

Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth

Polly R. Walker, Oliver G. Pybus, Andrew Rambaut, Edward C. Holmes*

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Received 14 January 2004; accepted 17 June 2004

Available online 7 December 2004

Abstract

The human immunodeficiency virus (HIV) pandemic continues to grow at an alarming rate, with a further 5 million new infections in 2003. Some 3.5 million of these were in sub-Saharan Africa, where ~70% of the world's HIV-positive population resides. In contrast, the spread of HIV in high-income countries has slowed since its discovery in the 1980s, and in regions such as Western Europe prevalence has decreased. Here, we employ coalescent methods to compare the epidemic growth rates of two subtypes of HIV-1 with differing epidemiological profiles: subtype C, which is dominant in sub-Saharan Africa and associated with heterosexual transmission, and subtype B, the main cause of AIDS in Western Europe and North America, and which was primarily transmitted through homosexual sex and injecting drug use. We show that although both subtypes emerged at approximately the same time (~1960), they have widely differing patterns of exponential population growth. At its current growth rate the epidemic of subtype C in sub-Saharan Africa is doubling every 2.4 years, which is approximately half the rate observed during the early stages of the subtype B epidemic in Western Europe and North America. However, the subtype C growth rate is still 5–10 times greater than that estimated for the blood-borne hepatitis C virus, supporting the hypothesis that sexual transmission has been primarily responsible for the HIV epidemic in sub-Saharan Africa.

© 2004 Elsevier B.V. All rights reserved.

Keywords: HIV-1; Population dynamics; Maximum likelihood; Heterosexual transmission; Coalescent theory

1. Introduction

There are currently over 29 million people in sub-Saharan Africa infected with the human immunodeficiency virus (HIV), accounting for ~70% of the global total (UNAIDS, 2003). This number is increasing at an alarming rate. In 2003, a further 5 million people became infected with HIV, and while there is some hope that the epidemic may be reaching a plateau in countries such as Senegal and Uganda, in most of Africa the spread of HIV is showing no signs of decelerating (UNAIDS, 2003).

Phylogenetic analysis allows classification of HIV type 1 (HIV-1) sequences into three groups: M, N and O. The M (main) group has been further subdivided into 14 subtypes and 16 established circulating recombinant forms (CRFs) (Kuiken et al., 2002). Subtype C has the highest global

prevalence, currently estimated at 56% of all HIV-1 infections worldwide, and is now the dominant subtype in sub-Saharan Africa (McCormack et al., 2002; Osmanov et al., 2002). Reported cases of other subtypes and CRFs remain relatively low in many parts of sub-Saharan Africa (Bredell et al., 2000). The main transmission route for subtype C viruses appears to be heterosexual contact (Buve et al., 2001; Van Harmelen et al., 1999; Williams and Gouws, 2001).

As well as its magnitude, the AIDS epidemic in sub-Saharan Africa is notable for the variation in prevalence and growth rates among localities. This variation is most strongly associated with factors impacting transmission efficiency such as circumcision and the presence of co-factor sexually transmitted infections, including herpes simplex virus type 2 (HSV-2) (Auvert et al., 2001; Buve et al., 2001; Buve et al., 2002). Socio-economic factors are also important in determining variation in HIV growth rates, although, notably, two of the wealthiest countries in southern

* Corresponding author. Tel.: +44 1865 281997; fax: +44 1865 310447.
E-mail address: edward.holmes@zoo.ox.ac.uk (E.C. Holmes).

Africa, Botswana and South Africa, have higher HIV prevalence than some of their poorer neighbours. In Botswana, the prevalence of HIV in women attending antenatal clinics rose by 6.4% in 2001 alone (UNAIDS, 2002a), and with 38.8% of adults between 15 and 49 infected it has the highest prevalence of HIV infection worldwide. The history of South Africa's HIV epidemic differs from other countries in sub-Saharan Africa in that the first HIV cases were observed in male homosexuals during the early 1980s and assigned to subtype B (Williams and Gouws, 2001). Until 1993, HIV prevalence in women attending antenatal clinics remained below 2%, but rose dramatically to 10.44% by 1995 (UNAIDS, 2002b). HIV prevalence in adults is currently estimated at 24.8%, but has reached 35% in some of the worst hit areas, such as KwaZulu-Natal (South African Department of Health, 2001), making it one of the fastest growing HIV epidemics in sub-Saharan Africa.

The subtype B epidemic in Western Europe and North America has a very different history. The first cases of HIV-1 were observed in homosexual men and Haitian immigrants in the US in the early 1980s (Gottlieb et al., 1981; Laverdiere et al., 1983). In the years that followed, subtype B spread extremely rapidly within specific risk groups such as homosexuals and injecting drug users (Selik et al., 1984). The majority of subtype B infections in the US are thought to be descended from infections introduced into the homosexual population in the 1970s (Foley et al., 2000). However, after this initial rapid epidemic growth, the rate of new infections began to slow during the 1990s, and AIDS incidence to steadily decrease from 1994, due to the success of community education and needle exchange programs as well as the introduction of highly active anti-retroviral therapy (HAART) (MAP, 2000). Subtype B currently accounts for 12.3% of infections globally, with the majority of these in developing regions such as South-East Asia and Latin America. In high-income countries HIV prevalence remains below 1% and the overall trend of the epidemic suggests a deceleration in growth rate over time (UNAIDS, 2003). Despite recent increases in the number of new infections in heterosexual populations, the subtype B epidemic in the industrialized world remains primarily associated with homosexual contact and injecting drug use (UNAIDS, 2003).

In recent years, various coalescent approaches have been developed to estimate key population parameters, most notably growth rates from gene sequence data. Under coalescent theory the distribution of coalescent times between sequences can be used to estimate the demographic history of that population (Griffiths and Tavaré, 1994; Kingman, 1982). Consequently, branch lengths within phylogenetic trees constructed under the assumption of a molecular clock contain information about changes in growth rate and population structure over time. In the context of viral infections, this enables the estimation of key demographic parameters such as the epidemic growth rate (r) and the effective number of infections, which in turn can

be used to make predictions about the future of that population (Pybus et al., 2001).

