


Origin and Phylodynamics of HIV-1 subtype C in South Africa

by

Eduan Wilkinson



*Dissertation presented for the degree
of Doctor of Medical Virology
in the Faculty of Medicine and Health Sciences
at Stellenbosch University*

Supervisor: Prof Susan Engelbrecht

Co-supervisor: Prof. Tulio de Oliveira

Date: December 2013

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature

Name in full

25 / 11 / 2013

Date

Copyright © 2013 Stellenbosch University

All rights reserved

Acknowledgements

I would hereby like to extend my fullest gratitude to the following people and institutions; without whom this thesis would not have been possible:

- Prof. Susan Engelbrecht, the supervisor of this study, for her expert advice, guidance and assistance throughout the course of this project.
- Prof. Tulio de Oliveira of the Africa Centre for Health and Population Studies, co-supervisor of this study, for all his expert advice, guidance and assistance with the phylogenetic analysis of the sequence information.
- To my parents, Edward and Elizabeth Wilkinson, for always believing in me and supporting me with my studies.
- To all the staff and students of the Division of Medical Virology, Department of Pathology, Faculty of Medicine and Health Sciences, Stellenbosch University, for their moral support during the course of this study.
- To all the staff and students of the Africa Centre for Health and Population Studies, University of KwaZulu-Natal, Mtubatuba for their moral support throughout the years.
- To my family and friends for all the support throughout the past couple of years.
- To the first African International Research Training Group (IRTG – 1522) between South Africa (NRF) and Germany (DFG) for the scholarship and all the associated members of the scholarship program.
- To the Harry Crossly Foundation, the Poliomyelitis Research Foundation (PRF), the National Research Foundation (NRF) of South Africa, as well as the Faculty of Medicine and Health of the University of Stellenbosch, for the generous financial support.

Abstract

The HIV epidemic in the past couple of decades has spread at an alarming rate throughout Southern Africa. Today the region accounts for roughly one third of all HIV infections, while prevalence rates in other areas of sub-Saharan Africa remain low. In the following study, sampled sequences from Cape Town, spanning over a 21-year period were used to investigate the epidemic history of HIV, which was compared to epidemic trends across Southern Africa.

Longitudinal sequence data sets were generated from stored patient samples from Cape Town through standard molecular techniques. Firstly, these sequences were used to estimate the date of origin of the HIV epidemic in Cape Town and to reconstruct a demographic history of the epidemic with advanced Bayesian inference methods. These analyses placed the estimated date of origin of the Cape Town epidemic around the mid 1960's with periods of strong epidemic growth observed during the mid 1980's and 1990's. Secondly, reference strains of HIV from Southern African countries were used to estimate the date of origin of the epidemic in the Southern African region. These analyses placed the date of origin of the epidemic in the Southern African region around the mid 1950's roughly ten years before the start of the epidemic in Cape Town/South Africa. These sequences were also used for the reconstruction of the demographic history of the epidemic in the region. A two phased growth in the HIV epidemic in the Southern African region was observed with exponential growth occurring in the mid 1980's and 1990's. Such findings are also supported by HIV prevalence estimates made by some of the leading HIV research centres and government health departments. Thirdly, a large number of homologous reference strains were used to establish the evolutionary relationship of HIV isolates from Cape Town with those from around the world. A close genetic relationship between Cape Town isolates with other South African and other Southern African isolates was observed in these analyses. Finally, large monophyletic clusters of Cape Town isolates, which was observed during the evolutionary inference, were further investigated. After detailed analyses it appears that these transmission clusters of HIV-1 have been in circulation amongst the infected population of Cape Town for several years or decades.

Opsomming

Die MIV-epidemie het in die afgelope paar dekades teen 'n snelspoed deur Suider-Afrika versprei. Een derde van die globale MIV-infeksies kom hiër voor terwyl ander dele van Afrika aansienlik minder infeksies aantoon. Verskeie studies skryf dit toe aan onder andere: manlike besnydenis, seksuele losbandigheid, migrasie en verskeie politieke faktore. Die MIV-epidemie in Suider-Afrika word deur 'n enkele sub tipe van die virus oorheers (nl. MIV Sub tipe C) terwyl ander subtipes sirkuleer deur die res van sub Sahara-Afrika. In die opeenvolgende studie word DNS-monsters uit Kaapstad (wat oor 'n 20 jaar tydperk strek) gebruik om die oorsprong en verloop van die epidemie te bestudeer. Die data van die Kaapstad epidemie word met die geskiedkundige verloop van die epidemie in Suider-Afrika vergelyk.

Deur gestoorde bloedmonsters van Kaapstad te gebruik, was DNS-datastelle gegeneer deur middel van standaard molekulêre tegnieke. Die DNS-monsters was eerstens gebruik om die evolusionêre oorsprong en verloop van die epidemie in Kaapstad te bepaal deur Bayesiaanse Markov-ketting Monte Carlo steekproefneming. Volgens die resultate het die epidemie sy oorsprong in die 1960's. Klein periodes van epidemiese groei kon waargeneem word gedurende die 1980's en -90's. Die bevindings is toe vergelyk met die geskiedkundige verloop van die epidemie in Suider-Afrika. Die Suider-Afrika epidemie se oorsprong en verloop was afgelei van DNS monsters wat verkry is van publieke databasisse en die gebruik van soortgelyke Bayesiaanse metodes. Die resultate van die ondersoek het bevind dat die epidemie in Suider-Afrika in die 1950's ontstaan het. In vergelyking toon dit 'n stadiger liniêre groei met kort periodes van eksponensiële groei. Verder is 'n standard filogenetiese analise onderneem om die evolusionêre verwantskap van die Kaapstad-monsters te bepaal met ander MIV sub tipe C isolate. Die filogenetiese steekproef toon dat die Kaapstad-monster baie nou verwant is aan ander isolate van Kaapstad, Suid-Afrika en Suider Afrika. Buiten hierdie bevindings was transmissie-bondels van MIV in Kaapstad ontdek. Na 'n deeglike verdere filogenetiese ondersoek blyk dit of die transmissie bondels al vir 'n paar dekades deur die geïnfecteerde populasie van Kaapstad sirkuleer.

Table of Contents

	Page
Abstract	4
Opsomming	5
List of abbreviations	7
List of Figures	13
List of Tables	18
Chapter 1: INTRODUCTION AND LITERATURE REVIEW	25
INTRODUCTION	25
LITERATURE REVIEW	26
AIM OF THE STUDY	60
Chapter 2: MATERIALS AND METHODS	64
Chapter 3: RESULTS	89
Chapter 4: DISCUSSION AND CONCLUSION	163
DISCUSSION	163
CONCLUSION	190
Chapter 5: REFERENCES	193
Chapter 6: APPENDIX	222

