Grant *et al. Retrovirology* 2022, **19**(1):28
https://doi.org/10.1186/s12977-022-00612-5

Retrovirology

RESEARCH

Open Access

A large population sample of African HIV genomes from the 1980s reveals a reduction in subtype D over time associated with propensity for CXCR4 tropism

Heather E. Grant[1*], Sunando Roy[2], Rachel Williams[3], Helena Tutill[2], Bridget Ferns[4], Patricia A. Cane[6], J. Wilson Carswell[7], Deogratius Ssemwanga[5], Pontiano Kaleebu[5], Judith Breuer[2] and Andrew J. Leigh Brown[1]

Abstract

We present 109 near full-length HIV genomes amplified from blood serum samples obtained during early 1986 from across Uganda, which to our knowledge is the earliest and largest population sample from the initial phase of the HIV epidemic in Africa. Consensus sequences were made from paired-end Illumina reads with a target-capture approach to amplify HIV material following poor success with standard approaches. In comparisons with a smaller 'intermediate' genome dataset from 1998 to 1999 and a 'modern' genome dataset from 2007 to 2016, the proportion of subtype D was significantly higher initially, dropping from 67% (73/109), to 57% (26/46) to 17% (82/465) respectively ($p < 0.0001$). Subtype D has previously been shown to have a faster rate of disease progression than other subtypes in East African population studies, and to have a higher propensity to use the CXCR4 co-receptor ("X4 tropism"); associated with a decrease in time to AIDS. Here we find significant differences in predicted tropism between A1 and D subtypes in all three sample periods considered, which is particularly striking the 1986 sample: 66% (53/80) of subtype D *env* sequences were predicted to be X4 tropic compared with none of the 24 subtype A1. We also analysed the frequency of subtype in the envelope region of inter-subtype recombinants, and found that subtype A1 is over-represented in *env*, suggesting recombination and selection have acted to remove subtype D *env* from circulation. The reduction of subtype D frequency over three decades therefore appears to be a result of selective pressure against X4 tropism and its higher virulence. Lastly, we find a subtype D specific codon deletion at position 24 of the V3 loop, which may explain the higher propensity for subtype D to utilise X4 tropism.

Keywords: HIV, Subtype D, East Africa, Co-receptor, Target-capture sequencing, Historic samples

Introduction

The main (M) group of HIV-1 viruses that cause AIDS can be categorised into distinct lineages or "subtypes" [64]. Evidence points to the epicentre of the HIV pandemic being Kinshasa [66] in the early part of the twentieth century [24], and it largely remained within the Democratic Republic of Congo for many decades, undergoing substantial recombination [40, 78]. Strong genetic bottlenecks created geographically [25] and phylogenetically distinct subtypes by the 1960s [80] which subsequently spread throughout Africa and into new susceptible populations across the rest of the world. Today we see this footprint in the global subtype distribution which varies considerably across different countries and risk groups [8].

*Correspondence: heather.grant@ed.ac.uk

[1] Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, UK
Full list of author information is available at the end of the article

There has been much speculation and interest in the possibility of phenotypic differences between subtypes (see review by [28] that may have contributed to any one subtype's relative success over another [23]. Disentangling the relative roles of genetic drift and selective adaptation in HIV lineages amongst hosts is difficult, not least because HIV transmission is heterogeneous, and some lineages may simply be amplified into bottlenecks by chance [62]. Furthermore, subtype comparisons are confounded by differences in mode of transmission (e.g. subtype B in men who have sex with men, [14], and viral characteristics that might be under selection, such as infectivity or virulence, are confounded by a range of factors, including host genetics, (particularly HLA types, [41]). Subtype comparisons within the same country, population, or cohort are therefore strengthened by the reduction of these factors [48]. In Uganda subtypes A1 and D have been co-circulating at high frequencies for many decades [82], in both general population cohorts [70] and high risk communities [6, 68], providing a rare opportunity to compare their impact directly.

Co-receptor tropism (the secondary receptor used alongside CD4) can be distinguished in cell-culture where "fast replicating" syncytium inducing (SI) viruses use CXCR4 (X4 tropic), and "slow" non syncytium inducing (NSI) viruses use CCR5 (R5 tropic) [16]. Fast replicating X4 viruses have long been associated with faster CD4 decline [45], and the risk of AIDS progression could be as much as 3.8 × higher [18], which in real terms translates to multiple years of additional lost life. Comparisons of R5 and X4 viruses at the V3 loop where tropism is largely determined (but not exclusively e.g. [73], indicate that positive amino acid charges at positions 11 and 25 are strongly predictive of X4 tropism (the '11/25 rule'; [79]). Currently, more sophisticated machine learning models are used to predict co-receptor tropism based on V3 amino acid training data (e.g., *geno2pheno*, [67]).

Subtype D has been shown consistently to progress to AIDS faster compared with other subtypes [10, 21, 37, 38, 42, 43, 69, 76]. It has also been reported that subtype D viruses are more likely to use CXCR4 co-receptors [36, 39, 74, 77], and that individuals infected with subtype D reach higher viral loads more rapidly [2].

HIV sequencing is important for use in detecting drug resistant mutations, but can also provide insights about epidemic size and diversity e.g. [70] or movement between key populations by phylogenetic analysis e.g. [6, 44]. Sequencing in East Africa up until 2013 had been limited mostly to consensus Sanger sequences of partial gene sequences of p24 or gp41 [49], with very little genome sequence data from the twentieth century, although partial *pol* sequencing has recently become more common since the roll out of antiretroviral therapy. The PANGEA project [59] aimed to rectify this for the twenty-first century and has obtained large datasets of near full-length sequences from Africa to provide more detailed phylogenetic information [82].

Samples from serological surveys conducted in early 1986 from hospitals and antenatal clinics in Uganda were re-discovered in storage in 2013 during the relocation of what were then the Public Health England laboratories at Porton Down. Standard clinical *pol* sequencing [11] was attempted with some limited success [82], and amplification and sequencing success with the PANGEA protocol was also limited (unpublished) due to the age of the samples. To overcome barriers in the face of considerable RNA degradation, we used target-capture techniques with baits designed to capture a wide variety of HIV-1M to recover 109 new near full-length and 37 partial genomes. This is a unique population dataset from the early African epidemic, shortly after AIDS was discovered from a decade where few HIV genomes are available, particularly from Africa.

