
# Origin, imports and exports of HIV-1 subtype C in South Africa: A historical perspective

Eduan Wilkinson, David Rasmussen, Oliver Ratmann, Tanja Stadler, Susan Engelbrecht, Tulio de Oliveira

PII: S1567-1348(16)30298-2
DOI: doi: 10.1016/j.meegid.2016.07.008
Reference: MEEGID 2837

To appear in:

Received date: 19 April 2016
Revised date: 7 July 2016
Accepted date: 8 July 2016

Please cite this article as: Wilkinson, Eduan, Rasmussen, David, Ratmann, Oliver, Stadler, Tanja, Engelbrecht, Susan, de Oliveira, Tulio, Origin, imports and exports of HIV-1 subtype C in South Africa: A historical perspective, (2016), doi: 10.1016/j.meegid.2016.07.008

**Origin, imports and exports of HIV-1 subtype C in South Africa: A historical perspective**

Eduan Wilkinson[a], David Rasmussen[b], Oliver Ratmann[c], Tanja Stadler[b], Susan Engelbrecht[d,e], Tulio de Oliveira[a,f,g]

[a]Africa Centre for Population Health, University of KwaZulu-Natal, Durban, 4041, Republic of South Africa

[b]ETH Zurich, Department of Biosystems Science and Engineering, 4058 Basel, Switzerland

[c]Imperial College London, School of Public Health, Department of Infectious Disease Epidemiology, London W2 1PG, United Kingdom

[d]Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, Western Cape Province, 7505, Republic of South Africa

[e]National Health Laboratory Services (NHLS), Tygerberg Coastal, Cape Town, 8000, Republic of South Africa

[f]College of Health Sciences, University of KwaZulu-Natal, Durban, 4041, Republic of South Africa

[g]Research Department of Infection, University College London, London WC1E 6BT, United Kingdom

e-mails: ewilkinson83@gmail.com (E.W.); david.rasmussen@bsse.ethz.ch (D.R.); oliver.ratmann@imperial.ac.uk (O.R.); tanja.stadler@bsse.ethz.ch (T.S.); susanen@sun.ac.za (S.E.); tuliodna@gmail.com (T.d.O.)

Author to whom correspondence should be addressed; e-mail: tuliodna@gmail.com

Wellcome Trust, Africa Centre for Population Health Laboratory, Room 135, 1st floor, Doris Duke Medical Research Institute (DDMRI) building, 719 Umbilo Road, Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Congella, Durban, KwaZulu-Natal, 4013, Republic of South Africa

**Abstract**

**Background:** While the HIV epidemic in South Africa had a later onset than epidemics in other southern African countries, prevalence grew rapidly during the 1990's when the country was going through socio-political changes with the end of Apartheid. South Africa currently has the largest number of people living with HIV in the world and the epidemic is dominated by a unique subtype, HIV-1 subtype C. This large epidemic is also characterized by high level of genetic diversity. We hypothesize that this diversity is due to multiple introductions of the virus during the period of change. In this paper, we apply novel phylogeographic methods to estimate the number of viral imports and exports from the start of the epidemic to the present.

**Methods:** We assembled 11,289 unique subtype C *pol* sequences from southern Africa. These represent one of the largest sequence datasets ever analyzed in the region. Sequences were stratified based on country of sampling and levels of genetic diversity were estimated for each country. Sequences were aligned and a maximum-likelihood evolutionary tree was inferred. Least-Squares Dating was then used to obtain a dated phylogeny from which we estimated the number of introductions into and exports out of South Africa using parsimony-based ancestral location reconstructions.

**Results:** Our results identified 189 viral introductions into South Africa with the largest number of introductions attributed to Zambia (n=109), Botswana (n=32), Malawi (n=26) and Zimbabwe (n=13). South Africa also exported many viral lineages to its neighbours. The bulk viral imports and exports appear to have occurred between 1985 and 2000, coincident with the period of socio-political transition.

**Conclusion:** The high level of subtype C genetic diversity in South Africa is related to multiple introductions of the virus to the country. While the number of viral imports and exports we identified was highly sensitive to the number of samples included from each country, they mostly clustered around the period of rapid political and socio-economic change in South Africa.

**Abbreviations:** HIV – Human Immunodeficiency Virus

**Keywords:** HIV-1 subtype C, South Africa, viral imports, viral exports, phylogeographic, genetic diversity

## 1.0 Introduction

In the 1900s there was an extensive system of migrant labour throughout southern Africa. This system recruited labourers from rural areas across the southern African region on a contract basis to work in mines and factories in larger urban centres. The labourers, the vast majority of whom were male, were housed in dormitories for several months at a time, completely separated from their family life in the rural countryside. As was previously suggested [Hargrove, 2008], the migrant labour system with the unnatural and forced separation of young men from their families and communities may have been central to the early spread of HIV-1 within the southern African context. Furthermore, political conflicts in the region (i.e. the Rhodesian Bush War, also known as the Zimbabwean War of Liberation, the South African-Angolan border war and the civil wars in Mozambique and Angola) may have been responsible for the initial seeding of HIV-1 strains in the region. However, the epidemic only started to spread to the general population once political tensions started to subside. This is evident in the rise of the number of HIV-1 cases following Zimbabwean independence after 1980, the end of the civil war in Mozambique in 1992, as well as the ending of Apartheid in South Africa in 1994 [UNAIDS, 2010].

In South Africa, the HIV epidemic began to grow rapidly in the early 1990s and by 2001 24.8% of the women at antenatal clinics were infected with HIV [Department of Health, 2002]. Currently, South Africa has the largest number of people living with HIV in the world and the epidemic is almost entirely dominated by subtype C (~98%) [Hemelaar et al. 2011]. The epidemic in South Africa is further characterized by high level of genetic diversity [Wilkinson et al. 2015]. This high HIV-1 prevalence and genetic diversity are surprising given the late onset of the epidemic in the country compared to other African countries [Abdool Karim, 2010]. We hypothesize that the extreme diversity and fast rate of spread were the result of multiple subtype C introductions into South Africa from

neighbouring countries, and that these introductions occurred during South Africa's political transition towards democracy.

To test our first hypothesis that there were many introductions of HIV-1 to South Africa, we used viral phylogenetic analyses to identify past viral introductions and to evaluate whether phylogenetically estimated dates of viral introductions were consistent with periods of socio-political change. Identifying viral introductions in southern Africa is challenging since comprehensive datasets of sequences coupled with good epidemiologically relevant data are typically not available in the public domain. We assembled the largest southern African HIV-1 subtype C sequence dataset to date in order to estimate the number of viral introductions into South Africa through time from other southern African countries, as well as the number of exports from South Africa. Our phylogeographic results strongly supported our second hypothesis that multiple introductions of HIV into South Africa occurred during the transition period between 1985 and 2000, and in fact reveal a much larger number of viral imports and exports than expected.