List of abbreviations

® - registered trade mark

°C – degrees Celsius

AIDS – Acquired Immunodeficiency Syndrome

ANC – African National Congress

ARV – antiretroviral

ASSA – Actuarial Society of South Africa

b – random branch lengths

BEAST – Bayesian Evolutionary Analysis Sampling Trees

BLAST – Basic Local Alignment Search Tool

bp – base pairs

BSP – Bayesian Skyline Plot

CA – California

CD4 – cluster of differentiation 4

CDC – Center for Disease Control and Prevention

cDNA – complimentary Deoxyribonucleic acid

cm – centimetre

CMV – Cytomegalovirus

Const – Constant population size tree prior

cpz – chimpanzee

CRF – circulating recombinant form

CSW – commercial sex worker

DDBJ – DNA Data Bank of Japan

DFG – Deutsche Forschungsgemeinschaft

DNA – Deoxyribonucleic acid

DNS – Deoksiribonukleïnesuur

dNTP – dinucleotide triphosphate

DRC – Democratic Republic of the Congo

EES – effective sample size

EMBL – European Molecular Biology Laboratory

env – envelope

EPS – Effective Population Size

ESS – Effective sample size

est – estimated mutation rate

et al., – et alia

F81 – Felsenstein 1981

fix – Fixed mutation rate

gag – glycoprotein

GOBICS – Göttingen Bioinformatics Compute Server

gor – gorilla

GRID – Gay-Related Immune Deficiency

GTR – General Time Reversible

HAART – highly active antiretroviral therapy

HCV – Hepatitis C Virus

HIV – Human Immunodeficiency Virus

HIV-1 – Human Immunodeficiency Virus Type 1

HIV-2 – Human Immunodeficiency Virus Type 2

HKY – Hasegawa, Kishino and Yano

HPD – highest posterior density

HREC - Human Research Ethics Committee

ID - Identification

Inc. - Incorporated

IRTG – International Research Training Group

IVDU – intravenous drug users

JC69 – Jukes and Cantor 1969

jpHMM – jumping profile Hidden Markov Model

K80 – Kimura 1980

kbp – kilo-base pairs

KS – Kaposi`s sarcoma

LAV – Lymphadenopathy associated virus

LTR – long terminal repeats

MCMC – Markov chain Monte Carlo

ME – Minimum Evolution

MgCl₂ – Magnesium Chloride

min – minutes

MIV – Menslike Immuniteitsgebreksvirus

MK – Umkhonto we Sizwe

ML – Maximum Likelihood

mM – millimolar

MMWR – Morbidity and Mortality Weekly Report

MP – Maximum Parsimony

MSM – Men who have sex with men

NCBI – National Center for Biotechnology Information

Ne – population size

ng - nanogram

NJ – Neighbor-Joining

NNI – nearest neighbor interchange

NRF – National Research Foundation

OUT – operational taxonomic unit

p – random substitution parameter

PCP – Pneumocystis carinii pneumonia

PCR – Polymerase Chain Reaction

PLTT – percentage lineages through time

pmol – picomole

pol - polymerase

PR – protease

PRF – Poliomyelitis Research Foundation

R_2 – coefficient of variation

relax – relaxed molecular clock assumption

RNA – Ribonucleic acid

RT – reverse transcriptase

RT-PCR – Reverse Transcriptase Polymerase Chain Reaction

sec – seconds

SIV – Simian Immunodeficiency Virus

smm – sooty mangabey monkey

SPR – subtree pruning and regrafting

SRD06 – Shapiro, Rambaut, Drummond 2006

sSA – sub-Saharan Africa

strict – strict molecular clock assumption

SWAPO – South West African Peoples Organization

TM – trade mark

tMRCa – time to the most recent common ancestor

Ts – Transitions

Tv – Transversions

UNAIDS – Joint United Nations program on HIV and AIDS

UPGMA – unweighted pair group method with arithmetic means

URF – unique recombinant form

USA – United States of America

UTU – unique taxonomic unit

UV – Ultraviolet

WHO – World Health Organization

WI - Wisconsin

μ l – microliter

μ M – micromolar

ZANU – Zimbabwe African National Union

ZAPU – Zimbabwe African Peoples Union

List of Figures

	Page
Figure 1.1: HIV prevalence rates across the Southern Africa region.	30
Figure 1.2: A diagrammatical representation of the genome layout of HIV-1.	37
Figure 1.3: Global distribution of different HIV-1 subtypes and circulating recombinant forms.	39
Figure 1.4: A breakdown of the basic steps involved in any phylogenetic investigation.	44
Figure 1.5: A diagrammatical breakdown of the most important nucleotide substitution models currently available.	47
Figure 1.6: Diagrammatical representation of a phylogenetic tree.	49
Figure 2.1: A step-by-step breakdown of the methodology that was used in the generation of the Cape Town data sets is illustrated in the different steps.	67
Figure 2.2: A diagrammatical breakdown of the phylogenetic methodology that was used during in order to answer the 4 main scientific questions in this study.	75
Figure 2.3: An illustration of a PhyloType file annotation.	82
Figure 3.1: The root-to-tip regression analysis of the <i>pol</i> Cape Town data set.	92
Figure 3.2: Bayesian skyline plot of Cape Town HIV-1 subtype C <i>gag</i> p24 data sets.	97
Figure 3.3: Bayesian skyline plot of Cape Town HIV-1 subtype C <i>pol</i> data sets.	98
Figure 3.4: Estimated percentage lineages through time for the <i>gag</i> p24 Cape Town data set.	100
Figure 3.5: Estimated percentage lineages through time for the <i>pol</i> Cape Town data set.	101

Figure 3.6: BSP of the Southern African <i>gag</i> p24 data set (excluding Cape Town).	110
Figure 3.7: BSP of the Southern African <i>pol</i> data set (excluding Cape Town).	111
Figure 3.8: BSP of the Southern African <i>gag</i> p24 data set (including Cape Town).	119
Figure 3.9: BSP of the Southern African <i>pol</i> data set (including Cape Town).	120
Figure 3.10: Large-scale ME-SPR tree of the <i>gag</i> p24 data set.	125
Figure 3.11: Large-scale ML tree with aLRT of the <i>gag</i> p24 data set.	126
Figure 3.12: Large-scale ME-SPR tree of the <i>gag-pol</i> concatenated data set.	128
Figure 3.13: Large-scale ML tree with aLRT of the <i>gag-pol</i> concatenated data set.	129
Figure 3.14: Large-scale ME-SPR tree of the <i>pol</i> data set.	131
Figure 3.15: Large-scale ML tree with aLRT of the <i>pol</i> data set.	132
Figure 3.16: Observed clustering of Cape Town isolates in the concatenated <i>gag-pol</i> ML.aLRT tree topology as was observed through the PhyloType analysis.	144
Figure 3.17: The five putative transmission clusters of Cape Town sequences that were observed through manual inspection of the <i>gag</i> p24 ML.aLRT tree topology.	147
Figure 3.18: The three putative transmission clusters of Cape Town sequences that were observed through manual inspection of the <i>pol</i> ML.aLRT tree topology.	149
Figure 3.19: A time resolved tree topology of the <i>gag</i> p24 Cape Town data set with the 5 different clusters.	159
Figure 3.20: Time resolved tree topology of the <i>pol</i> Cape Town data sets with the 3 different clusters.	160

Figure 4.1: Major migratory routes across Southern African during the second half of the 20th century.	188
Figure 6.1: A diagrammatical representation of a molecular clock analysis that was performed in Path-O-Gen on the final <i>gag</i> p24 Cape Town data set.	236
Figure 6.2: Convergence in the trace file for the run Const.strict.est.2 in the <i>gag</i> p24 Southern African (excluding Cape Town isolates) data set.	243
Figure 6.3: Convergence in the trace file for the run BSP.relax.est.1 in the <i>gag</i> p24 Southern African (including Cape Town isolates) data set.	243
Figure 6.4: Convergence in the trace file for the run Const.relaxed.fix.1 in the concatenated <i>gag-pol</i> Southern African (including Cape Town isolates) data set.	244
Figure 6.5: Convergence in the trace file for the BSP.relax.est.1 model parameter run in the <i>pol</i> Southern African (including Cape Town isolates) data set.	244
Figure 6.6: NJ-tree topology of the <i>gag</i> .cluster.1 data set with bootstrap resampling.	245
Figure 6.7: ME-tree topology of the <i>gag</i> .cluster.1 data set with bootstrap resampling.	246
Figure 6.8: ML-tree topology of the <i>gag</i> .cluster.1 data set with aLRT.	247
Figure 6.9: ML-tree topology of the <i>gag</i> .cluster.1 data set with bootstrap resampling.	248
Figure 6.10: Bayesian tree topology of the <i>gag</i> .cluster.1 data set.	249
Figure 6.11: NJ-tree topology of the <i>gag</i> .cluster.2 data set with bootstrap resampling.	250
Figure 6.12: ME-tree topology of the <i>gag</i> .cluster.2 data set with bootstrap resampling.	251
Figure 6.13: ML-tree topology of the <i>gag</i> .cluster.2 data set with aLRT.	252
Figure 6.14: ML-tree topology of the <i>gag</i> .cluster.2 data set with bootstrap resampling.	253
Figure 6.15: Bayesian tree topology of the <i>gag</i> .cluster.2 data set.	254

Figure 6.16: NJ-tree topology of the <i>gag</i> .cluster.3 data set with bootstrap resampling.	255
Figure 6.17: ME-tree topology of the <i>gag</i> .cluster.3 data set with bootstrap resampling.	256
Figure 6.18: ML-tree topology of the <i>gag</i> .cluster.3 data set with aLRT.	257
Figure 6.19: ML-tree topology of the <i>gag</i> .cluster.3 data set with bootstrap resampling.	258
Figure 6.20: Bayesian tree topology of the <i>gag</i> .cluster.3 data set.	259
Figure 6.21: NJ-tree topology of the <i>gag</i> .cluster.4 data set with bootstrap resampling.	260
Figure 6.22: ME-tree topology of the <i>gag</i> .cluster.4 data set with bootstrap resampling.	261
Figure 6.23: ML-tree topology of the <i>gag</i> .cluster.4 data set with aLRT.	262
Figure 6.24: ML-tree topology of the <i>gag</i> .cluster.4 data set with bootstrap resampling.	263
Figure 6.25: Bayesian tree topology of the <i>gag</i> .cluster.4 data set.	264
Figure 6.26: NJ-tree topology of the <i>gag</i> .cluster.5 data set with bootstrap resampling.	265
Figure 6.27: ME-tree topology of the <i>gag</i> .cluster.5 data set with bootstrap resampling.	266
Figure 6.28: ML-tree topology of the <i>gag</i> .cluster.5 data set with aLRT.	267
Figure 6.29: ML-tree topology of the <i>gag</i> .cluster.5 data set with bootstrap resampling.	268
Figure 6.30: Bayesian tree topology of the <i>gag</i> .cluster.5 data set.	269
Figure 6.31: NJ-tree topology of the <i>pol</i> .cluster.1 data set with bootstrap resampling.	270
Figure 6.32: ME-tree topology of the <i>pol</i> .cluster.1 data set with bootstrap resampling.	271
Figure 6.33: ML-tree topology of the <i>pol</i> .cluster.1 data set with aLRT.	272
Figure 6.34: ML-tree topology of the <i>pol</i> .cluster.1 data set with bootstrap resampling.	273

