192 days). The ART regimen comprised of zidovudine, lamivudine, and nevirapine. Most participants lived in Ratodero, the epicentre of the outbreak, while the remainder were from other areas of Larkana district and neighbouring districts (**Table 1**).

TABLE 1 | Characteristics of study participants.

Category/variable	Total no. (%)
Age (years)	
0–5	248 (77%)
5–10	58 (18%)
10–15	15 (5%)
Sex	
Male	208 (65%)
Female	113 (35%)
Location	
Ratodero, Larkana	132 (41%)
Outside Ratodero but within Larkana district	141 (44%)
Shikarpur district	44 (13.7%)
Jafarabad district	1 (0.3%)
Khairpur district	1 (0.3%)
Nawabshah district	1 (0.3%)
ART History	
Naïve	50 (15.6%)
Experienced	271 (84.4%)
ART duration (268)	
<30 days	93 (28.97%)
>30–180 days	175 (54.5%)
HCVand HBV co-infections	
HBV positive	75 (23.4%)
HCV positive	26 (8%)
Maternal HIV-1 status	
HIV-1 positive mother	28 (8.7%)
Mother's HIV-1 status unknown	4 (1.2%)

Pakistani HIV-1 Sequence Dataset and HIV-1 Subtypes and CRFs

Overall, we analysed 532 HIV-1 partial *pol* sequences in the Pakistani dataset including outbreak sequences (N = 344) and previously published sequences (N = 188). HIV-1 CRF02_AG (N = 338, 63.5%) dominated the outbreak, followed by sub-subtype A1 (N = 149, 28.0%). Additional subtypes found in the outbreak were subtype C (N = 20, 3.8%), subtype G (N = 8, 1.5%), CRF35_AD (N = 7, 1.3%), subtype B (N = 5, 0.9%), and subtype D (N = 5, 0.9%, **Figure 1**).

IDENTIFICATION OF PAKISTANI-SPECIFIC HIV-1 TRANSMISSION CLUSTERS

ML phylogenies were reconstructed independently for CRF02_AG (N = 338) and sub-subtypes A1 (N = 149), which were the most prevalent HIV-1 strains in the outbreak. The final reference dataset comprised of 310 non-Pakistani reference sequences for CRF02_AG, and 382 for sub-subtype A1 remained. Overall, 291 (86.1% CRF02_AG outbreak sequences) and 59 (40.0% sub-subtype A1 outbreak sequences) sequences formed 17 supported Pakistani clusters (size range: 2–283 sequences per cluster). These 17 clusters included 9 dyads (52.9% of all clusters), six networks (33.3%), and two large clusters (11.8%, **Table 2** and **Figures 2A,B**). Sub-subtype A1 clusters were more common (N = 13, 76.5%) as compared to the CRF02_AG clusters (N = 4, 23.5%). Of the 344 outbreak sequences, 312 (90.7%) were found in four distinct clusters. More specifically, 283 sequences were found in one large CRF02_AG cluster (**Figure 2A**), two sequences in a CRF02_AG dyad (**Figure 2A**), 22 sequences in one large sub-subtype A1 cluster (**Figure 2B**), and four sequences in a sub-subtype A1 network (**Figure 2B**). HIV-1 clusters of outbreak sequences showed no evidence of mixing with other HIV-1 risk groups of Pakistan. Instead, the Pakistani sequences that were not part of the outbreak formed an exclusive Pakistani cluster of non-outbreak sequences (**Table 2**).

Estimation of Time to the Most Recent Common Ancestor (tMRCA)

The tMRCAs were estimated for the two large clusters (one CRF02_AG and one subtype A1 cluster) comprising of the outbreak sequences. The median tMRCA of the CRF02_AG cluster was estimated to 2016 (95% higher posterior density [HPD] interval: 2015–2017), and the tMRCA of the sub-subtype A1 cluster was estimated to be 2016 (95% HPD interval: 2015–2018). In addition, the divergence time between the outbreak sequences and that of other high-risk groups in Pakistan, such as PWID, was dated to the year 1999 (95% HPD interval: 1994–2010) for the CRF02_AG cluster,

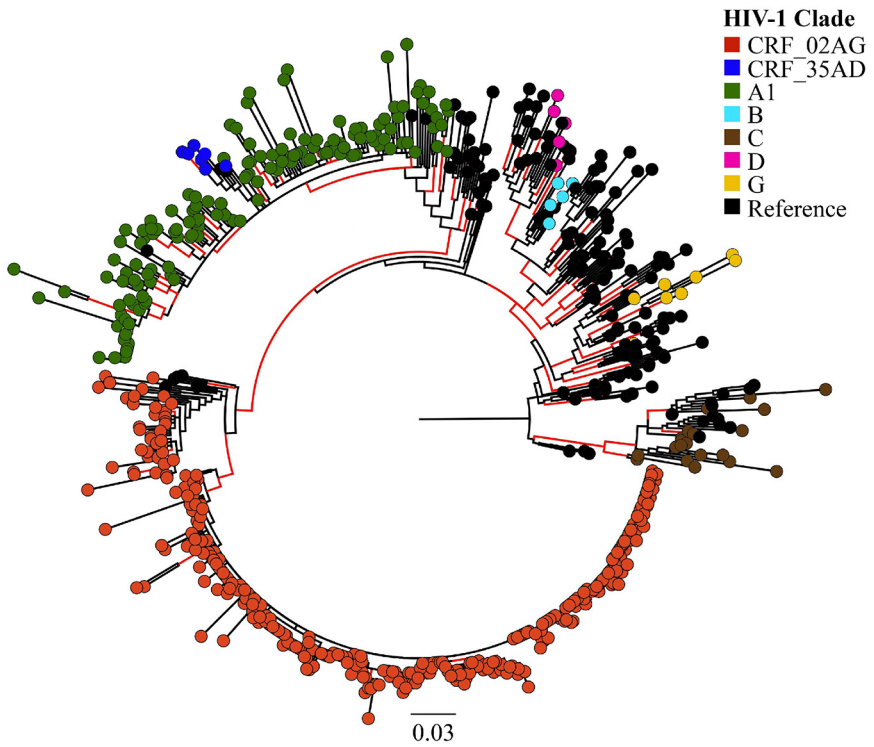


FIGURE 1 | Maximum likelihood (ML)-based subtyping of HIV-1 sequences: ML-based phylogeny/subtyping of 532 HIV-1 sequences (including 344 outbreak sequences from Larkana-Pakistan). Branch tips on the phylogenetic tree are coloured according to HIV CRF/subtype (Red: CRF_02AG; Blue: CRF_35AD, Green: subtype A1; Sky Blue: subtype B; Brown: subtype C; Pink: subtype D; Yellow: subtype G; Black: reference sequences from the Los Alamos HIV database). Scale bar units are nucleotide substitutions per site.

TABLE 2 | The number of Pakistani transmission clusters by cluster size, risk group, and shared drug resistance mutations.

Cluster name	HIV-1 subtype	Number of Tips	Transmission group	Number with shared DRM
Cluster_A1_1	A1	22	Paediatric	RT:E138A; *N* = 21
Cluster_A1_2	A1	3	Paediatric	RT:E138A; *N* = 3
Cluster_A1_3	A1	2	Unknown	
Cluster_A1_4	A1	2	MSM/unknown	
Cluster_A1_6	A1	2	PWID	
Cluster_A1_7	A1	2	Unknown	
Cluster_A1_8	A1	2	Unknown	
Cluster_A1_9	A1	2	Unknown	
Cluster_A1_10	A1	9	PWID/unknown	
Cluster_A1_11	A1	2	Unknown	
Cluster_A1_12	A1	5	Unknown	
Cluster_A1_15	A1	3	MSM/unknown	
Cluster_A1_16	A1	3	CWSW/unknown	
Cluster_CRF02AG_1	CRF02_AG	283	Paediatric	RT:K219Q; *N* = 5, RT:K103N; *N* = 19, RT:E138A; *N* = 7, RT:V17; *N* = 10
Cluster_CRF02AG_2	CRF02_AG	4	Paediatric	RT:E138A; *N* = 1, RT:V17; *N* = 2
Cluster_CRF02AG_3	CRF02_AG	2	Unknown	
Cluster_CRF02AG_4	CRF02_AG	2	Unknown	

Abbreviations: MSM, men who have sex with men; PWID, people who inject drugs; CWSW, female sex worker; Unknown, sequences lacking information on risk group or transmission route; Outbreak, sequences collected from the 2019 Larkana outbreak; RT, HIV-1 reverse transcriptase gene.

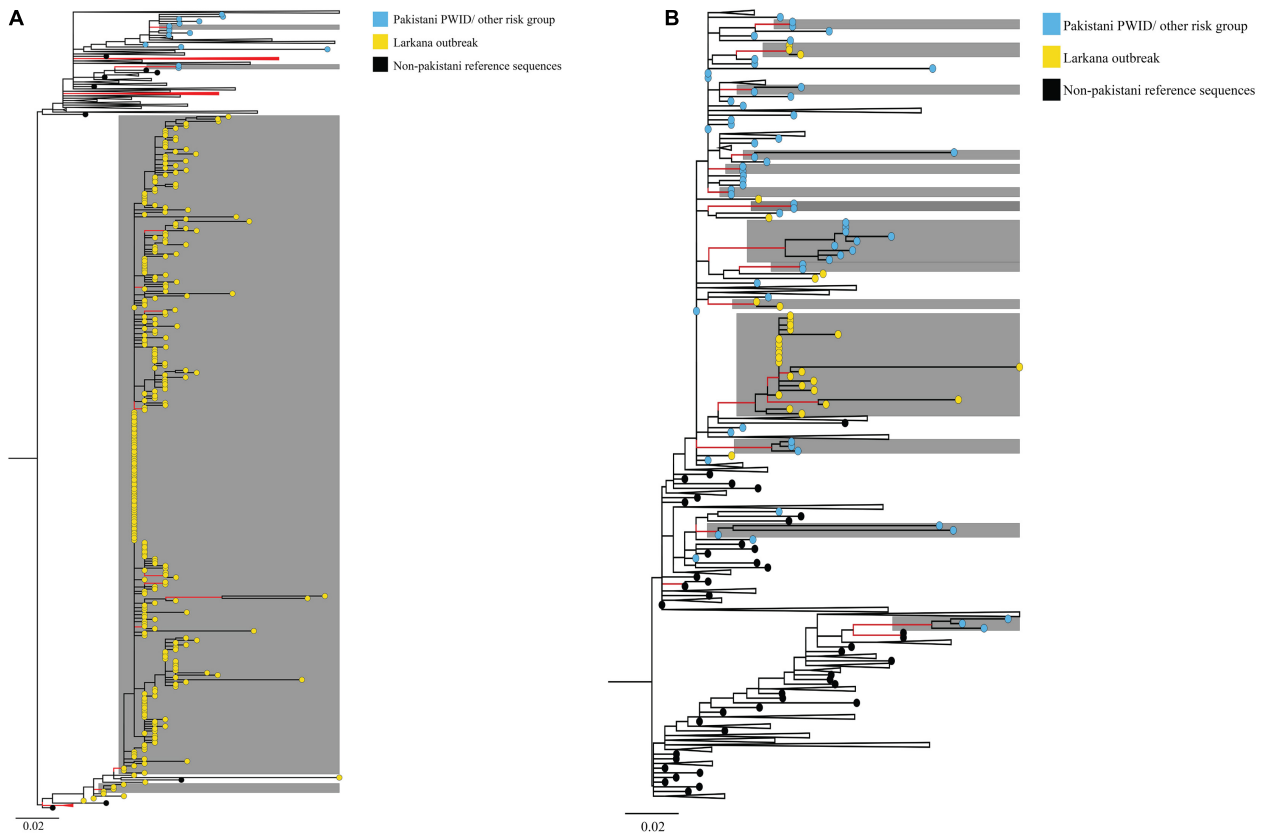


FIGURE 2 | Maximum-likelihood trees summarising clustering patterns of different risk groups in Pakistani: Trees represent **(A)** CRF02_AG, and **(B)** sub-subtype A1 transmission clusters, respectively. Each phylogeny is rooted at the midpoint, and branches are arranged in increasing node order. Branches with aLRT-SH support ≥0.9 are coloured red. Monophyletic clusters with aLRT-SH support ≥0.9 and which have ≥80% sequences from Pakistan are highlighted in grey. To enhance cluster visualisation, some branches containing either reference sequences or Pakistani sequences that are not part of clusters have been collapsed (shown as black triangles, with the recent end of the triangle indicating the latest sampling date). Branch tips within respective clusters are coloured as per cluster risk group (Bluish-green: MSM; Sky blue: Pakistani PWID/other risk groups IDU; Yellow: Larkana paediatric sequences; and Black: Non-Pakistani Reference sequences). Scale bars represent the genetic distance in substitutions per site in both phylogenies. As an overview, among CRF02_AG sequences, whereas PWID and other risk groups formed small clusters (size range 2–4 sequences per cluster), paediatric sequences from the Larkana outbreak formed one large cluster (*N* = 283 sequences). Likewise, among subtype A1 sequences, PWID and individuals from other risk groups formed several small clusters (size range, 2–9 sequences per cluster), whilst paediatric sequences from the Larkana outbreak formed one large cluster (*N* = 22 sequences).

and 2004 (95% HPD interval: 1998–2014) for sub-subtype A1 cluster (**Figure 3**).

Drug Resistance Mutation Analysis

Drug resistance mutation analysis was conducted on both ART-naïve (15.6%) and ART- experienced (84.4%) sequences from the outbreak. Among ART naïve individuals, 15 (30%) had drug resistance mutations (DRM); the most common of which were the Reverse Transcriptase (RT):E138A (8.0%) and RT:K219Q (8.0%) mutations. Among treatment-experienced individuals, the most common mutations were RT:E138A (12.92%), RT:K219Q (8.86%), and RT:K103N (6.64%). The DRMs RT:E138A and RT:K103N confer resistance against non-nucleoside reverse transcriptase inhibitors (NNRTI) rilpivirine and efavirenz, respectively, while RT:K219Q is associated with resistance against nucleoside reverse transcriptase inhibitors (NRTI) zidovudine. Similarly, DRM PI:N88D, associated with resistance against protease inhibitors (PI), such as atazanavir/ritonavir, and tipranavir/ritonavir was observed in two treatment-experienced participants, while DRMs PI:M46L, PI:D30N, PI:N83D, PI:K43T, PI:G73S, PI:L33F were seen in one treatment-experienced individual each (**Table 3**). No DRM against protease inhibitors was observed in ART-naïve individuals. Analysis also showed that 114 (42%) patients with DRMs belonged to the drug-experienced group, while, 15 (30%) belonged to the ART-naïve group (**Table 3**).