## 2.0 Methods

### 2.1 Ethics

The sequences from the Division of Medical Virology, Tygerberg National Health Laboratory Services (NHLS), were obtained with a waiver of informed consent from the Health Research Ethics Committee (HREC) of Stellenbosch University (IRB0005239), reference number N11/02/054. This HREC complies with the South Africa National Health Act No 612003 and the United States code of Federal Regulations title 45 Part 46. The investigation also complies with the South African National Health Act No 612003 and abides by the ethical norms and principles for research as established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the South African National Department of Health (NDoH) Guidelines.

### 2.2 Sampling

We obtained all partial polymerase (*pol*) HIV-1 sequences from South Africa and other southern African countries. We included all countries of the Southern African Development Community (SADC) with the exclusion of the three Indian Ocean island member nations (the Seychelles, Mauritius & Madagascar). The SADC region includes the following countries: Democratic Republic of the Congo (DRC), Angola, Tanzania, Zambia, Malawi, Mozambique, Zimbabwe, Botswana, Namibia, Swaziland, Lesotho and South Africa. Even though the DRC is not classified geographically as a southern African nation but as a central African nation. However, we included the DRC in our analyses, as it is the most probable origin of the most recent common ancestor (tMRCA) of the global HIV-1 subtype C pandemic [Faria et al. 2014].

We retrieved all subtype C isolates from SADC nations available in the HIV-1 Los Alamos National Laboratory (LANL) database (www.hiv.lanl.gov). In total, we could identify 6,049 unique HIV-1 sequences downloaded from LANL that included the *pol* region, positions 2100 to 3300 relative to HXB2 (date of access – 4 February, 2016). These sequences were aligned to 5,240 unique sequences from the Tygerberg HIV-1 drug resistance, The Tygerberg cohort contains previously unpublished sequence data from seven provinces in South Africa, sampled between 2006 and 2015. Van Zyl and colleagues previously described the methodology used to genotype these drug resistance isolates [Van Zyl et al. 2008].

In total, 11,289 unique subtype C *pol* sequences, along with a homologous section of the HXB2 reference strain [Wong-Staal et al. 1985], were aligned in ClustalW v 2.0 [Larkin et al. 2007] using a QuickTree method to speed up the alignment process. The alignment was manually edited in Geneious v 8.0.5 until a codon alignment was obtained. Major HIV drug resistance sites were removed from the alignment prior to any analysis (source code available @ https://github.com/olli0601/).

### 2.3 Genetic Diversity

The main subtype C dataset (n=11,289) was stratified into geographical datasets based on each sample's country of origin. The overall mean genetic distance and standard error

(SE) for each geographical dataset was estimated in MEGA v 6.0 [Tamura et al. 2013] with the use of the Kimura 2 parameter subsitutional model [Kimura M, 1980], a gamma shape parameter of 0.5 [Yang, 1994] and 1,000 bootstrap replicates [Felsenstein, 1985; Efron et al. 1996]. The mean genetic diversity and standard error for each geographical dataset was recorded and plotted in R with the ggplot2 package.

In order to assess the robustness of the genetic diversity estimates against varying sample sizes, we generated smaller geographical datasets by randomly selecting 20, 50, 100 and 200 taxa from the original geographical datasets. Genetic diversity estimates of these smaller subsampled datasets were performed using the method described above to determine if the initial estimates were robust to our sampling strategy.

**2.4 Phylogenetic and Phylogeographic inference**

A maximum likelihood (ML) phylogenetic tree was constructed using FastTree 2 [Price et al. 2010] for the complete alignment containing 11,290 southern African sequences and the HXB2 reference, a subtype B isolate. For phylogenetic reconstruction, sequences were assumed to evolve under a General Time Reversible (GTR) model [Tavaré, 1986] with gamma-distributed rate heterogeneity across sites. The ML tree was then rooted using HXB2 as an outgroup. In the resulting tree, one South African sequence (PS045.FJ199626) was more divergent to all other subtype C sequences than HXB2 and was subsequently removed. An additional set of bootstrap trees was obtained by running FastTree 2 on 100 bootstrapped alignments created by resampling sites with replacement from the original alignment.

To obtain dated phylogenies, we converted branch lengths into units of real calendar time using Least-Squares Dating [To et al. 2016]. We were unable to reliably infer a molecular clock rate without making strong assumptions about the tree height and root time. We therefore used a fixed molecular clock rate of $2.0 \times 10^{-3}$, which lies in the center of the rates previously reported for subtype C [Dalai et al. 2009, Abecasis et al. 2009, de Oliveira et al. 2010, Bello et al. 2012, Jung et al. 2012, Wilkinson et al. 2015].

For our phylogeographic analysis, we first reconstructed the ancestral location of each internal node in the rooted ML tree using Fitch's parsimony algorithm [Fitch, 1979]. In particular, we use Sankoff's dynamic programming version of Fitch's original algorithm, which finds the most parsimonious reconstruction in linear time relative to the number of taxa [Sankoff, 1975]. Equally parsimonious reconstructions were randomly resolved in each replicate analysis (see sensitivity analysis below).

The reconstructed ancestral location of each node was then used to identify introductions into South Africa, as well as from South Africa to its neighbors. Introductions into a country were assumed to occur whenever a node *n* reconstructed to be in that country had a parent node reconstructed to be outside of that country. We then took the date/height of node *n* to be our best estimate of the introduction time. Note that an introduction event does not necessarily define a monophyletic clade, as its possible that lineages descending from an introduction event may have subsequently moved elsewhere. Using this approach, we identified all unique introduction events into and out of South Africa present in the ML phylogeny. We then repeated this analysis on all 100 bootstrap phylogenies to construct bootstrapped confidence intervals on the number and timing of introduction events.

**2.5 Sensitivity analyses**

Our sequence data contained many more samples from South Africa than other southern African nations, and our sampling coverage varied widely among them. It is therefore likely that we underestimated the number of introductions into South Africa from less well-sampled nations and overestimated the number of exported cases from South Africa relative to imported cases. We therefore first performed a sensitivity analysis where we systematically varied the number of samples taken from each potential source nation by subsampling. To do this, we randomly removed between 10 and 90% of samples from each potential source by pruning these samples from the full ML tree. We then reran the maximum parsimony reconstruction on this pruned tree and re-estimated the number of introductions from each potential source. This subsample-then-prune procedure was repeated 10 times at each subsampling frequency for each potential source.

Finally, we ran a sensitivity analysis with an equal number of samples from each nation to see if any potential source contributed disproportionally to the South African epidemic after controlling for sampling. Of all nations that we identified as potential major sources, Zimbabwe was the least well sampled (n=268). We therefore randomly down-sampled all other nations to a sample size of 268, except for Tanzania (n=168), which was not identified as a major source. We then randomly sampled an equal number of sequences from South Africa as from all other countries combined (n=1508). This random down sampling was repeated 100 times. For each replicate, we pruned all other sequences from the phylogeny, reran the maximum parsimony analysis and then recounted the number of introductions from each potential source.