Figure 6.35: Bayesian tree topology of the <i>pol.cluster.1</i> data set.	274
Figure 6.36: NJ-tree topology of the <i>pol.cluster.2</i> data set with bootstrap resampling.	275
Figure 6.37: ME-tree topology of the <i>pol.cluster.2</i> data set with bootstrap resampling.	276
Figure 6.38: ML-tree topology of the <i>pol.cluster.2</i> data set with aLRT.	277
Figure 6.39: ML-tree topology of the <i>pol.cluster.2</i> data set with bootstrap resampling.	278
Figure 6.40: Bayesian tree topology of the <i>pol.cluster.2</i> data set.	279
Figure 6.41: NJ-tree topology of the <i>pol.cluster.3</i> data set with bootstrap resampling.	280
Figure 6.42: ME-tree topology of the <i>pol.cluster.3</i> data set with bootstrap resampling.	281
Figure 6.43: ML-tree topology of the <i>pol.cluster.3</i> data set with aLRT.	282
Figure 6.44: ML-tree topology of the <i>pol.cluster.3</i> data set with bootstrap resampling.	283
Figure 6.45: Bayesian tree topology of the <i>pol.cluster.3</i> data set.	284

List of Tables

	Page
Table 1.1: Total number of documented AIDS cases per country and the estimated number of AIDS cases per million inhabitants for several Southern African countries by the early 1990's.	32
Table 1.2: Estimated national population level HIV sero-prevalence trends for several Southern African countries calculated from antenatal clinic data.	32
Table 1.3: Summary of the various methods of tree construction.	43
Table 1.4: A representation of a distance matrix calculated for a data set of eight taxa or sequences.	50
Table 2.1: List of chemicals and commercial products used in the study.	65
Table 2.2: Equipment used to perform sample analysis.	65
Table 2.3: Software programs and online analytical tool that were used in the analysis of sequence information.	66
Table 2.4: Specific inclusion and exclusion criteria for sample selection.	67
Table 2.5: Cycling conditions for the PreNested and Nested <i>gag</i> p24 PCR assays.	70
Table 2.6: Cycling conditions for the PreNested and Nested protease- <i>pol</i> PCR assays.	70
Table 2.7: Cycling conditions for the PreNested and Nested reverse transcriptase- <i>pol</i> PCR assays.	71
Table 2.8: Sequencing primers and the annealing temperatures that were used for the sequencing of amplified products.	72
Table 3.1: The estimated tMRCA and mutation rates for the Cape Town <i>gag</i> p24 data set.	94

Table 3.2: The estimated tMRCA and mutation rates for the Cape Town <i>gag-pol</i> concatenated data set.	95
Table 3.3: The estimated tMRCA and mutation rates for the Cape Town <i>pol</i> data set.	96
Table 3.4: Estimated tMRCA and mutation rates for the Southern African <i>gag p24</i> data set.	103
Table 3.5: Inferred tMRCA of the various countries in the Southern African only <i>gag p24</i> data set.	104
Table 3.6: The estimated tMRCA and mutation rates for the Southern African only <i>gag-pol</i> concatenated data set.	105
Table 3.7: Inferred tMRCA of the various countries in the Southern African only <i>gag-pol</i> concatenated data set.	107
Table 3.8: The estimated tMRCA and mutation rates for the Southern African <i>pol</i> data set (excluding Cape Town).	108
Table 3.9: Inferred tMRCA of the various countries in the <i>pol</i> Southern African data set.	109
Table 3.10: The estimated tMRCA and mutation rates for the entire Southern Africa <i>gag p24</i> data set.	113
Table 3.11: Inferred tMRCA of the various countries in the entire Southern African <i>gag p24</i> data set (including Cape Town).	114
Table 3.12: The estimated tMRCA and mutation rates for the entire Southern Africa <i>gag-pol</i> concatenated data set (including Cape Town).	115
Table 3.13: Inferred tMRCA of the various countries in the Southern African <i>gag-pol</i> concatenated data set, including sequence information from Cape Town.	116
Table 3.14: The estimated tMRCA and mutation rates for the Southern Africa <i>pol</i> data set (including Cape Town).	117

Table 3.15: Inferred tMRCA of the various countries in the Southern African <i>pol</i> data set (including Cape Town).	118
Table 3.16: The complete <i>gag</i> p24 data set for the large-scale phylogenetic inference.	121
Table 3.17: The complete <i>gag-pol</i> concatenated data set for the large-scale phylogenetic inference.	122
Table 3.18: The complete <i>pol</i> data set for the large-scale phylogenetic inference.	123
Table 3.19: Results of the PhyloType analysis of the <i>gag</i> p24 ME-SPR tree topology.	134
Table 3.20: Results of the PhyloType analysis of the <i>gag</i> p24 ML.aLRT tree topology.	135
Table 3.21: PhyloType analysis of <i>gag-pol</i> ME-SPR phylogeny based on temporal and geographical criteria.	137
Table 3.22: PhyloType analysis of <i>gag-pol</i> ML.aLRT phylogeny based on temporal and geographical criteria.	138
Table 3.23: PhyloType analysis of <i>pol</i> ME-SPR phylogeny based on spatiotemporal criteria.	140
Table 3.24: PhyloType analysis of <i>pol</i> ML.aLRT phylogeny based on spatiotemporal criteria.	142
Table 3.25: The five putative transmission clusters of Cape Town sequences that were found through manual assessment of the <i>gag</i> p24 ML.aLRT tree topology.	146
Table 3.26: The three putative transmission clusters of Cape Town sequences that were found through manual assessment of the <i>pol</i> ML.aLRT tree topology.	148
Table 3.27: Summary of the clustering analyses for the five putative <i>gag</i> p24 clusters of Cape Town sequences, as well as the results of the clustering analyses for the Indian and Brazilian clades.	151

Table 3.28: Summary of the clustering analyses for the three putative <i>pol</i> clusters of Cape Town sequences, as well as the results of the clustering analyses for the Indian and Brazilian clades.	153
Table 3.29: The estimated tMRCA's of the various <i>gag</i> p24 transmission clusters of Cape Town sequences.	155
Table 3.30: The estimated tMRCA's of the various <i>pol</i> transmission clusters of Cape Town sequences.	157
Table 6.1: Composition of the Cape Town <i>gag</i> p24 data set.	223
Table 6.2: Composition of the Cape Town <i>gag-pol</i> concatenated data set.	229
Table 6.3: Composition of the Cape Town <i>pol</i> data set.	231
Table 6.4: Total number of taxa that were used in each of the various data sets for the Bayesian inference of the Southern African HIV-1 subtype C epidemic.	237
Table 6.5: Results of Bayes factor comparison for the Cape Town <i>gag</i> p24 data set.	238
Table 6.6: Results of Bayes factor comparison for the Cape Town concatenated <i>gag-pol</i> data set.	238
Table 6.7: Results of Bayes factor comparison for the Cape Town <i>pol</i> data set.	239
Table 6.8: Results of Bayes factor comparison for the Southern African <i>gag</i> p24 data set, excluding sequence data from the original Cape Town data set.	239
Table 6.9: Results of Bayes factor comparison for the concatenated Southern African <i>gag-pol</i> data set, excluding sequence data from the original Cape Town data set.	240
Table 6.10: Results of Bayes factor comparison for the Southern African <i>pol</i> data set, excluding sequence data from the original Cape Town data set.	240

Table 6.11: Results of Bayes factor comparison for the Southern African <i>gag</i> p24 data set, including sequence data from the original Cape Town data set.	241
Table 6.12: Results of Bayes factor comparison for the concatenated Southern African <i>gag-pol</i> data set, including sequence data from the original Cape Town data set.	241
Table 6.13: Results of Bayes factor comparison for the Southern African <i>pol</i> data set, including sequence data from the original Cape Town data set.	242

CHAPTER ONE - TABLE OF CONTENTS

	Page
INTRODUCTION	25
LITERATURE REVIEW	26
1.1. The history of the Human Immunodeficiency Virus	26
1.1.1 The start of the global HIV pandemic	26
1.1.2 The origin of HIV	28
1.1.3 The history of the epidemic in Southern Africa	30
1.2 HIV transmission and prevalence trends	33
1.2.1 HIV transmission risk and male circumcision	33
1.2.2 HIV transmission risk and concurrency	34
1.2.3 HIV transmission risk and the effect of migration	34
1.2.4 HIV transmission risk and the role of political factors	35
1.3 The genomic organization of the structural genes of HIV-1	36
1.4 Genetic diversity of HIV-1	38
1.5 HIV-1 subtype diversity in Africa	40
1.6 Phylogenetic analysis and HIV	42
1.6.1 An introduction to phylogenetics	42
1.6.2 Retrieving of relevant genetic information	44
1.6.3 Sequence alignments	45
1.6.4 Nucleotide substitution models	46
1.6.5 Phylogenetic inference	48
1.6.5.1 Distance based methods of tree inference	49
1.6.5.1.1 UPGMA	50
1.6.5.1.2 Fitch-Margoliash method	51