Next, sequences with any DRMs were analysed for clustering. Among the 78 sequences from the Larkana outbreak with prevalent drug resistance mutations, the DRM RT:K219Q was shared between five sequences whereof all clustered together in cluster_CRF02_AG_1 (**Tables 2,3**). The DRM RT:M184V was found in two sequences – both in cluster_CRF02_AG_1. The DRM RT:K103N was found among 21 sequences, whereof 19 sequences were found in cluster (cluster_CRF02_AG_1). The two remaining sequences were not part of any cluster. The DRM RT:E138A was found in 38 sequences of different subtypes

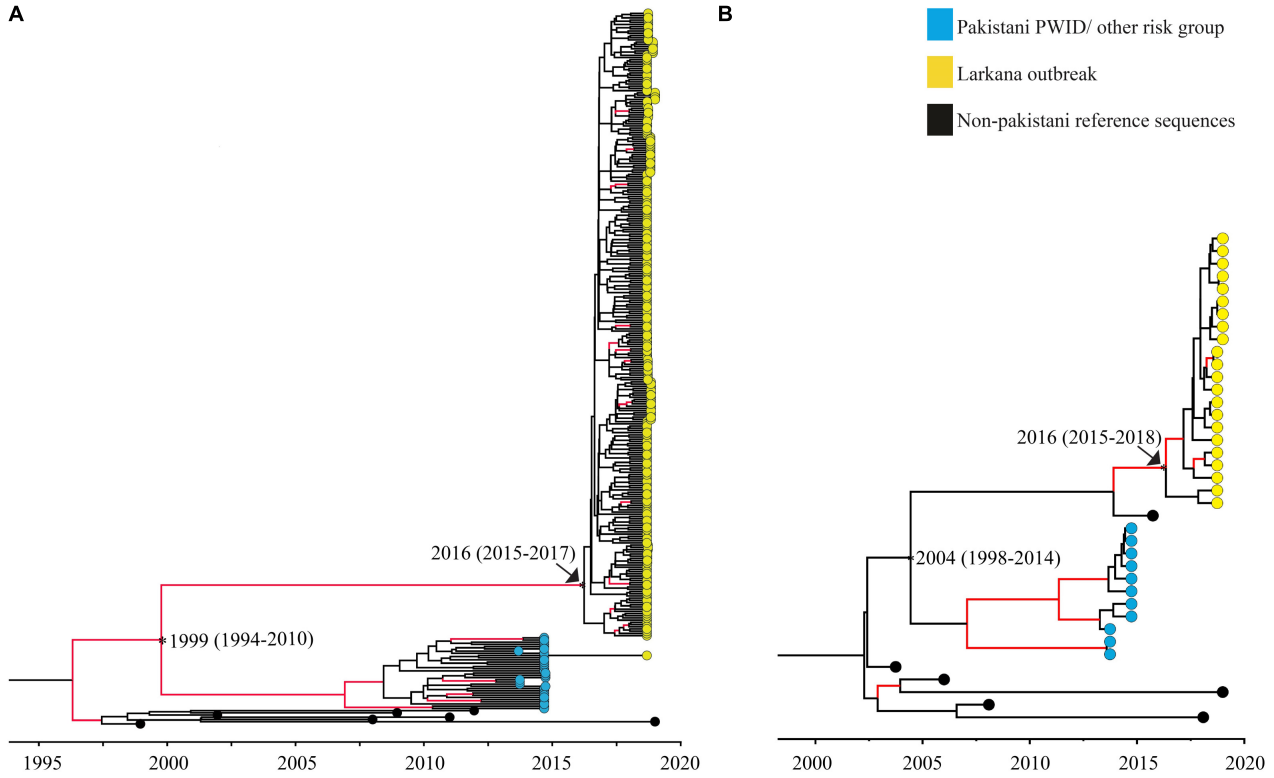


FIGURE 3 | Maximum clade credibility trees used to date clusters: Maximum clade credibility (MCC) trees used to determine the time to the most recent common ancestor of the Pakistani clusters. Trees represent **(A)** the large CRF02_AG outbreak cluster, and **(B)** the large sub-subtype A1 outbreak cluster, respectively. Nodes representing divergence (tMRCA; median and 95% HPD estimates) between the Larkana outbreak and prevailing epidemic sequences (representing PWID and other risk groups) are highlighted as asterisks only. Nodes representing the tMRCA of all outbreak sequences per sub-type/CRF have been highlighted using both an arrow and an asterisk. Branches with posterior value ≥0.9 are coloured red.

including 21 sequences in cluster_A1_1, three sequences in cluster A1_2, seven sequences in cluster_CRF02_AG_1, one sequence in cluster_CRF02AG_2, and six sequences that were not part of a cluster. The DRM RT:V179L was distributed among 10 sequences in cluster_CRF02AG_1, and two sequences in cluster_CRF02AG_2 (**Tables 2,3**).

DISCUSSION

In this study, we determined the HIV-1 subtype distribution, phylodynamics, and presence of HIV-1 drug-resistance mutations of the 2019 HIV-1 outbreak among children in Larkana, Pakistan. Seventeen distinct clusters were found among the Larkana outbreak sequences, indicating that HIV-1 was introduced from multiple sources rather than from a single source, as previously suggested (Arif, 2019; Siddiqui et al., 2020). A similar large-scale nosocomial HIV-1 outbreak was reported in Libya 1998–1999 (Visco-Comandini et al., 2002), where a monophyletic HIV-1 CRF02_AG cluster was identified among children visiting the El-Fatih Children's hospital in Benghazi. The HIV-1 transmission in the Libyan children was suggested to originate from contaminated intravenous injections (although not from blood or blood products) (Visco-Comandini et al., 2002). Similarly, the HIV-1 transmissions in the 2019 Larkana outbreak were strongly associated with visits to both public and private sector facilities, but not with a single healthcare facility, and with receipt of infusions, injections and blood transfusions (Mir et al., 2020b), implying transmission through poor infection control practices. Some of the children had HIV-1 positive mothers, raising the possibility of mother-to-child HIV-1 transmission. However, the possibility of vertical transmission could not be verified due to unavailability of maternal samples. Taken together, our results indicate that the Larkana outbreak was not the result of a single-source transmission from one health care practitioner, but may have resulted from through multiple sources at different health facilities. Moreover, and as previously suggested, our results indicate that poor infection prevention control is still present in Larkana (Altaf, 2018).

The HIV-1 subtype analysis showed a high prevalence of CRF02_AG and sub-subtype A1. This is consistent with previous reports indicating sub-subtype A1 as the dominant circulating subtype in Pakistan (Khan et al., 2018), whereas CRF02_AG has shown increasing prevalence more recently (Chen et al., 2016). A study reporting on the molecular epidemiology of HIV-1 in Pakistan suggested that sub-subtype A1 was introduced in Pakistan 1989 (95% HPD: 1984–1994) (Chen et al., 2016).

TABLE 3 | Classification of the Drug resistance mutations.

Drug		Mutation	Naïve, *N* (%) (*n* = 50)	Experienced, *N* (%) (*n* = 271)	Mutation classification	Drug associated with resistance
Reverse Transcriptase Inhibitors (RTIs)	NRTIs	K219Q	4 (8.00)	1 (0.37)	Major	AZT
		M184V	0 (0.00)	2 (0.74)	Major	ABC, 3TC
		M184I	0 (0.00)	1 (0.37)	Major	ABC, 3TC
		L210W	0 (0.00)	1 (0.37)	Major	AZT
		K70R	0 (0.00)	2 (0.74)	Major	ABC, TDF, AZT
		Y115F	0 (0.00)	1 (0.37)	Major	TDF, ABC
		A98G	0 (0.00)	3 (1.11)	Minor	NVP, EFV
	NNRTIs	K103N	2 (4.00)	19 (7.0)	Major	EFV, NVP
		K101E	0 (0.00)	1 (0.37)	Major/Minor	NVP, EFV
		E138A	4 (8.00)	35 (12.92)	Major/Minor	ETR, RPV
		E138K	1 (2.00)	0 (0)	Major/Minor	EFV, NVP
		E138G	0 (0.00)	1 (0.37)	Major/Minor	EFV, NVP
		V179L	2 (4.00)	10 (3.69)	Major	EFV, NVP
		V179F	1 (2.00)	0 (0)	Minor	EFV, NVP
		Y181C	0 (0.00)	1 (0.37)	Major	EFV, NVP
		G190A	0 (0.00)	2 (0.74)	Major/Minor	EFV, NVP
		V106I	1 (2.00)	2 (0.74)	Minor	NVP
		F227C	0 (0.00)	1 (0.37)	Major/Minor	EFV, NVP
Protease Inhibitors (PIs)	Major	L234I	0 (0.00)	1 (0.37)	Minor	DOR
		H221Y	0 (0.00)	1 (0.37)	Major	EFV, NVP
		M46L	0 (0.00)	1 (0.37)	Minor/Major	LPV/r
		D30N	0 (0.00)	1 (0.37)	Major	NFV
		N88D	0 (0.00)	2 (0.74)	Minor	ATV/r, SQV/r, NFV
	Minor	N83D	0 (0.00)	1 (0.37)	Minor/Major	ATV/r, TPV/r
		G73S	0 (0.00)	1 (0.37)	Minor	LPV/r
		L33F	0 (0.00)	1 (0.37)	Minor	LPV/r

The mutation classification column, indicates mutations that are classified as major or minor DRMs.
The last column shows the association of a mutation with resistance to a particular drug, as indicated in the Stanford drug resistance database and/or International AIDS Society (IAS) 2019 report, respectively, while bold drug names indicate the major resistance by IAS 2019.
The abbreviations of the antiretroviral drugs are as follows: NRTI; zidovudine (AZT), lamivudine (3TC), abacavir (ABC), tenofovir (TDF), etravirine (ETR), rilpivirine (RPV). NNRTI; efavirenz (EFV), nevirapine (NVP), doravirine (DOR). PI: lopinavir/ritonavir (LPV/r), nelfinavir (NFV), atazanavir/ritonavir (ATV/r), tipranavir/ritonavir (TPV/r), Saquinavir/ritonavir (SQV/r). % = percentage of total participants.

After this introduction, sub-subtype A1 disseminated rapidly to become the dominant HIV-1 strain (Chen et al., 2016). However, no information exists about the introduction of CRF02_AG in Pakistan, although it has been shown that the prevalence of HIV-1 CRF02_AG infections are increasing – especially in high-risk populations (Cholette et al., 2020). It is therefore possible that HIV-1 CRF02_AG become the dominant HIV-1 strain in Pakistan over time.

In our analysis, the tMRCA for the two large clusters were both estimated to 2016 (CRF02_AG: 95% HPD interval: 2015–2017; A1: 95% HPD interval: 2015–2018), suggesting ongoing HIV-1 transmissions several years prior to the 2019 outbreak. Interestingly, the estimated tMRCA of the two main clusters coincides with the time of the previously reported HIV-1 outbreak in Larkana in 2016, which also occurred in a nosocomial setting (Brenner et al., 2007; Siddiqui et al., 2020). Furthermore, the majority (80%) of children identified in the outbreak had stage 3 or 4 disease (the moderately and severely symptomatic stage, respectively, mostly associated with chronic to acute-chronic infection) (Weinberg and Kovarik, 2010; Altaf et al., 2016), indicating that some of the HIV-1 infections identified in the 2019 outbreak occurred a few years prior to 2019. Moreover, no reference sequence from the global HIV-1 epidemic was found in the Pakistani clusters, suggesting a localised HIV-1 epidemic. It is possible that the transmission may have been ongoing within healthcare settings, and the active screening programme implemented by the provincial AIDS Control Program in response to the initial diagnosis of HIV-1 in 14 children in 2019 identified these infections.

Phylogenetic analyses demonstrated no mixing between the outbreak sequences and sequences previously obtained from other high-risk groups in Pakistan. This was further supported by dating of the outbreak and other Pakistani clusters, indicating that the splitting time points between the outbreak sequences and that of other high-risk groups in Pakistan occurred between 1994 and 2012 (combined HPD interval). This suggests that HIV-1 most likely were introduced in Larkana between 10 and 27 years ago, and that the HIV-1 transmissions are now localised to certain regions of Larkana. It is possible that HIV-1 arose from nosocomial routes through

the widespread poor infection prevention control practices in healthcare settings and contaminated blood (Cotton and Rabie, 2020; Mir et al., 2020a).

DRM analysis showed the presence of multiple DRMs associated with resistance against reverse transcriptase inhibitors, some of which were shared among sequences in the identified clusters. The presence of shared DRMs may indicate transmission of drug-resistant strains in outbreak sequences. The most prevalence DRM identified among both ART-experienced and ART-naïve individuals was RT:E138A which confers a high-level resistance to NNRTIs such as Rilpivirine (Jeulin et al., 2014). A high prevalence of this mutation has been seen previously in ART-experienced PLWH in Pakistan (Shah et al., 2011) and when found in the ART-naïve population, it may be present due to transmission from non-compliant ART-experienced individuals. The presence of DRMs, especially associated with resistance to zidovudine and nevirapine, two out of three that are drugs part of the ART regimen given to the children, may lead to ART failure (correlating with failure to suppress the viral load), and increase the possibility that these individuals may acquire severe form of disease (Gupta-Wright et al., 2020). A second DRM common in the outbreak sequences was the RT:K219N mutation, a thymidine analogue mutation associated with potential low-level resistance against zidovudine (Rhee et al., 2003). Presence of DRMs and spread of strains containing the mutation may lead to first line medications becoming obsolete and present challenges due to limited availability of second-line medications.

The strengths of the study are a relatively good sample size, active collection of samples during the outbreak and a comprehensive phylogenetic and phylodynamic analysis of the Larkana outbreak to identify subtype distribution and evolutionary relationship between sequences. The limitations include amplification of a single gene (*pol*) only and short sequence length. While the parent case-control study showed that visits to both private and public sector health facilities and higher frequency of injections was associated with HIV-1 infection, we were unable to correlate the dyads and clusters with geographical data. The lack of samples of all biological mothers, where the mother was also HIV-positive, precluded investigation of the role of mother-to-child transmission. Future studies based on HIV-1 sequences sampled from different years (and HIV-1 risk groups) could shed light on the effectiveness of ART programs in Pakistan and provide an even more detailed picture of the Larkana outbreak.

In conclusion, our study findings showed that the Larkana paediatric outbreak did not originate from a single source and is likely a consequence of ongoing transmission within the high-risk groups of Larkana and introduced into the exposed individuals at risk of acquiring through poor infection prevention control practices in healthcare settings. Furthermore, the presence of multiple drug resistance mutations in the strains circulating in Larkana, especially to first-line ART drugs, is worrying as it limits treatment options. Large-scale transmission of resistant strains can hamper Pakistan's efforts to achieve the 90-90-90 goal (Maddali et al., 2016). These findings highlight not only the urgent need to improve blood safety and infection prevention control, but also the need for comprehensive molecular epidemiological studies and molecular surveillance to understand the distribution of different genotypes as well as origin, transmission, and drug resistance patterns.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, MN698251, MN698252, MN698253, MN698255, MN698256, MN698257, MN698258, MN698259, MN698260, MN698261, MN698262, MN698263, MN698264, MN752136, MN752137, and MT748850–MT749178.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Aga Khan University Ethics Review Committee (ERC #2019-1536-4200). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SA, RF, and FM conceived the study. SA, GN, DS, and WR performed the experiments. SA and GN wrote the first draft. SFM, AS, ShS, SM, SaS, SA, RF, PK, JE, and FM were involved in outbreak investigation. GN, DS, and AN cleaned the data and prepared the final datasheets. SA, GN, and DS have equal contributions in all experiments and data analysis. JE, RF, and FM had an equal contribution in project supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Higher Education Commission (grant no. 5217/Sindh/NRPU/R&D/HEC/2016), Pakistan Science Foundation [grant no. PSF/Res/S-AKU/Med (488)], and World Health Organization (grant 2019/969219-0). RF was funded by the Wellcome Trust through a Senior Fellowship in Clinical Science (206316/Z/17/Z). GN acknowledges support from the Sub-Saharan African Network for TB/HIV-1 Research Excellence (SANTHE), a DELTAS Africa Initiative (grant #DEL-15-006). The DELTAS Africa Initiative was an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (grant #107752/Z/15/Z) and the United Kingdom Government.

REFERENCES

Abi-Habib, M., and Masood, S. H. I. V. (2019). *Scandal Puts Focus on Pakistan's Health Care System*. New York, NY: The New York Times.

Altaf, A. (2018). Delays and gaps in HIV programmes in Pakistan. *Lancet HIV* 5, e678–e679.

Altaf, A., Pasha, S., Vermund, S. H., and Shah, S. A. (2016). A second major HIV outbreak in Larkana, Pakistan. *J. Pak. Med. Assoc.* 66, 1510–1511.

Arif, F. (2019). HIV crisis in Sindh, Pakistan: the tip of the iceberg. *Lancet Infect. Dis.* 19, 695–696. doi: 10.1016/s1473-3099(19)30265-8

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., and Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evolut.* 29, 2157–2167. doi: 10.1093/molbev/mss084

Baqi, S., Nabi, N., Hasan, S. N., Khan, A. J., Pasha, O., Kayani, N., et al. (1998). HIV antibody seroprevalence and associated risk factors in sex workers, drug users, and prisoners in Sindh, Pakistan. *J. Acquir. Immune Deficien. Syndrom. Hum. Retrovirol.* 18, 73–79. doi: 10.1097/00042560-199805010-00011

Brenner, B. G., Roger, M., Routy, J. P., Moisi, D., Ntemgwa, M., Matte, C., et al. (2007). High rates of forward transmission events after acute/early HIV-1 infection. *J. Infect. Dis.* 195, 951–959. doi: 10.1086/512088

Chen, Y., Hora, B., DeMarco, T., Shah, S. A., Ahmed, M., Sanchez, A. M., et al. (2016). Fast dissemination of new HIV-1 CRF02/A1 recombinants in Pakistan. *PLoS One* 11:e0167839. doi: 10.1186/s12864-018-5380-8

Cholette, F., Joy, J., Pelcat, Y., Thompson, L. H., Pilon, R., Ho, J., et al. (2020). HIV-1 phylodynamic analysis among people who inject drugs in Pakistan correlates with trends in illicit opioid trade. *PLoS One* 15:e0237560. doi: 10.1371/journal.pone.0237560

Cotton, M. F., and Rabie, H. (2020). HIV outbreak in children in Pakistan: localised or more widespread? *Lancet Infect. Dis.* 20:269. doi: 10.1016/s1473-3099(19)30746-7

Dawson, L., Benbow, N., Fletcher, F. E., Kassaye, S., Killelea, A., Latham, S. R., et al. (2020). Addressing Ethical Challenges in US-Based HIV Phylogenetic Research. *J. Infect. Dis.* 222, 1997–2006. doi: 10.1093/infdis/jiaa107

de Mendoza, C. (2019). UNAIDS Update Global HIV Numbers. *AIDS Rev.* 21, 170–171.