Matlab source code for performing the maximum parsimony reconstruction, identifying introductions and running the sensitivity analysis is available @ https://github.com/davidrasm.

## 3.0 Results

### 3.1 Sample dataset

Our final dataset included 11,290 HIV-1 subtype C sequences, 5,240 of which were unpublished sequences from the Tygerberg HIV-1 drug resistance cohort and 6,049 of which were public HIV-1 sequences downloaded from LANL (www.hiv.lanl.gov). Previously unpublished sequences have been submitted to GenBank with the following accession numbers: (**Awaiting accession numbers from GenBank**). All of the sequences were confirmed as HIV-1 subtype C by two automated subtyping methods, Rega V3 and COMET. A temporal and geographic breakdown of the dataset is provided in **Figure 1**.

### 3.2 Genetic diversity

We divided the sequences into countries of origin and used the datasets to estimate the genetic diversity of each country. The South African dataset was the most diverse, with a mean genetic diversity of 7,45% (SE 0.3%). The DRC and Botswana also had high levels

of HIV-1 genetic diversity (**Figure 1**). Zambia and Zimbabwe had the lowest mean genetic diversity of all of the analyzed geographical datasets. Swaziland, Malawi, Mozambique, Angola and Tanzania had intermediate levels of genetic diversity (**Figure 1**). In order to determine if genetic diversity was influenced by sample sizes, we sub-sampled each countries' dataset. The results of the sub-sampled datasets were similar (data not shown).

### 3.3 Phylogenetic analysis

We constructed an ML phylogeny for all 11,290 samples contained in the final alignment and then estimated branch lengths in terms of calendar time using Least-Squares Dating. All of the southern African subtype C sequences shared an MRCA between 1960 and 1970, with the majority of the different sub-clades arising between 1980 and 1995 (**Figure 2**). The DRC was identified as the origin of the southern African subtype C, with a DRC sequence clustering basal to the entire southern African subtype C clade. A few early diverging lineages from Zambia, Tanzania and Mozambique fell between the DRC sequences and all other southern African samples. The time-calibrated tree contained many divergent lineages that coalesced in the early 1980's, giving the tree a deep comb-like backbone (**Figure 2**).

Consistent with the high levels of genetic diversity observed, lineages sampled in South Africa were distributed widely throughout the entire subtype C phylogeny. The maximum parsimony reconstruction of ancestral locations (**Figure 2**) suggested that HIV-1 subtype C likely entered South Africa in the mid to late 1980s, and that these early introductions allowed several large sub-epidemics that persisted locally in South Africa. Some of these early introductions appear to have subsequently seeded smaller sub-epidemics in neighbouring countries such as Botswana, Malawi, Mozambique and Zimbabwe.

In order to determine the number of introductions of HIV-1 subtype C into South Africa, we performed a maximum parsimony reconstruction analysis using the ML tree. Overall, our introduction analysis revealed a total of 189 introductions into South Africa. The

temporal distribution of introductions by inferred source country is shown in **Figure 3 (a-f)**. Most of the introductions were from Zambia (n=109), Botswana (n=32), Malawi (n=26) and Zimbabwe (n=13). It is important to note that even though the DRC was identified as the MRCA for the whole southern Africa HIV-1 subtype C epidemic, no direct viral introductions from the DRC into South Africa were detected.

The majority of the introductions into South Africa occurred between 1985 and 2000, a time of political change in the country. Rerunning the introduction analysis on 100 bootstrap phylogenies revealed a temporal distribution of introductions largely consistent with the pattern inferred from the single ML phylogeny **(Figure 3a-f)**. During the same period, South Africa was also the source of many exports of HIV-1 subtype C to other countries in the region. We identified a total of 393 exports from South Africa to other countries. The temporal distribution of these exports with bootstrapped confidence intervals is shown in **Figure 3 (g-l)**. Overall, these results support a scenario of high exchange of subtype C strains in the region, which is consistent with circular migration patterns.

### 3.4 Sensitivity analyses

The number of introductions identified from each source country was found to correlate highly with the number of sequences sampled from each country, with the exception of Tanzania. A sensitivity analysis, in which we systematically varied the number of samples included from each country by down-sampling, showed that the number of introductions grew linearly with the number of samples included from each location, with little apparent saturation for even the most well-sampled countries (**Figure 4 a-f**). The number of introductions we identified is, therefore, likely to be an extreme underestimate of the true number. Likewise, the number of viral introductions exported by South Africa to other countries also grows linearly with the number of samples included from South Africa (**Figure 4 g-l**).

Furthermore, while most introductions into South Africa were attributed to Zambia before accounting for sampling biases, no country contributed significantly more

introductions than any other once we included an even number of samples from each country (**Figure 5**). With even sampling, Malawi was identified as the largest contributor, followed closely by Zimbabwe and Zambia. As previously mentioned and discussed below, the large number of introductions identified from these countries is consistent with historical patterns of (circular) migration and political/economic exchange with South Africa.

## 4.0 Discussion

The genetic diversity analysis showed that South Africa is home to one of the most genetically diverse subtype C epidemics of all countries in our dataset. Estimates from other southern African countries indicate that the subtype C epidemics in the DRC and Botswana have similar genetic diversity levels compared to the South African dataset (**Figure 1**). As the most likely origin of the global subtype C pandemic, it is reasonable that we find a high degree of genetic diversity in the DRC. In addition, given the small number of subtype C sequences available (n=25), it is possible that the subtype C epidemic in the DRC is more diverse than we report here.

Our analyses further suggested that most other southern African nations have slightly less genetically diverse HIV epidemics than South Africa and Botswana, with the exception of Zimbabwe, which was significantly less diverse. South Africa is home to the largest number of people living with HIV in the world. Furthermore, our results indicated a large number of viral introductions into South Africa from various southern African countries, some of which also experience high levels of genetic diversity. The genetic diversity of the epidemic in Botswana can be explained on the basis of the very high levels of HIV prevalence, bidirectional migration due to its economic prosperity and its central geographical position within the southern African context. The low level of genetic diversity in the Zimbabwean epidemic was unexpected. Our current understanding of the HIV epidemic in Zimbabwe, based on the findings from Dalai and colleagues [Dalai et al. 2009], is that the epidemic in the country was caused by multiple introductions of HIV-1 from neighbouring countries following the end of the Rhodesian bush war and the move to full independence in 1980 (**Supplementary Figure 1**). Following independence, the

Zimbabwean government halted all bilateral labour policies with South Africa in protest against Apartheid. Therefore, Zimbabwean nationals were excluded from the migrant labour system, which has been central to the dissemination and spread of HIV-1 subtype C variants across the southern African region [Hargrove, 2008]. It is therefore possible that the genetic diversity within Zimbabwe has been limited to the initial introductions following political change in the early 1980s, with few new introductions occurring since then.