Methods

Sample preparation

Serum samples were collected from across Uganda between January and May 1986, including as part of a serological survey of HIV prevalence in different populations [13]. Samples were sent to Porton Down in the UK for antibody testing in 1986 and were subsequently stored there at − 80 °C. After their rediscovery they were passed to the PANGEA project in 2013.

In the current work, 168 HIV positive samples which had been identified by ELISA were RNA extracted with the QIAamp viral RNA mini kit (Qiagen). A target-capture approach [20] developed for samples with low concentrations or degraded RNA virus genomes was adopted. Thus 120 base pair capture baits were designed with an in-house pipeline to target the whole HIV genome, using 2635 reference genomes covering global subtype and CRF diversity (baits licensed to Agilent no. 5191-6709, SureSelectXT CD Pan HIV1). cDNA libraries were constructed with SuperScript IV Reverse Transcriptase (Invitrogen) followed by NEB Second Strand cDNA Synthesis before using the SureSelectXT Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library. This included a pre-capture PCR step during library preparation; followed by bait hybridization and a capture step with streptavidin beads to enrich for HIV fragments; and a post-capture indexing PCR. Paired end sequencing was carried out with the Illumina MiSeq v2 500 cycle kit.

Sequence assembly

Trimming, adapter removal, and quality checking of reads was performed with TrimGalore, cutadapt and FastQC [3, 47, 53], using a minimum Phred score of 30. Mapping to reference genomes was done with the Burrow-Wheeler Aligner MEM algorithm [52] and the samtools and bcftools libraries [19], firstly to 170 reference genomes (encompassing a wide range of subtype and CRF diversity) to identify the best genotype, and then to the best reference for a single reference assembly. A visual assessment in Geneious Prime 2022.0.1 (www.geneious.com) was carried out to check for good coverage across the genome, or any dips that might indicate an inter-subtype recombinant sequence. If this was the case, the multi-reference BAM files were examined, or an alternative de-novo assembly with HAPHPIPE and SPAdes was attempted [5, 29]. Either the single reference assembly (or de-novo assembly if improvement could be found) was then fed into the HAPHPIPE framework for fine tuning with three rounds of iterative improvement. Coverage statistics and vcf files were produced for each and finally a consensus sequence with a minimum of 10 × coverage at every base pair position was generated using GATK [54] within HAPHPIPE.

'Intermediate' and 'modern' datasets

In addition to the newly generated 'historical' dataset, a collection (n = 46) of genomes from the Rakai district (Uganda) in 1998 and 1999 provided an 'intermediate set' [34], whilst a 'contemporary set' was taken from the MRC/UVRI PANGEA genome collection (n = 465) sampled in Central Uganda between 2007 and 2016 (described fully in [30].

Subtyping and co-receptor prediction

All genomes were subtyped with the full genome version of SCUEAL [46]. All sequences were subjected to co-receptor prediction using the *geno2pheno* co-receptor tool [67] first by aligning the V3 loop by eye and extracting the amino acid sequence in Geneious Prime 2022.0.1 (www.geneious.com). The inter-subtype recombinant genomes (unique recombinant forms; URFs from all three datasets with a clear A1 or subtype D majority (over 70% the length of *env* as determined by SCUEAL breakpoints and clearly covering the V3 loop) were included. Subtype level consensus amino acid sequences were found and Shannon's entropy of the two were calculated then compared with the Entropy-Two tool from the Los Alamos Database (https://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html).

To investigate the origin of a deletion at position 24 in the V3 loop, additional data from the oldest subtype B and D envelope sequences were obtained from the Los Alamos National Laboratory (www.hiv.lanl.gov) for comparison. A BEAST [72] phylogeny was constructed (see [31], and an ancestral state reconstruction for presence or absence of the position 24 deletion by parsimony was then carried out with the R package 'castor' [71] and plotted in ggtree [83].

Results

Historical sequences

HIV specific baits were used in a target-capture step to enrich HIV genetic material before Illumina MiSeq sequencing to generate a paired-end read dataset of 109 near full-length consensus sequences with a minimum of 10 × coverage at every position. In addition to these near full-length consensus genomes (> 8000 bp from *gag* to *nef*), 37 partial sequences (> 1000 bp) were generated (a 65% genome recovery success from 168 samples, or 87% partial sequence recovery, Additional file 4: Table S1).

Average coverage spanned from × 27 to × 1769, with no significant difference found between subtypes or between subtypes and inter-subtype recombinants (Fig. 1). This method is considerably more sensitive than without the target-capture step; in 2014 some of these samples were subjected to the PANGEA protocol [27] with modest success, generating 5 near full-length genomes and 17 partial genomes, (a success rate of 5% and 22% respectively from a 96-sample plate; data not shown).

Of the 109 consensus genomes, 90 had some basic location information. The majority are from the "Central" region (n = 55) which includes Kampala and hospitals within Kampala, Rubaga (n = 12), Mulago (n = 8) and Nsambya (n = 7), and unidentified antenatal clinics (n = 2). A further 31 genomes were recovered from Kitovu Hospital (Masaka District), 4 from Lacor hospital in Gulu in northern Uganda, one from a hospital in Jinja (80 km East of Kampala) (see Additional file 5: Table S2 and map of Uganda in Additional file 1: Fig. S1).

Change in subtype frequency

The overall subtype distribution of the 109 historical 1986 genome set was as follows: 73 subtype D (67.0%); 17 subtype A1 (15.6%); 1 subtype C (0.9%) and 17 inter-subtype recombinants composed of A1, D (15.6%) and 1 composed of A1, C and D (0.9%). All inter-subtype recombinants had a unique recombination pattern (Additional file 2: Fig. S2). Subtype D was the most prevalent in all sampling locations, but particularly prevalent in Kitovu Hospital in Masaka, 130 km to the southwest of Kampala where 27/31 (87%) of genomes and 3/5 (60%) partial genomes were subtype D. The SCUEAL designated subtype distribution for the 'intermediate' genome dataset was 26 D (56.5%); 7 A1 (15.2%); and 13 inter-subtype recombinants containing A1, D, and C (28.2%).