In an early application of coalescent theory to the study of HIV-1, the epidemic was shown to be growing exponentially at a constant rate world-wide (Holmes et al., 1995). Subsequent analyses of subtypes A and B supported the hypothesis of a relatively constant exponential growth rate, although demographic differences between subtypes were observed (Grassly et al., 1999; Holmes et al., 1999; Pybus et al., 1999, 2000). In particular, subtype B showed a remarkably rapid increase in growth rate typical of the introduction of an infection into a fully connected network of susceptible individuals, or a "standing network" of transmission, hence infection rapidly reaches saturation within that population. Indeed, the subtype B sequences used in these studies were mainly taken from North America during the early stages of the epidemic, when it was predominantly associated with populations of homosexuals and injecting drug-users. In contrast, subtype A from Africa showed a slower exponential growth rate and a less sudden expansion, although the most recent coalescent analysis of sequences from central Africa indicates that the exponential growth rate in this region may have increased (Yusim et al., 2001).

Given the power of coalescent methods, it is particularly important to determine the growth trajectory of HIV-1 subtype C, which is primarily heterosexually transmitted and has become the most globally widespread variant in the last 10 years. In this study, we apply a coalescent-based method to derive demographic information about the subtype C epidemic in sub-Saharan Africa and to determine whether there are regional differences in this growth rate. Due to the size of their epidemics, we focus on Botswana and South Africa. We compare these results to an equivalent analysis conducted on subtype B sequence data from high-income countries sampled between 1983 and 2000. The coalescent method we employed allows the testing of alternative models of population growth within a maximum likelihood framework and provides estimates of the confidence intervals (CI) around the parameter estimates (Pybus et al., 2000). We also consider a diverse set of epidemic growth models. In particular, we assess how well the subtype B and C sequence data fits models of constant, exponential, logistic and generalized logistic (constant-exponential-constant) population growth (Pybus and Rambaut, 2002) and also estimate epidemic doubling times. Furthermore, we take account of different sampling times of the sequences analyzed by using "tip-dated" molecular clock trees (Rambaut, 2000), rather than assuming that all sequences were sampled contemporaneously.

2. Materials and methods

2.1. Sequence data

All HIV-1 gene sequences were taken from the Los Alamos HIV Sequence Database (<http://hiv-web.lanl.gov/>;

Table 1
Estimated substitution rates for HIV-1 subtypes B and C

Subtype	Gene	Position ^a	Length (nt)	<i>N</i>	Date range	Geographic sampling (<i>N</i>)	π^b	μ^c (CI)
C	<i>env</i>	6225–8792	2451	32	1986–2000	Asia (7), Central/West Africa (11), Southern Africa (12), South America (2)	0.1275	0.00394 (0.00312, 0.00479)
	<i>gag</i>	790–2289	1656	29	1986–2000	Asia (8), Central/West Africa (3), Southern Africa (18)	0.0858	0.00150 (0.00005, 0.00295)
	RT	2580–3485	906	38	1986–2001	Asia (7), Europe (1), Central/West Africa (5), South America (3), Southern Africa (21)	0.0602	0.00066 (0.00030, 0.00127)
B	<i>env</i>	6225–8792	2451	39	1983–1997	Asia (1), Australasia (6), Europe (6), North America (26)	0.1068	0.003084 (0.002514, 0.003684)
	<i>gag</i>	790–2289	1617	41	1983–2000	Asia (1), Australasia (8), Europe (10), North America (17), South America (5)	0.0629	0.00191 (0.00153, 0.00231)
	RT	2581–3497	927	41	1983–2000	Asia (1), Australasia (8), Europe (10), North America (17), South America (5)	0.0465	0.00095 (0.00062, 0.00130)

^a Gene positions are given according to HIV-1 strain HXB2 numbering.

^b Mean pairwise genetic distance.

^c Number of nucleotide substitutions per site per year.

Kuiken et al., 2002) and aligned manually using the Se–Al program (<http://evolve.zoo.ox.ac.uk/software.html>). Importantly, regions of uncertain alignment, most notably the hypervariable regions in the *env* gene, were removed prior to analysis. Two sets of data were collated, which differed according to the analytical task at hand. Table 1 describes the sequences used for the estimation of substitution rates for both subtypes B and C. To increase the accuracy of this rate estimation, these sequences were selected to maximize the geographic distribution and the range of sampling times (the geographical distribution of subtype B and C *env* sequences analyzed is shown in Fig. 1). The entire *gag* (capsid), *env* (envelope) and partial *pol* (reverse transcriptase, RT) genes were used. The subtype B sequences were selected from a broad time range encompassing the years 1983–2000 and representing high-income countries from the Americas, Western Europe and Australia. A phylogenetic analysis of these sequences (not shown) suggests that they have a common epidemiological history, so we are justified in

treating them collectively. These subtype B sequences were also used to estimate population dynamics. However, a second data set was used to estimate population dynamics in subtype C, chosen to maximize the number of samples from each geographic location (Table 2). These subtype C sequences were largely sampled from Botswana (prefix “BW”) and South Africa (prefix “ZA”), which were analyzed separately. The *env*, *gag* and RT sequences from Botswana were derived from full-length genome sequences and from the same set of infected individuals, whereas each gene in the South African data set was derived from different patient groups, with little overlap. Because phylogenetic analyses (not shown) showed some degree of common ancestry between these countries, we also pooled all sequences from sub-Saharan Africa (designated “SSAC”). This data set comprised all sequences from Botswana, South Africa, plus additional sequences available from Zambia (*env* *N* = 4, *gag* *N* = 2, RT *N* = 2), Malawi (*env* *N* = 4), and the East African Community (RT *N* = 9).