Debyser, Z., Van Wijngaerden, E., Van Laethem, K., Beuselinck, K., Reynders, M., De Clercq, E., et al. (1998). Failure to quantify viral load with two of the three commercial methods in a pregnant woman harboring an HIV type 1 subtype G strain. *AIDS Res. Hum. Retroviruses* 14, 453–459. doi: 10.1089/aid.1998.14.453

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evolut.* 22, 1185–1192. doi: 10.1093/molbev/msi103

Esbjörnsson, J., Mild, M., Audelin, A., Fonager, J., Skar, H., Bruun Jørgensen, L., et al. (2016). HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. *Virus Evolut.* 2:vew010. doi: 10.1093/ve/vew010

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evolut.* 30, 713–724. doi: 10.1093/molbev/mss265

Gill, M. S., Lemey, P., Suchard, M. A., Rambaut, A., and Baele, G. (2020). Online Bayesian Phylodynamic Inference in BEAST with Application to Epidemic Reconstruction. *Mol. Biol. Evol.* 37, 1832–1842. doi: 10.1093/molbev/msaa047

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systemat. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Gupta, S., Taylor, T., Patterson, A., Liang, B., Bullard, J., Sandstrom, P., et al. (2017). A Robust PCR Protocol for HIV Drug Resistance Testing on Low-Level Viremia Samples. *Biomed. Res. Int.* 2017:4979252.

Gupta-Wright, A., Fielding, K., van Oosterhout, J. J., Alufandika, M., Grint, D. J., Chimbayo, E., et al. (2020). Virological failure, HIV-1 drug resistance, and early mortality in adults admitted to hospital in Malawi: an observational cohort study. *Lancet HIV* 7, e620–e628.

Hasan, Z., Shah, S., Hasan, R., Rao, S., Ahmed, M., Stone, M., et al. (2018). Late diagnosis of human immunodeficiency virus infections in high-risk groups in Karachi, Pakistan. *Int. J. STD AIDS* 2018:956462418785264.

Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J., and Esbjörnsson, J. (2017). Defining HIV-1 transmission clusters based on sequence data. *AIDS* 31:1211. doi: 10.1097/qad.0000000000001470

Jeulin, H., Foissac, M., Boyer, L., Agrinier, N., Perrier, P., Kennel, A., et al. (2014). Real-life rilpivirine resistance and potential emergence of an E138A-positive HIV strain in north-eastern France. *J. Antimicrob. Chemother.* 69, 3095–3102. doi: 10.1093/jac/dku256

Khan, S., Zahid, M., Qureshi, M. A., Mughal, M. N., and Ujjan, I. D. (2018). HIV-1 genetic diversity, geographical linkages and antiretroviral drug resistance among individuals from Pakistan. *Archiv. Virol.* 163, 33–40. doi: 10.1007/s00705-017-3564-1

Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J., and Wertheim, J. O. (2018). HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Mol. Biol. Evol.* 35, 1812–1819. doi: 10.1093/molbev/msy016

Kuiken, C., Korber, B., and Shafer, R. W. (2003). HIV sequence databases. *AIDS Rev.* 5, 52–61.

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Maddali, M. V., Gupta, A., and Shah, M. (2016). Epidemiological impact of achieving UNAIDS 90-90-90 targets for HIV care in India: a modelling study. *BMJ Open* 6:e011914. doi: 10.1136/bmjopen-2016-011914

Mir, F., Mahmood, F., Siddiqui, A. R., Baqi, S., Abidi, S. H., Kazi, A. M., et al. (2020a). HIV infection predominantly affecting children in Sindh, Pakistan, 2019: a cross-sectional study of an outbreak. *Lancet Infect. Dis.* 20, 362–370. doi: 10.1016/s1473-3099(19)30743-1

Mir, F., Nathwani, A. A., Simms, V., Abidi, S. H., Siddiqui, A. R., Hotwani, A., et al. (2020b). Investigation of an extensive outbreak of HIV infection among children in Sindh, Pakistan: a case-control study. *Lancet HIV* 10:e036723. doi: 10.1136/bmjopen-2019-036723

Nazziwa, J., Faria, N. R., Chaplin, B., Rawizza, H., Kanki, P., Dakum, P., et al. (2020). Characterisation of HIV-1 Molecular Epidemiology in Nigeria: Origin, Diversity, Demography and Geographic Spread. *Sci. Rep.* 10:3468.

Nduva, G. M., Hassan, A. S., Nazziwa, J., Graham, S. M., Esbjörnsson, J., and Sanders, E. J. (2020). HIV-1 Transmission Patterns Within and Between Risk Groups in Coastal Kenya. *Sci. Rep.* 10:6775.

Raees, M. A., Abidi, S. H., Ali, W., Khanani, M. R., and Ali, S. (2013). HIV among women and children in Pakistan. *Trends Microbiol.* 21, 213–214. doi: 10.1016/j.tim.2012.12.005

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolut.* 2:vew007. doi: 10.1093/ve/vew007

Rhee, S. Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 31, 298–303. doi: 10.1093/nar/gkg100

Romero-Severson, E., Nasir, A., and Leitner, T. (2020). What Should Health Departments Do with HIV Sequence Data? *Viruses* 12:1018.

Shah, S., Xing, H., Altaf, A., Chen, B., Liao, L., Jia, Y., et al. (2011). Antiretroviral drug resistance mutations among treated and treatment-naive patients in Pakistan: diversity of the HIV type 1 pol gene in Pakistan. *AIDS Res. Hum. Retroviruses* 27, 1277–1282. doi: 10.1089/aid.2010.0324

Siddiqui, A. R., Ali Nathwani, A., Abidi, S. H., Mahmood, S. F., Azam, I., Sawani, S., et al. (2020). Investigation of an extensive outbreak of HIV infection among children in Sindh, Pakistan: protocol for a matched case-control study. *BMJ Open* 10:e036723. doi: 10.1136/bmjopen-2019-036723

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data

integration using BEAST 1.10. *Virus Evolut.* 4:vey016. doi: 10.1093/ve/vey016

Tariq, U., Parveen, A., Akhtar, F., Mahmood, F., Ali, S., and Abidi, S. H. (2018). Emergence of circulating recombinant form 56_cpx in Pakistan. *AIDS Res. Hum. Retroviruses* 34, 1002–1004. doi: 10.1089/aid.2018.0128

Taylor, B. S., Sobieszczyk, M. E., McCutchan, F. E., and Hammer, S. M. (2008). The challenge of HIV-1 subtype diversity. *N. Engl. J. Med.* 358, 1590–1602. doi: 10.1056/nejmra0706737

van Zyl, G., Bale, M. J., and Kearney, M. F. (2018). HIV evolution and diversity in ART-treated patients. *Retrovirology* 15:14.

Visco-Comandini, U., Cappiello, G., Liuzzi, G., Tozzi, V., Anzidei, G., Abbate, I., et al. (2002). Monophyletic HIV type 1 CRF02-AG in a nosocomial outbreak in Benghazi, Libya. *AIDS Res. Hum. Retroviruses* 18, 727–732. doi: 10.1089/088922202760072366

Waheed, Y., and Waheed, H. (2017). Pakistan needs to speed up its human immunodeficiency virus control strategy to achieve targets in fast-track acquired immune deficiency syndrome response. *World J. Virol.* 6, 46–48. doi: 10.5501/wjv.v6.i2.46

Weinberg, J. L., and Kovarik, C. L. (2010). The WHO Clinical Staging System for HIV/AIDS. *Virtual Mentor.* 12, 202–206. doi: 10.1001/virtualmentor.2010.12.3.cprl1-1003

Wensing, A. M., Calvez, V., Ceccherini-Silberstein, F., Charpentier, C., Günthard, H. F., Paredes, R., et al. (2019). 2019 update of the drug resistance mutations in HIV-1. *Topics Antiviral Med.* 27:111.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Abidi, Nduva, Siddiqui, Rafaqat, Mahmood, Siddiqui, Nathwani, Hotwani, Shah, Memon, Sheikh, Khan, Esbjörnsson, Ferrand and Mir. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

viruses

MDPI

Review

The Role of Phylogenetics in Discerning HIV-1 Mixing among Vulnerable Populations and Geographic Regions in Sub-Saharan Africa: A Systematic Review

George M. Nduva [1,2,†], Jamirah Nazziwa [1,†], Amin S. Hassan [1,2], Eduard J. Sanders [2,3] and Joakim Esbjörnsson [1,3,*]

[1] Department of Translational Medicine, Lund University, 205 02 Malmö, Sweden; george.makau_nduva@med.lu.se (G.M.N.); jamirah.nazziwa@med.lu.se (J.N.); ahassan@kemri-wellcome.org (A.S.H.)
[2] Kenya Medical Research Institute (KEMRI)-Wellcome Trust Research Programme, Kilifi 80108, Kenya; ESanders@kemri-wellcome.org
[3] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, The University of Oxford, Oxford OX1 2JD, UK
* Correspondence: joakim.esbjornsson@med.lu.se
† Authors with equal contribution.

Abstract: To reduce global HIV-1 incidence, there is a need to understand and disentangle HIV-1 transmission dynamics and to determine the geographic areas and populations that act as hubs or drivers of HIV-1 spread. In Sub-Saharan Africa (sSA), the region with the highest HIV-1 burden, information about such transmission dynamics is sparse. Phylogenetic inference is a powerful method for the study of HIV-1 transmission networks and source attribution. In this review, we assessed available phylogenetic data on mixing between HIV-1 hotspots (geographic areas and populations with high HIV-1 incidence and prevalence) and areas or populations with lower HIV-1 burden in sSA. We searched PubMed and identified and reviewed 64 studies on HIV-1 transmission dynamics within and between risk groups and geographic locations in sSA (published 1995–2021). We describe HIV-1 transmission from both a geographic and a risk group perspective in sSA. Finally, we discuss the challenges facing phylogenetic inference in mixed epidemics in sSA and offer our perspectives and potential solutions to the identified challenges.

Keywords: HIV-1; phylogenetics; mixed epidemics; Sub-Saharan Africa; transmission dynamics

check for updates

Citation: Nduva, G.M.; Nazziwa, J.; Hassan, A.S.; Sanders, E.J.; Esbjörnsson, J. The Role of Phylogenetics in Discerning HIV-1 Mixing among Vulnerable Populations and Geographic Regions in Sub-Saharan Africa: A Systematic Review. *Viruses* **2021**, *13*, 1174. https://doi.org/10.3390/v13061174

Academic Editor: Bluma G. Brenner

Received: 25 May 2021
Accepted: 10 June 2021
Published: 19 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Molecular phylogenetic approaches have evolved into powerful tools in understanding pathogens and how they cause disease in human populations [1]. Based on genetic relatedness between pathogen strains, these studies have been coupled with epidemiological data to decipher transmission events in infected hosts [2]. This approach has several applications and has been used to understand the geographic distribution of a large number of pathogens (e.g., reconstructing the 2009 global spread of human influenza A H1N1 pandemic and, more recently, characterising the emergence and global spread of SARS-CoV-2) [3–6]. Phylogenetics and phylodynamics have also been used to reconstruct and date the emergence and early spread of HIV-1, to assess epidemic growth dynamics, to determine HIV-1 genetic diversity and the prevalence of antiretroviral resistance mutations, to infer putative transmission events, and to determine evolutionary rates and spread of specific HIV-1 strains [7–15]. In recent years, the phylogenetic analysis of HIV-1 sequences from individuals with known risk behaviour and/or geographic location has become a powerful tool to identify sources of infections that potentially could be targeted to reduce HIV-1 incidence [13,16–22]. However, most of these studies have been conducted in well-resourced countries with HIV-1 epidemics that are likely to differ in transmission dynamics

compared with epidemics in Sub-Saharan Africa (sSA). For example, the HIV-1 epidemic in North America and Europe is concentrated amongst key populations—defined by UNAIDS as men who have sex with men (MSM), female and male sex workers (FSW/MSW), transgender people, people who inject drugs (PWID), and prisoners and other incarcerated people [23]. In contrast, the epidemic in sSA is mostly spread out among heterosexuals (HET, presumed heterosexuals) who have lower HIV-1 incidence and prevalence compared to key populations [24–26]. The term "general population" has been used where risk assessment data are not available for HET in sSA [27]. For ease in comparison between studies in this review, the term HET is used to refer to populations not belonging to either HIV-1 key populations or HIV-1 vulnerable populations (i.e., adolescent girls in sSA, orphans, street children, people with disabilities, and migrant or mobile workers as defined by UNAIDS (including miners, fishing communities, and long-distance truck drivers)) [23]. Additionally, the HIV-1 epidemic in sSA has revealed extensive geographic heterogeneity with HIV-1 hotspots (i.e., geographic areas and HIV-1 key populations with high HIV-1 incidence and prevalence) [25].

Targeting HIV-1 control strategies to HIV-1 hotspots has been proposed as a feasible approach to reduce the global HIV-1 incidence [28–30]. However, targeting likely needs to be guided by an in-depth understanding of the molecular epidemiology and drivers of local epidemics [31]. Existing reviews have summarised phylogenetic studies of transmission in concentrated epidemics and have offered perspectives on how transmission dynamics in the mixed epidemics of sSA could be assessed by phylogenetics [32]. In a review article from 1999, Dennis et al. reported high levels of both new and existing infections in high-risk populations in Kenya, South Africa, and Uganda [32]. In another review article from the African context, the PANGEA (Phylogenetics Additionally, Networks for Generalized Epidemics in Africa) consortium proposed to use phylogenetics to identify characteristics of individuals or groups most likely to be at risk of infection or at risk of infecting others [24]. However, to date, no published reviews have explicitly assessed the role of phylogenetics in discerning the mixing between geographies and populations in sSA. In this review, we provide an overview of the contribution of phylogenetic inference in dissecting HIV-1 mixing between geographic areas with varying HIV-1 prevalence, as well as HIV-1 mixing between key populations and HET in sSA.

2. Materials and Methods

2.1. Systematic Literature Review

2.1.1. Information Sources

An exhaustive search of the PubMed database (https://pubmed.ncbi.nlm.nih.gov/ (accessed on 12 March 2021)) was carried out by analysing peer-reviewed research articles on HIV-1 phylodynamics in sSA published in English in 1995–2021. Review articles, book chapters, editorials, and articles published in other languages were excluded from the search.

2.1.2. Search Strategy

First, we determined keywords and MeSH terms that could be used to identify research articles where phylogenetic approaches have been used to understand HIV-1 transmission in sSA. The MeSH terms (HIV-1) AND (Africa) were used to select HIV-1 articles from African countries. The keywords "phylogenetic analysis" OR "phylodynamics" OR "evolution" OR "phylogeny" OR "molecular epidemiology" OR "transmission" were used to widen the scope and to ensure that all relevant research articles were included. Filters on the year of publication, language, and article type were applied to refine the search.

2.1.3. Selection Process

Two investigators carried out the selection process independently. The articles were manually screened, first by title, then by abstract, to assess relevance based on our eligibility criteria (i.e., description of HIV-1 mixing within and between geographic regions and risk

groups). Any discordance between the two independent reviewers on the eligibility of articles was resolved through discussions for a consensus.

2.1.4. Data Extraction and Data Analysis

Shortlisted articles were imported into EndNote X8 (Clarivate, Philadelphia, PA, USA) for further management and to compile the information presented in this review.

3. Results

Based on a literature search done on 12 March 2021, 2722 articles were identified. Among these, 357 articles were not in English or involved nonhuman subjects, 2000 were ineligible by title review, 85 were ineligible by abstract review, and 216 were ineligible by a full-text review as they did not address HIV-1 transmission dynamics from a phylogenetic perspective (Figure 1). Sixty-four articles were considered eligible for full-text review, including 29 articles assessing geographic dispersion (Table 1) and 35 assessing HIV-1 mixing between HIV-1 populations in sSA.