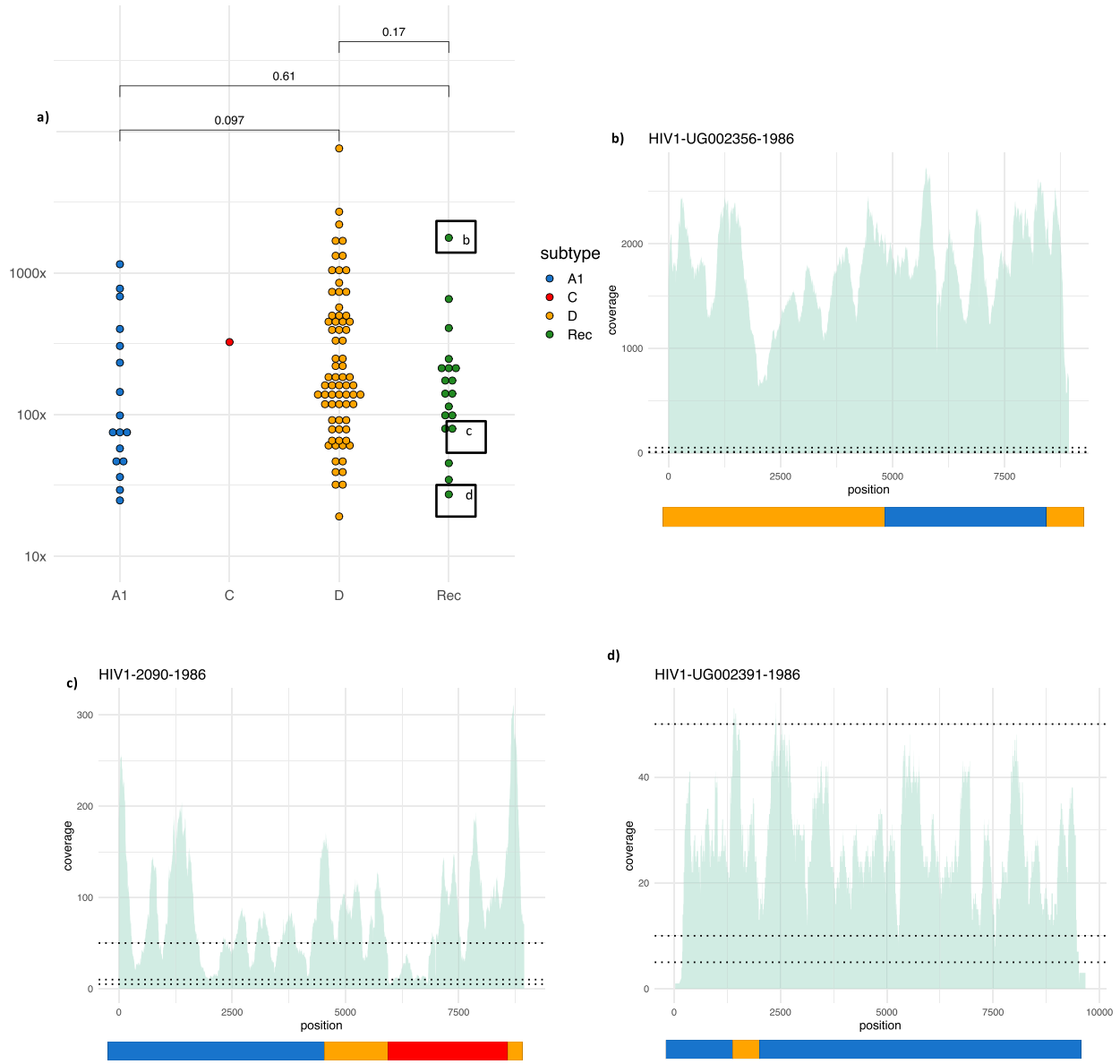

Fig. 1 Average read coverage across the genome for each genome assembly, and three example coverage plots of inter-subtype recombinant genomes from historical samples. Panel **a** gives the average coverage per genome assembly by subtype (with base 10 scale), which shows no significant difference between subtypes A1 and D, or between either subtype and the inter-subtype recombinants. Panel **b** shows the coverage plot of the inter-subtype recombinant with the highest average coverage (× 1769), panel **c** shows the only A1,C,D inter-subtype recombinant genome (average × 80), and **d** the lowest coverage inter-subtype recombinant (average × 27). The dotted horizontal lines show 5 × , 10 × and 50 × and the SCUEAL designated subtype and breakpoints are shown as schematic bars underneath each plot (subtype A1 in blue, D in orange, and C in red)

The 'modern set' had the distribution: 82 D (17.6%); 3 C (0.6%); 143 A1 (30.8%) and 232 inter-subtype recombinants (49.9%). The proportional change of genome level subtype over the three periods is illustrated in Fig. 2a. Combining other subtypes with recombinants, the relative frequencies of A1 and D, are significantly different in these three time periods ($\chi^2 = 122.68$, df = 4, $p < 0.0001$), and show a significant linear trend for reduction in subtype D genome frequency over time (Cochran Armitage, $Z = -10.861$, $p < 0.0001$).

Furthermore, the frequencies of subtypes A1 and D within the URF envelopes were also assessed (Fig. 2b). The majority subtype within the envelope region of all URFs was determined based on SCUEAL-estimated