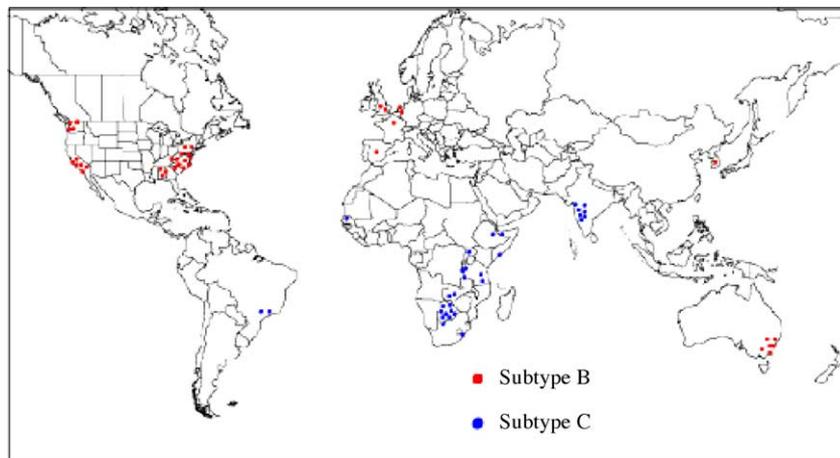


Fig. 1. A sample distribution map for subtypes B and C sequences used in the estimation of nucleotide substitution rates for the HIV-1 *env* genes. Geographical distributions for *gag* and RT data are similar (Table 1).

Table 2
Population dynamic estimates for HIV-1 subtypes B and C

Subtype	Geographic region	Gene	N	Length (nt)	π^a	T_{mrc} ^b (CI)	r^c (CI)	λ^d (CI)	Best-fit Model	–ln L ^e	–ln L	–ln L
										E	L	CEC
C	SSAC ^f	<i>env</i>	87	2451	0.129	1950 (1938–1958)	0.266 (0.235, 0.297)	2.607 (2.945, 2.331)	E	657.99	658.53	658.53
		<i>gag</i>	71	1656	0.095	1953 (1942–1961)	0.273 (0.233, 0.316)	2.536 (2.974, 2.193)	E	436.36	436.39	436.39
		RT	81	918	0.062	1956 (1933–1968)	0.274 (0.236, 0.335)	2.530 (2.943, 2.070)	E	556.25	556.25	556.25
	Botswana	<i>env</i>	51	2451	0.136	1951 (1940–1959)	0.228 (0.195, 0.259)	3.044 (3.548, 2.680)	E	271.28	271.28	271.28
		<i>gag</i>	51	1656	0.098	1957 (1946–1964)	0.266 (0.219, 0.317)	2.608 (3.170, 2.188)	E	295.82	295.82	295.82
		RT	48	921	0.069	1952 (1925–1965)	0.235 (0.193, 0.280)	2.951 (3.595, 2.473)	E	302.81	302.81	302.81
	South Africa	<i>env</i>	40	2541	0.128	1958 (1948–1965)	0.297 (0.248, 0.347)	2.337 (2.793, 2.000)	E	261.37	261.37	261.37
		<i>gag</i>	18	1656	0.088	1968 (1960–1973)	0.405 (0.288, 0.530)	1.713 (2.405, 1.309)	E	91.77	92.65	92.68
		RT	97	918	0.059	1958 (1936–1970)	0.235 (0.206, 0.266)	2.944 (3.368, 2.601)	E	682.77	682.77	682.77
B	High income countries	<i>env</i>	39	2979	0.107	1960 (1952–1966)	0.479 (0.386, 0.577)	1.446 (1.796, 1.202)	E	219.40	219.40	219.40
		<i>gag</i>	41	1629	0.063	1962 (1953–1969)	0.483 (0.363, 0.609)	1.437 (1.910, 1.139)	L	247.84	249.89	251.44
		RT	41	927	0.047	1959 (1937–1970)	0.522 (0.408, 0.641)	1.327 (1.697, 1.082)	L	273.03	280.71	281.26

^a Mean pairwise genetic distance.

^b Date of most recent common ancestor estimated from tip-dated tree node depth/substitution rate.

^c Maximum likelihood parameter estimates for r (growth rate, years⁻¹).

^d λ (epidemic doubling time, years).

^e Log likelihood (–ln L) values for each exponential growth model tested: exponential (E), logistic (L) and generalized logistic (CEC).

^f SSAC data set includes all samples listed for South Africa and Botswana plus additional sequences available for *env* and *gag* gene region. These are *env* (Zambia $N = 4$, Malawi $N = 4$), *gag* (Zambia $N = 2$).

For both subtypes, only one sequence was taken from individual patients and known recombinants (identified as intersubtype recombinants within the Los Alamos Database) were excluded in accordance with the assumptions of the coalescent model (i.e., that sampling is random and no recombination has occurred). Additionally, sequences with anomalously long-branch lengths suggesting a recombinant ancestry were also removed. All subtype C sequences used in this analysis were from patients who were naïve to antiretroviral treatment at the time of sampling (Novitsky et al., 1999; Gordon et al., 2003; Pillay et al., 2002). Subtype B sequences were predominantly sampled prior to the widespread use of effective antiretroviral therapy. Moreover, any natural selection associated with antiviral treatment affects a small number of sites within the RT gene and so is unlikely to have a large effect on either the substitution rate estimates averaged across the entire gene or on the overall inter-host tree topology.

2.2. Inferring phylogenetic trees

Maximum likelihood (ML) trees for each data set were estimated using the PAUP* package (Swofford, 2003). We

employed the general time reversible (GTR) model of nucleotide substitution, incorporating maximum likelihood estimates of base composition, the proportion of invariant sites (I) and the shape parameter (α) of a gamma distribution (Γ) model of among-site rate variation (with four rate categories). This model was chosen over a codon position rate model as it consistently gave much higher likelihoods. All parameter values are available from the authors on request. ML trees were estimated under this model using TBR branch swapping.

For the coalescent analysis, the variable rate ML trees described above were used to produce molecular clock (constant rate) trees taking into account the sampling times of the sequences in question (“tip-dated” trees—see below) using the RHINO program (available at <http://evolve.zoo.ox.ac.uk/software.html>). In this case, substitution parameters were re-estimated under the assumption of a molecular clock and rooted using outgroups determined in comparison to a single sequence from a different subtype (for example, a subtype C sequence in the case of the subtype B tree). Although subtype B tip-dated trees were co-estimated with substitution rates, the subtype C tip-dated trees were scaled with the substitution

rates estimated from subtype B (see Section 3 for explanation).