1.6.5.1.3 Neighbor-Joining	51
1.6.5.1.4 Minimum-Evolution	52
1.6.5.2 Character based methods of tree inference	53
1.6.5.2.1 Maximum-Parsimony	53
1.6.5.2.2 Maximum-likelihood	54
1.6.5.2.3 Bayesian methods of tree inference	55
1.6.5.3 The problem of finding the best tree topology	56
1.6.6 The implication of a molecular clock	57
1.6.7 The coalescent theory in modern phylodynamics	59
AIM OF THE STUDY	60

CHAPTER ONE

INTRODUCTION

The human immunodeficiency virus (HIV) is arguably one of the most devastating and serious health problems facing humanity today. HIV was introduced into human populations through a zoonotic transmission of a similar virus called simian immunodeficiency virus (SIV), which is commonly found in non-human primates throughout sub-Saharan Africa (sSA) [Hahn *et al.*, 2000]. This virus was transmitted to human populations roughly a century ago in areas of Central Africa [Korber *et al.*, 2000; Worobey *et al.*, 2008]. Since the transmission of the virus to humans, HIV spread to various geographical locations before we became aware of the virus existence in the early 1980's. Shortly after the first cases of HIV/AIDS were reported in the United States of America (USA) and Europe, similar cases started to be documented amongst individuals in countries from Central and Eastern Africa [Clumeck *et al.*, 1984; Serwadda *et al.*, 1985]. These new cases of HIV-1 amongst Africans, in comparison to the epidemic in Europe and the USA, were largely documented amongst heterosexual individuals.

Southern African nations, with the exception of South Africa who reported early cases of HIV/AIDS amongst urban homosexual men, only started to report their first cases of HIV a couple of years after the countries in Central and East Africa [Hira *et al.*, 1989; Reeve, 1989; Ingstad, 1990; Vuylsteke *et al.*, 1993; Phits'ane, 1994; Ojo and Delaney, 1997]. While the HIV epidemics in Central and East African countries have remained relatively stable since the outbreak of the epidemic, prevalence trends in Southern African countries have grown almost exponentially and have only stabilized in recent years [UNAIDS, 2012]. Currently, UNAIDS estimates that roughly one out of every three people living with HIV/AIDS in the world resides in the Southern African region [UNAIDS, 2012].

Another stark contradiction between the HIV-1 epidemics in the sSA region are that the HIV-1 epidemics in countries in Central and Eastern Africa are caused by a multitude of viral HIV-1 subtypes, while the epidemic in Southern Africa is predominantly caused by a single subtype called HIV-1 subtype C (www.hiv.lanl.gov). Today, HIV-1 subtype C accounts for just over half of all the HIV-1 infections in the world [Santos and Soares, 2010]. This may be due to the overwhelming prevalence of HIV-1 subtype C in the most severely affected nations of Southern African. However, large HIV-1 subtype C epidemics are also found in other areas of the world such as: South America (largely in the south-eastern region of Brazil), East Africa (Ethiopia,

Kenya, and Somalia), the Indian sub-continent, South-East Asia, and in the Far East [Santos and Soares, 2010]. To date, several studies have been conducted on the evolutionary history and dynamics of the HIV-1 subtype C epidemics in some countries such as; Malawi [Travers *et al.*, 2004], Brazil [Bello *et al.*, 2008], Zimbabwe [Dalai *et al.*, 2009], Ethiopia [Tully and Wood, 2010], the United Kingdom [de Oliveira *et al.*, 2010], India [Shen *et al.*, 2011], Senegal [Jung *et al.*, 2012], and Angola [Afonso *et al.*, 2012]. In addition to these studies, which focused on the HIV-1 epidemics within countries, two other studies investigated the evolutionary history of the global HIV-1 subtype C epidemic [Travers *et al.*, 2004; Novitsky *et al.*, 2010] and the regional HIV-1 subtype C epidemic in East Africa [Delatorre and Bello, 2012] respectively. Such studies provide valuable insight into the origin and growth of these epidemics.

However, no evolutionary investigation has been conducted on the HIV-1 subtype C epidemic within South Africa. In the following study the evolutionary history, including the estimated date of origin of the South African epidemic, as well as the phylodynamic aspects of the epidemic, will be investigated. Due to the close relation of the HIV-1 subtype C epidemic in South Africa to those in other Southern African countries, it is of importance to compare the evolutionary trends of the epidemic in South Africa with those in other Southern African countries. Such a study will greatly enhance our current understanding of the HIV-1 subtype C epidemic in South African and within the larger Southern African region.

LITERATURE REVIEW

In the following section the history of the HIV pandemic, the genomic organization of HIV-1, the genetic diversity, factors influencing HIV transmission, as well as phylogenetic methods commonly used in the analysis of HIV, will be briefly reviewed.

1.1. The history of the Human Immunodeficiency Virus

1.1.1 The start of the global HIV pandemic

In the early months of 1981, several homosexual men presented with a variety of unusual symptoms at different hospitals and clinics throughout the USA. These men suffered from opportunistic infections such as; *Pneumocystis jiroveci* pneumonia (which historically is known as *Pneumocystis carinii* pneumonia or PCP), oral thrush, high viral loads for cytomegalovirus (CMV), and a relatively malignant cancer called Kaposi`s sarcoma (KS). Close investigation also revealed that the majority of the men had very low T-cell counts, which indicated an immune

dysfunction. These opportunistic infections, all coinciding in otherwise healthy young men, prompted doctors to submit a paper to be included in the Center for Disease Control and Prevention's (CDC) *Morbidity and Mortality Weekly Report (MMWR)* weekly newsletter [MMWR – June 1981].

These symptoms all coincided in people from the same demographic and social background, and their association with a compromised immune system, in particular lower levels of T-cells [Friedman-Kien, 1981], led doctors to believe that they were dealing with a new unknown disease. In these early days doctors called this new disease GRID, or Gay-Related Immune Deficiency, but by the end of the year similar cases of the disease were starting to appear in the heterosexual population [Brennan and Durack, 1981].

The first group in the general heterosexual population to present with these unusual symptoms were intravenous drug users or (IVDU's) [Masur *et al.*, 1981]. The second group was young Haitian immigrants in the USA [MMWR – July 1982]. Shortly after that, reports of the disease were documented amongst haemophiliacs who had been treated with blood and other blood products [MMWR – December 1982a]. By the end of 1982 reports of the disease in new born babies, who were born to IVDU mothers, were documented [MMWR – December 1982b]. The occurrence of the disease in non-homosexual individuals meant that the acronym GRID was no longer appropriate. A new term for this illness, Acquired Immune Deficiency Syndrome or AIDS, was suggested in July of 1982 at a meeting in Washington D.C. [MMWR – September 1982]. AIDS turned out to be an appropriate name because when people acquired the condition, it led to a deficiency within the host immune system, and because it was a syndrome, with a wide range of possible manifestations, rather than a single disease. Over the following years, other countries, particularly in Europe and Africa, started to report their first cases of AIDS [Vilaseca *et al.*, 1982; Rozenbaum *et al.*, 1982; Clumeck *et al.*, 1984; Weller *et al.*, 1984; Serwadda *et al.*, 1985].

In May of 1983, Professor Luc Montagnier and his team at the Pasteur Institute in Paris reported that they had isolated a new retrovirus from the lymph node of a patient suffering from AIDS. The French team named the new isolated virus LAV for Lymphadenopathy-Associated Virus [Barre-Sinoussi *et al.*, 1983]. The findings of the French team were confirmed by research teams in the USA [Levy *et al.*, 1984; Gallo *et al.*, 1984, Schüpbach *et al.*, 1984; Sarngadharan *et al.*, 1984]. Ratner and co-workers independently confirmed that these new viruses, which were

isolated by the French and American researchers, were similar to one another and also published the first fully sequenced genome of the virus [Ratner *et al.*, 1985].

After 30 years since the first reported cases of AIDS appeared in the USA, more than 60 million people have been infected with the virus worldwide and roughly 25 million people have died from HIV/AIDS related illnesses [Merson *et al.*, 2008]. Current estimates put the total number of people infected in the world today around 35 million people. Every year an additional 2,8 million people become infected with the virus and roughly 1,8 million HIV/AIDS related deaths occur. The sSA region has been hit the hardest where 28 million are infected with the virus. The Southern African region accounts for the highest HIV prevalence rates on the African continent (roughly one third of the global burden). The epidemic in the sSA region is driven largely by heterosexual transmission of the virus [UNAIDS, 2012].