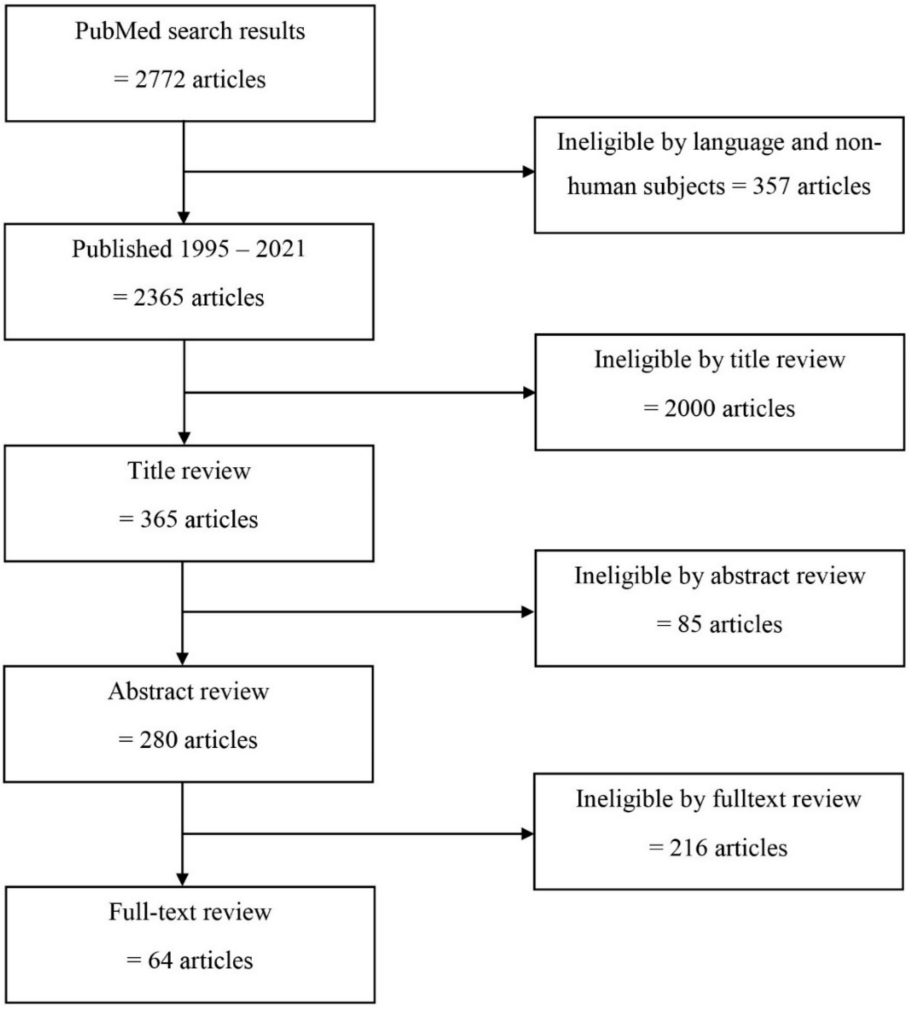


Figure 1. Study flowchart. Overview of the inclusion and exclusion of articles assessed in this review.

Table 1. Summary of HIV-1 phylogenetic studies in sSA from geographical context.

Country	HIV-1 Subtype	Estimated Date of Introduction	Summary of the Main Findings	PMID [1]
Central and West African countries				
Angola	F1	1958 (1934–1973)	Spread from DRC, derived from a single founder event. "Pure" F1 variants are most common in Angola.	19386115
	F1	1983 (1978–1989)	The Angolan civil war was associated with a wave of emigration and a phase of negative migratory outflow during 1960–1980.	22484759
	C	1978 (1973–1985) 1979 (1973–1985) 1983 (1977–1990) 1990 (1982–1997) 1994 (1989–1998) 2005 (2002–2008)	HIV-1 subtype C epidemic in Angola originated from multiple independent introductions from Burundi, Zambia, Zimbabwe, and South Africa. The civil war (1974–2002) may have contributed to the emergence of the HIV-1 epidemic in Angola.	22634597
	J, H	Not available	HIV-1 subtypes J and H seem to have been present in Angola since at least 1993.	27098898
	Group M	1978 (1975–1985)	The majority of sequences sampled in 2008–2010 in Luanda clustered together which is consistent with a locally fuelled epidemic.	25479241
Cameroon	CRF02_AG	1973 (1972–1975)	Two distinct lineages of CRF02_AG seem to have ignited in the urban centre of Cameroon. Ethnographic data suggests that well-supported HIV-1 migration was related to chance exportation events rather than by sustained human migratory flows.	21565285
	CRF02_AG	1976 (1966–1984) 1976 (1968–1986) 1979 (1953–1989)	Three monophyletic variants were identified and emerged in the mid-1970's and spread slowly over 30 years. Continuous exchange of HIV-1 strains between Cameroon and other African countries.	21453131
DRC [2]	A1, C, D	The 1960s	HIV-1 subtype C origin was estimated to originate in Mbuji-Mayi in the 1950s and subtypes A1, D originated in Kinshasa. The earliest dispersal events of subtype C occurred in a mining region close to Mbuji-Mayi and Lubumbashi. Subtype C spread at least three-fold faster than other subtypes circulating in Central and East Africa.	31809523
DRC [2], RC [3]	Group M	1920 (1909–1930)	Kinshasa estimated to be the origin of the HIV-1 group M pandemic. HIV-1 spread to Brazzaville in the Republic of the Congo, and Lubumbashi and Mbuji-Mayi in the 1930s, which were better connected to Kinshasa, indicating a critical role of mobility networks in the early spread and establishment of the HIV-1 epidemic from the epicentre.	25278604
DRC [2], RC [3]			General Eastward and Southward trends in the spread of HIV-1 from the Kinshasa–Brazzaville and the Pointe-Noire areas to other population centres.	27798403

Table 1. *Cont.*

Country	HIV-1 Subtype	Estimated Date of Introduction	Summary of the Main Findings	PMID [1]
Guinea-Bissau	CRF02_AG	1981 (1974–1986) 1976 (1968–1982) 1980 (1974–1984)	Multiple introductions of CRF02_AG 1976–1981, and a single introduction of sub-subtype A3 in 1979 (median estimates). HIV-1 was introduced into the urban centre (the Capital Bissau) from where it spread to rural areas.	21365013
	A3	1979 (1972–1984) 1981 (1975–1985) 1979 (1960–1988)		
Nigeria	G	1975 (1969–1982)	Urban areas (Abuja and Lagos) were the major hubs of HIV-1 transmission in Nigeria. HIV-1 first emerged and expanded within large urban centres before migrating to smaller rural areas.	32103028
	CRF02_AG	1963 (1948–1974) 1970 (1960–1980) 1960 (1947–1974)		
	CRF43_02G	1971 (1952–1983)		
East and Southern African countries				
Botswana	C	1996–2002	Presence of multiple phylogenetically distinct HIV-1 subtype C variants (subepidemics) circulating in Mochudi with limited lifespans and temporal dominance. None of the sequences from a rural community of Mochudi clustered with non-Botswana sequences.	26616041 24349005
Ethiopia	C	1965 (1959–1973)	Reconstruction of the epidemic history in Ethiopia revealed that subtype C likely originated from a single lineage in the late 1960s.	20539092
	C	1980	Evidence of clustering between Gondar sequences and sequences from East Africa.	30304061
Kenya	A1	1985–2012	Kilifi sequences clustered closely with sequences from Kenya and other parts of Africa, including West Africa. HIV-1 has been introduced in coastal Kenya multiple times.	32317722
South Africa	C	1960 (1956–1964)	Johannesburg was identified as the hub of HIV-1 dissemination in South Africa. The central region of KwaZulu-Natal was identified as the most likely ancestral location for HIV-1transmission in South Africa for 2 of 14 variants.	26574165
	C	1979–1992	The HIV-1 epidemic in South Africa is suggested to have multiple, parallel subepidemics spreading in the country at the same time.	30804361
	C	1990–2000	Early HIV-1 epidemic dynamics in KwaZulu-Natal were largely driven by external introductions.	30555720
Uganda	A1	1960 (1950–1968)	Ugandan epidemics originated in rural Southwestern Uganda with subsequent spread to other locations without any substantial HIV-1 introductions into this location suggesting that emerging infections from this low-incidence location are mostly from within the region.	25724670
	D	1973 (1970–1977)		33182587

Table 1. *Cont.*

Country	HIV-1 Subtype	Estimated Date of Introduction	Summary of the Main Findings	PMID [1]
Beyond borders				
West and Central Africa	CRF02_AG	1980 (1978–1981)	CRF02_AG originated from Cameroon from where it spread to other Central and West African countries.	27063411
West and Central Africa	CRF02_AG	1967 (1961–1974) West African	Five different CRF02_AG variants, four of which were restricted to Cameroon and one that grew out into West Africa.	27180893
West and Central Africa	CRF11_cpx	1957 (1950–1966)	Cameroon as the epicentre of dissemination of CRF11_cpx to Central African Republic, Chad, Gabon, and Equatorial Guinea.	27852214
West Africa	CRF06_cpx	1979 (1970–1985)	Burkina Faso was the hub of dissemination of CRF06_cpx to Mali, Nigeria, and the rest of western Central Africa.	23343915
West and Central Africa	G	1974 (1966–1981) 1979 (1973–1984)	Subtype G epidemic clustered into two clusters according to sequence location, i.e., either West or Central Africa. Sequences from West Africa were further subdivided into two large monophyletic clusters that were nested within the Central African variant.	24918930
East Africa	C	1962 (1942–1975)	Subtype C sequences from East Africa (Burundi, Ethiopia, Kenya, Tanzania, and Uganda) formed one large monophyletic cluster separate from sequences from Southern Africa.	22848653 29884822
East Africa	A1 D	1948 (1958–1967)	Both subtype A1 and subtype D were suggested to have spread exponentially during the 1970s.	19644346
East and Southern Africa	C	Not available	The largest number of HIV-1 introductions into South Africa came from Zambia, followed by Botswana, Malawi, and Zimbabwe between 1985 and 2000, a period of mass inward immigration from neighbouring countries into South Africa.	27421210
Zimbabwe	C	1972 (1979–1981)	Multiple cross-border independent introductions of subtype C HIV-1 into Zimbabwe between 1979 and 1981.	19770693

[1] PMID: PubMed identifier or PubMed unique identifier; [2] DRC: The Demographic Republic of Congo; [3] RC: The Republic of the Congo.

3.1. HIV-1 Molecular Transmission Networks in sSA: What Has Been Done from a Geographic Perspective?

The HIV-1 epidemic in sSA is driven by different HIV-1 subtypes, often geographically restricted [33,34]. Between 2010 and 2015, about 99% of the HIV-1 infections in Southern African countries and Ethiopia were subtype C, whereas the dominating HIV-1 strains in East Africa were sub-subtype A1 and subtype D [35]. The epidemic in West Africa is mainly driven by the circulating recombinant form (CRF) 02_AG, sub-subtype A3, and subtype G [12,35,36]. In contrast, the HIV-1 epidemic in Central Africa is more complex and diverse, and most HIV-1 subtypes have been found in this region [35,37]. In this review, we grouped the assessment of HIV-1 transmissions in sSA into two geographic regions according to the UNAIDS classification (https://aidsinfo.unaids.org/ (accessed on 20 January 2021)): East and Southern Africa; and West and Central Africa (Figure 2).

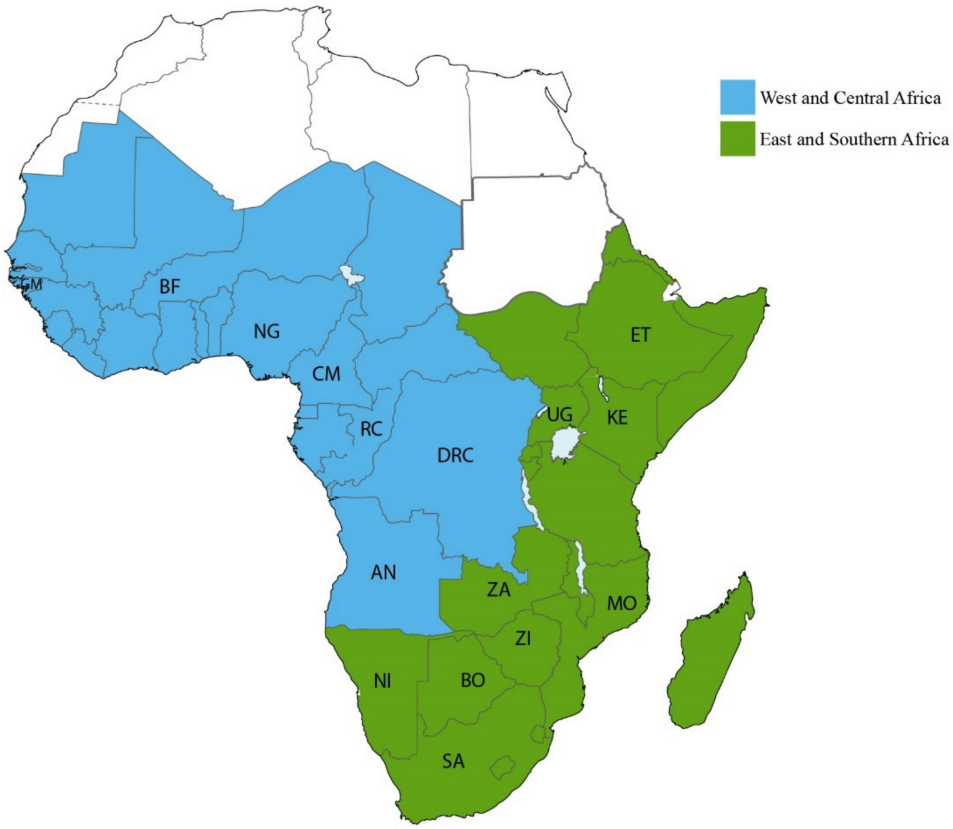


Figure 2. Subregions of Sub-Saharan Africa. A map showing different subregions of Sub-Saharan Africa as defined by UNAIDS. Countries belonging to Central and West Africa (N = 25) are coloured blue whereas countries belonging to Eastern and Southern Africa (N = 24) are coloured green. Where published information on HIV-1 transmission is available, the country code is included in the map. Countries belonging to Central and West Africa include Angola (AN), Benin, Burkina Faso, Cameroon (CM), Cape Verde, Chad, Central African Republic, Republic of the Congo (RC), Côte D'Ivoire, Democratic Republic of Congo (DRC), Equatorial Guinea, Gabon, The Gambia, Ghana, Guinea, Guinea-Bissau (GM), Liberia, Mali, Mauritania, Niger, Nigeria (NG), Saint Helena, Senegal, Sierra Leone, and Togo. Countries belonging to Eastern and Southern Africa include Burundi, Botswana (BO), Comoros, Djibouti, Ethiopia (ET), Eritrea, Kenya (KE), Lesotho, Madagascar, Malawi, Mauritius, Mozambique, Réunion, Namibia (NI), Rwanda, Seychelles, Somalia, Somaliland, Tanzania, South Africa (SA), Eswatini (former Swaziland), Uganda (UG), Zambia (ZA), and Zimbabwe (ZI).

3.1.1. HIV-1 Transmission in West and Central African Countries

Nigeria is the most populous country in sSA. A recent study by Nazziwa et al. used 1442 *pol* sequences (collected in 1999–2013) from four geopolitical zones in Nigeria (Southwest, North Central, Northeast, and Northwest) to reconstruct HIV-1 transmission dynamics [12]. Phylogeographic analyses suggested that HIV-1 first emerged and expanded within the large urban centres (Lagos and Abuja), before migrating to smaller and more rural areas. Abuja, the capital city of Nigeria, was estimated to be the geographical origin of both subtype G and CRF02_AG in Nigeria. In addition, the analysis indicated that one single introduction resulted in the main Nigerian subtype G epidemic (time to the most recent common ancestor, tMRCA; 1987). In contrast, the CRF02_AG had multiple introductions which expanded into larger subepidemics (tMRCAs; 1974, 1972 and 1961) [12]. Another study in Guinea-Bissau by Esbjörnsson et al. (based on 82 Guinean HIV-1 *env* sequences collected 1993–2008) found that the dominating HIV-1 strains were CRF02_AG and sub-subtype A3. In line with the study by Nazziwa et al., both subepidemics originated in the capital before dispersing out to smaller and more rural areas [36]. Interestingly, although the two HIV-1 strains were introduced into the country around the same time (median estimates 1976–1981), the phylogeographic analysis suggested that the CRF02_AG

strain started to migrate to more rural areas almost instantly after being introduced into the Capital Bissau. In contrast, sub-subtype A3 was estimated to have circulated within the capital for approximately 10 years before migrating to more rural areas of the country. The underlying reasons for this, however, remain to be determined.