2.3. Estimating substitution rates

As our coalescent method requires phylogenetic estimation under a molecular clock, it is necessary to know the viral substitution rate (μ), calculated as the number of nucleotide substitutions per site per year. This rate was previously estimated for the HIV-1 group M *env* genes by plotting branch lengths by sampling time on phylogenetic trees and estimating a regression line, with confidence intervals generated using bootstrapping (Korber et al., 2000). We used a similar method to obtain an initial graphical approximation of the clock-like behavior of subtype B and C by plotting root-to-tip distances through time on ML trees. In this method, a rooted phylogeny is estimated (as described above) and a linear regression is constructed using the sampling times of each tip against its genetic distance to the root, according to the model:

$$E[d_{\text{root},i}] = \mu(t_i - t_{\text{root}}) = \mu t_i - m t_{\text{root}}$$

where t_{root} is the unknown time to the root of the tree (in years), t_i is the time of tip i , and μ is the substitution rate estimated by the gradient of the linear regression of $d_{\text{root},i}$ against t_i .

To obtain more accurate substitution rates for each data set we used the “tip-date” model available in the RHINO program. This provides a maximum likelihood estimate of the substitution rate under the assumption of a constant rate of evolution across all branches and incorporating sequences sampled at different times—the “single rate dated tip” (SDRT) model (Rambaut, 2000). As before, the GTR + I + Γ model of nucleotide substitution was used and confidence intervals were determined using a likelihood ratio statistic. Using these substitution rates we also calculated the time to the most recent common ancestor (T_{mrca}) in years using the equation $T_{\text{mrca}} = \delta/\mu$, where δ is the root height of the molecular clock tree, and μ is the substitution rate per site per year as estimated above. Finally, we also tested (using a standard likelihood ratio test in RHINO) the molecular clock assumption in each data set by comparing the log likelihood of the variable rate ML trees with those obtained under the SDRT model, which assumes a constant rate of substitution.

2.4. Estimating demographic history

We used a statistical method based on coalescent theory to estimate demographic parameters for HIV-1 subtypes B and C (Pybus et al., 2000). Using the GENIE program (Pybus and Rambaut, 2002) we first constructed skyline plots under a chronological time-scale employing the substitution rates estimated previously. Skyline plots provide a graphical and non-parametric estimate of the change in effective population size during each coalescent

interval and hence outline the demographic history of the population through time. We then compared, sequentially, the models of constant, exponential, logistic and generalized logistic (constant–exponential–constant) population growth. The exponential model is a special case of the logistic model, the latter of which describes a curve with an exponential phase followed by a deceleration in growth rate. The best-fit model chosen was that with the highest likelihood, provided a difference in log likelihood greater than 1.92 was observed (Wilks, 1938). This is considered a significant evidence to reject a simple model, such as exponential growth, by a more complex one such as logistic growth. For each data set we also estimated the exponential growth rates (r), with approximate lower and upper confidence limits computed using the likelihood ratio statistic. Growth rate estimates were in turn used to calculate the time taken for the epidemic to double in size (λ) in years, using the relation $\lambda = \ln(2)/r$.

3. Results

3.1. Rates and dates of HIV-1 subtype B and C evolution

Regression plots of root-to-tip distances against time for HIV-1 subtypes B and C suggest that the sequence data sets used in this analysis have a strong temporal structure (Fig. 2). However, in the maximum likelihood analysis, the variable and tip-dated trees had significantly different likelihoods (Table 2), hence rejecting a strict molecular clock. This indicates that some rate variation is present in these data and that this should be considered when estimating divergence times. The shallower sampling date range for subtype C results in larger confidence intervals around maximum likelihood rate estimates (Table 1) and lower concordance between genes in estimated times of origin of this data set (Fig. 2). By comparison, the subtype B data is more evenly distributed through time and shows good concordance between different genes.

Maximum likelihood estimates of substitution rate were obtained separately for all *env*, *gag* and RT data sets using RHINO (Table 1). For both subtypes, substitution rates for *env* and *gag* were in the range of $2.51\text{--}4.79 \times 10^{-3}$ and $0.05\text{--}2.95 \times 10^{-3}$ substitutions per site per year, respectively. The lowest substitution rate was in RT at between $0.30\text{--}1.30 \times 10^{-3}$ substitutions per site per year. Most notably, confidence intervals overlap for each rate estimate between the two subtypes, thereby revealing no major subtype-specific differences in substitution rate across genes. However, the confidence intervals for subtype C, most notably in the *gag* data, are very large due to the uneven distribution through time of sequences available. As such, point estimates for the subtype C substitution rate are likely to be less reliable than those obtained for subtype B. For this reason, we used the substitution rates from subtype B to estimate the divergence times in the subtype C data sets.