Our results indicate that the bulk of viral introductions into South Africa and exports from South Africa occurred during a period of socio-political change in the country (1980-2000). The release of political prisoners, the unbanning of political parties, the abolition of Apartheid laws and the adoption of an interim constitution marked this period leading up to the end of Apartheid in South Africa. It appears that multiple introductions during this time period seeded the epidemic, giving rise to smaller sub-epidemics co-circulating within the country. This is evident in the time resolved phylogeny (**Figure 2),** where large South Africa specific clusters can be seen. The estimated time to MRCAs for these South African clusters corresponds to this period of social-political changes.

The association between political changes, the returning of individuals displaced by conflicts and the spread of HIV has been previously described [Becker et al. 2008, UNAIDS 2015]. Countries undergoing socio-political changes may be at increased risk of the spread of HIV. This may be caused by several factors: (1) increase movement between countries as political exiles return home to their families, (2) increases in trade and labour migrations between countries during times of peace, and (3) increases in fertility rates amongst the population following periods of conflict (baby-boom). The results presented here and those of other studies in the southern African region [Dalai et al. 2009; Wilkinson et al. 2015] support this trend of HIV spread following conflict resolutions in the region.

The extensive circular labour migration system within the southern African region is by far the single biggest contributing factor to the spread of HIV within the region. Due to

the importance of SADC migrants to the South African economy, the number of SADC nationals living and working within the country has been monitored closely since the formation of the Union in 1910. The best source of data on the number of SADC migrants comes from national census data [StatsSA, 2011]. In the late 1980s and early 1990s, as conflicts in other southern African nations came to an end and the campaign against Apartheid in South Africa was gaining ground, the number of SADC migrants started to increase. After the elections in South Africa in 1994 and the countries reintroduction into SADC, the number of migrants grew rapidly. Between 1985 and 2011, SADC migrants increased by 500% from 300,000 to an estimated 1,5 million. Though these figures represent official government statistics, the true number of migrants is almost certainly higher.

The large number of viral imports and exports within the South African context is consistent with an increase number of migrants living and working in South Africa as well as a period of rapid socio-political transformation in the country. What is surprising is that we uncovered few viral exchange events in the period between 2000 and 2010 and only a mean of one viral introduction in the period from 2010 till the present. Given the scale of migration and human mobility in the southern African region at the moment, it is reasonable to assume that viral exchange continues. It is possible that the early viral exchange events observed between 1985 and 2000 could have seeded large epidemics co-circulating within countries and that the epidemic stabilized as it became more generalized. The tree structure observed in **Figure 2** would support such a hypothesis. However, it is more likely that this is simply an artifact of the asymmetrical temporal and geographic structure of our dataset (**Figure 1**) and that continued sampling will identify new viral exchange events.

Both maximum parsimony and more complex likelihood-based phylogeographic methods are highly biased by uneven sampling across locations. Nevertheless, we chose to use parsimony-based ancestral state reconstructions. We chose parsimony because it is computationally efficient enough to perform replicate analyses on different subsampled but still large datasets, and therefore to explore the sensitivity of our results to uneven sampling. Our sensitivity analyses suggest that the number of viral introductions, both in

and out of South Africa, is highly dependent on the number of samples included from other southern African countries (**Figure 4**). These results further indicate no sign of saturation as the number of estimated introductions from each country continues to increase in an almost linear fashion with increasing numbers of samples. The exception to this is Tanzania, which suggests that the subtype C epidemic in Tanzania is more integrated into the east African subtype C epidemic than the southern African epidemic.

To further investigate the effect of uneven sampling, we performed a second sensitivity analysis where we randomly sampled an even number of sequences from all southern African countries. Notably, the relative number of viral introductions from different source countries became far more uniform, with the bulk of introductions being attributed to Malawi, Zimbabwe and Zambia rather than just Zambia (**Figure 5**). Our work, therefore, highlights the need for new phylogeographic methods that either explicitly model differential sampling across populations [Kuehnert et al. 2016] or are more naturally robust to sampling biases [De Maio et al. 2015]. Given these limitations, it is possible that the number of viral introductions observed is only a conservative estimate. However, the timing of the viral introductions we identified should still be representative of their true temporal distribution even if we have systematically underestimated the absolute number of introductions.

## 5.0 Conclusion

In conclusion, our results underscore the extensive genetic diversity of the HIV-1 subtype C epidemic in South Africa and the heterogeneity of genetic diversity in the southern African region in general. The extreme genetic diversity in South Africa has been driven by multiple introductions of the virus from other southern African countries into the country, coinciding with a period of socio-political transition (1985-2000). Our results also indicate that during the same time period South Africa became a major source of viral introductions into other southern African countries. This import-export dynamic is closely associated with the extensive circular migrant system within the southern African context, which facilitated the early spread of the virus within the region. These findings

underscore the need for better infection control and prevention in countries going through political and socio-economic transitions.

**Funding**

The research of T.d.O. and E.W. is funded through a Medical Research Council flagship grant from the Republic of South Africa (MRC-RFA-UFSP-01-2013/UKZN HIVEPI) and by a Royal Society Newton Advanced Fellowship to T.d.O. The Bill and Melinda Gates Foundation funded O.R. through the PANGEA-HIV initiative.

**Authors Contributions**

T.d.O., E.W., D.R., O.R. and T.S. designed the study and E.W., D.R. and O.R. performed the inferential analyses. S.E. and T.d.O. provided the sequences from the drug resistance cohort, which was used in the inference. E.W. and D.R. wrote the manuscript and produced all figures. All the authors reviewed the manuscript prior to submission.

**Potential conflicts of interests**

The authors have no competing financial interests to declare.

**Acknowledgements**

We thank Mathilda Claassen for her excellent contribution towards generating most of the Tygerberg sequences. We also acknowledge Graeme Jacobs and Gert van Zyl for adding additional information to the Tygerberg HIV-1 drug resistance cohort and database.

**References**

- Abdool Karim, S.S., and Abdool Karim, Q., 2010. HIV/AIDS in South Africa. 2nd ed, Cambridge University Press. Chapter/Section 1: Birth of a rapidly growing epidemic.

- Abecasis, A.B., Vandamme, A-M., Lemey, P. 2009. Quantifying differences in the Temp of Human Immunodeficiency Virus Type 1 Subtype Evolution. Journal of Virology. 83(24): 12917-12924.

- Becker JU, Theodosis C, Kulkarni R. HIV/AIDS, conflict and security in Africa: rethinking relationships. 2008. Journal of the International AIDS Society, 11:3. DOI: 10.1186/1758-2652-11-3

- Bello, G., Zanotto, P.M.A., Imarino, I., Gräf, T., Pinto, A.R., Couto-Fernandez, J.C., Morgado, M.G. 2012. Phylogeographic Analysis of HIV-1 Subtype C Dissemination in Southern Brazil. PLoS ONE. DOI: 10.1371/journal.pone.0035649.