1.1.2 The origin of HIV

HIV is a member of the lentivirus subfamily of retroviruses (Retroviridae) [Desrosiers *et al.*, 1989]. Recent phylogenetic studies have shown that the virus has been circulating amongst human populations for decades before humans became aware of the virus's existence in the early 1980's [Korber *et al.*, 2000; Lemey *et al.*, 2003; Lemey *et al.*, 2004; Worobey *et al.*, 2008; Wertheim and Worobey, 2009]. In addition to the overwhelming phylogenetic evidence, several possible retrospective cases of HIV before the 1980's have also been identified [Nahmias *et al.*, 1986; Frøslash *et al.*, 1988; Garry *et al.*, 1988].

In 1985, 1172 blood plasma samples, dated between 1950 and 1980 from several African countries, were tested for HIV-1 antibodies. Only one of the plasma samples, dating back to 1959 from a patient from the Central African country of Zaire, now the Democratic Republic of the Congo (DRC), tested strongly positive for HIV-1 antibodies in four different assays [Nahmias *et al.*, 1986]. More recently the research team of Dr Worobey recovered a number of old specimens of HIV from the DRC dating back to 1960 [Worobey *et al.*, 2008]. These samples provide credible evidence that the disease has been in humans for some time and also suggests that the epidemic might have originated in Africa.

The true nature and origin of HIV-1 has been a subject of intense study and debate since the discovery of the virus in 1983. The first major breakthrough in the race to uncover the origin of the virus came in the discovery of similar viruses found amongst non-human primates [Daniel *et al.*, 1985]. These simian viruses were collectively termed simian immunodeficiency viruses

(SIVs) with a suffix to denote the particular species of origin [Sharp and Hahn, 2011]. The most important finding was the discovery that chimpanzees and sooty mangabey monkeys harboured SIV strains, which genetically and antigenically were closely related to HIV-1 and HIV-2 respectively [Hirsch *et al.*, 1989; Huet *et al.*, 1990]. These close relationships between HIV and SIV strains provided the first credible evidence that HIV/AIDS emerged in humans as a consequence of cross-species infections with lentiviruses from different primate species [Sharp *et al.*, 1994]. Subsequent molecular studies of HIV-1 and its three groups (HIV-1 Group M, N and O), with a wide range of SIV sequences, confirmed that HIV-1 arose through zoonotic transmissions of SIV_{cpz} from chimpanzees (*Pan troglodytes*) to humans [Santiago *et al.*, 2002; Worobey *et al.*, 2004; Keele *et al.*, 2006]. Likewise it has been shown that HIV-2 is more closely related to SIV_{smm} sequences commonly found in sooty mangabeys (*Cercocebus atys*) monkeys [Apetrei *et al.*, 2005; Apetrei and Marx, 2005]. In recent years another HIV-1 variant was identified in a Cameroonian migrant who was living in France at the time. Genetic investigation of this new HIV-1 variant revealed a close relationship with SIV_{gor}, a Simian Immunodeficiency Virus commonly found in western lowland Gorillas (*Gorilla gorilla*) from Central Africa [Van Heuverswyn *et al.*, 2007]. This newly discovered variant was designated to a new tentative group, HIV-1 group P [Vallari *et al.*, 2011].

The timing of the zoonotic events that led to the rise of HIV in human populations has been a question for debate. With the use of advanced molecular clock techniques, independent researchers estimated the time of the most recent common ancestor (tMRCA) of HIV-1 group M in Central Africa at 1908 (with a confidence interval between 1884 – 1924) [Korber *et al.*, 2000]. Similar analyses of sequence data from HIV-1 groups N and O have dated the time of origin for these groups at 1963 (1948 – 1977) [Wertheim and Worobey, 2009] and 1920 (1890 – 1940) respectively [Lemey *et al.*, 2004]. Phylogenetic investigation of HIV-2 sequence data has dated the time of origin of HIV-2 groups A and B at 1932 (1906 – 1955) and 1935 (1907 – 1961) respectively [Lemey *et al.*, 2003].

These dating methods have come under serious scrutiny in the past as some analysts believe that viral recombination, which is a common occurrence in the HIV genome, might seriously confound such phylogenetic analysis, but recent work on the subject suggest that recombination (though leading to increased variance) is not likely to systematically bias the results of such analysis [Worobey *et al.*, 2008].

1.1.3 The history of the epidemic in Southern Africa

Southern Africa is in the grip of the most devastating HIV/AIDS epidemic in the world. According to the latest UNAIDS and WHO estimates roughly one-third of the people living with HIV/AIDS in the world resides in the Southern African region [UNAIDS, 2012]. To put things further into perspective 9 out of the top ten countries for sero-prevalence of HIV-1 in the world are Southern African nations (Figure 1.1). Of the estimated 15 million HIV infected individuals in the Southern African region roughly 6 million live in South Africa [UNAIDS, 2012].

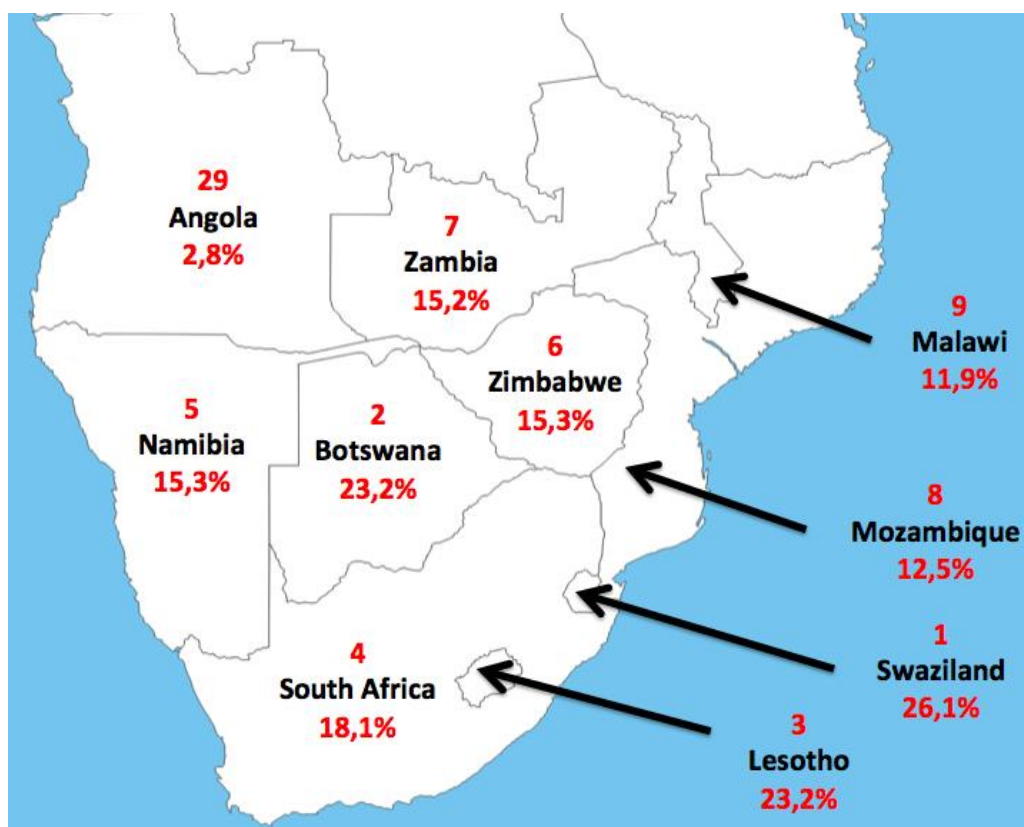


Figure 1.1: HIV prevalence rates across the Southern Africa region. Each of the Southern African nations and its global position for HIV prevalence is indicated on the map. [Figure authors own art work; data used from UNAIDS 2012]

In 1982, the first reported cases of AIDS in the Southern African region were documented in South Africa, amongst two airline workers who worked on regular international flights [Ras *et al.*, 1983]. During December 1984, HIV-1 was first isolated at Tygerberg Academic Hospital from two patients, diagnosed with the clinical condition known as AIDS [Becker *et al.*, 1985]. These early cases were concentrated in the major urban areas of Johannesburg, Cape Town and Durban. The majority of these early cases occurred in high-risk groups, such as men who have sex with men (MSM) [Sher, 1989]. An HIV-1 sero-prevalence study amongst the MSM

community in Johannesburg in the late 1980's indicated that there was a prevalence of 12.6% [Sher, 1989]. These early cases of HIV/AIDS in South Africa were almost exclusively associated with the MSM population, but later AIDS cases were documented amongst haemophiliacs [Sher, 1989].

In 1984, Zambia became the second Southern African nation to report its first confirmed case of AIDS [Hira *et al.*, 1989]. The following year the nations of Malawi, Botswana, and Zimbabwe reported their first cases of HIV/AIDS [Reeve, 1989; Ingstad, 1990]. By 1986, Lesotho, Swaziland, Angola, Namibia, and Mozambique also reported their first cases of AIDS [Vuylsteke *et al.*, 1993; Phits'ane, 1994; Ojo and Delaney, 1997]. The majority of these HIV/AIDS cases in other Southern African countries contrasted sharply with the early documented cases from South Africa as they were diagnosed amongst individuals in the general heterosexual population. Only in 1987 did South Africa report its first documented cases of AIDS amongst heterosexual individuals [Sher, 1989]. These individuals were from rural areas of South Africa working as migrant labourers in the gold mines around Johannesburg at the time, where they came into close contact with foreign migrant labourers from other Southern African countries.