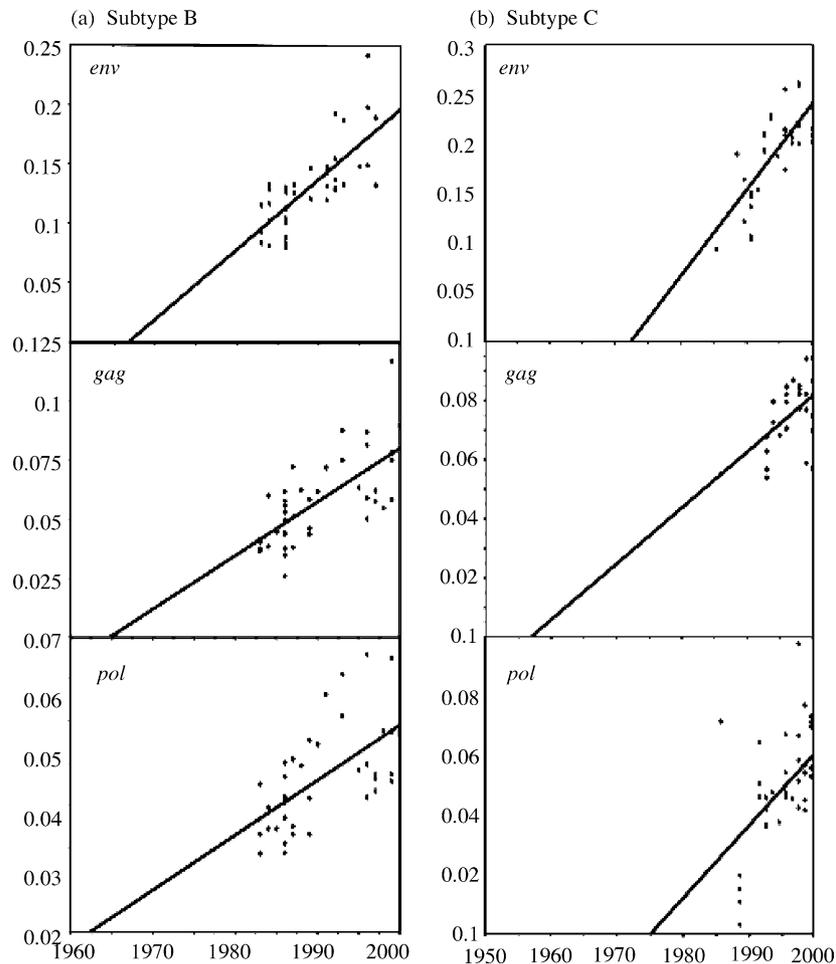


Fig. 2. Linear regression plots, providing an estimate of the nucleotide substitution rate per site per year (μ), of root-to-tip distance against sampling time for the *env*, *gag* and RT genes from HIV-1 subtypes B (a) and C (b).

With these rates of nucleotide substitution, we were able to estimate the times of the most recent common ancestors (T_{mrca}) for HIV-1 subtypes B and C (Table 2). For both subtypes, confidence intervals around T_{mrca} estimates overlap, supporting a similar emergence date of the two epidemics in the late 1950s to early 1960s. Within subtype C, confidence intervals overlap between South Africa and Botswana, despite a later documented expansion of the epidemic in South Africa compared to its neighbours (Williams et al., 2001). This further supports the view that there have been multiple introductions of HIV-1 into South Africa, as suggested by the phylogenetic trees.

3.2. Demographic history of HIV-1 subtypes B and C

We found the best-fit model of population growth and maximum likelihood estimate of the epidemic growth rate r (plus upper and lower confidence limits) for each tip-dated tree using GENIE. In turn, these were used to estimate the epidemic doubling time (λ) as described above. Population growth models of constant, exponential, logistic and generalized logistic (constant–exponential–constant) were

compared for all data sets (Table 2). We also present a graphical representation of epidemic history using skyline plots, scaled in years using the subtype B substitution rate (Fig. 3a–d).

In the sub-Saharan subtype C data sets, a model of exponential population growth was the best fit to the data, with a constant rate of growth over the entire epidemic history (Fig. 3a). Although the logistic growth model had a slightly higher likelihood than the exponential growth model in the *gag* and *env* genes, this was insufficient to reject the latter model. Moreover, no significant differences were found in r and λ between the *env*, *gag* and RT genes (Table 2). Over the course of the epidemic we can, therefore, estimate that the sub-Saharan Africa subtype C had a constant exponential growth rate of 0.271 year^{-1} (0.247 and 0.337) or a doubling time of 2.558 years (2.818 and 1.996), averaged across all three genes.

In contrast, the *gag* and RT data from subtype B data show a significant logistic trend; that is, the epidemic growth rate has been slowing over time (Fig. 3b). In particular, the epidemic appeared to start slowing in growth rate during the early 1990s. For the *env* gene, fewer recently sampled