- Dalai, S.C., de Oliveira, T., Harkins, G.W., Kassaye, S.G., Lint, J., Manasa, J., Johnston, E., Katzenstein, D., 2009. Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabawe. AIDS. 23(18): DOI: 10.1097/QAD.0b13e3283320ef3.

- De Maio, N., Wu, C.H., O'Reilly, K.M., Wilson, D., 2015. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. PLoS Genetics. 11(8): e1005421.

- de Oliveira, T., Pillay, D., Gifford, R.J., for the UK Collaborative Group on HIV Drug Resistance. 2010. The HIV-1 Subtype C Epidemic in South America Is Linked to the United Kingdom. PLoS ONE. DOI: 10.1371/journal.pone.0009311.

- Department of Health – Republic of South Africa. Summary Report, 2002: National HIV and Syphilis Antenatal sero-prevalence survey in South Africa. 2002. (http://www.gov.za/sites/www.gov.za/files/hivsyphilis_0.pdf).

- Efron, B., Halloran, E., Holmes, S., 1996. Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences. 93(23): 13429-13434.

- Faria, N.R., Rambaut, A., Suchard, M.A., Baele, G., Bedford, T., Ward, M.J., Tatem A.J., Sousa, J.D., Arinaminpathy N, Pépin, J., Posada, D., Peeters, M., Pybus, O.G., Lemey, P., 2014. The early spread and epidemic ignition of HIV-1 in human populations. Science. 346(6205): 56-61.

- Felsenstein, J., 1985 Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39(4): 783-791.

- Fitch, W.M., 1979. Cautionary remarks on using gene expression events in parsimony procedures. Systematic Zoology. 28: 375-379.

- Hargrove, J., 2008. Migration, mines and mores: the HIV epidemic in southern Africa. South African Journal of Science. 104: 53-61.

- Hemelaar, J., Gouws, E., Ghys, P.D., Osmanov, S., WHO-UNAIDS Network for HIV Isolation and Characterisation., 2011. Global trends in molecular epidemiology of HIV-1 during 2000-2007. AIDS. 25(5): 679-89.

- Jung, M., Leye, N., Vidal, N., Fargette, D., Diop, H., Kane, C.T., Gascuel, O., Peeters, M. 2012. The Origin and Evolutionary History of HIV-1 Subtype C in Senegal. PLoS ONE. DOI: 10.1371/journal.pone.0033579.

- Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 16: 111-120.

- Kühnert, D., Stadler, T., Vaughan, T. G., & Drummond, A. J. 2016. Phylodynamics with migration: A computational framework to quantify population structure from genomic data. Molecular Biology and Evolution, published online.

- Larkin, M.A., Blackshields, G., Brown, N.P., Cehnna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson,

J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23(21): 2947-2948.

- Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One. DOI: 10.1371/journal.pone.0009490

- Sankoff, D., 1975. Minimal mutation trees of sequences. SIAM Journal of Applied Mathematics. 28: 35-42.

- StatsSA 2011 - Statistics South Africa (http://www.statssa.gov.za/publications/P03014/P030142011.pdf)

- Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution. 30: 2725-2729.

- Tavaré, S., 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences. 17: 57-86

- To, T-H., Jung, M., Lycett, S., Gascuel, O., 2016. Fast Dating Using Least-Squares Criteria and Algorithms. Systematic Biology. 65(1): 82-97.

- UNAIDS report. Securing an AIDS Free Future: Practical Lessons about Security and AIDS in Conflict and Post-Conflict Settings. 2015. ISBN: 978-92-9173-990-5. Available at: http://www.unaids.org/sites/default/files/media_asset/JC2402_UNAIDS_CASE_STUDY_en_0.pdf

- UNAIDS, 2010. Global report, UNAIDS report on the global AIDS epidemic 2010. Available at: http://www.unaids.org/sites/default/files/en/media/unaids/contentassets/documents/unaidspublication/2010/20101123_globalreport_en%5b1%5d.pdf

- Van Zyl, G.U., Claassen, M., Engelbrecht, S., Laten, J.D., Cotton, M.F., Theron, G.B., Preiser, W., 2008. Zidovudine with nevirapine for the prevention of HIV mother-to-child transmission reduces nevirapine resistance in mothers from the Western Cape, South Africa. Journal of Medical Virology. 80, 942–946.

- Wilkinson, E., Engelbrecht, S., de Oliveira, T., 2015. History and origin of the HIV-1 subtype C epidemic in South Africa and the greater southern African region. Scientific Reports. 5:16897. DOI 10.1038.srep16897.

- Wong-Staal, F., Gallo, R.C., Haseltine, W., Chang, N.T., Ghrayeb, J., Papas, T.S., Josephs, S.F., Lautenberger, J.A., Pearson, M.L., Petteway S.R Jr.., Ivanoff, L., Baumeister, K., et al., 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature. 313(6000): 277-84.

- Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequence with variable rates over sites: Approximate methods. Journal of Molecular Evolution. 39, 306-314.


**Figure legends**

**Figure 1:** Breakdown of the large subtype C dataset. The graph in the top left hand corner presents the temporal distribution of samples. The graph at the bottom left hand corner gives a breakdown of the number of sequences by country. The graph in the top right hand corner represents the temporal distribution of samples bye country. The graph in the bottom right hand corner represents the estimated mean genetic diversity of each geographical dataset.

**Figure 2:** Dated maximum likelihood phylogeny reconstructed from all 11,290 southern African subtype C sequences. Lineages are colored according to ancestral locations inferred using maximum parsimony.

**Figure 3:** Estimated viral introductions into and out of South Africa identified from parsimony-based ancestral location reconstructions. **(a-f)** The temporal distribution of

estimated introductions imported into South Africa by inferred source country. The error bars represent the 95% bootstrapped confidence intervals computed by rerunning the introduction analysis on a set of 100 bootstrap trees. **(g-l)** The temporal distribution of exports from South Africa into neighboring SADC countries.

**Figure 4:** The sensitivity of identified introductions to the number of samples included from each country. **(a-f)** Introductions imported into South Africa with varied subsampling fractions from each source country. **(g-l)** Introductions exported from South Africa into other countries with varied subsampling fractions from South Africa. Subsampling fraction refers to the fraction of samples included out of the total number of samples available from each country. Error bars show the standard deviation from the mean number of introductions identified from 10 subsampling replicates at each sampling fraction.

**Figure 5:** Estimated viral introductions into South Africa by source country after accounting for biased sampling by subsampling oversampled locations. Error bars show the standard deviation from the mean number of introductions identified from 100 replicates of the subsampling procedure.