In Cameroon, Véras et al. used 291 HIV-1 CRF02_AG *pol* sequences collected in 1996–2007 with geographic information system (GIS) data to assess HIV-1 transmission dynamics. Seventy percent of the sequences were found in three distinct clusters, suggesting several subepidemics with different origins [38]. Southern Cameroon has denser human mobility networks compared to the rest of the country; a large cluster comprising sequences from southern Cameroon was identified, suggesting that human mobility may play a role in increasing HIV-1 transmission. In another Cameroonian study, based on 336 HIV-1 *gag*, *pol*, and *env* sequences collected in 1996–2004, two HIV-1 CRF02_AG Cameroonian clusters were identified [39]. Interestingly, both clusters were estimated to have originated in Yaoundé, the capital of Cameroon, before spreading to the Littoral and West regions of Cameroon and remote areas in the South and East. In a study in the Democratic Republic of the Congo (DRC), Faria et al. used a phylogeographic approach to analyse 346 HIV-1 *pol* sequences (subtypes A1, C and D) collected in 2008 from four locations—the capital Kinshasa, Matadi (West DRC), Mbuji-Mayi (Central DRC), and Lubumbashi (South DRC) [9]. Mbuji-Mayi was suggested as the origin of the subtype C epidemic, whereas the origin for subtypes A1 and D was Kinshasa. The study also indicated that several group M HIV-1 strains had spread from the DRC to other countries. In another study, analysis of *env* C2V3 sequences collected at multiple sites in the DRC (from Bwamanda in North, Kisangani and Mbuji-Mayi in Central, the Capital Kinshasa, Lubumbashi and Likasi in South), and the Republic of the Congo (from the Capital city Brazzaville, and Porte-Noire in West) suggested that HIV-1 dispersed from Kinshasa to Brazzaville, as well as from Bwamanda and Kisangani [8,40]. The authors suggested that good transport connectivity and human mobility linked to mining activities may have been involved in the rapid expansion of HIV-1 spread between Kinshasa, Brazzaville, Lubumbashi, and Mbuji-Mayi [8].

The HIV-1 epidemic in Angola is one of the most diverse in sSA, and all HIV-1 group M subtypes and several CRFs have been identified here [37,41,42]. In a study by Bártolo et al., 364 HIV-1 *pol* sequences collected in 1993–2010 from Luanda and seven other provinces in Angola were analysed. The results indicated that 36% of the sequences formed relatively small Angolan clusters. Seventy-four percent of the sequences in the identified clusters were from Luanda, indicating extensive local transmission and much lower transmission (24% of the clusters) beyond the capital city of Luanda [42].

3.1.2. HIV-1 Transmission in East and Southern Africa Countries

The HIV-1 epidemic in Kenya is diverse and has had multiple and separate introductions [10,43,44]. Hue and colleagues performed a phylogeographic analysis based on 153 sequences collected in Kilifi county in 2008–2009 together with published Kenyan sequences to investigate how HIV-1 transmission in rural Coastal Kenya related to the region [43]. It was observed that 73% of the HIV-1 sub-subtype A1 sequences from Kilifi clustered with sequences from other areas in Kenya and that there was substantial clustering with strains from other East African countries, such as Uganda and Tanzania, possibly indicating HIV-1 transmission links between these countries. HIV-1 transmission in Uganda has been well studied, especially in rural Southwestern Uganda (which is suggested to be the geographic origin of HIV-1 sub-subtype A1 and subtype D in Uganda) [45]. To understand HIV-1 transmission dynamics in Uganda, Ssemwanga et al. used 3796 HIV-1 *pol* sequences collected between 2003 and 2015 from Southwestern, Central, and Eastern Uganda [46]. HIV-1 subtype A infections were more common in Central Uganda, whereas subtype D infections were more common in Southwestern Uganda. The study also found a high proportion of localized clustering among sequences from Southwestern Uganda and significant virus export from this region to other regions. However, no virus introductions

into this region were observed. In another study, Yebra et al. used 162 HIV-1 *pol* sequences collected in 2005–2010 from Kampala, Masaka, and Entebbe and 414 previously published *pol* sequences from Rakai, Kampala, and Entebbe and observed that HIV-1 subtype D initially spread from the rural Southwest, then to the Capital Kampala, before spreading to areas around Lake Victoria [45]. In Ethiopia, the HIV-1 epidemic is dominated by two phylogenetically distinct subtype C types—the Ethiopian HIV-1 C'-ET, and the East African HIV-1 C-EA [47]. Arimide et al. used 301 HIV-1 *pol* sequences collected in 2003–2013 from Gondar (Northern Ethiopia) to define and understand the transmission dynamics among these variants. The study showed that the C-EA sequences in Gondar clustered with sequences from other East African countries and that multiple introductions of the South African subtype C (C-SA) were observed in Gondar [47].

In Southern Africa, the mass migration of people into South Africa from neighbouring countries has been suggested to have an impact on the local HIV-1 epidemic [48,49]. To understand the transmission dynamics of HIV-1 within South Africa and its neighbouring countries, Wilkinson and colleagues analysed 15257 HIV-1 subtype C southern African sequences [50]. The analysis indicated that Johannesburg and KwaZulu-Natal were the main epicentres of HIV-1 dissemination in South Africa. Viruses from KwaZulu-Natal spread to the Northern regions close to the Mozambican and Swaziland borders, and to Johannesburg, whereas viruses from Johannesburg spread to KwaZulu-Natal, Kimberly, Bloemfontein, Mpumalanga, and Western and Eastern Cape. Another study quantified the contribution of local transmission and external introductions to the HIV-1 incidence specifically in KwaZulu-Natal [51]. Phylodynamic analysis of 1068 HIV-1 *pol* sequences collected in 2011–2014 in KwaZulu-Natal together with 11,289 subtype C sequences from Southern African countries revealed multiple HIV-1 introductions into KwaZulu-Natal from other locations in South Africa and neighbouring countries. The majority of the virus introductions in this study occurred in the early stages of the South African HIV-1 epidemic during the 1990s, where human movements played a role in driving the epidemic and sustaining high HIV-1 incidence in KwaZulu-Natal. In addition, 35% of new infections in KwaZulu-Natal were due to HIV-1 imports from other regions. To understand the structure of the local HIV-1 epidemic in periurban Botswana, Novitsky et al. analysed 2219 HIV-1 *env* sequences (785 sequences from Mochudi, 190 sequences from other locations in Botswana, and 1244 non-Botswana sequences) [52]. Close clustering of sequences originating from Mochudi suggested that the HIV-1 epidemic in Mochudi was dominated by locally circulating HIV-1 variants. Moreover, none of the Mochudi sequences clustered with non-Botswana sequences.

3.1.3. HIV-1 Transmission beyond Borders

Few studies have investigated the geographic mixing of HIV-1 between different African regions.

To shed light on the dissemination of HIV-1 CRF02_AG in Central and West Africa, Yebra et al. applied phylodynamic analysis to 1247 HIV-1 *env* and 1478 HIV-1 *pol* sequences collected 1984–2013 from 19 African countries. The analysis indicated that CRF02_AG originated from Cameroon from where it spread to other Central and West African countries [53]. To further characterise the CRF02_AG epidemic in West and Central Africa, Mir et al. used 2246 HIV-1 *pol* sequences collected 1990–2013 from 20 African countries [54]. The study indicated that the current CRF02_AG diversity resulted from the spread of a small number of founder strains from Central to West Africa in the period of 1960–1980. The study identified five different CRF02_AG variants, four of which were restricted to Cameroon and one that grew out into West Africa. In addition, other phylogeographic studies have indicated Cameroon as the epicentre of the dissemination of HIV-1 CRF11_cpx to Central African Republic, Chad, Gabon, and Equatorial Guinea. However, it has also been suggested that CRF06_cpx spread from Burkina Faso to Mali, Nigeria, and the rest of West-Central Africa [55,56].

A phylogeographic study of the dissemination routes of HIV-1 subtype G in West and Central Africa by Delatorre et al., using 305 HIV-1 *pol* sequences collected in 1992–2011 from 11 countries, showed that the African subtype G epidemic could be divided into two subepidemics according to sequence location, i.e., West and Central Africa [57]. Sequences from West Africa were further subdivided into two large monophyletic clusters that were nested within the Central African variant. One of the Western African variants emerged from Nigeria and spread to Benin, Cameroon, Equatorial Guinea, Ghana, and Senegal. The other West African variant emerged from Togo and/or Ghana from which it spread to Nigeria and then to Benin, Cameroon, Gabon, and Senegal [57].

To reconstruct the HIV-1 transmission dynamics of subtype C in East Africa, Delatorre et al. analysed 1981 *pol* sequences collected in 1990–2010 from 13 countries in Central, East, and Southern Africa [58]. Subtype C sequences from East Africa (Burundi, Ethiopia, Kenya, Tanzania, and Uganda) formed one large monophyletic cluster separate from sequences from Southern Africa. In addition to the East African C variant, another monophyletic cluster exclusive to Ethiopia was observed. The East Africa subtype C cluster disseminated from Burundi and later spread to other East African countries where local epidemics were established [58]. A later study including sequences collected in recent years (2013–2016) showed that most of the East African subtype C sequences still clustered into one monophyletic variant, consistent with strong interconnectivity between population centres across the East African region, which has likely fostered the rapid growth of the HIV-1 subtype A1, C, and D epidemic [59,60].

A comparative genetic analysis of HIV-1 subtypes A1, C, and D using 8701 *pol* sequences collected in 1996–2011 from DRC, Burundi, Kenya, Rwanda, Tanzania, and Uganda by Faria et al. indicated that subtypes A1 and D originated from DRC and that sequences from the same regions clustered closely together [9]. Additionally, 80% of total transmissions occurred within national borders and only 20% of transmissions were due to cross-border virus movements. Furthermore, Rwanda, DRC, and Tanzania were identified as the main exporters of subtype C in the Central and Eastern Africa region, whereas Uganda was the source of subtypes A1 and D.

To understand how human migration has influenced HIV-1 diversity and spread in Southern Africa, Wilkinson et al. performed a phylogeographic analysis of 11,289 sequences collected from DRC, Tanzania, Zambia, Malawi, Mozambique, Zimbabwe, Botswana, Namibia, Swaziland, Lesotho, and South Africa. The study showed that the high level of subtype C diversity in South Africa was linked to multiple HIV-1 introductions into the country [49]. Zambia, Botswana, Malawi, and Zimbabwe contributed to most of the HIV-1 introductions into South Africa between 1985 and 2000. However, South Africa also contributed to HIV-1 export to its neighbouring countries. HIV-1 mixing between Zimbabwe and other neighbouring countries (South Africa, Botswana, Zambia, Malawi, Mozambique, and Tanzania) has also been reported in a study by Dalai et al. [61]. Moreover, subtype C sequences from Southern and Central Africa have been shown to cluster closely together but separate from other subtype C sequences from other parts of the world, suggesting strong HIV-1 panmixia in Southern Africa [48,62].

3.1.4. Conclusion Phylogeographic Linkages in sSA

In summary, the HIV-1 epidemics in West and Central Africa seem to have emerged and expanded within urban areas before spreading to rural areas—possibly driven by human mobility [12,36,39,42]. In other instances, HIV-1 mixing between rural and urban locations, as well as across national borders, has also been observed [9,42,43,47]. Some HIV-1 subepidemics appear to be localized in specific communities where HIV-1 mixing with neighbouring communities is not observed [54]. In contrast, in other settings localized HIV-1 subepidemics serve as important sources of HIV-1 infection to neighbouring communities [47,48]. Furthermore, human migration linked to economic activities such as mining and fishing may contribute to increased HIV-1 transmission [9,49,63].

3.2. The Role of HIV-1 Key and Vulnerable Populations in Mixed HIV-1 Epidemics: A Risk Groups Perspective

The early HIV-1 epidemic in sSA was exclusively defined as heterosexual and involving FSW and long-distance truck drivers [64–66]. The role of other HIV-1 key populations such as MSM and PWID was not apparent and this, coupled with ethical-legal hurdles, led to the exclusion of these key populations from early HIV-1 responses in sSA [67–69]. HIV-1 key populations in sSA are strongly affected by legal and social stigma, where risk behaviour associated with these populations (e.g., same-sex behaviour) are often criminalized [70]. As a consequence, individuals in these populations often withhold risk information, which results in limited HIV-1 research involving key populations [27]. Additionally, there is evidence of overlapping sexual networks and phylogenetic linkages between HIV-1 key populations and HET, which may have implications for the dynamics of HIV-1 spread [44,65]. Figure 3 summarises HIV-1 prevalence estimates among HIV-1 key and vulnerable populations relative to HET in different regions of sSA. In general, HIV-1 key populations have higher HIV-1 prevalence compared to HET in all sSA countries (with the exception of MSM in Eswatini, Malawi, Botswana, and Guinea Bissau).

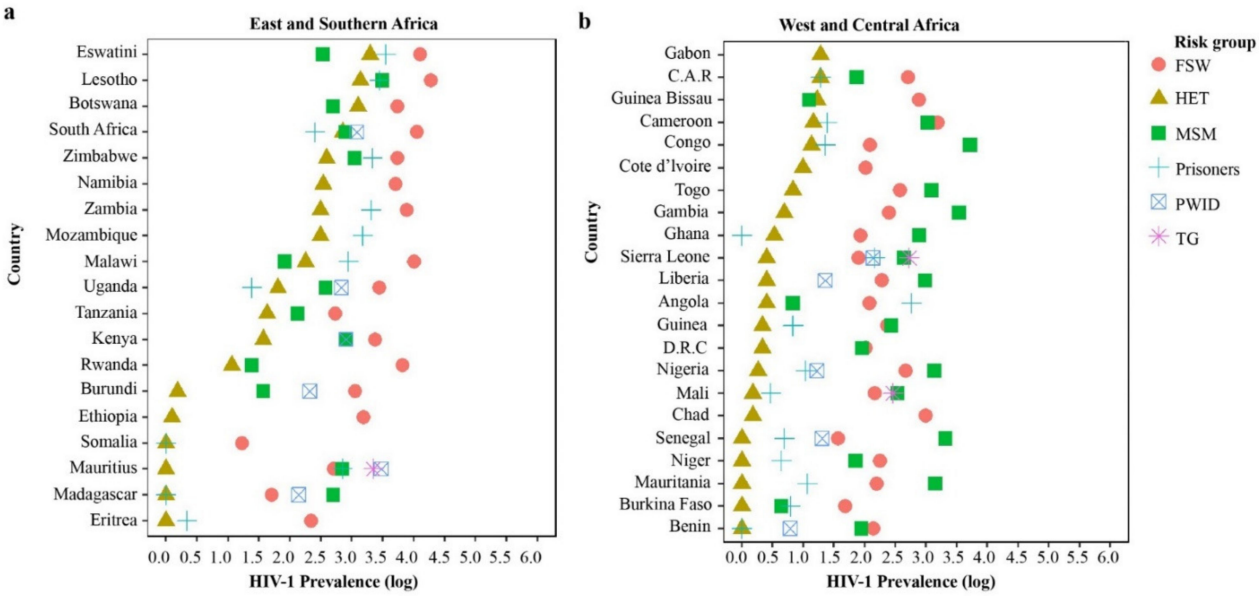


Figure 3. HIV-1 prevalence in different risk groups in sub-Saharan Africa (sSA). A comparison of national estimates of HIV-1 prevalence in the heterosexuals (HET) and among vulnerable populations in sSA as reported by UNAIDS in 2020 (https://aidsinfo.unaids.org/ (accessed on 20 January 2021)). East and Southern African (**a**), and West and Central African (**b**) regions were grouped together, respectively. The countries in each region were arranged in increasing HIV-1 prevalence among (HET), and HIV-1 prevalence data have been transformed into a log scale on the x-axis. Different risk groups are coloured as shown in the legend (Red: female sex workers; Brown: HET; Green: men having sex with men; Sky Blue: prisoners; Dark Blue: PWID; and Pink: transgender persons).

3.2.1. HIV-1 Phylogenetic Linkages Involving Heterosexual Transmission

Data on HIV-1 phylogenetic linkages involving HIV-1 key and vulnerable populations are not available in most sSA countries. Consequently, and albeit largely under sampled, most studies investigating HIV-1 networks have focused on HET transmission. In Botswana, a phylogenetic analysis of 1247 HIV-1 subtype C *env* sequences (collected in 2010–2013) by Novitsky et al. found 233 clusters, the majority of which were HET transmission pairs, and where the largest cluster involved 18 individuals [71]. This study was conducted in Mochudi, a periurban community, and it proposed tracking HIV-1 transmission clusters at the community level and extinguishing them, one by one, through targeted interventions. In a similar setting in South Africa, another analysis by Sivay et al. reported

partial transmission chains constructed among young women attending high school in rural South Africa (in the context of missed sampling among males in the community) and revealed a stable local epidemic with no evidence of super-spreading events or large networks [72]. In addition, this study also showed that recent HIV-1 transmissions may play a key role in driving the local HIV-1 epidemic. In Zambia, a phylogenetic analysis of 149 married couples by Trask et al. showed that the majority (87%) of couples were epidemiologically linked [73]. However, 13% of pairs in the study had distantly related viruses, suggesting possible extramarital HIV-1 transmission. In Rwanda, a phylogenetic analysis of men and high-risk women in the context of multiple heterosexual partnerships by Rusine et al. identified only three potentially linked transmission pairs [74]. In the DRC, a study by Rubio-Garrido et al. using 165 newly generated HIV-1 *pol* sequences representing adults and majority paediatric individuals found only four clusters, one of which had sequences from children with no epidemiological links, indicating under sampling in otherwise denser networks [75]. Small clusters have also been observed in Ethiopia by Arimide et al., where data on MSM and PWID is not available, but the epidemic among FSW and HET is acknowledged [47]. Overall, multiple studies in all geographic regions of Sub-Saharan Africa have characterised HIV-1 phylogenetic linkages involving heterosexual transmission. The majority of these studies identified only small clusters, highlighting challenges in sampling coverage which results in many missing links in otherwise large networks [76].