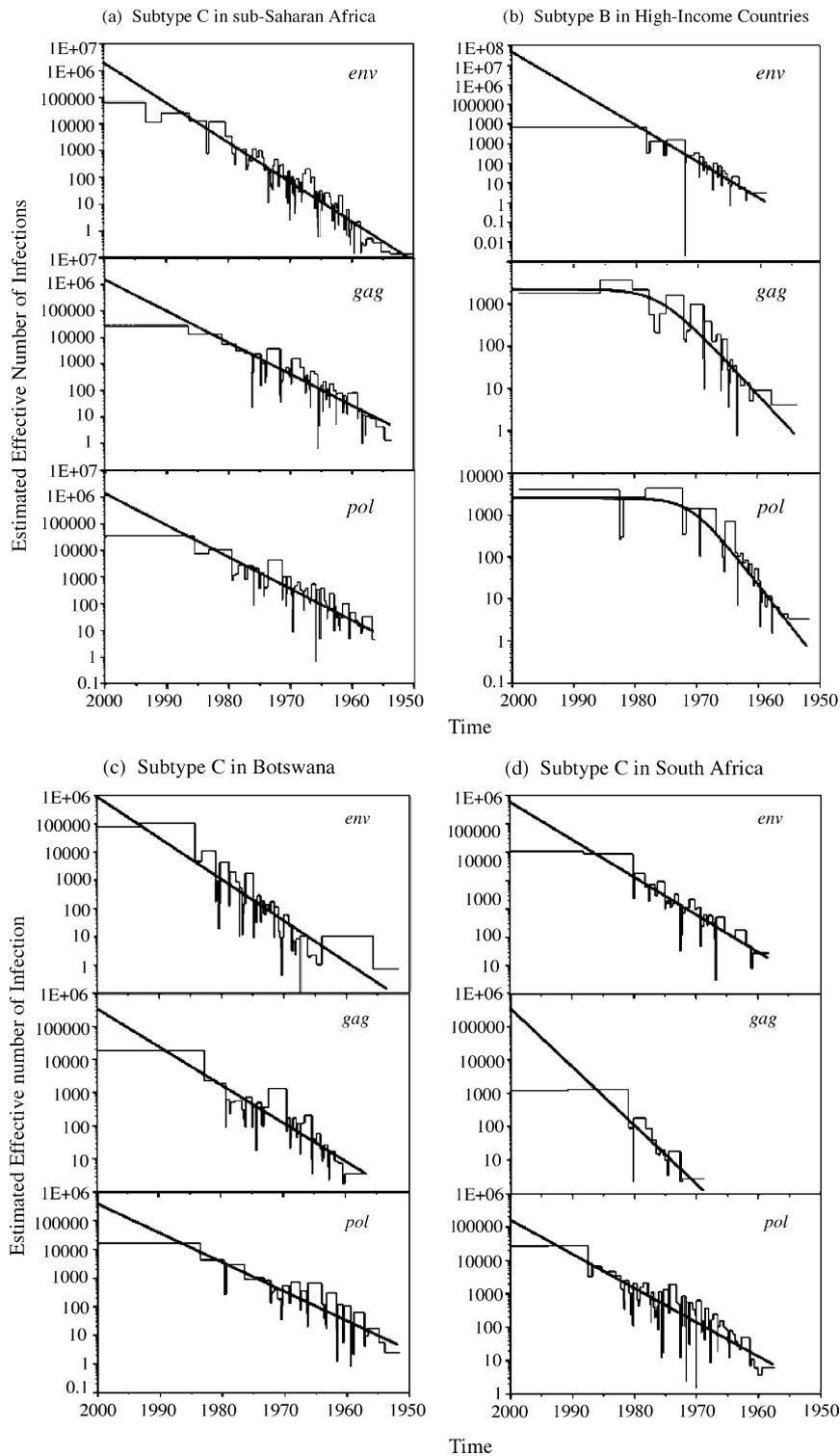


Fig. 3. Skyline plots showing demographic histories of HIV-1 subtypes C (a) and B (b), and subtype C in Botswana (c) and South Africa (d). Maximum likelihood demographic models are superimposed (thick lines). The vertical axes represent the estimated effective number of infections on a logarithmic scale.

sequences were included in the analysis, with the most recently sampled sequence from 1997. In this case, there is no significant evidence for a logistic trend. No significant differences were found in estimates of r or λ between the *env*, *gag* and RT genes, and in all the genes analysed the

subtype B epidemic shows a significantly faster growth rate than subtype C. The average growth rate for subtype B is 0.495 year^{-1} (0.386 and 0.609), which is equivalent to doubling time of 1.403 years (1.801 and 1.141), almost twice that of subtype C.

In the case of subtype C we also analysed population dynamics within particular countries. The demographic history of subtype C in Botswana shows a constant exponential growth since its origin in 1955–1960, with no evidence for a logistic trend (Fig. 3c). Estimates of r and λ do not differ significantly between *gag*, *env* and RT genes, and give an average r of 0.243 year^{-1} (0.202 and 0.285). Similarly, all data sets from South Africa show exponential population growth and in both *env* and RT estimates of r and λ have overlapping confidence intervals, *env* at 0.297 (0.248 and 0.347) and RT at 0.235 year^{-1} (0.206 and 0.266), giving an average doubling time of 2.991 years (2.483 and 3.673). Hence, there are no significant differences between these parameter estimates and those of Botswana. However, the *gag* gene gives a growth rate (r) of 0.405 year^{-1} (0.288 and 0.530), and a λ of 1.713 years (1.309 and 2.405), which is significantly higher than that estimated from ZA-RT, any of the Botswana data sets, and the average for all the sub-Saharan Africa sequences.

4. Discussion

4.1. Differences in populations dynamics between HIV-1 subtypes B and C

Our comparison of the population dynamics of HIV-1 subtypes B and C has revealed that although their epidemics began at a similar time, the subtype B epidemic in Western Europe and the America in the 1980s spread twice as fast as the current subtype C epidemic in sub-Saharan Africa. More precisely, subtype B appears to have been introduced into these geographic regions during 1950s to early 1960s and experienced a sudden rapid expansion over the first 15 years of the epidemic, followed by a more recent decline in growth rate. In comparison, subtype C has grown at a constant rate since the beginning of the epidemic around 1960 to the present day, and shows no evidence of a recent slow-down. Importantly, our coalescent analysis is able to illustrate the expansion of the epidemics in both regions *prior* to the identification of HIV-1 and the introduction of widespread serological testing.

The expansion of the subtype B epidemic in high-income countries has decelerated considerably in recent years, as shown by prevalence statistics (MAP, 2000). We have been able to illustrate this logistic trend by including new contemporary sequences for the *gag* and RT genes that were unavailable to earlier population dynamic analyses (Holmes et al., 1995; Grassly et al., 1999; Holmes et al., 1999; Pybus et al., 1999). However, in the coalescent analysis of *env* gene, which only includes data up to 1995, the logistic trend is not strongly significant. This may be a product of sampling, as a recent study found such a logistic trend in the epidemic in the United States using samples taken prior to 1995 (Robbins et al., 2003).

The skyline plots and r estimates for the full sub-Saharan Africa subtype C set show consistent trends within the given

confidence intervals of the substitution rate for the *gag*, RT and *env* genes. That there are overlapping confidence intervals for each data set also indicates that there is no significant difference between the estimates of doubling time, which is expected if substitution rates are accurate and recombination rates are comparable among different genes. In sum, our coalescent analysis reveals a constant exponential growth rate over time in the subtype C epidemic in southern Africa, with no strong evidence for a recent deceleration in growth, although it will be important to return to this issue with more recently sampled data.

We suggest that the differences between the growth rates of HIV-1 subtypes B and C during the exponential growth phase are more likely to reflect large differences in transmission networks, and particularly the distribution of times between transmissions than differences in intrinsic infectivity. In particular, the rapid population growth of subtype B supports previous studies which suggested a very rapid expansion within a standing network, associated with closely linked high-risk groups such as injecting drug users or homosexual networks with high rates of partner exchange (Holmes et al., 1999). This is significant in light of recent controversial studies that have proposed that no more than ~35% of HIV-1 infections in Africa can be attributed to sexual transmission (Gisselquist and Potterat, 2003). If parenteral transmission had played a substantial part in the subtype C epidemic in sub-Saharan Africa, we might expect to see a more sudden rise in population growth, similar to that observed in subtype B. Another key prediction of the parenteral transmission hypothesis is that the epidemic history of HIV-1 in sub-Saharan Africa should be similar to that of other blood-borne pathogens such as hepatitis C virus (HCV) (Walker et al., 2003). However, the growth rate of HIV-1 subtype C estimated here is 5–10 times *greater* than that estimated previously for HCV types 4 and 6, endemically transmitted in Africa and Asia, and twice that of HCV subtypes 1a and 1b (Pybus et al., 2001). This difference in growth rates is difficult to reconcile with the hypothesis that HIV transmission is mainly parenteral, especially as HCV is more infectious than HIV via this route (Jagger et al., 2002). Indeed, the high variation in partner exchange rates, low reported condom use (Craiel et al., 1995) and high prevalence of other sexually transmitted infections, particularly HSV-2 (Avert et al., 2001), make sub-Saharan Africa particularly susceptible to a heterosexually transmitted HIV epidemic.