By the end of the 1980's the number of documented cases of AIDS had risen sharply in several Southern African countries. By the beginning of the 1990's, Zambia, Malawi, and Zimbabwe, had reported more than 25 000 diagnosed AIDS cases (Table 1.1) [Wiseman, 1998]. However, when the number of diagnosed AIDS cases is adjusted for population size other countries such as Namibia also showed a surprisingly high infection rate amongst the population [Wiseman, 1998]. This was in the early years of the global HIV pandemic when the number of clinically diagnosed cases was used to assess the magnitude of the epidemic in a given country. Given the general assumption of a 10 year latency period from infection to the development of AIDS like symptoms, two hypothetical conclusions can be made about HIV/AIDS in the Southern African region: (1) that the number of infected individuals were far greater than the total number of patients diagnosed with AIDS by the start of the 1990's and (2) that HIV/AIDS must have been circulating amongst the population of Southern Africa since at least the early 1970's or possibly earlier.

Table 1.1: Total number of documented AIDS cases per country and the estimated number of AIDS cases per million inhabitants for several Southern African countries by the early 1990's.

[Data adopted from Wiseman, 1998]

Country	Diagnosed AIDS cases	Rates / Million
Zambia	29 734	3 457
Namibia	510	3 188
Malawi	31 857	3 185
Zimbabwe	25 332	2 367
Botswana	1 415	1 010
Swaziland	413	516
Lesotho	479	252
South Africa	3 210	82
Angola	608	64
Mozambique	826	54

Rates / Million = Total number of diagnosed AIDS cases per million inhabitants of each country.

By the early 1990's international and national health officials started to estimate sero-prevalence trends of HIV by testing women attending antenatal clinics [García-Calleja *et al.*, 2006]. These prevalence trends are then used to estimate the population level of HIV prevalence for each country or region with the use of complex mathematical models and algorithms such as the ASSA 2003 model from the Actuarial Society of South Africa [ASSA, 2003]. HIV prevalence rate estimates for various Southern African countries are listed in Table 1.2.

Table 1.2: Estimated national population level HIV sero-prevalence trends for several Southern African countries calculated from antenatal clinic data. The data is presented as the total sero-prevalence of HIV in percentage of the total national population [UNAIDS, 2012].

Country Name	1992	1995	1998	2001	2004	2007
Angola	1,0	1,6	1,8	1,9	1,9	1,9
Botswana	7,3	16,6	24,1	26,3	25,8	25,1
Lesotho	3,3	14,3	23,4	24,5	23,8	23,5
Malawi	10,5	13,9	14,7	13,8	12,5	11,4
Mozambique	2,1	4,1	6,7	9,4	11,0	11,4
Namibia	3,0	7,1	12,5	16,1	16,2	14,3
South Africa	1,8	6,1	12,9	17,1	18,1	18,0
Swaziland	4,4	10,6	18,5	23,6	25,5	25,8
Zambia	14,2	15,0	14,7	14,3	14,0	13,7
Zimbabwe	17,2	25,1	26,3	23,7	19,8	16,1

From the epidemiological data available it is clear that the Southern African region has experienced two major growth periods in HIV prevalence. During the 1970's, when HIV/AIDS was still unknown to us, and the 1980's, increases in HIV prevalence rates were the highest in the northern most countries of Southern Africa (e.g. Zambia, Zimbabwe and Malawi). During the 1990's a second wave in the HIV epidemic took place with massive increases in HIV prevalence rates in the southern most countries of the region (e.g. Namibia, Botswana, South Africa,

Lesotho, Swaziland and Mozambique). The HIV epidemic in the northern most countries started to peak during the late 1990's after which HIV prevalence rates started to decline significantly (Table 1.2). This has been most evident in Zimbabwe where HIV prevalence trends declined from its peak at 26,3% in 1998 to 15,3% by 2010 [Halperin *et al.*, 2011]. While HIV prevalence rates in the northern most parts of the Southern African region started to decline, HIV prevalence rates continued to increase throughout the 1990's in the other Southern African nations and only started to stabilize early in the 21st century [Gouws *et al.*, 2008].

With the massive roll out of antiretroviral (ARV) treatment throughout Southern African countries it is expected that more and more infected individuals will continue to live longer. Therefore, in the age of highly active antiretroviral therapy (HAART) HIV prevalence rates will remain very high, as people live longer, and the incidence (number of new infections per year) of HIV may become a more important measure to assess the epidemic.

1.2 HIV transmission and prevalence trends

Since the outbreak of the HIV-1 epidemic in sSA, HIV prevalence rates have increased unevenly across the region. During the 1980's HIV prevalence rates were the highest amongst Central and East African nations. During the 1990's HIV prevalence rates had increased dramatically in the Southern African nations, while HIV prevalence rates had remained fairly stable in other areas of the continent and in some cases even decreased significantly (e.g. in Uganda). This led to major discrepancies in the prevalence of HIV across the sSA region. This large north-south discrepancy in HIV prevalence trends has been a major focal point of the international HIV research community. Several theories have been proposed that could account for the large discrepancies in HIV prevalence in the region as discussed in the following sections [Kreiss *et al.*, 1986; Plummer *et al.*, 1991; Allen *et al.*, 1993; Yamaguchi, *et al.*, 1994; Decosas *et al.*, 1995; Wollants *et al.*, 1995; Janssens *et al.*, 1997; Williams *et al.*, 2000; Glynn *et al.*, 2001; Buvé *et al.*, 2002; Quinn and Overbaugh, 2005; Abu-Raddad *et al.*, 2006; Baggaley *et al.*, 2010; Paxton, 2010; Fenwick, 2012].

1.2.1 HIV transmission risk and male circumcision

One of the major epidemiological focal points has been the large discrepancy in the practice of male circumcision. The relationship between male circumcision and personal risk for HIV has been extensively investigated in the past [Halperin and Bailey, 1999; Drain *et al.*, 2006; Dinh *et al.*, 2011]. The potential benefits of male circumcision in reducing the risk of HIV infection were

raised as early as 1989 [Bongaarts *et al.*, 1989]. Since then this relationship has become one of the most contested fields in the HIV scientific community. Today, it is generally regarded by the majority of the scientific community that uncircumcised men are at a higher risk for HIV infection. This is due to the abundance of Langerhans' cells in the foreskin, which provides a larger cell receptor for HIV entry and infection [Dinh *et al.*, 2011]. The incidence of HIV amongst uncircumcised men is on average 8 times higher than in circumcised men [Williams *et al.*, 2006].

However, given the large controversy surrounding the role of male circumcision in the global HIV pandemic it is still too early to establish what role male circumcision has played in the shaping of the Southern African HIV epidemic. Additionally, the benefits of male circumcision are limited and do not pose any direct benefit to women in sSA, which are the most severely affected group [Baeten *et al.*, 2009].

1.2.2 HIV transmission risk and concurrency

Another major epidemiological focal point has been concurrency between partners in the Southern African region. Several studies have shown that even though the average number of lifetime sexual partners of Southern Africans may be less than for individuals in other parts of the world there is an on-going practice of having multiple sexual partners over an extended period of time amongst these individuals [Morris and Mirjam, 1997; Mah and Helperin, 2010; Beyrer *et al.*, 2010]. This is generally referred to as sexual concurrency and is a common practice amongst Southern African individuals. The cultural practice of having multiple sexual partners, all of whom may also have multiple sexual partners, increases the sexual network within a small community and may therefore play an important role in the spread of any sexually transmitted disease, not only HIV. Although this topic has extensively been researched in the past, recent reviews suggest that the large-scale concurrency throughout Southern Africa cannot alone account for the discrepancies in HIV prevalence trends throughout the sSA region [Lurie and Rosenthal, 2010; Knopf and Morris, 2012].

1.2.3 HIV transmission risk and the effect of migration

One of the biggest epidemiological focal points surrounding the epidemic in Southern Africa has been the effect of migration on the spread of the HIV epidemic in the region. South Africa, and Southern Africa to a lesser extent, has experienced a large degree in seasonal migration of labour compounded by political and economic factors, throughout the 20th century.