3.2.2. HIV-1 Phylogenetic Linkages among MSM

In the context of homosexual transmission, phylogenetic studies in East Africa have demonstrated extensive clustering among MSM [10,44,77,78]. Studies in Coastal Kenya, have demonstrated extensive clustering of HIV-1 *pol* sequences from men who have sex with men only (MSM only) and bisexual men, suggesting that bisexual MSM may link infections across different risk groups although such linkages may only be modest as observed in Coastal Kenya [10,44]. In West Africa, studies in Nigeria have observed clustering among MSM in a cohort involving a majority (62%) bisexual men [79,80]. Relatively large clusters (with up to 15 individuals per cluster) have been found in this cohort. Interestingly, 37% of bisexual men in this cohort were in clusters involving MSM [80]. In this cohort, clustering between newly infected MSM and previously diagnosed MSM has been reported, indicating ongoing transmission among MSM, the majority of whom were not in treatment and did not report consistent condom use. Elsewhere in the region, phylogenetic clustering analysis of 67 Senegalese MSM (of whom 80% reported to be married) identified 15 transmission clusters, three of which involved MSM from multiple regions in Senegal, indicating linked MSM networks with a wide geographic presence [81]. High numbers of MSM having female sex contacts and exclusive clustering among Senegalese MSM have also been reported by Ndiaye and colleagues [82]. Although cross-risk groups linkages between MSM and HET were not reported in either of these studies in West Africa, such mixing could be expected, to some extent, as has been observed in East Africa [44]. Overall, although MSM in the majority of cohorts in sSA often report being married or having female sex partners, phylogenetic evidence of HIV-1 transmission often reveals MSM exclusive clusters and only a few clusters involving HIV-1 sequences from MSM and HET, suggesting limited mixing.

3.2.3. HIV-1 Phylogenetic Linkages among PWID

Phylogenetic studies involving PWID in sSA are exceedingly rare and have only been reported at a subnational scale in Kenya [44,78]. Nduva et al. used 658 sequences to investigate HIV-1 phylogenetic linkages involving MSM, FSW, PWID, and HET in Coastal Kenya [44]. Whereas MSM, FSW, and HET were found in several small clusters (indicating introduction from multiple sources), the vast majority of PWID sequences were found in one large PWID-exclusive cluster suggesting introduction from one single source and long-term gradual spread within the PWID in Coastal Kenya. Phylodynamic analysis

of PWID sequences in this study suggested that HIV-1 infections had increased steadily among PWID since the date of origin in 1987. Additionally, unlike in previous studies (non-African) where PWID sequences clustered with exceptionally low genetic diversity, the genetic diversity among PWID in the Coastal Kenyan cluster was high [17,21,83]. The reason for this could be long times between infection and sampling dates and/or low sampling density among PWID in the region. Overall, studies so far suggest separate transmission for PWID with limited overlap between other key populations. However, more research is warranted as the molecular epidemiology of PWID in sSA is largely understudied.

3.3. Phylogenetic Analysis to Examine HIV-1 Mixing between Risk Groups

Very few studies in sSA have investigated HIV-1 linkages involving individuals belonging to different risk groups. In Southern Africa, Bártolo et al. assessed HIV-1 phylogenetic linkages in Angola using 364 HIV-1 *pol* sequences collected in 1993–2010 and identified 48 transmission clusters (size range: two to seven) [42]. More than half of the clustering sequences did not have risk group information. However, three clusters involving mixing between MSM and females were identified, suggesting HIV-1 genetic mixing between HET and MSM. In South Africa, Wilkinson and colleagues detected phylogenetic mixing between HET and MSM, where linkages involving two MSM (infected through homosexual contact) and an incarcerated man (infected in a prison setting) were found within a large cluster dominated by HET (including female individuals) [84]. HIV-1 mixing involving bisexual MSM and HET has also been reported in Cape Town, South Africa [85].

In West Africa, phylogenetic intermixing of HIV-1 variants between HET women and MSM has also been documented in Senegal, where sequences from HET females were found among MSM clusters [86]. Another study has reported on the intermixing of HIV-1 between MSM and HET in Togo [87]. The authors describe extensive clustering among 79 MSM, where at least 40% of MSM were found in recent transmission chains of two to seven sequences, and where almost half (49%) of MSM were found in one major CRF02_AG cluster, indicating infections within a close network. Additionally, in this study, a comparison of 950 published HIV-1 sequences from HET, perinatally infected infants, and MSM indicated HIV-1 mixing between MSM and HET because strains from infants and HET females were found among MSM-dominated clusters.

In East Africa, two studies in Kenya have reported limited mixing between key populations and HET [44,77]. Bezemer et al. found only one single transmission pair of an MSM and a known HET female partner in Coastal Kenya—indicating infrequent HIV-1 mixing between MSM and HET in Coastal Kenya [77]. A follow-up study by Nduva et al. used a larger sample size to study mixing between MSM, PWID, FSW, and HET in Coastal Kenya and found that only 7% of the clusters had MSM and HET sequences, indicating limited mixing between MSM and HET in Coastal Kenya [44].

In Uganda, phylogenetic clustering has been studied among Lake Victoria's fishing communities (considered an HIV-1 vulnerable population) [88–90], and HIV-1 mixing between fishing communities and HET residing in in-land regions has been reported [90,91]. Grabowski et al. showed that HIV-1 diversity is similar both within and between fishing communities and with HET in surrounding communities [91]. In a different study, phylodynamic analysis of sequences from FSW, fishing communities, and HET identified only a few small clusters of exclusively HET individuals [45]. Although the sample size and the sampling coverage were low, no mixing between risk groups was observed. However, in the context of missed sampling of sex partners of FSW, a study in Kampala observed clustering among FSW, suggesting infection from the same source—possibly linked to frequent partner exchange among FSW [92]. Overall, multiple studies have provided evidence of HIV-1 phylogenetic linkages between HIV-1 key populations and HET in Sub-Saharan Africa. A common observation in most of these studies is clustering between HET females and MSM, in addition to the expected links between HET and FSW owing to sex work.

HIV-1 mixing appears to be at relatively low rates across the region (although this has been difficult to quantify empirically because of the dearth of HIV-1 sequence data from MSM, FSW, and PWID).

3.4. Phylogenetic Analysis to Examine Sources and Direction of HIV-1 Transmission between HIV-1 Key and Vulnerable Populations and HET in sSA

Few studies have investigated the directionality in HIV-1 transmission involving different risk groups in sSA. Recent phylogenetic analyses have shown that fishing communities do not serve as a source of HIV-1 infection to much larger populations with lower HIV-1 prevalence in Uganda [46,90,93]. In Senegal, Nascimento et al. showed that 3.2% of infections in HET females were acquired from MSM, whereas 0.3% infections among MSM were acquired from HET females [94]. In Nigeria, a phylodynamic analysis of HIV-1 *pol* sequences from MSM and HET females by Volz et al. estimated a 9.1% virus flow from MSM to HET females and 0.2% HIV-1 transmissions from HET females to MSM [95].

Dennis et al. evaluated HIV-1 phylogenetic and behavioural characteristics among 45 newly diagnosed and acutely infected HIV-1 individuals (index partners) and their referred HET partners in Malawi [96]. None of the 45 index partners were closely linked phylogenetically. However, most index partners were linked with their chronically infected HET partners, highlighting the contribution of chronic infections to new HIV-1 transmissions. Another phylogenetic study by Jennes et al. analysed 46 HIV-1 concordant positive HET couples in Dakar, Senegal, to understand the dynamics and risk factors of within-couple HIV-1 transmissions [97]. The analysis showed that male partners were the most likely index partners (and hence the source of infection) to married women.

Phylogenetic studies have also revealed the role of age-disparate HET relationships in perpetuating local HIV-1 transmission in Sub-Saharan Africa [46,98,99]. In Uganda, Ssemwanga et al. found that HET individuals older than 25 years were more likely to appear in phylogenetic clusters than younger individuals [48]. This study suggested that high-risk HET behaviour involving older individuals living with HIV-1 may drive recurring new infections. In Botswana, a country-wide study involving 6078 sequences by Novitsky et al. identified 984 phylogenetically distinct clusters, revealing complex HIV-1 phylogenetic linkages with mixing between different communities and geographic regions [99]. This study suggested that HIV-1 may first be transmitted from older women to middle aged men, followed by transmission from these men to young women. This HIV-1 transmission cycle had been described earlier in KwaZulu-Natal, South Africa, where HIV-1 is first transmitted from women aged 25–40 years to men aged 25–40 years who would then transmit to girls and young women (15–25 years) [98]. Overall, research has shown that key populations may contribute a modest fraction of infections to the HET population and that key populations may be a sink and not the major source of infections in the mixed epidemic. Further research is needed to reveal the drivers of the HIV-1 epidemic in sSA [90,100].

4. Perspectives, Challenges, and Potential Solutions with Phylogenetic Inference in sSA

First, most sequence-based studies in sSA have focused on transmitted drug resistance, and more phylogenetic studies dissecting how HIV-1 in different populations mix and spread are warranted. Second, there is a need to incorporate mobility networks into the phylogenetic spatiotemporal models to quantify the movement patterns and links between urban and rural communities more precisely. Although these mobility methodologies have been developed and used to quantify the impact of human mobility on malaria transmission in different African countries, including Kenya and Madagascar, their application in deciphering HIV-1 transmission is limited [101–103]. While these phylogeographic models can reveal and quantify the movement of viruses between locations, they are limited in the in-depth determination of how and where virus transmission has occurred without additional information, e.g., on human movement. Residents in a community may get infected while living or travelling outside their homes, and such external introductions

could be further disentangled by combining movement and migration data with virus data. However, foreseeable hurdles include obtaining mobility data from telecommunication companies as well as individual rights protection issues. Third, many phylogenetic studies in sSA have been constrained by low sampling and limited geographic coverage. This limits the extent to which the entirety of HIV-1 transmission dynamics in a country may be characterised. A low sampling density generally results in missing links and smaller clusters of HIV-1 sequences and may therefore limit the reliability of phylogenetic evidence in guiding policy decisions [76]. A potential solution to this problem would be for studies in sSA to aim to increase sampling efforts to achieve larger and proportional sample coverage across all risk groups and geographic locations. Another related challenge is skewed sampling between risk groups and locations resulting in the over representation of some populations and, as a result, a bias in the phylogenetic assessment of transmission dynamics and trait linkage. In the absence of dense sampling, some insights may be accomplished through subsampling available datasets relative to HIV-1 prevalence per risk group or geographic location for proportional representation, albeit with a loss of links due to exclusion of some sequences [8,9,12]. Fourth, a substantial number of published sequences lack information on patient demographics, sampling location, and sampling date, hence limiting their use in phylogeographic studies. In the case of published sequences lacking risk data in sSA, such sequences could be assumed to have been collected from HET individuals (the dominant route of HIV-1 transmission in sSA) [44]. Thereafter, based on phylogenetic clustering, the probable risk group for nodes within a cluster with inadequate annotation may be deduced from association with nodes with a known risk group—as was done to identify potential nondisclosed MSM (self-reported HET men who clustered only with men) in the United Kingdom [104]. With the establishment of the PANGEA consortium (although no data on the contribution of MSM, FSW, or PWID to the epidemic have been reported), a more homogenous and dense sampling from the participating countries may improve and strengthen the limitations of phylodynamic methods [24]. Finally, a potential limitation of our literature search is that it was restricted to studies available only in the PubMed database (https://pubmed.ncbi.nlm.nih.gov/ (accessed on 12 March 2021)). It is therefore possible that some studies were not assessed in our analysis.

5. Conclusions

Determining the drivers of the HIV-1 epidemic may be important to guide targeted HIV-1 prevention [29]. Phylogenetic methods could help in characterising such drivers but rely on the availability of large numbers of sequences obtained from well-characterised cohorts. Where these criteria have been achieved (e.g., in European and Northern American settings with dense sampling among infected individuals and patient demographics), phylogenetic studies have provided useful information for HIV-1 prevention [13,16,18,20,21,104–106]. Low sampling density is a constant limitation to phylogenetic studies in Africa, and the shortage of HIV-1 sequences from key and vulnerable populations has limited our understanding of the contribution of these populations to the HIV-1 epidemic in sSA. Where data involving populations that are at high risk for HIV-1 infection (such as young girls and fishing communities) are available in sSA, phylogenetic characterisation of sources and directionality of HIV-1 transmission involving these vulnerable populations has been achieved [63,90,93,98,99]. Likewise, if HIV-1 sequences from HIV-1 key populations (i.e., MSM, PWID, and FSW) are made available, phylogenetic studies may guide understanding HIV-1 transmission dynamics and contemporary drivers in these populations. Phylogenetic studies analysing densely sampled and well-characterised HIV-1 key and vulnerable populations sampled in recent years from multiple geographic locations may play a key role in identifying patterns that could be useful in informing HIV-1 prevention strategies in sSA. Overall, although limited, available data from different studies suggest that epidemics among MSM and PWID are more separated and could thus be targeted to reduce population-level incidence. Given that limited HIV-1 sequence data

in Africa may continue to present a challenge in the unforeseen future, there is a need to develop statistical and or phylogenetic models that could control for missed sampling.

Author Contributions: Conceptualization, J.E., G.M.N. and J.N.; formal analysis and figure design, G.M.N. and J.N.; writing—original draft preparation, G.M.N. and J.N.; writing—review and editing, G.M.N., J.N., J.E., A.S.H. and E.J.S. All authors have read and agreed to the published version of the manuscript.

Funding: G.M.N. is funded through the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant #DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant #107752/Z/15/Z] and the UK government. This work was also supported in part by funding from the Swedish Research Council (grant #2016-01417) and the Swedish Society for Medical Research (grant #SA-2016). IAVI's support is made possible by the generous support of the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) through the United States Agency for International Development (USAID). The full list of IAVI donors is available at www.iavi.org. J.N. is funded by the Swedish Research Council (grant # 2016-01417) and the Medical Faculty at Lund University. The views expressed in this review article are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, IAVI, PEPFAR, USAID, or the United States Government, Swedish Research Council, or the UK government. This report was published with permission from the Kenya Medical Research Institute (KEMRI).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analysed in this study. Data sharing does not apply to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Field, N.; Cohen, T.; Struelens, M.J.; Palm, D.; Cookson, B.; Glynn, J.R.; Gallo, V.; Ramsay, M.; Sonnenberg, P.; MacCannell, D. Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): An extension of the STROBE statement. *Lancet Infect. Dis.* **2014**, *14*, 341–352. [CrossRef]
2. Riley, L.W. *Molecular Epidemiology of Infectious Diseases: Principles and Practices*; American Society for Microbiology: Washington, DC, USA, 2004.
3. Lemey, P.; Suchard, M.; Rambaut, A. Reconstructing the initial global spread of a human influenza pandemic: A Bayesian spatial-temporal model for the global spread of H1N1pdm. *PLoS Curr.* **2009**, *1*. [CrossRef]
4. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **2020**, *5*, 1403–1407. [CrossRef] [PubMed]
5. Worobey, M.; Pekar, J.; Larsen, B.B.; Nelson, M.I.; Hill, V.; Joy, J.B.; Rambaut, A.; Suchard, M.A.; Wertheim, J.O.; Lemey, P. The emergence of SARS-CoV-2 in Europe and North America. *Science* **2020**, *370*, 564–570. [CrossRef] [PubMed]
6. Lemey, P.; Hong, S.L.; Hill, V.; Baele, G.; Poletto, C.; Colizza, V.; O'Toole, Á.; McCrone, J.T.; Andersen, K.G.; Worobey, M. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.* **2020**, *11*, 1–14. [CrossRef]
7. Bruhn, C.A.; Audelin, A.M.; Helleberg, M.; Bjorn-Mortensen, K.; Obel, N.; Gerstoft, J.; Nielsen, C.; Melbye, M.; Medstrand, P.; Gilbert, M.T.; et al. The origin and emergence of an HIV-1 epidemic: From introduction to endemicity. *AIDS* **2014**, *28*, 1031–1040. [CrossRef]
8. Faria, N.R.; Rambaut, A.; Suchard, M.A.; Baele, G.; Bedford, T.; Ward, M.J.; Tatem, A.J.; Sousa, J.D.; Arinaminpathy, N.; Pépin, J. The early spread and epidemic ignition of HIV-1 in human populations. *Science* **2014**, *346*, 56–61. [CrossRef]
9. Faria, N.R.; Vidal, N.; Lourenco, J.; Raghwani, J.; Sigaloff, K.C.E.; Tatem, A.J.; van de Vijver, D.A.M.; Pineda-Peña, A.C.; Rose, R.; Wallis, C.L.; et al. Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa. *PLoS Pathog.* **2019**, *15*, e1007976. [CrossRef]
10. Hassan, A.S.; Esbjornsson, J.; Wahome, E.; Thiong'o, A.; Makau, G.N.; Price, M.A.; Sanders, E.J. HIV-1 subtype diversity, transmission networks and transmitted drug resistance amongst acute and early infected MSM populations from Coastal Kenya. *PLoS ONE* **2018**, *13*, e0206177. [CrossRef]