4.2. HIV-1 subtype C growth rates in Botswana and South Africa

The three data sets for subtype C corroborate one another well; estimates of doubling time are similar between genes and countries and are also in keeping with an epidemic doubling time of 1.9 years calculated from HIV-1 age prevalence data in South Africa (Gouws et al., 2002; Williams et al., 2001). However, country-specific estimates for

Botswana and South Africa are less reliable (although they are generally congruent) as they represent much smaller sample sizes. In the case of South Africa, growth rate estimates from RT and *env* are not significantly different, although the *gag* data set shows a significantly higher growth rate than that of RT. As the sequences used were sampled from different individuals in each gene, they may reflect HIV dynamics in different populations. Specifically, in the *gag* data set, 14 of 18 sequences were derived from one cohort, sampled in Cape Town between 1998 and 1999 (Engelbrecht et al., 2001). In contrast, the RT sequences were derived from more geographically disparate sources, and therefore reflect a more countrywide estimate of epidemic growth. This suggests that although there are no differences between epidemic growth patterns in South Africa and Botswana as a whole, there may be “hotspots” of higher transmission in urban areas. For other genes there are no significant differences in r , λ or in the best-fit model of growth rate between Botswana and South Africa, and the sequences cluster together on phylogenetic trees despite having different geographical origins. This supports the idea that the HIV-1 epidemics in Botswana and South Africa have a common history and so might be expected to show a similar demographic trend.

The coalescent method used here assumes random sampling, the existence of a molecular clock and no recombination. Although violation of these assumptions may introduce biases, the fact that there is strong similarity in the results obtained from different genes, and to other sources of data, shows that our key findings are robust. In particular, the fact that our estimates of epidemic doubling time are conservative compared to those deduced from HIV incidence data (Williams et al., 2001) suggests that any undetected recombination has not introduced a significant bias into the analyses.

In conclusion, we reveal that the HIV-1 subtype C epidemic in sub-Saharan Africa has grown at a slower rate than the initial epidemic spread of subtype B in high-income countries. Furthermore, within Botswana and South Africa, countrywide estimates of growth rate are equivalent, although there is some evidence for a more rapid growth in urban areas in South Africa, which may strongly contribute to the rise of the South African AIDS epidemic. Most disturbingly, the growth rate of subtype C in southern Africa shows no evidence of declining.

Acknowledgement

This work was supported by the Wellcome Trust Biodiversity programme.

References

Auvert, B., Ballard, R., Campbell, C., Carael, M., Carton, M., Fehler, G., Gouws, E., MacPhail, C., Taljaard, D., Van Dam, J., Williams, B.G.,

2001. HIV infection among youth in a South African mining town is associated with herpes simplex virus-2 seropositivity and sexual behaviour. *AIDS* 15, 885–898.
- Bredell, H., Hunt, G., Morgan, B., Tiemessen, C.T., Martin, D.J., Morris, L., 2000. Identification of HIV type 1 intersubtype recombinants in South Africa using *env* and *gag* heteroduplex mobility assays. *AIDS Res. Hum. Retro.* 16, 493–497.
- Buve, A., Bishikwabo-Nsarhaza, K., Mutangadura, G., 2002. The spread and effect of HIV-1 infection in sub-Saharan Africa. *Lancet* 359, 2011–2017.
- Buve, A., Carael, M., Hayes, R.J., Auvert, B., Ferry, B., Robinson, N.J., Anagonou, S., Kanhonou, L., Laourou, M., Abega, S., Akam, E., Zekeng, L., Chege, J., Kahindo, M., Rutenberg, N., Kaona, F., Musonda, R., Sukwa, T., Morison, L., Weiss, H.A., Laga, M., 2001. The multi-centre study on factors determining the differential spread of HIV in four African cities: summary and conclusions. *AIDS* 15, S127–S131.
- Carael, M., Cleland, J., Deheneffe, J.C., Ferry, B., Ingham, R., 1995. Sexual behaviour in developing countries: implications for HIV control. *AIDS* 9, 1171–1175.
- Engelbrecht, S., de Villiers, T., Sampson, C.C., zur Megede, J., Barnett, S.W., Janse van Rensburg, E., 2001. Genetic analysis of the complete *gag* and *env* genes of HIV type 1 subtype C primary isolates from South Africa. *AIDS Res. Hum. Retro.* 17, 1533–1547.
- Foley, B., Pan, H., Buchbinder, S., Delwart, E.L., 2000. Apparent founder effect during the early years of the San Francisco HIV type 1 epidemic (1978–1979). *AIDS Res. Hum. Retro.* 16, 1463–1469.
- Gisselquist, D., Potterat, J.J., 2003. Heterosexual transmission of HIV in Africa: an empiric estimate. *Int. J. STD AIDS* 14, 162–173.
- Gordon, M., De Oliveira, T., Bishop, K., Coovadia, H.M., Madurai, L., Engelbrecht, S., Janse van Rensburg, E., Mosam, A., Smith, A., Cassol, S., 2003. Molecular characteristics of human immunodeficiency virus type 1 subtype C viruses from KwaZulu-Natal South Africa: implications for vaccine and antiretroviral control strategies. *J. Virol.* 77, 2587–2599.
- Gottlieb, M.S., Schroff, R., Schanker, H.M., Weisman, J.D., Fan, P.T., Wolf, R.A., Saxon, A., 1981. *Pneumocystis carinii* pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *N. Engl. J. Med.* 305, 1425–1431.
- Gouws, E., Williams, B.G., Sheppard, H.W., Enge, B., Karim, S.A., 2002. High incidence of HIV-1 in South Africa using a standardized algorithm for recent HIV seroconversion. *J. Acquir. Immune. Defic. Syndr.* 29, 531–535.
- Grassly, N.C., Harvey, P.H., Holmes, E.C., 1999. Population dynamics of HIV-1 inferred from gene sequences. *Genetics* 151, 427–438.
- Griffiths, R.C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond. B* 344, 403–410.
- Holmes, E.C., Nee, S., Rambaut, A., Garnett, G.P., Harvey, P.H., 1995. Revealing the history of infectious disease epidemics using phylogenetic trees. *Phil. Trans. R. Soc. Lond. B* 349, 33–40.
- Holmes, E.C., Pybus, O.G., Harvey, P.H., 1999. The molecular population dynamics of HIV-1. In: Crandall, K.A. (Ed.), *The Evolution of HIV*. John Hopkins University Press, pp. 177–207.
- Jagger, J., Puro, V., De Carli, G., 2002. Occupational transmission of hepatitis C virus. *JAMA* 288, 1469.
- Kingman, J.F.C., 1982. On the genealogy of large populations. *J. Appl. Probab.* 19A, 27–43.
- Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., Bhattacharya, T., 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* 288, 1789–1796.
- Kuiken, C.L., Foley, B., Freed, E., Hahn, B., Korber, B., Marx, P.A., McCutchan, F., Mellors, J.W., Wolinsky, S., (Eds.), 2002. *HIV Sequence Compendium 2002*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 03-3564.
- Laverdiere, M., Tremblay, J., Lavallee, R., Bonny, Y., Lacombe, M., Boileau, J., Lachapelle, J., Lamoureux, C., 1983. AIDS in Haitian immigrants and in a Caucasian woman closely associated with Haitians. *Can. Med. Assoc. J.* 129, 1209–1212.