The relationship between HIV and migrancy are particularly complex and may be influenced by several factors. Levels of migrant labour are particularly high, but not exclusively confounded, amongst rural men. In a study done amongst young rural individuals it was found that the level of migrant labourers was as high as 60% for men and 30% for rural women [Williams *et al.*, 2000]. However, most men from rural areas tend to migrate over longer distances and for extended periods of time than their female counterparts. This result in large disparities in the gender ratio in areas that experiencing high levels of seasonal migration (in both the inwards and outwards settings). Areas with large disparities in the gender ratio are particularly vulnerable to HIV, when compared with areas experiencing lower levels of migrancy [Williams *et al.*, 2000]. In a study conducted amongst migrant labourers during the 1990's, only a small proportion of men indicated that they had a regular partner that was living in the area they were working in at the time [Williams *et al.*, 2000]. Due to this separation from their regular sexual partners, many of these men often engage in casual sexual relationships or solicit the use of a commercial sex worker (CSW) in the surrounding community [Decosas *et al.*, 1995]. Those migrant labourers who become infected in urban areas may pass the virus on to their partners when they return to their rural place of origin. However, data also suggests that the route of transmission may also occur from the rural to the urban setting. In a seroprevalence study that was conducted amongst discordant migrant couples it was found that 40% of the female companions were infected with the virus but not their male spouses [Lurie, 2000].

The extent of migrant labour in other areas of sSA, outside of the Southern African region, is not well known and less understood. It can be hypothesized that Southern Africa, due to its relative economic prosperity in comparison with the rest of the sSA region may experience higher rates of migrant labour. Therefore migration, compounded by economic and political factors, in Southern Africa may provide a relative good explanation as to how HIV has spread so far and so quickly across the region.

1.2.4 HIV transmission risk and the role of political factors

In recent years another major epidemiological focal point within the HIV research community has been the effect of political factors on the spread of HIV/AIDS in various countries or regions [Mendelson and Carballo, 2001]. Traditionally the consensus of the international HIV research community has been that political instability (e.g. wars, conflict, or civil unrest) could significantly increase the transmission and spread of HIV amongst local communities [Mock *et al.*, 2004]. This view was largely based on factors that could increase personal risk to infection

during such times. For instance, during conflict situations the government may be limited in its ability to provide basic medical services [Mock *et al.*, 2004]. This inability may lead to increases in other sexually transmitted infections, malnutrition and malaria, as well as a decrease in the efficacy of HIV prevention programs, all of which have been linked to increased susceptibility to HIV infection. Additionally, conflict situations may lead to an increase deterioration of family structures and social values [Asenkeyne *et al.*, 2002]. Similarly, conflict may also lead to increases in gender-based violence [Wollants *et al.*, 1995; Mendelson and Carballo, 2001]. These are all factors that may increase personal risk for HIV infection.

However, in recent years several studies have been published which report no significant correlation between combat situations and increases in HIV prevalence amongst the population [Spiegel *et al.*, 2007; Paxton, 2010]. Even more surprising in some of these studies was that the researchers found more significant increases in HIV prevalence trends in politically more stable countries and countries who experience major political change towards a more democratic and representative form of government [Paxton, 2010]. This could potentially prove to be a key factor in the HIV epidemic in the Southern Africa region, as the region experienced the biggest increases in HIV prevalence trends following the political stabilization of the region during the 1980's and early 1990's (e.g. the end of Apartheid, the independence of Namibia, the end of South African military involvement in southern Angola, the end of the Mozambican civil war, and the end of the Rhodesian war of independence).

1.3 The genomic organization of the structural genes of HIV-1

The genome of the human immune virus is approximately 9.2-kilo base pairs (kbp) long, which is flanked by long terminal repeats (LTR's) on both sides [Krebs *et al.*, 2001]. Within the genome, there are several open reading frames, which code for the various proteins of the virus. As with most other retroviruses the genome of HIV-1 encodes for three main types of proteins: *gag*, *pol* and *env* [Gelderblom, 1997]. HIV however does carry a large number of accessory genes that are smaller in size such as *tat*, *rev*, *nef*, *vif*, *vpr* and *vpu* (for HIV-1) or *vpx* (in the case of HIV-2) all of which have different functions [Cullen, 1998]. A diagrammatic representation of the genomic layout of HIV-1 is indicated in Figure 1.2.

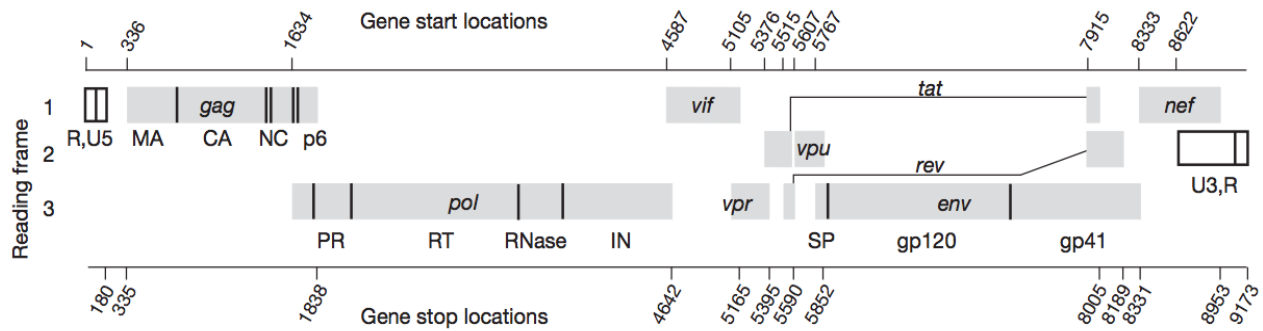


Figure 1.2: A diagrammatical representation of the genome layout of HIV-1. All three reading frames with all the most important genes are shown. All start and stop coordinates of genes on the diagram corresponds to that of the HXB2 reference strain [Adapted from Watts *et al.*, 2009].

The *gag* gene codes for the capsid of the virus [Göttlinger, 2001]. The *gag* gene, which is roughly 1500 base pairs (bp) long, is transcribed in one single fragment, which is then spliced into the various polyproteins [Göttlinger, 2001]. The *gag* p24 part of the gene makes up the viral capsid whereas the *gag* p6 and *gag* p7 parts code for the nucleocapsid and *gag* p17 provides a protective matrix [Henderson *et al.*, 1992].

The *pol* gene is a common feature of retroviruses [Coffin *et al.*, 1997]. As with the *gag* gene, *pol* is transcribed in a single protein, which is then spliced into the four functional polyproteins: *reverse transcriptase*, the *RNase*, the *integrase* and the *protease* [Wlodawer *et al.*, 1989]. The function of the reverse transcriptase gene is to transcribe the viral RNA to double stranded DNA [Kohlstaedt *et al.*, 1993]. The *protease* gene is responsible for the cleaving/splicing of large protein segments of *gag*, *pol*, *env*, and *nef* into the separate functional units [Nicholson *et al.*, 1995]. The *integrase* fragment of the *pol* gene is responsible for the integration of the double stranded viral DNA into the host cells genome [Lodi *et al.*, 1995].

The *env* gene encodes for a precursor protein, gp 160, which is spliced by the host cellular enzymes into the two functional proteins gp 120 and gp 41 [Wyatt *et al.*, 1998; Kim *et al.*, 1998]. Env gp120 is exposed on the surface of the viral envelope and binds the virus to the CD4 receptors on the surface of any target cells [Wyatt *et al.*, 1998]. The glycoprotein gp41 is non-covalently bound to gp120, and facilitates the second step of viral entry into the target cells [Hunter, 1997; Kim *et al.*, 1998]. The gp41 is originally found inside the viral envelope, but when gp120 binds to the CD4 receptor, gp120 undergoes a conformational change causing gp41 to become exposed on the viral envelope, where it can assist in the fusion of the virus with the host cell [Wyatt *et al.*, 1998].

1.4 Genetic diversity of HIV-1

HIV is characterized by a high degree of genetic variation driven by a wide range of factors, such as the lack of a proofreading ability by its reverse transcriptase [Preston *et al.*, 1988], the rapid turnover time of HIV-1 *in vivo* [Perelson *et al.*, 1996], host selective pressures [Rambaut *et al.*, 2004], and recombination events in dually infected patients [Rambaut *et al.*, 2004]. The rate of sequence variation across the genome of HIV varies, with the highest degree of sequence variation in the *env* gene, intermediate amounts in the *gag* and a low degree in the *pol* gene [Shankarappa *et al.*, 1999].

Genetic classification of HIV is based on a phylogenetic system, which means that viral isolates are grouped into a subtype based on their inferred evolutionary relationship [Butler *et al.*, 2007], rather than on other characteristics such as serological reactivity, phenotype, co-receptor usage and many other possible biological characteristics, which are routinely used for the classification of other viruses. This method sets HIV subtype classification apart from other older viral pathogens where serological subtyping is the norm.

HIV-1 group M is the major epidemic strain of HIV today and has spread across the world. HIV-1 group M can be divided into 9 main subtypes (Subtypes A, B, C, D, F, G, H, J, and K). Several of these subtypes can be divided into sub-subtypes such as subtype A (A1, A2, and A3) and F (F1 and F2) [Santos and Soares, 2010]. Viral recombination also forms a major part of the genetics of HIV. To date several circulating recombinant forms (CRF) and unique recombinant forms (URF) have been identified, the most important of which are CRF01_AE and CRF02_AG [Murphy *et al.*, 1993; Carr *et al.*, 1996; Carr *et al.*, 1998].