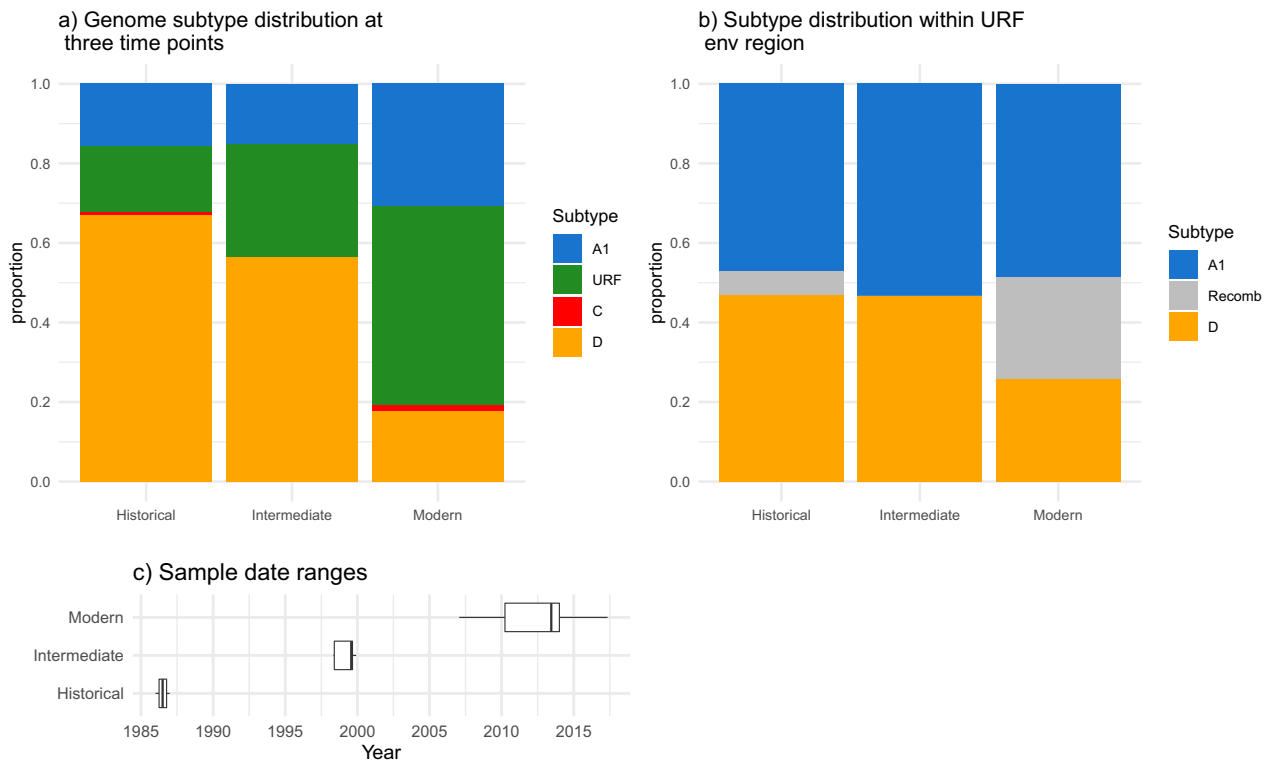

Fig. 2 **a** Ugandan subtype distribution from full-genome sequences at three time periods. Bars show subtype proportions in the 1986 'historical' genome set described here n = 109, the 'intermediate set' from Harris et al. (n = 46), and the 'modern' dataset from PANGEA (n = 465), **b** the subtype within URF envelope region over the same three time periods, and **c** the sampling date range of each period as box plots

breakpoints. A threshold of 70% over the length of the *env* gene, including the V3 loop, was used to classify *env* as D, A1, or a recombinant. The relative frequency of subtype D also falls in the URF envelope region over time (Cochran Armitage, Z = − 1.9225, p = 0.027), and considering all time periods together, there were many more URFs with subtype A1 envelopes (n = 131), than URFs with subtype D envelopes (n = 76), significantly different to the expected frequency from the genome level ($\chi^2 = 45.973$, df = 3, p < 0.0001).

Co-receptor usage

The machine learning application *geno2pheno* [51] was used to predict virus co-receptor tropism of all V3 sequences in the three datasets (Additional file 6: Table S3). Adopting a 5% false positivity rate threshold, there is a significantly higher proportion of CXCR4 coreceptor usage by subtype D compared with A1 during all three periods (Table 1). Of the historical set, 66% (53/80) of subtype D envelope sequences were predicted to be X4 tropic whilst none (0/24) were predicted to be X4 tropic for the A1 sequences ($\chi^2 = 29.8$, df = 1, p < 0.0001). Of the intermediate genomes, 33% (11/33) of subtype D were X4 tropic compared with none (0/13) of subtype A1 ($\chi^2 = 4.01$, df = 1, p = 0.04), and of the modern day 49% (70/143) subtype D were X4 tropic, compared with 5% (13/256) for subtype A1 ($\chi^2 = 104.5$, df = 1, p < 0.0001).

Subtype specific differences in V3 loop at the amino acid level

We used the Los Alamos Entropy-Two tool which uses randomisation with replacement to test for differences in entropy between subtypes. In total, 14 positions were significantly more entropic in Subtype D than A1 (including the crucial positions of 11 and 25), while three sites were more entropic in subtype A1 than D (positions 19, 22, and 24), see Table 2a.

The consensus length of subtype A1 was 35 codons, whilst that for Subtype D was 34 codons, due to a deletion at position 24 in the majority of both historic (94%; 68/72), and modern-day, (90%; 73/81) subtype D. Whilst the deletion 24 is found in the vast majority of Ugandan subtype D sequences, it is found only in some subtype D outgroup sequences, and not found in the Subtype B consensus (Table 2b). By mapping this deletion onto a subtype B/D phylogeny, we suggest that a deletion arose before the introduction of subtype D in Uganda, but also independently in some other subtype D lineages

Table 1 Co-receptor tropism predictions for subtypes D and A1 adopting the 5% false discovery rate from *geno2pheno*. Distinction is made between V3 sequences from genomes containing only one subtype and URFs

Subtype	Historic 1986			Intermediate 1998/9			Modern 2007-2016		
	X4	R5	Proportion X4	X4	R5	Proportion X4	X4	R5	Proportion X4
D (genome)	46	26	53/80 (66%)	9	17	11/33 (33%)	44	38	70/143 (49%)
D env (URF)	7	1		2	5		26	35	
A1 (genome)	0	16	0/24 (0%)	0	5	0/13 (0%)	5	136	13/256 (5%)
A1 env (URF)	0	8		0	8		8	107	
Other	0	1	0%	0	0	0%	5	55	8%
Total	53	52	50%	11	35	24%	88	371	19%

Table 2 **a** Consensus V3 amino acid sequences of subtypes A1 and D from Uganda with pairwise entropy comparison at each site, and **b** V3 sequences of outgroup sequences to subtype D in Uganda

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

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

(see ancestral state reconstruction in Additional file 3: Fig. S3). In the 'modern' subtype D dataset, there are a number of additional changes in the V3 loop, including a further codon deletion at position 23 in many sequences, confirming a distinctive difference in the behaviour of this region of *env* in this subtype (see alignment files in Additional files 7 and 8).