- MAP, 2000. Durban MAP symposium report: the status and trends of the HIV/AIDS epidemics in the world, pp. 19–20.
- McCormack, G.P., Glynn, J.R., Crampin, A.C., Sibande, F., Mulawa, D., Bliss, L., Broadbent, P., Abarca, K., Ponnighaus, J.M., Fine, P.E., Clewley, J.P., 2002. Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi. *J. Virol.* 76, 12890–12899.
- Osmanov, S., Pattou, C., Walker, N., Schwarlander, B., Esparza, J., 2002. WHO-UNAIDS network for HIV isolation and characterization: Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year. *J. Acquir. Immune. Defic. Syndr.* 29, 184–190.
- Novitsky, V.A., Montano, M.A., McLane, M.F., Renjifo, B., Vannberg, F., Foley, B.T., Ndung'u, T.P., Rahman, M., Makhema, M.J., Marlink, R., Essex, M., 1999. Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. *J. Virol.* 73, 4427–4432.
- Pillay, C., Bredell, H., McIntyre, J., Gray, G., Morris, L., 2002. HIV-1 subtype C reverse transcriptase sequences from drug-naïve pregnant women in South Africa. *AIDS Res. Hum. Retro.* 18, 605–610.
- Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C., Harvey, P.H., 2001. The epidemic behaviour of the hepatitis C virus. *Science* 292, 2323–2325.
- Pybus, O.G., Holmes, E.C., Harvey, P.H., 1999. The mid-depth method and HIV-1: a practical approach to testing hypotheses of viral epidemic history. *Mol. Biol. E. vol.* 16, 953–959.
- Pybus, O.G., Rambaut, A., 2002. GENIE: estimating demographic history from molecular phylogenies. *Bioinformatics* 18, 1404–1405.
- Pybus, O.G., Rambaut, A., Harvey, P.H., 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155, 1429–1437.
- Rambaut, A., 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 16, 395–399.
- Robbins, K.E., Lemey, P., Pybus, O.G., Jaffe, H.W., Youngpairaj, A.S., Brown, T.M., Salemi, M., Vandamme, A., Kalish, M.L., 2003. U.S. Human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J. Virol.* 77, 6359–6366.
- Selik, R.M., Haverkos, H.W., Curran, J.W., 1984. Acquired immune deficiency syndrome (AIDS) trends in the United States (1978–1982). *Am. J. Med.* 76, 493–500.
- South African Department of Health, 2001. National HIV and syphilis seroprevalence of women attending public antenatal clinics in South Africa—2001. Directorate Health Systems Research.
- Swofford, D.L., 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, MA.
- UNAIDS, 2003. AIDS Epidemic Update 2003, (<http://www.unaids.org/Unaid/EN/Resources/Publications/Corporate+publications/AIDS+epidemic+update++December+2003.asp>).
- UNAIDS, 2002a. Epi fact sheets by country—2002 updates: Botswana, (<http://www.unaids.org/hivaidsinfo/statistics/>).
- UNAIDS, 2002b. Epi fact sheets by country—2002 updates: South Africa, (<http://www.unaids.org/hivaidsinfo/statistics/>).
- Van Harmelen, J.H., Van der Ryst, E., Loubser, A.S., York, D., Madurai, S., Lyons, S., Wood, R., Williamson, C., 1999. A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations. *AIDS Res. Hum. Retro.* 15, 395–398.
- Walker, P.R., Worobey, M., Rambaut, A., Holmes, E.C., Pybus, O.G., 2003. Sexual transmission of HIV in Africa. *Nature* 422, 679.
- Wilks, S.S., 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Rev. Math. Stat.* 9, 60–62.
- Williams, B.G., Gouws, E., Wilkinson, D., Karim, S.A., 2003–2016. Estimating HIV incidence rates from age prevalence data in epidemic situations. *Stat. Med.* 20 .
- Williams, B.G., Gouws, E., 2001. 2001 The epidemiology of human immunodeficiency virus in South Africa. *Phil. Trans. R. Soc. Lond. B* 356, 1007–1086.
- Yusim, K., Peeters, M., Pybus, O.G., Bhattacharya, T., Delaporte, E., Mulanga, C., Muldoon, M., Theiler, J., Korber, B., 2001. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. *Phil. Trans. R. Soc. Lond. B* 356, 855–866.