**Figure 6**: The estimated number of SADC migrants living and working within South Africa through time from 1911-2011. These estimates represent official government census data estimates. The area shaded in red represents the period in which our phylogeographic reconstruction identified the most viral imports and exports (1985-2000), coinciding with a period of rapid growth in the number of migrants.

**Supplementary Figure 1:** Timeline of historical events. On the left side, historical events from other southern African countries are shown, while on the right hand side, historical events from South Africa are shown. Events in bold typeface relate to HIV events within the region.

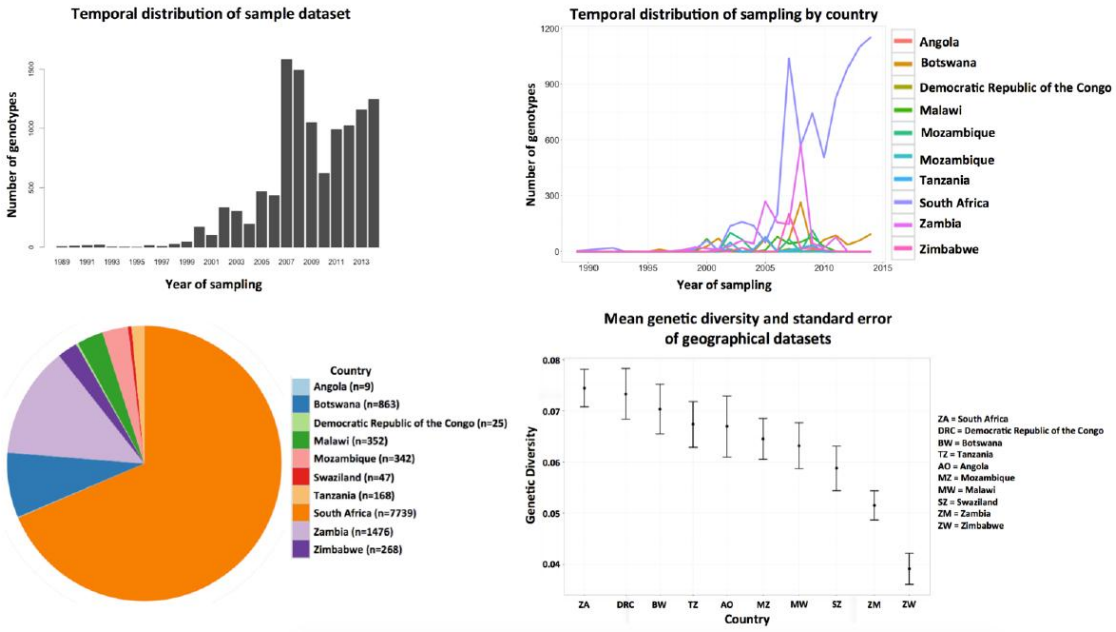
Temporal distribution of sample dataset
Number of genotypes
Year of sampling
Temporal distribution of sampling by country
Number of genotypes
Year of sampling
Angola
Botswana
Democratic Republic of the Congo
Malawi
Mozambique
Mozambique
Tanzania
South Africa
Zambia
Zimbabwe
Country
Angola (n=9)
Botswana (n=863)
Democratic Republic of the Congo (n=25)
Malawi (n=352)
Mozambique (n=342)
Swaziland (n=47)
Tanzania (n=168)
South Africa (n=7739)
Zambia (n=1476)
Zimbabwe (n=268)
Mean genetic diversity and standard error of geographical datasets
Genetic Diversity
Country
ZA = South Africa
DRC = Democratic Republic of the Congo
BW = Botswana
TZ = Tanzania
AO = Angola
MZ = Mozambique
MW = Malawi
SZ = Swaziland
ZM = Zambia
ZW = Zimbabwe

Fig. 1

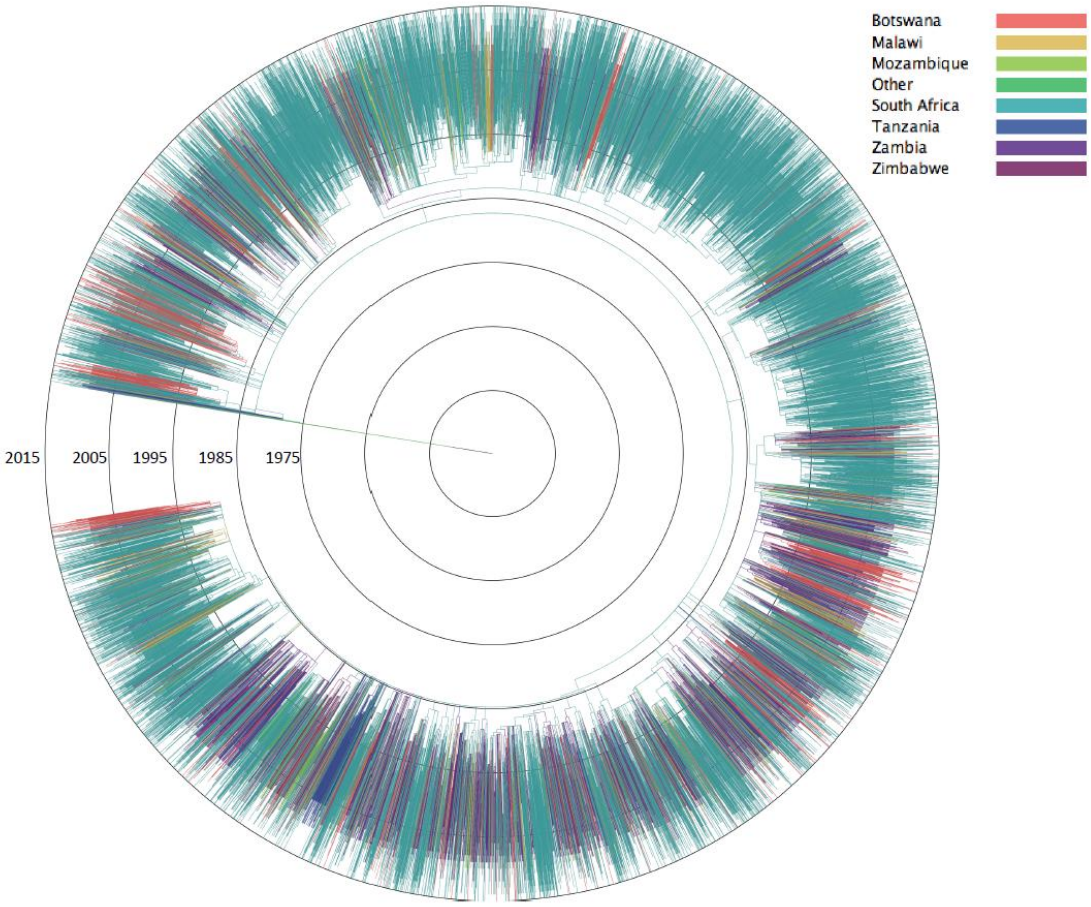
Botswana
Malawi
Mozambique
Other
South Africa
Tanzania
Zambia
Zimbabwe
2015
2005
1995
1985
1975