Discussion

Here we describe a population sample of 109 HIV genomes from the early stages of the epidemic in Uganda, and a period where very few HIV genomes are available globally. Most sequences from the early years of the epidemic are now retrospectively obtained by amplification of material from preserved serum or tissue. For example, the oldest sequence fragment to date (ZR59 from the DRC) was obtained from a 1959 plasma sample, but unfortunately, only a few hundred base pairs were sequenced [84] due to its degraded nature, and the limitations of the technology at the time of sequencing. Two well-known isolates (MAL and ELI; [1]) were the first full-length genome sequences generated from African samples obtained contemporaneously, but following passage in cell culture, which would have rapidly accumulated lab-induced changes [58]. We show here that target-capture with next generation sequencing can work well on highly degraded serum samples from over 30 years ago, and without cell passage induced errors. Yamaguchi et al. [81] have also successfully employed similar target-capture methods, obtaining genomes from a wide range of subtypes including from 1987 and the DRC. More recently "jackhammer" techniques recovered a 1966 genome sequence of a subtype C virus, where target-capture methods failed [33]. New sequencing

technologies now mean that we are increasingly limited more by the availability of preserved virus material rather than method sensitivity to recover sequences from old samples.

In this historical dataset, most genomes are 'pure' subtypes, consistent with the sample being taken during an early point in the Ugandan epidemic when the two subtypes had not co-existed for very long. However, we do find 18 inter-subtype recombinant forms, all of which have a unique pattern, representing at least 18 independent co-infection or super-infection events with different subtypes. Dual infection and recombination between these two subtypes was therefore occurring well before 1986. There is now an extremely high prevalence of unique recombinant forms in Uganda [12, 30, 50], without any evidence of a major circulating recombinant form. This is not unexpected within a generalised epidemic of such large scale and network complexity involving two subtypes at similar prevalence [7, 63], which has not experienced any obvious bottlenecks.

Ugandan cohort studies have been of global interest because unusually, the generalised epidemic provides a natural experiment for directly comparing the phenotypes of two distinct HIV subtypes. These cohort studies have consistently found subtype D to be more virulent than subtype A1, with faster drops in CD4 counts and more rapid progression to AIDS [38, 42, 69]). A faster rate of progression in Subtype D has been confirmed in neighbouring Tanzania [76] and in the UK [21].

There is an extensive literature on the subject of differences in virulence between viral strains, often framed in terms of viral load [9, 26, 35]. Viral load is a well-known predictor of HIV virulence [56]. However, cohort studies often report no significant difference in viral load between subtypes e.g. [10], and it appears that differences in viral load cannot explain differences in mortality risk between subtypes A1 and D [4, 55], suggesting that the "subtype D effect" contributes to virulence even after accounting for differences in viral load [22].

Like viral load, co-receptor usage is also well known to be associated with virulence in HIV [45, 65]. We found a significant co-receptor usage difference between subtypes D and A1, confirming what has been previously reported by studies with smaller sample sizes [36, 39]. Any observation made about co-receptor changes over time at the population level would be confounded by the disease stage of patients, since co-receptor switching is associated with advanced disease stage [16, 45], and many of the 1986 patients would have been experiencing severe AIDS, while many of the modern patients had access to antiretroviral therapies. Taking each of the three time points independently however, we found consistently that subtype D is more likely to be X4 tropic than subtype A1. This difference was particularly stark in the 1986 dataset where 66% (53/80) of subtype D envelopes had X4 tropic viruses compared to 0/24 subtype A1 envelopes. The high proportion of X4 tropism in subtype D may not be surprising given that the majority of the 1986 samples came from late-stage AIDS patients in hospitals, but this was true of *both subtypes*, and none of subtype A1 had an X4 tropism prediction.

Uganda is one of the best sampled countries in East Africa, and these are some of the largest African HIV genome datasets available, but even so, the data here represent only a tiny fraction of the Ugandan epidemic. Additional samples would lend more power to our findings, particularly if they could be stratified into regions and risk groups which are subject to some heterogeneity by subtype (see [6, 68, 70]. Any subtype specific amplification bias can be assumed absent in the historical dataset, since baits were designed with all HIV-1M diversity (2635 reference genomes) and there was no significant difference in read depth between the two subtypes. In the modern dataset, the near full-length genomes and partial genomes had a comparable subtype distribution [30] again suggesting the absence of preferential subtype specific amplification. For the intermediate [34] dataset however, samples underwent cell passage before nested PCR, which may have preferentially amplified X4 viruses and introduced artefacts [57, 75].

Previously, a change in relative proportion of the two subtypes has been shown using sequence fragments of *gag* and gp41 coding regions in a single district (Rakai), between 1994 and 2002 [17] and also between 1993 and 2012 [49]. We support these findings by showing an even more dramatic drop in subtype D, in near full-length genomes, sustained over a longer time period (1986–2016), and over a wider geographical area. All HIV subtype D genes decreased in frequency over time, but this was particularly true in *env*. We looked at URF genomes containing either A1 or D *env* segments and found that subtype D was under-represented compared to subtype A1 at the genome level. This suggests selection has acted with the help of recombination to preferentially include V3 loops with a higher propensity to be R5 tropic (subtype A1) over those with a propensity to be X4 tropic (subtype D) in URFs.

Finally, we show a subtype specific difference at the amino acid level of subtype D, where subtype D has higher entropy at many positions including the key positions 11 and 25, and find a deletion at position 24 which was likely present during the bottleneck of the expansion of subtype D into Uganda from the DRC. This deletion does not always confer X4 tropism, and the R5 phenotype is still predicted by *geno2pheno* for many sequences with the deletion. Instead, we propose that the inherited

"sequence space" of subtype D alters the mutational pathways available, pre-disposing subtype D to X4 tropism since there are a multitude of diverse mutational pathways that V3 loops can take to "switch" from R5 to X4 during the course of infection [61]. Interestingly, some authors have reported that the reverse is true for subtype C: which has a lower propensity to utilise CXCR4 [60], because it requires additional mutations to reach X4 tropism [15].

In the 1990s, Uganda mounted a concerted national effort from the highest levels in government down to grass roots which helped to encourage large scale behavioural changes [32]. Once the epidemic was no longer growing, HIV variants would have come under a selective pressure to lengthen the time to AIDS, thereby increasing their effective reproduction and expanding the exposure window [26]. We propose that selection acted against viruses most likely to encode the X4 phenotype, and favored those with the R5 phenotype. Differences in co-receptor switching propensity is therefore a very compelling explanation for the dramatic reduction in subtype D over time, and its association with more rapid disease progression as reported by various Ugandan cohort studies from the twentieth century.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12977-022-00612-5.