11. Hassan, A.S.; Mwaringa, S.M.; Obonyo, C.A.; Nabwera, H.M.; Sanders, E.J.; Rinke de Wit, T.F.; Cane, P.A.; Berkley, J.A. Low prevalence of transmitted HIV type 1 drug resistance among antiretroviral-naive adults in a rural HIV clinic in Kenya. *AIDS Res. Hum. Retrovir.* **2013**, *29*, 129–135. [CrossRef]
12. Nazziwa, J.; Faria, N.R.; Chaplin, B.; Rawizza, H.; Kanki, P.; Dakum, P.; Abimiku, A.; Charurat, M.; Ndembi, N.; Esbjörnsson, J. Characterisation of HIV-1 Molecular Epidemiology in Nigeria: Origin, Diversity, Demography and Geographic Spread. *Sci. Rep.* **2020**, *10*, 3468. [CrossRef]
13. Poon, A.F.; Gustafson, R.; Daly, P.; Zerr, L.; Demlow, S.E.; Wong, J.; Woods, C.K.; Hogg, R.S.; Krajden, M.; Moore, D. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: An implementation case study. *Lancet HIV* **2016**, *3*, e231–e238. [CrossRef]
14. Rodger, A.J.; Cambiano, V.; Bruun, T.; Vernazza, P.; Collins, S.; Degen, O.; Corbelli, G.M.; Estrada, V.; Geretti, A.M.; Beloukas, A. Risk of HIV transmission through condomless sex in serodifferent gay couples with the HIV-positive partner taking suppressive antiretroviral therapy (PARTNER): Final results of a multicentre, prospective, observational study. *Lancet* **2019**, *393*, 2428–2438. [CrossRef]
15. Rodger, A.J.; Cambiano, V.; Bruun, T.; Vernazza, P.; Collins, S.; Van Lunzen, J.; Corbelli, G.M.; Estrada, V.; Geretti, A.M.; Beloukas, A. Sexual activity without condoms and risk of HIV transmission in serodifferent couples when the HIV-positive partner is using suppressive antiretroviral therapy. *JAMA* **2016**, *316*, 171–181. [CrossRef] [PubMed]
16. Vasylyeva, T.I.; Liulchuk, M.; Friedman, S.R.; Sazonova, I.; Faria, N.R.; Katzourakis, A.; Babii, N.; Scherbinska, A.; Thézé, J.; Pybus, O.G.; et al. Molecular epidemiology reveals the role of war in the spread of HIV in Ukraine. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 1051–1056. [CrossRef]
17. Esbjörnsson, J.; Mild, M.; Audelin, A.; Fonager, J.; Skar, H.; Bruun Jørgensen, L.; Liitsola, K.; Björkman, P.; Bratt, G.; Gisslén, M. HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. *Virus Evol.* **2016**, *2*, vew010. [CrossRef]
18. Volz, E.M.; Ionides, E.; Romero-Severson, E.O.; Brandt, M.G.; Mokotoff, E.; Koopman, J.S. HIV-1 transmission during early infection in men who have sex with men: A phylodynamic analysis. *PLoS Med.* **2013**, *10*, e1001568. [CrossRef]
19. Brenner, B.G.; Roger, M.; Routy, J.P.; Moisi, D.; Ntemgwa, M.; Matte, C.; Baril, J.G.; Thomas, R.; Rouleau, D.; Bruneau, J.; et al. High rates of forward transmission events after acute/early HIV-1 infection. *J. Infect. Dis.* **2007**, *195*, 951–959. [CrossRef]
20. Ratmann, O.; Van Sighem, A.; Bezemer, D.; Gavryushkina, A.; Jurriaans, S.; Wensing, A.; De Wolf, F.; Reiss, P.; Fraser, C. Sources of HIV infection among men having sex with men and implications for prevention. *Sci. Transl. Med.* **2016**, *8*, 320ra2. [CrossRef]
21. Sallam, M.; Esbjörnsson, J.; Baldvinsdóttir, G.; Indriðason, H.; Björnsdóttir, T.B.; Widell, A.; Gottfreðsson, M.; Löve, A.; Medstrand, P. Molecular epidemiology of HIV-1 in Iceland: Early introductions, transmission dynamics and recent outbreaks among injection drug users. *Infect. Genet. Evol.* **2017**, *49*, 157–163. [CrossRef]
22. Ragonnet-Cronin, M.; Lycett, S.J.; Hodcroft, E.B.; Hué, S.; Fearnhill, E.; Brown, A.E.; Delpech, V.; Dunn, D.; Leigh Brown, A.J. Transmission of Non-B HIV Subtypes in the United Kingdom Is Increasingly Driven by Large Non-Heterosexual Transmission Clusters. *J. Infect. Dis.* **2016**, *213*, 1410–1418. [CrossRef]
23. World Health Organization. Consolidated guidelines on HIV prevention, diagnosis, treatment and care for key populations: 2016 update. In *Consolidated Guidelines on HIV Prevention, Diagnosis, Treatment and Care for Key Populations: 2016 Update*; WHO: Geneva, Switzerland, 2016.
24. Abeler-Dörner, L.; Grabowski, M.K.; Rambaut, A.; Pillay, D.; Fraser, C. PANGEA-HIV 2: Phylogenetics and networks for generalised epidemics in Africa. *Curr. Opin. HIV AIDS* **2019**, *14*, 173. [CrossRef]
25. Dwyer-Lindgren, L.; Cork, M.A.; Sligar, A.; Steuben, K.M.; Wilson, K.F.; Provost, N.R.; Mayala, B.K.; VanderHeide, J.D.; Collison, M.L.; Hall, J.B.; et al. Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature* **2019**, *570*, 189–193. [CrossRef] [PubMed]
26. Tanser, F.; de Oliveira, T.; Maheu-Giroux, M.; Bärnighausen, T. Concentrated HIV sub-epidemics in generalized epidemic settings. *Curr. Opin. HIV AIDS* **2014**, *9*, 115. [CrossRef] [PubMed]
27. Makofane, K.; van der Elst, E.M.; Walimbwa, J.; Nemande, S.; Baral, S.D. From general to specific: Moving past the general population in the HIV response across sub-Saharan Africa. *J. Int. AIDS Soc.* **2020**, *23*, e25605. [CrossRef] [PubMed]
28. Joint United Nations Programme on HIV/AIDS (UNAIDS). UNAIDS. Global AIDS Report. 2020:384. Seizing the Moment 7/6/2020. Available online: https://www.unaids.org/sites/default/files/media_asset/2020_global-aids-report_en.pdf (accessed on 12 May 2020).
29. Anderson, S.-J.; Cherutich, P.; Kilonzo, N.; Cremin, I.; Fecht, D.; Kimanga, D.; Harper, M.; Masha, R.L.; Ngongo, P.B.; Maina, W. Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: A modelling study. *Lancet* **2014**, *384*, 249–256. [CrossRef]
30. Kelly, S.L.; Martin-Hughes, R.; Stuart, R.M.; Yap, X.F.; Kedziora, D.J.; Grantham, K.L.; Hussain, S.A.; Reporter, I.; Shattock, A.J.; Grobicki, L. The global Optima HIV allocative efficiency model: Targeting resources in efforts to end AIDS. *Lancet HIV* **2018**, *5*, e190–e198. [CrossRef]
31. Wilson, D.; Halperin, D.T. "Know your epidemic, know your response": A useful approach, if we get it right. *Lancet* **2008**, *372*, 423–426. [CrossRef]

32. Dennis, A.M.; Herbeck, J.T.; Brown, A.L.; Kellam, P.; de Oliveira, T.; Pillay, D.; Fraser, C.; Cohen, M.S. Phylogenetic studies of transmission dynamics in generalized HIV epidemics: An essential tool where the burden is greatest? *J. Acquir. Immune Defic. Syndr.* **2014**, *67*, 181–195. [CrossRef]
33. Lihana, R.W.; Ssemwanga, D.; Abimiku, A.; Ndembi, N. Update on HIV-1 diversity in Africa: A decade in review. *AIDS Rev.* **2012**, *14*, 83–100.
34. Bbosa, N.; Kaleebu, P.; Ssemwanga, D. HIV subtype diversity worldwide. *Curr. Opin. HIV AIDS* **2019**, *14*, 153–160. [CrossRef]
35. Hemelaar, J.; Elangovan, R.; Yun, J.; Dickson-Tetteh, L.; Fleminger, I.; Kirtley, S.; Williams, B.; Gouws-Williams, E.; Ghys, P.D.; Abimiku, A.L.G.; et al. Global and regional molecular epidemiology of HIV-1, 1990–2015: A systematic review, global survey, and trend analysis. *Lancet Infect. Dis.* **2019**, *19*, 143–155. [CrossRef]
36. Esbjörnsson, J.; Mild, M.; Månsson, F.; Norrgren, H.; Medstrand, P. HIV-1 molecular epidemiology in Guinea-Bissau, West Africa: Origin, demography and migrations. *PLoS ONE* **2011**, *6*, e17025. [CrossRef] [PubMed]
37. Hemelaar, J.; Loganathan, S.; Elangovan, R.; Yun, J.; Dickson-Tetteh, L.; Kirtley, S. Country Level Diversity of the HIV-1 Pandemic between 1990 and 2015. *J. Virol.* **2020**, *95*, e01580–e01620. [CrossRef] [PubMed]
38. Véras, N.M.; Santoro, M.M.; Gray, R.R.; Tatem, A.J.; Lo Presti, A.; Olearo, F.; Cappelli, G.; Colizzi, V.; Takou, D.; Torimiro, J.; et al. Molecular epidemiology of HIV type 1 CRF02_AG in Cameroon and African patients living in Italy. *Aids Res. Hum. Retrovir.* **2011**, *27*, 1173–1182. [CrossRef]
39. Faria, N.R.; Suchard, M.A.; Abecasis, A.; Sousa, J.D.; Ndembi, N.; Bonfim, I.; Camacho, R.J.; Vandamme, A.M.; Lemey, P. Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **2012**, *12*, 453–460. [CrossRef]
40. Gill, M.S.; Tung Ho, L.S.; Baele, G.; Lemey, P.; Suchard, M.A. A Relaxed Directional Random Walk Model for Phylogenetic Trait Evolution. *Syst. Biol.* **2016**, syw093. [CrossRef]
41. Bártolo, I.; Calado, R.; Borrego, P.; Leitner, T.; Taveira, N. Rare HIV-1 Subtype J Genomes and a New H/U/CRF02_AG Recombinant Genome Suggests an Ancient Origin of HIV-1 in Angola. *AIDS Res. Hum. Retrovir.* **2016**, *32*, 822–828. [CrossRef] [PubMed]
42. Bártolo, I.; Zakovic, S.; Martin, F.; Palladino, C.; Carvalho, P.; Camacho, R.; Thamm, S.; Clemente, S.; Taveira, N. HIV-1 diversity, transmission dynamics and primary drug resistance in Angola. *PLoS ONE* **2014**, *9*, e113626. [CrossRef]
43. Hué, S.; Hassan, A.S.; Nabwera, H.; Sanders, E.J.; Pillay, D.; Berkley, J.A.; Cane, P.A. HIV type 1 in a rural coastal town in Kenya shows multiple introductions with many subtypes and much recombination. *AIDS Res. Hum. Retrovir.* **2012**, *28*, 220–224. [CrossRef]
44. Nduva, G.M.; Hassan, A.S.; Nazziwa, J.; Graham, S.M.; Esbjörnsson, J.; Sanders, E.J. HIV-1 Transmission Patterns Within and Between Risk Groups in Coastal Kenya. *Sci. Rep.* **2020**, *10*. [CrossRef] [PubMed]
45. Yebra, G.; Ragonnet-Cronin, M.; Ssemwanga, D.; Parry, C.M.; Logue, C.H.; Cane, P.A.; Kaleebu, P.; Brown, A.J. Analysis of the history and spread of HIV-1 in Uganda using phylodynamics. *J. Gen. Virol.* **2015**, *96*, 1890–1898. [CrossRef] [PubMed]
46. Ssemwanga, D.; Bbosa, N.; Nsubuga, R.N.; Ssekagiri, A.; Kapaata, A.; Nannyonjo, M.; Nassolo, F.; Karabarinde, A.; Mugisha, J.; Seeley, J.; et al. The Molecular Epidemiology and Transmission Dynamics of HIV Type 1 in a General Population Cohort in Uganda. *Viruses* **2020**, *12*, 1283. [CrossRef] [PubMed]
47. Arimide, D.A.; Abebe, A.; Kebede, Y.; Adugna, F.; Tilahun, T.; Kassa, D.; Assefa, Y.; Balcha, T.T.; Björkman, P.; Medstrand, P. HIV-genetic diversity and drug resistance transmission clusters in Gondar, Northern Ethiopia, 2003-2013. *PLoS ONE* **2018**, *13*, e0205446. [CrossRef] [PubMed]
48. Wilkinson, E.; Engelbrecht, S.; de Oliveira, T. History and origin of the HIV-1 subtype C epidemic in South Africa and the greater southern African region. *Sci. Rep.* **2015**, *5*, 16897. [CrossRef] [PubMed]
49. Wilkinson, E.; Rasmussen, D.; Ratmann, O.; Stadler, T.; Engelbrecht, S.; de Oliveira, T. Origin, imports and exports of HIV-1 subtype C in South Africa: A historical perspective. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **2016**, *46*, 200–208. [CrossRef]
50. Wilkinson, E.; Junqueira, D.M.; Lessells, R.; Engelbrecht, S.; van Zyl, G.; de Oliveira, T.; Salemi, M. The effect of interventions on the transmission and spread of HIV in South Africa: A phylodynamic analysis. *Sci. Rep.* **2019**, *9*, 2640. [CrossRef] [PubMed]
51. Rasmussen, D.A.; Wilkinson, E.; Vandormael, A.; Tanser, F.; Pillay, D.; Stadler, T.; De Oliveira, T. Tracking external introductions of HIV using phylodynamics reveals a major source of infections in rural KwaZulu-Natal, South Africa. *Virus Evol.* **2018**, *4*. [CrossRef]
52. Novitsky, V.; Bussmann, H.; Logan, A.; Moyo, S.; van Widenfelt, E.; Okui, L.; Mmalane, M.; Baca, J.; Buck, L.; Phillips, E.; et al. Phylogenetic relatedness of circulating HIV-1C variants in Mochudi, Botswana. *PLoS ONE* **2013**, *8*, e80589. [CrossRef]
53. Yebra, G.; Kalish, M.L.; Leigh Brown, A.J. Reconstructing the HIV-1 CRF02_AG and CRF06_cpx epidemics in Burkina Faso and West Africa using early samples. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **2016**, *46*, 209–218. [CrossRef]
54. Mir, D.; Jung, M.; Delatorre, E.; Vidal, N.; Peeters, M.; Bello, G. Phylodynamics of the major HIV-1 CRF02_AG African lineages and its global dissemination. *Infect. Genet. Evol.* **2016**, *46*, 190–199. [CrossRef]
55. Delatorre, E.; Bello, G. Spatiotemporal dynamics of the HIV-1 CRF06_cpx epidemic in western Africa. *AIDS* **2013**, *27*, 1313–1320. [CrossRef] [PubMed]
56. Delatorre, E.; Bello, G. Time-scale of minor HIV-1 complex circulating recombinant forms from Central and West Africa. *BMC Evol. Biol.* **2016**, *16*. [CrossRef] [PubMed]