Fig. 2

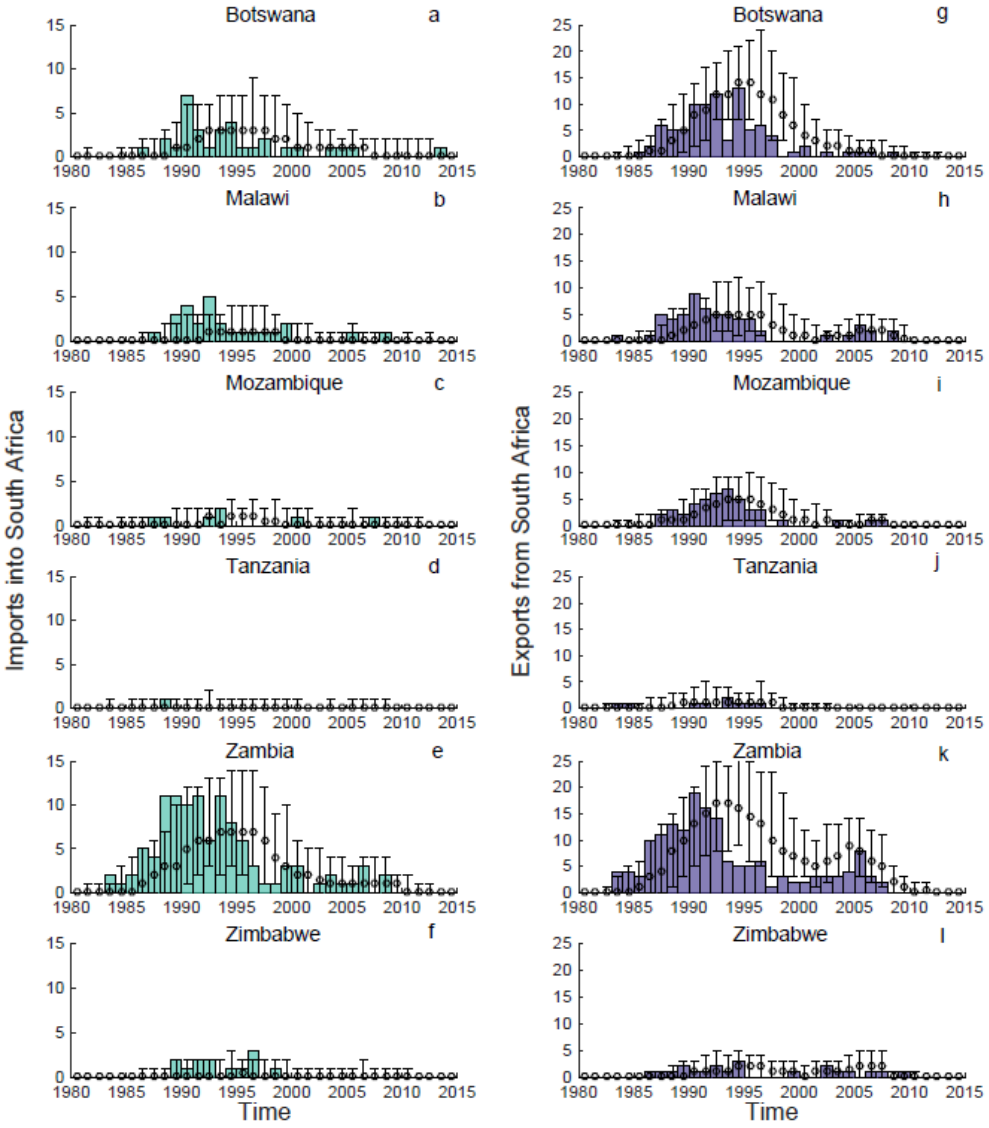
Botswana
a
Malawi
b
Mozambique
c
Tanzania
d
Zambia
e
Zimbabwe
f
Botswana
g
Malawi
h
Mozambique
i
Tanzania
j
Zambia
k
Zimbabwe
l
Imports into South Africa
Exports from South Africa
Time
Time

Fig. 3

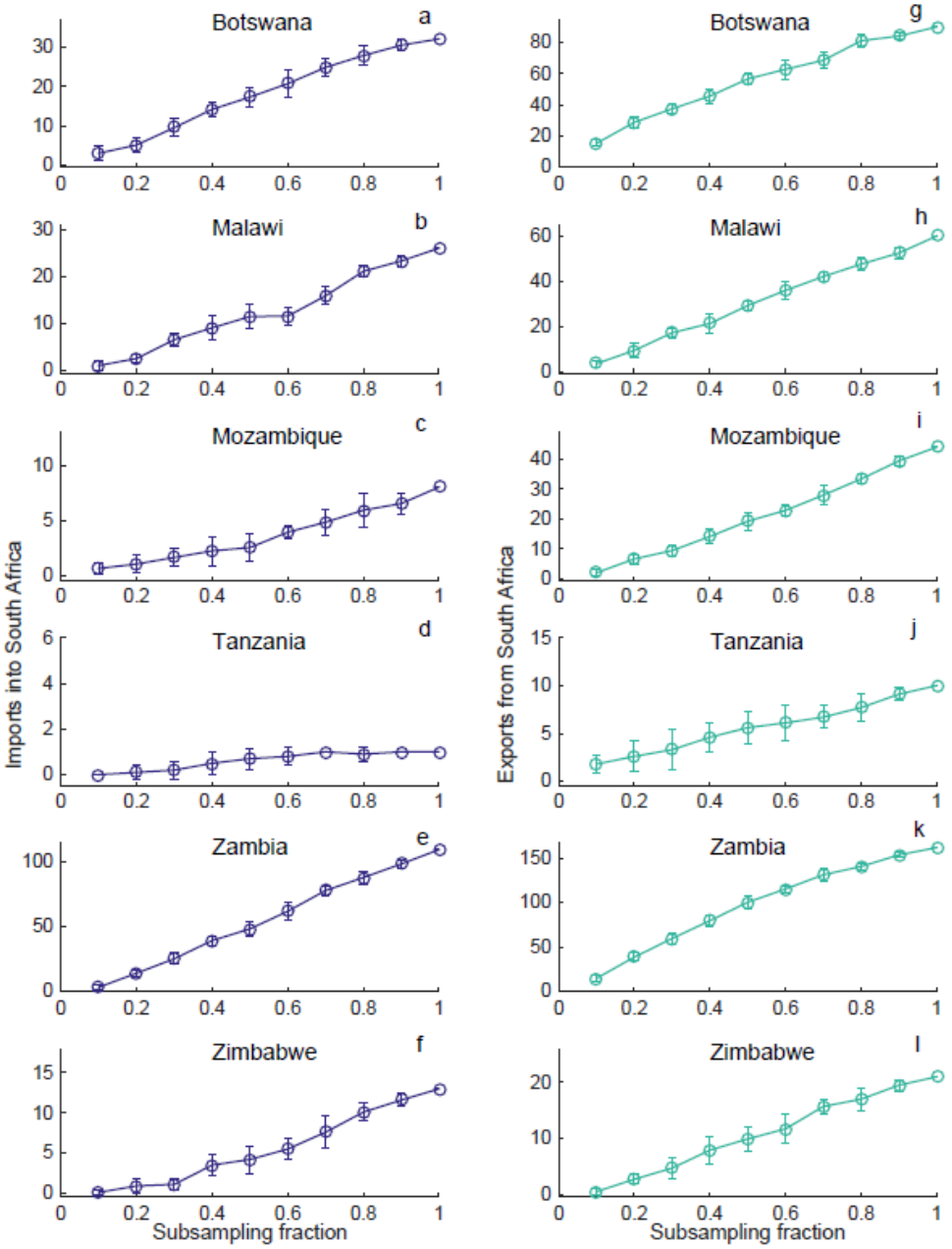
Botswana
a
Malawi
b
Mozambique
c
Tanzania
d
Zambia
e
Zimbabwe
f
Botswana
g
Malawi
h
Mozambique
i
Tanzania
j
Zambia
k
Zimbabwe
l
Imports into South Africa
Exports from South Africa
Subsampling fraction
Subsampling fraction