Additional file 1: Fig. S1. Map of Uganda showing the largest towns and cities including the sampling locations Kampala, Masaka, Jinja, and Gulu.

Additional file 2: Fig. S2. SCUEAL assessment of the 18 inter-subtype recombinants from historical samples. All had a unique recombination pattern. Subtype D fragments shown in orange, A1 in blue, subtype C in red.

Additional file 3: Fig. S3. A phylogenetic tree of subtype D (and outgroup subtype B) constructed with *env* and BEAST. Tip labels show presence or absence of a deletion at position 24 in the V3 loop, and an ancestral state reconstruction using parsimony is mapped onto the tree nodes.

Additional file 4: Table S1. Information about the 109 genomes and 37 partial sequences including SCUEAL subtype, read depth, location, and Genbank numbers

Additional file 5: Table S2. Frequency of HIV genomes by subtype recovered from each sampling location.

Additional file 6: Table S3. Amino acid sequences of all V3 loops from all three datasets and *geno2pheno* predictions.

Additional file 7: Subtype A1 fasta file alignment of V3 sequences at amino acid level.

Additional file 8: Subtype D fasta file alignment of V3 sequences at amino acid level.

Acknowledgements
We thank the MRC/UCLH-BRC funded UCL Pathogen Genomics Unit for their skilled assistance. Thanks also go to Katie Atkins, Julian Villabona-Arenas, Emma Pujol-Hodge, and James Baxter for helpful discussions

Author contributions
JWC, PC, DS and PK provided samples for this study which were tested and extracted by BF. JB developed the target-capture sequencing approach, performed by HT and RW, while SR designed the RNA baits and bioinformatic assembly pipeline. HG carried out all other analyses. HG and ALB conceived the study and wrote the manuscript which all authors reviewed. All authors read and approved the final manuscript.

Funding
HG was supported by the MRC Precision Medicine Doctoral Training Programme. ALB was supported through the PANGEA-HIV consortium with support provided by the Bill and Melinda Gates Foundation (OPP1084362), and by NIH (GM110749). PK and DS acknowledge support from the UKRI/MRC and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement. SR, RW and HT were funded by the EU FP7 PATHSEEK Grant from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 304875. JB received funding from the NIHR UCL/UCLH biomedical Research Centre.

Availability of data and materials
Near full-length genome consensus sequences have been deposited in GenBank (numbers OP039379:OP039487), and partial sequences (OP39488:OP039526), and read data are available on request. SCUEAL (full-length genome version) is available from (https://github.com/veg/hyphy-analyses).

Declarations

Ethics approval and consent to participate
Ethics approval was granted from the Uganda National Council for Science and Technology dated 14th September 2015 reference HS 1432, the Uganda Virus Research Institute Research Ethics Committee dated June 27th 2017 reference GC/127/17/06/428, and the School of Biological Sciences Ethics Committee dated 12th June 2018 reference ajlbrown-0002.

Consent for publication
All authors have reviewed and agreed to publication of this work.

Competing interests
The authors declare no competing interests.

Author details
[1]Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, UK. [2]Division of Infection and Immunity, University College London, London, UK. [3]UCL Great Ormond Street Institute of Child Health, London, UK. [4]Department of Virology, University College London Hospitals NHS Foundation Trust, London, UK. [5]Medical Research Council (MRC)/Uganda Virus Research Institute (UVRI) and London School of Hygiene and Tropical Medicine (LSHTM) Uganda Research Unit, Entebbe, Uganda. [6]Salisbury, UK. [7]London, UK.

Received: 3 August 2022 Accepted: 12 November 2022
Published: 13 December 2022

References

1. Alizon M, et al. Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. Cell. 1986;46(1):63–74. https://doi.org/10.1016/0092-8674(86)90860-3.
2. Amornkul PN, et al. Disease progression by infecting HIV-1 subtype in a seroconverter cohort in sub-Saharan Africa. AIDS. 2013;27(17):2775–86. https://doi.org/10.1097/QAD.0000000000000012.
3. Andrews S. FastQC: A quality control tool for high throughput sequence data. [Online] 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
4. Baeten JM, et al. HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. J Infect Dis. 2007;195(8):1177–80. https://doi.org/10.1086/512682.

5. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77. https://doi.org/10.1089/cmb.2012.0021.
6. 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. 2019. https://doi.org/10.1038/s41598-018-37458-x.
7. Bbosa N, et al. Phylogenetic and demographic characterization of directed HIV-1 transmission using deep sequences from high-risk and general population cohorts/groups in Uganda. Viruses. 2020;12(3):1–21. https://doi.org/10.3390/v12030331.
8. Bbosa N, Kaleebu P, Ssemwanga D. HIV subtype diversity worldwide. Curr Opin HIV AIDS. 2019;14(3):153–60. https://doi.org/10.1097/COH.0000000000000534.
9. Blanquart F, et al. A transmission-virulence evolutionary trade-off explains attenuation of HIV-1 in Uganda. eLife. 2016;5:1–32. https://doi.org/10.7554/elife.20492.
10. Bousheri S, et al. Infection with different HIV subtypes is associated with CD4 activation-associated dysfunction and apoptosis. J Acquir Immune Defic Syndr. 2009;52(5):548–52. https://doi.org/10.1097/QAI.0b013e3181c1d456.
11. Cane P. HIV drug resistance testing. Methods Mol Biol. 2011;665:123–32. https://doi.org/10.1007/978-1-60761-817-1_8.
12. Capoferri AA, et al. Recombination analysis of near full-length HIV-1 sequences and the identification of a potential new circulating recombinant form from Rakai, Uganda. AIDS Res Hum Retroviruses. 2020. https://doi.org/10.1089/aid.2019.0150.
13. Carswell JW. HIV infection in healthy persons in Uganda. AIDS. 1987;1(4):223–7.
14. Chalmet K, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. BMC Infect Dis. 2010. https://doi.org/10.1186/1471-2334-10-262.
15. Coetzer M, et al. Extreme genetic divergence is required for coreceptor switching in HIV-1 subtype C. J Acquir Immune Defic Syndr. 2011;56(1):9–15. https://doi.org/10.1097/QAI.0b013e3181f63906.
16. Connor RI, et al. Change in coreceptor use correlates with disease progression in HIV-1 infected individuals. J Exp Med. 1997;185(2):621–8. https://doi.org/10.3141/1543-18.
17. Conroy SA, et al. Changes in the distribution of HIV type 1 subtypes D and A in Rakai District, Uganda Between 1994 and 2002. AIDS Res Hum Retroviruses. 2010;26(10):1087–91. https://doi.org/10.1089/aid.2010.0054.
18. Daar ES, et al. Baseline HIV type 1 coreceptor tropism predicts disease progression. Clin Infect Dis. 2007;45(5):643–9. https://doi.org/10.1086/520650.
19. Danecek P, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):1–4. https://doi.org/10.1093/gigascience/giab008.
20. Depledge DP, et al. Specific capture and whole-genome sequencing of viruses from clinical samples. PLoS ONE. 2011;6(11). https://doi.org/10.1371/journal.pone.0027805.
21. Easterbrook PJ, et al. Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy. J Int AIDS Soc. 2010;13(1):1–9.
22. Eller MA, et al. HIV type 1 disease progression to AIDS and death in a rural Ugandan cohort is primarily dependent on viral load despite variable subtype and T-cell immune activation levels. J Infect Dis. 2015;211(10):1574–84. https://doi.org/10.1093/infdis/jiu646.
23. Essex M. Human immunodeficiency viruses in the developing world. Adv Virus Res. 1999;53(C):71–88. https://doi.org/10.1016/S0065-3527(08)60343-7.
24. Faria NR, et al. The early spread and epidemic ignition of HIV-1 in human populations. Science. 2014;346(6205):56–61. https://doi.org/10.1126/science.1256739.The.
25. Faria NR, et al. Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa. PLoS Pathog. 2019;15(12):1–23. https://doi.org/10.1371/journal.ppat.1007976.
26. Fraser C, et al. Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. Proc Natl Acad Sci USA. 2007;104(44):17441–6. https://doi.org/10.1073/pnas.0708559104.
27. Gall A, et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. J Clin Microbiol. 2012;50(12):3838–44. https://doi.org/10.1128/JCM.01516-12.
28. Geretti AM. HIV-1 subtypes: epidemiology and significance for HIV management. Curr Opin Infect Dis. 2006;19:1–7. https://doi.org/10.1097/01.qco.0000200293.45532.68.
29. Gibson KM, et al. Validation of variant assembly using haphpipe with next-generation sequence data from viruses. Viruses. 2020. https://doi.org/10.3390/v12070758.
30. Grant HE, et al. Pervasive and non-random recombination in near full-length HIV genomes from Uganda. Virus Evolution. 2020;6(1):1–12. https://doi.org/10.1093/ve/veaa004.
31. Grant HE. Characterisation of the Ugandan HIV epidemic with full-length genome sequence data from 1986 to 2016. Edinburgh: University of Edinburgh; 2022.
32. Green EC, et al. Uganda's HIV prevention success: the role of sexual behavior change and the national response. AIDS Behav. 2006;10(4):335–46. https://doi.org/10.1007/s10461-006-9073-y.
33. Gryseels S, et al. A near full-length HIV-1 genome from 1966 recovered from formalin-fixed paraffin-embedded tissue. Proc Natl Acad Sci USA. 2020. https://doi.org/10.1073/pnas.1913682117.
34. Harris M, et al. Among 46 near full length HIV type 1 genome sequences from Rakai District, Uganda, subtype D and AD recombinants predominate. AIDS Res Hum Retroviruses. 2002;18(17):1281–90. https://doi.org/10.1089/088922202320886325.
35. Hodcroft E, et al. The contribution of viral genotype to plasma viral setpoint in HIV infection. PLoS Pathog. 2014. https://doi.org/10.1371/journal.ppat.1004112.
36. Huang W, et al. Coreceptor tropism in human immunodeficiency virus type 1 subtype D: High prevalence of CXCR4 tropism and heterogeneous composition of viral populations. J Virol. 2007;81(15):7885–93. https://doi.org/10.1128/jvi.00218-07.
37. Kaleebu P, et al. Relationship between HIV-1 Env subtypes A and D and disease progression in a rural Ugandan cohort. AIDS. 2001;15(3):293–9. https://doi.org/10.1097/00002030-200102160-00001.
38. Kaleebu P, et al. Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. J Infect Dis. 2002;185(9):1244–50. https://doi.org/10.1086/340130.
39. Kaleebu P, et al. Relation between chemokine receptor use, disease stage, and HIV-1 subtypes A and D: results from a rural Ugandan cohort. J Acquir Immune Defic Syndr. 2007;45(1):28–33. https://doi.org/10.1097/QAI.0b013e3180385aa0.
40. Kalish ML, et al. Recombinant viruses and early global HIV-1 epidemic. Emerg Infect Dis. 2004;10(7):1227–34. https://doi.org/10.3201/eid1007.030904.
41. Kaslow R, et al. Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. Nat Med. 1996;2(4):405–11.
42. 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. J Infect Dis. 2008;197(5):707–13. https://doi.org/10.1086/527416.
43. Kiwanuka N, et al. HIV-1 viral subtype differences in the rate of CD4+ T-Cell decline. J Acquir Immune Defic Syndr. 2010;54(2):180–4. https://doi.org/10.1097/QAI.0b013e3181c98fc0.HIV-1.
44. Kiwuwa-Muyingo S, 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(10):1–23. https://doi.org/10.1371/journal.pone.0185818.
45. Koot M, et al. Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. Ann Intern Med. 1993;118(9):681–8. https://doi.org/10.7326/0003-4819-118-9-199305010-00004.
46. Kosakovsky Pond SL, et al. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. PLoS Comput Biol. 2009;5(11):1–21. https://doi.org/10.1371/journal.pcbi.1000581.
47. Krueger F. TrimGalore. [Online] 2020. https://github.com/FelixKrueger/TrimGalore.
48. Kuritzkes DR. HIV-1 subtype as a determinant of disease progression. J Infect Dis. 2008;197(5):638–9. https://doi.org/10.1086/527417.