57. Delatorre, E.; Mir, D.; Bello, G. Spatiotemporal Dynamics of the HIV-1 Subtype G Epidemic in West and Central Africa. *PLoS ONE* **2014**, *9*, e98908. [CrossRef] [PubMed]
58. Delatorre, E.O.; Bello, G. Phylodynamics of HIV-1 subtype C epidemic in east Africa. *PLoS ONE* **2012**, *7*, e41904.
59. Mir, D.; Gräf, T.; Esteves de Matos Almeida, S.; Pinto, A.R.; Delatorre, E.; Bello, G. Inferring population dynamics of HIV-1 subtype C epidemics in Eastern Africa and Southern Brazil applying different Bayesian phylodynamics approaches. *Sci. Rep.* **2018**, *8*, 8778. [CrossRef]
60. Gray, R.R.; Tatem, A.J.; Lamers, S.; Hou, W.; Laeyendecker, O.; Serwadda, D.; Sewankambo, N.; Gray, R.H.; Wawer, M.; Quinn, T.C.; et al. Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. *AIDS* **2009**, *23*, F9–F17. [CrossRef]
61. Dalai, S.C.; de Oliveira, T.; Harkins, G.W.; Kassaye, S.G.; Lint, J.; Manasa, J.; Johnston, E.; Katzenstein, D. Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe. *AIDS* **2009**, *23*, 2523–2532. [CrossRef]
62. Afonso, J.M.; Morgado, M.G.; Bello, G. Evidence of multiple introductions of HIV-1 subtype C in Angola. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **2012**, *12*, 1458–1465. [CrossRef]
63. Grabowski, M.K.; Lessler, J.; Bazaale, J.; Nabukalu, D.; Nankinga, J.; Nantume, B.; Ssekasanvu, J.; Reynolds, S.J.; Ssekubugu, R.; Nalugoda, F. Migration, hotspots, and dispersal of HIV infection in Rakai, Uganda. *Nat. Commun.* **2020**, *11*, 1–12. [CrossRef]
64. Kreiss, J.K.; Koech, D.; Plummer, F.A.; Holmes, K.K.; Lightfoote, M.; Piot, P.; Ronald, A.R.; Ndinya-Achola, J.O.; D'Costa, L.J.; Roberts, P.; et al. AIDS virus infection in Nairobi prostitutes. Spread of the epidemic to East Africa. *N. Engl. J. Med.* **1986**, *314*, 414–418. [CrossRef] [PubMed]
65. Smith, A.D.; Tapsoba, P.; Peshu, N.; Sanders, E.J.; Jaffe, H.W. Men who have sex with men and HIV/AIDS in sub-Saharan Africa. *Lancet* **2009**, *374*, 416–422. [CrossRef]
66. Rakwar, J.; Lavreys, L.; Thompson, M.L.; Jackson, D.; Bwayo, J.; Hassanali, S.; Mandaliya, K.; Ndinya-Achola, J.; Kreiss, J. Cofactors for the acquisition of HIV-1 among heterosexual men: Prospective cohort study of trucking company workers in Kenya. *AIDS* **1999**, *13*, 607–614. [CrossRef]
67. Biggar, R. The AIDS problem in Africa. *Lancet* **1986**, *327*, 79–83. [CrossRef]
68. Smith, A.D.; Muhaari, A.D.; Agwanda, C.; Kowuor, D.; van der Elst, E.; Davies, A.; Graham, S.M.; Jaffe, H.W.; Sanders, E.J. Heterosexual behaviours among men who sell sex to men in coastal Kenya. *AIDS* **2015**, *29* (Suppl. 3), S201–S210. [CrossRef]
69. Parker, R.; Khan, S.; Aggleton, P. Conspicuous by their absence? Men who have sex with men (msm) in developing countries: Implications for HIV prevention. *Crit. Public Health* **1998**, *8*, 329–346. [CrossRef]
70. Carroll, A.; Mendos, L.R. *State Sponsored Homophobia 2017: A World Survey of Sexual Orientation Laws: Criminalisation, Protection and Recognition*; International Lesbian, Gay, Bisexual, Trans and Intersex Association (ILGA): Geneva, Switzerland, 2017; pp. 26–191.
71. Novitsky, V.; Kühnert, D.; Moyo, S.; Widenfelt, E.; Okui, L.; Essex, M. Phylodynamic analysis of HIV sub-epidemics in Mochudi, Botswana. *Epidemics* **2015**, *13*, 44–55. [CrossRef]
72. Sivay, M.V.; Hudelson, S.E.; Wang, J.; Agyei, Y.; Hamilton, E.L.; Selin, A.; Dennis, A.; Kahn, K.; Gomez-Olive, F.X.; MacPhail, C.; et al. HIV-1 diversity among young women in rural South Africa: HPTN 068. *PLoS ONE* **2018**, *13*, e0198999. [CrossRef]
73. Trask, S.A.; Derdeyn, C.A.; Fideli, U.; Chen, Y.; Meleth, S.; Kasolo, F.; Musonda, R.; Hunter, E.; Gao, F.; Allen, S.; et al. Molecular epidemiology of human immunodeficiency virus type 1 transmission in a heterosexual cohort of discordant couples in Zambia. *J. Virol.* **2002**, *76*, 397–405. [CrossRef]
74. Rusine, J.; Jurriaans, S.; van de Wijgert, J.; Cornelissen, M.; Kateera, B.; Boer, K.; Karita, E.; Mukabayire, O.; de Jong, M.; Ondoa, P. Molecular and phylogeographic analysis of human immuno-deficiency virus type 1 strains infecting treatment-naive patients from Kigali, Rwanda. *PLoS ONE* **2012**, *7*, e42557. [CrossRef]
75. Rubio-Garrido, M.; González-Alba, J.M.; Reina, G.; Ndarabu, A.; Barquín, D.; Carlos, S.; Galán, J.C.; Holguín, Á. Current and historic HIV-1 molecular epidemiology in paediatric and adult population from Kinshasa in the Democratic Republic of Congo. *Sci. Rep.* **2020**, *10*, 18461. [CrossRef]
76. Novitsky, V.; Moyo, S.; Lei, Q.; DeGruttola, V.; Essex, M. Impact of sampling density on the extent of HIV clustering. *AIDS Res. Hum. Retrovir.* **2014**, *30*, 1226–1235. [CrossRef]
77. Bezemer, D.; Faria, N.R.; Hassan, A.; Hamers, R.L.; Mutua, G.; Anzala, O.; Mandaliya, K.; Cane, P.; Berkley, J.A.; Rinke de Wit, T.F. HIV Type 1 transmission networks among men having sex with men and heterosexuals in Kenya. *AIDS Res. Hum. Retrovir.* **2014**, *30*, 118–126. [CrossRef] [PubMed]
78. Osman, S.; Lihana, R.W.; Kibaya, R.M.; Ishizaki, A.; Bi, X.; Okoth, F.A.; Ichimura, H.; Lwembe, R.M. Diversity of HIV type 1 and drug resistance mutations among injecting drug users in Kenya. *AIDS Res. Hum. Retrovir.* **2013**, *29*, 187–190. [CrossRef] [PubMed]
79. Billings, E.; Kijak, G.H.; Sanders-Buell, E.; Ndembi, N.; O'Sullivan, A.M.; Adebajo, S.; Kokogho, A.; Milazzo, M.; Lombardi, K.; Baral, S.; et al. New Subtype B Containing HIV-1 Circulating Recombinant of sub-Saharan Africa Origin in Nigerian Men Who Have Sex with Men. *J. Acquir. Immune Defic. Syndr.* **2019**, *81*, 578–584. [CrossRef] [PubMed]
80. Li, Y.; Liu, H.; Ramadhani, H.O.; Ndembi, N.; Crowell, T.A.; Kijak, G.; Robb, M.L.; Ake, J.A.; Kokogho, A.; Nowak, R.G.; et al. Genetic clustering analysis for HIV infection among MSM in Nigeria: Implications for intervention. *AIDS* **2020**, *34*, 227–236. [CrossRef] [PubMed]
81. Ndiaye, H.D.; Toure-Kane, C.; Vidal, N.; Niama, F.R.; Niang-Diallo, P.A.; Dièye, T.; Gaye-Diallo, A.; Wade, A.S.; Peeters, M.; Mboup, S. Surprisingly High Prevalence of Subtype C and Specific HIV-1 Subtype/CRF Distribution in Men Having Sex With Men in Senegal. *J. Acquir. Immune Defic. Syndr.* **2009**, *52*, 249–252. [CrossRef]

82. Ndiaye, H.D.; Tchiakpe, E.; Vidal, N.; Ndiaye, O.; Diop, A.K.; Peeters, M.; Mboup, S.; Toure-Kane, C. HIV type 1 subtype C remains the predominant subtype in men having sex with men in Senegal. *AIDS Res. Hum. Retrovir.* **2013**, *29*, 1265–1272. [CrossRef]
83. Skar, H.; Axelsson, M.; Berggren, I.; Thalme, A.; Gyllensten, K.; Liitsola, K.; Brummer-Korvenkontio, H.; Kivela, P.; Spangberg, E.; Leitner, T.; et al. Dynamics of Two Separate but Linked HIV-1 CRF01_AE Outbreaks among Injection Drug Users in Stockholm, Sweden, and Helsinki, Finland. *J. Virol.* **2011**, *85*, 510–518. [CrossRef]
84. Wilkinson, E.; Engelbrecht, S.; de Oliveira, T. Detection of transmission clusters of HIV-1 subtype C over a 21-year period in Cape Town, South Africa. *PLoS ONE* **2014**, *9*, e109296. [CrossRef]
85. Middelkoop, K.; Rademeyer, C.; Brown, B.B.; Cashmore, T.J.; Marais, J.C.; Scheibe, A.P.; Bandawe, G.P.; Myer, L.; Fuchs, J.D.; Williamson, C.; et al. Epidemiology of HIV-1 Subtypes Among Men Who Have Sex With Men in Cape Town, South Africa. *J. Acquir. Immune Defic. Syndr.* **2014**, *65*, 473–480. [CrossRef] [PubMed]
86. Leye, N.; Vidal, N.; Ndiaye, O.; Diop-Ndiaye, H.; Wade, A.S.; Mboup, S.; Delaporte, E.; Toure-Kane, C.; Peeters, M. High frequency of HIV-1 infections with multiple HIV-1 strains in men having sex with men (MSM) in Senegal. *Infect. Genet. Evol.* **2013**, *20*, 206–214. [CrossRef] [PubMed]
87. Konou, A.A.; Vidal, N.; Salou, M.; Anato, S.; Singo-Tokofaï, A.; Ekouevi, D.K.; Pitché, P.; Prince-David, M.; Delaporte, E.; Peeters, M.; et al. Genetic diversity and transmission networks of HIV-1 strains among men having sex with men (MSM) in Lomé, Togo. *Infect. Genet. Evol.* **2016**, *46*, 279–285. [CrossRef]
88. Nazziwa, J.; Njai, H.F.; Ndembi, N.; Birungi, J.; Lyagoba, F.; Gershim, A.; Nakiyingi-Miiro, J.; Nielsen, L.; Mpendo, J.; Nanvubya, A.; et al. Short communication: HIV type 1 transmitted drug resistance and evidence of transmission clusters among recently infected antiretroviral-naive individuals from Ugandan fishing communities of Lake Victoria. *AIDS Res. Hum. Retrovir.* **2013**, *29*, 788–795. [CrossRef] [PubMed]
89. Kiwuwa-Muyingo, S.; Nazziwa, J.; Ssemwanga, D.; Ilmonen, P.; Njai, H.; Ndembi, N.; Parry, C.; Kitandwe, P.K.; Gershim, A.; Mpendo, J.; et al. HIV-1 transmission networks in high risk fishing communities on the shores of Lake Victoria in Uganda: A phylogenetic and epidemiological approach. *PLoS ONE* **2017**, *12*, e0185818. [CrossRef]
90. 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. Phylogeography of HIV-1 suggests that Ugandan fishing communities are a sink for, not a source of, virus from general populations. *Sci. Rep.* **2019**, *9*, 1–8. [CrossRef]
91. 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. The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: Evidence from spatial clustering, phylogenetics, and egocentric transmission models. *PLoS Med.* **2014**, *11*, e1001610. [CrossRef]
92. Ssemwanga, D.; Ndembi, N.; Lyagoba, F.; Bukenya, J.; Seeley, J.; Vandepitte, J.; Grosskurth, H.; Kaleebu, P. HIV type 1 subtype distribution, multiple infections, sexual networks, and partnership histories in female sex workers in Kampala, Uganda. *Aids Res. Hum. Retrovir.* **2012**, *28*, 357–365. [CrossRef]
93. Ratmann, O.; Kagaayi, J.; Hall, M.; Golubchick, T.; Kigozi, G.; Xi, X.; Wymant, C.; Nakigozi, G.; Abeler-Dörner, L.; Bonsall, D. Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: A population-based study in Rakai, Uganda. *Lancet HIV* **2020**, *7*, e173–e183. [CrossRef]
94. Nascimento, F.F.; Baral, S.; Geidelberg, L.; Mukandavire, C.; Schwartz, S.R.; Turpin, G.; Turpin, N.; Diouf, D.; Diouf, N.L.; Coly, K.; et al. Phylodynamic analysis of HIV-1 subtypes B, C and CRF 02_AG in Senegal. *Epidemics* **2020**, *30*, 100376. [CrossRef]
95. Volz, E.M.; Ndembi, N.; Nowak, R.; Kijak, G.H.; Idoko, J.; Dakum, P.; Royal, W.; Baral, S.; Dybul, M.; Blattner, W.A.; et al. Phylodynamic analysis to inform prevention efforts in mixed HIV epidemics. *Virus Evol.* **2017**, *3*, vex014. [CrossRef]
96. Dennis, A.M.; Cohen, M.S.; Rucinski, K.B.; Rutstein, S.E.; Powers, K.A.; Pasquale, D.K.; Phiri, S.; Hosseinipour, M.C.; Kamanga, G.; Nsona, D.; et al. Human Immunodeficiency Virus (HIV)-1 Transmission Among Persons With Acute HIV-1 Infection in Malawi: Demographic, Behavioral, and Phylogenetic Relationships. *Clin. Infect. Dis.* **2019**, *69*, 853–860. [CrossRef]
97. Jennes, W.; Kyongo, J.K.; Vanhommerig, E.; Camara, M.; Coppens, S.; Seydi, M.; Mboup, S.; Heyndrickx, L.; Kestens, L. Molecular epidemiology of HIV-1 transmission in a cohort of HIV-1 concordant heterosexual couples from Dakar, Senegal. *PLoS ONE* **2012**, *7*, e37402. [CrossRef]
98. De Oliveira, T.; Kharsany, A.B.; Gräf, T.; Cawood, C.; Khanyile, D.; Grobler, A.; Puren, A.; Madurai, S.; Baxter, C.; Karim, Q.A. Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: A community-wide phylogenetic study. *Lancet HIV* **2017**, *4*, e41–e50. [CrossRef]
99. Novitsky, V.; Zahralban-Steele, M.; Moyo, S.; Nkhisang, T.; Maruapula, D.; McLane, M.F.; Leidner, J.; Bennett, K.; Wirth, K.E. Mapping of HIV-1C Transmission Networks Reveals Extensive Spread of Viral Lineages Across Villages in Botswana Treatment-as-Prevention Trial. *J. Infect. Dis.* **2020**, *222*, 1670–1680. [CrossRef] [PubMed]
100. Grabowski, M.K.; Lessler, J. Phylogenetic insights into age-disparate partnerships and HIV. *Lancet HIV* **2017**, *4*, e8–e9. [CrossRef]
101. Ihantamalala, F.A.; Herbreteau, V.; Rakotoarimanana, F.M.J.; Rakotondramanga, J.M.; Cauchemez, S.; Rahoilijaona, B.; Pennober, G.; Buckee, C.O.; Rogier, C.; Metcalf, C.J.E.; et al. Estimating sources and sinks of malaria parasites in Madagascar. *Nat. Commun.* **2018**, *9*. [CrossRef]
102. Wesolowski, A.; Eagle, N.; Tatem, A.J.; Smith, D.L.; Noor, A.M.; Snow, R.W.; Buckee, C.O. Quantifying the Impact of Human Mobility on Malaria. *Science* **2012**, *338*, 267–270. [CrossRef]

103. Okano, J.T.; Sharp, K.; Valdano, E.; Palk, L.; Blower, S. HIV transmission and source-sink dynamics in sub-Saharan Africa. *Lancet. HIV* **2020**, *7*, e209–e214. [CrossRef]
104. Ragonnet-Cronin, M.; Hué, S.; Hodcroft, E.B.; Tostevin, A.; Dunn, D.; Fawcett, T.; Pozniak, A.; Brown, A.E.; Delpech, V.; Brown, A.J.L. Non-disclosed men who have sex with men in UK HIV transmission networks: Phylogenetic analysis of surveillance data. *Lancet HIV* **2018**, *5*, e309–e316. [CrossRef]
105. Fisher, M.; Pao, D.; Brown, A.E.; Sudarshi, D.; Gill, O.N.; Cane, P.; Buckton, A.J.; Parry, J.V.; Johnson, A.M.; Sabin, C. Determinants of HIV-1 transmission in men who have sex with men: A combined clinical, epidemiological and phylogenetic approach. *AIDS* **2010**, *24*, 1739–1747. [CrossRef] [PubMed]
106. Kouyos, R.D.; Von Wyl, V.; Yerly, S.; Böni, J.; Taffé, P.; Shah, C.; Börgisser, P.; Klimkait, T.; Weber, R.; Hirschel, B. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J. Infect. Dis.* **2010**, *201*, 1488–1497. [CrossRef] [PubMed]