Fig. 4

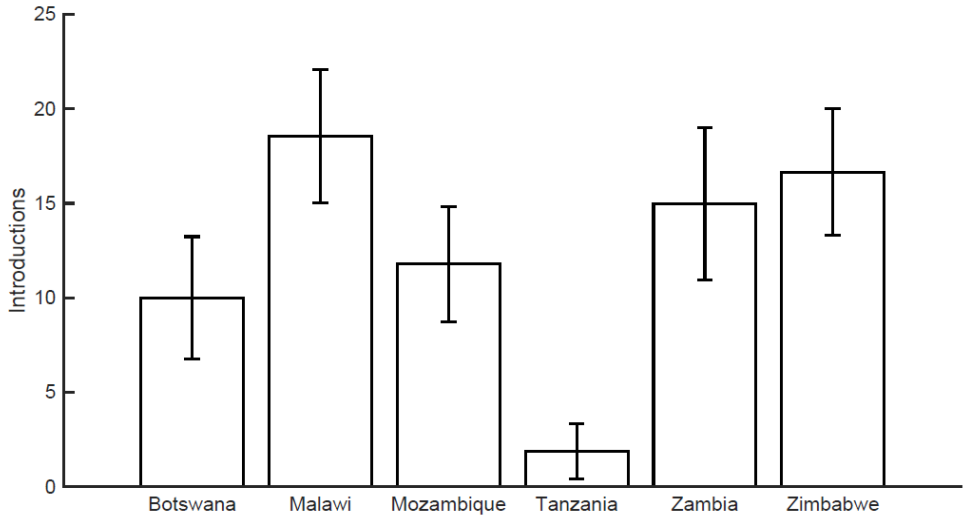
25
20
15
10
5
0
Introductions
Botswana
Malawi
Mozambique
Tanzania
Zambia
Zimbabwe

Fig. 5

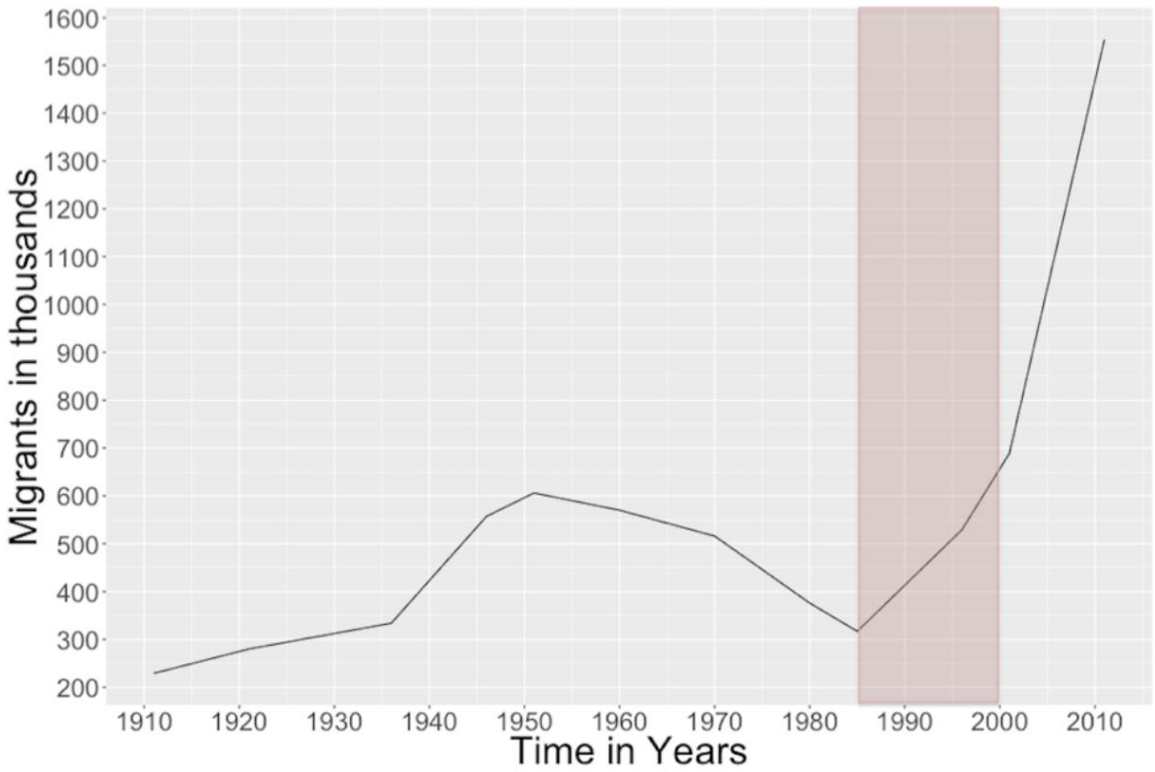
Migrants in thousands
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
Time in Years

Fig. 6

## Highlights

- The HIV-1 subtype C epidemic in South Africa is characterized by very high genetic diversity.
- Most viral introductions of HIV into South Africa occurred between 1985 and 2000.
- During the same time period South Africa became a major source of viral exports to other southern African countries.
- The period of viral exchange between southern African countries coincide with a period of socio-political change in the region and is closely linked to the extensive circular labour migration system, which played a central role in the early spread of the virus in the region.
- Our phylogeographic analyses are sensitive to the number of sequence included from different countries.