49. Lamers SL, et al. HIV-1 subtype distribution and diversity over 18 years in Rakai, Uganda. AIDS Res Hum Retroviruses. 2020;36(6):522–6. https://doi.org/10.1089/aid.2020.0062.
50. Lee GQ, et al. Prevalence and clinical impacts of HIV-1 intersubtype recombinants in Uganda revealed by near-full-genome population and deep sequencing approaches. AIDS. 2017;31(17):2345–54. https://doi.org/10.1097/QAD.0000000000001619.
51. Lengauer T, et al. Bioinformatics prediction of HIV coreceptor usage. Nat Biotechnol. 2007;25(12):1407–10. https://doi.org/10.1038/nbt1371.
52. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. https://arXiv.org/abs/1303.3997v2 [q-bio.GN].
53. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. 2011;17(1):10–2.
54. McKenna A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303. https://doi.org/10.1101/gr.107524.110.20.
55. McPhee E, et al. Short communication: the interaction of HIV set point viral load and subtype on disease progression. AIDS Res Hum Retroviruses. 2019;35(1):49–51. https://doi.org/10.1089/aid.2018.0165.
56. Mellors JW, et al. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. Science. 1996;272(5265):1167–70. https://doi.org/10.1126/science.272.5265.1167.
57. Meyerhans A, et al. Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. Cell. 1989;58(5):901–10. https://doi.org/10.1016/0092-8674(89)90942-2.
58. Peden K, Emerman M, Montagnier L. Changes in growth properties on passage in tissue culture of viruses derived from infectious molecular clones of HIV-1LAI, HIV-1MAL, and HIV-1ELI. Virology. 1991;185(2):661–72. https://doi.org/10.1016/0042-6822(91)90537-L.
59. Pillay D, et al. PANGEA-HIV: Phylogenetics for generalised epidemics in Africa. Lancet Infect Dis. 2015;15(3):259–61. https://doi.org/10.1016/S1473-3099(15)70036-8.
60. Pollakis G, et al. Phenotypic and genotypic comparisons of CCR5- and CXCR4-tropic human immunodeficiency virus type 1 biological clones isolated from subtype C-infected individuals. J Virol. 2004;78(6):2841–52. https://doi.org/10.1128/jvi.78.6.2841-2852.2004.
61. Poon AFY, et al. Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data. PLoS Comput Biol. 2012. https://doi.org/10.1371/journal.pcbi.1002753.
62. Rambaut A, et al. The causes and consequences of HIV evolution. Nat Rev Genet. 2004;5(1):52–61. https://doi.org/10.1038/nrg1246.
63. Ratmann O, et al. Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda. Lancet HIV. 2020;7(3):e173–83. https://doi.org/10.1016/S2352-3018(19)30378-9.
64. Robertson DL, et al. HIV-1 nomenclature proposal. Science. 2000;288(5463):55. https://doi.org/10.1126/science.288.5463.55d.
65. Schuitemaker H, Van'Wout AB, Lusso P. Clinical significance of HIV-1 coreceptor usage. J Transl Med. 2010;9(1):1–17. https://doi.org/10.1186/1479-5876-9-S1-S5.
66. Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. Cold Spring Harb Perspect Med. 2011. https://doi.org/10.1101/cshperspect.a006841.
67. Sing T, et al. Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. Antivir Ther. 2007;12(7):1097–106. https://doi.org/10.1177/135965350701200709.
68. Ssemwanga D, et al. 'HIV type 1 subtype distribution, multiple infections, sexual networks, and partnership histories in female sex workers in Kampala, Uganda. AIDS Res Hum Retroviruses. 2012. https://doi.org/10.1089/aid.2011.0024.
69. Ssemwanga D, et al. Effect of HIV-1 subtypes on disease progression in rural Uganda: a prospective clinical cohort study. PLoS ONE. 2013. https://doi.org/10.1371/journal.pone.0071768.
70. Ssemwanga D, et al. The molecular epidemiology and transmission dynamics of HIV type 1 in a general population cohort in Uganda. Viruses. 2020;12(11):1–17. https://doi.org/10.3390/v12111283.
71. Stilianos L, Doebeli M. Efficient comparative phylogenetics on large trees. Bioinformatics. 2017. https://doi.org/10.1093/bioinformatics/btx701.
72. Suchard MA, et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution. 2018;4(1):1–5. https://doi.org/10.1093/ve/vey016.
73. Thielen A, et al. Improved prediction of HIV-1 coreceptor usage with sequence information from the second hypervariable loop of gp120. J Infect Dis. 2010;202(9):1435–43. https://doi.org/10.1086/656600.
74. Tscherning C, et al. Differences in chemokine coreceptor usage between genetic subtypes of HIV-1. Virology. 1998;241(2):181–8. https://doi.org/10.1006/viro.1997.8980.
75. Vartanian JP, et al. Selection, recombination, and G to A hypermutation of human immunodeficiency virus type 1 genomes. J Virol. 1991;65(4):1779–88. https://doi.org/10.1128/jvi.65.4.1779-1788.1991.
76. Vasan A, et al. Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype. Clin Infect Dis. 2006;42(6):843–52. https://doi.org/10.1086/499952.
77. Wambui V, et al. Predicted HIV-1 coreceptor usage among Kenya patients shows a high tendency for subtype D to be CXCR4 tropic. AIDS Res Ther. 2012;9:1–7. https://doi.org/10.1186/1742-6405-9-22.
78. Ward MJ, et al. Estimating the rate of intersubtype recombination in early HIV-1 group M strains. J Virol. 2013;87(4):1967–73. https://doi.org/10.1128/JVI.02478-12.
79. de Wolf F, et al. Syncytium-inducing and non-syncytium-inducing capacity of human immunodeficiency virus type 1 subtypes other than B: phenotypic and genotypic characteristics. AIDS Res Hum Retroviruses. 1994;10(11):1387–400. https://doi.org/10.1089/aid.1994.10.1387.
80. Worobey M, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature. 2008;455(7213):661–4. https://doi.org/10.1038/nature07390.
81. Yamaguchi J, et al. Universal target capture of HIV sequences from NGS libraries. Front Microbiol. 2018; pp. 1–13. https://doi.org/10.3389/fmicb.2018.02150.
82. Yebra G, et al. Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. In: Scientific Reports. Nature Publishing Group; 2016, pp. 1–6. https://doi.org/10.1038/srep39489.
83. Yu G, et al. Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1):28–36. https://doi.org/10.1111/2041-210X.12628.
84. Zhu T, et al. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature. 1998;391:594–7.

Publisher's Note