In some cases, HIV-1 subtypes can be linked to a specific geographical region or epidemiological risk group. These distribution patterns of HIV-1 are either the consequence of accidental trafficking or due to a prevalent route of transmission, which results in a strong advantage for a specific subtype to become dominant within a certain region or country [Santos and Soares, 2010]. Molecular epidemiology studies have shown that, with the exception of the sSA region where most of the existing HIV-1 Group M subtypes and recombinants can be found, there is a specific geographic-demographic distribution pattern (Figure 1.3) for HIV-1 subtypes [Santos and Soares, 2010].

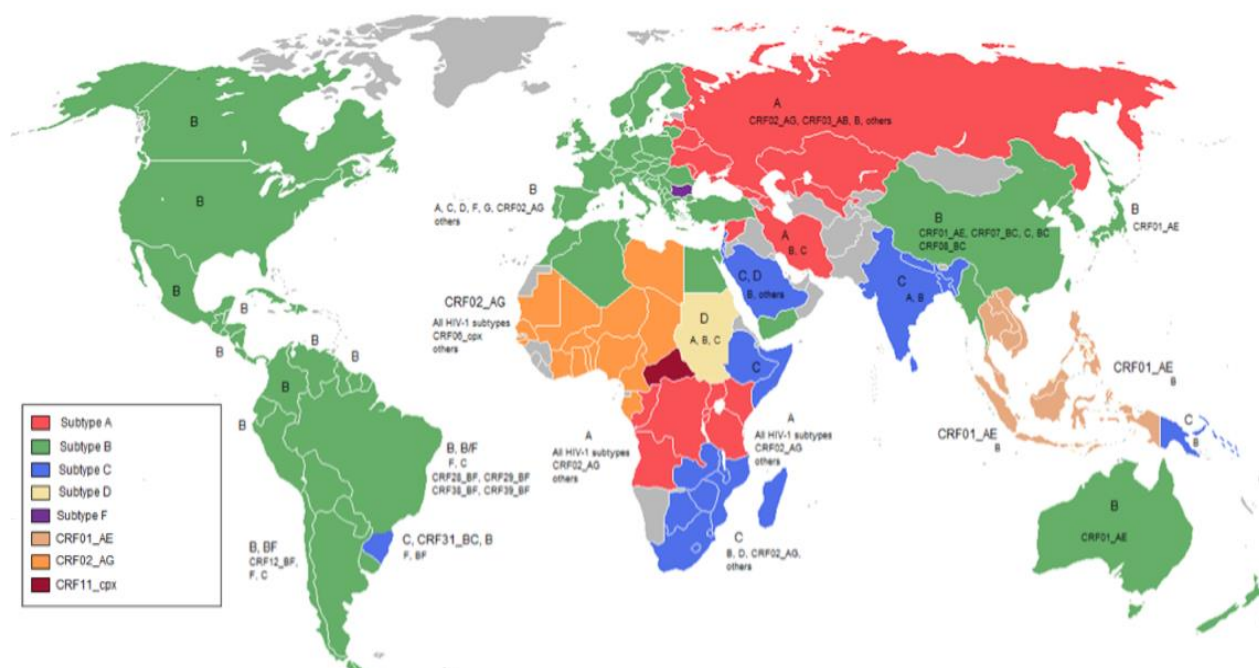


Figure 1.3: Global distribution of different HIV-1 subtypes and circulating recombinant forms. Subtype B dominates the epidemic in the Americas, Australia, Europe, and the Eastern parts of Asia. Subtype C dominates the regional HIV-1 epidemics in Southern Africa, the Horn of Africa, and the Indian sub-continent and the south-eastern parts of Brazil. Subtype A is the most prominent viral form of HIV in the countries of the former Soviet Union and in parts of Central and East Africa. Additionally, the circulating recombinant forms CRF01_AE and CRF02_AG are the most prevalent viral forms in South East Asia and West Africa respectively. [Adopted from Santos and Soares, 2010].

The latest epidemiological data indicates that the most prevalent genetic variants of HIV-1 are subtypes A, B, and C, with HIV-1 subtype C accounting for just over half of the global HIV-1 infection burden [Santos and Soares, 2010].

HIV-1 subtype A for example is the most prevalent form of HIV-1 in areas of Central and Eastern Africa and within member countries of the former Soviet Union [Santos and Soares, 2010]. In the east and central African countries this viral subtype of HIV-1 is predominantly spread via heterosexual contact whereas in the countries of the former Soviet Union the predominant mode of infection is via intravenous drug users [Ingram, 1996; Naganawa *et al.*, 2002]. Areas where HIV-1 subtype A co-circulate with other viral genetic forms of HIV-1 have seen the rise of recombinant viral forms containing fragments of both subtype A and the other subtypes of HIV-1. Examples of these areas of co-circulation are Eastern Europe where AB viral recombinants (CRF03_AB) have emerged [Liitsola *et al.*, 1998; Lukashov *et al.*, 1999].

HIV-1 subtype B is the major genetic clade of HIV-1 circulating in Western Europe, the Americas, Japan, and Australia, but it can also be found in large numbers in countries in

Southeast Asia, Northern African, and the Middle East and amongst the homosexual populations within South Africa and the Russian Republic [Santos and Soares, 2010]. HIV-1 subtype B was the first major subtype of HIV to be isolated and characterized. It was introduced from Central Africa into the Caribbean country of Haiti, where it subsequently spread to the MSM population of the US and the rest of the industrialized world [Gilbert *et al.*, 2007].

HIV-1 subtype C is the predominant form of HIV-1 in the countries of Southern Africa (excluding migrants from other areas of Africa or the homosexual population of South Africa), the Horn of Africa (Ethiopia, Eritrea and Somalia) and the Indian subcontinent [Santos and Soares, 2010]. Smaller subtype C epidemics has also emerged in areas of China and Brazil where this particular viral form is predominantly associated with intravenous drug users [Santos and Soares, 2010].

1.5 HIV-1 subtype diversity in Africa

The African continent is home to nearly two thirds of the people living with HIV/AIDS in the world [UNAIDS, 2012]. The central African region is widely accepted as the place of origin of HIV-1 [Vidal *et al.*, 2000]. A large amount of genetic variation has occurred within the HIV-1 genome since the emergence of the virus in the region and a multitude of subtypes and circulating recombinant forms can be found within the region [Rambaut *et al.*, 2004]. This degree of genetic diversity can be attributed to the founding of small isolated regional epidemics, which diverged over the course of several years in isolation. The distribution and occurrence of different subtypes within the African population are not linked to particular lifestyle habits, as they would be in other parts of the world. However, certain geographical regions within the African continent are home to particular subtypes, which is possibly due to founder effects.

In the central part of Africa a large degree of HIV-1 genetic variation can be found. This is largely due to the fact that the central African region is the place of origin of the global HIV-1 pandemic [Vidal *et al.*, 2000]. In the eastern parts of the African continent, HIV-1 subtype A1 and D are the most prevalent forms of HIV circulating amongst the infected population, though HIV-1 subtype C has in recent years become increasingly important [Torques *et al.*, 1999; Tapia *et al.*, 2003; Lwembe *et al.*, 2009; Nofemela *et al.*, 2011].

Recombinant form CRF02_AG is the most important genetic variant of HIV circulating in the west-central region of the African continent in countries such as Cameroon, Gabon, the Congo and Nigeria. Since the emergence of CRF02_AG this viral form has become important in the

global epidemiology of the virus and today accounts for up to 4.6% of global infections. Initially the recombinant form CRF02_AG was described as a divergent lineage within HIV-1 subtype A1 (based on partial *gag* and *env* sequence data), but after full genome analysis of these isolates was obtained it was recognized as a complex mosaic virus of alternating subtype A1 and G fragments [Imamichi *et al.*, 2009; Ajoge *et al.*, 2011]. CRF02_AG has also been identified in North African countries (Libya), which can be attributed to the increased movement of people from central-west Africa, via Northern Africa to Europe [Myriam *et al.*, 2001; de Oliveira *et al.*, 2006].

In other Northern African countries (Egypt, Algeria, Tunisia, and Morocco) HIV-1 subtype B is the most prevalent form of HIV circulating in the infected population [Myriam *et al.*, 2001].

The HIV-1 epidemic in Southern Africa is largely dominated by HIV-1 subtype C where it accounts for more than 90% of all the infections in the region. In the largest country in the region, South Africa, there are two very distinct epidemics, which only seldom intermingle. HIV-1 subtype B and to a lesser extent subtype D viral forms are largely associated with infections amongst the homosexual epidemic in South Africa [Engelbrecht *et al.*, 1994; Becker *et al.*, 1995; Loxton *et al.*, 2005]. HIV-1 subtype C however, is overwhelmingly associated with the heterosexual epidemic in South Africa, where it accounts for nearly 95% of all infections [Williamson *et al.*, 1995; van Harmelen *et al.*, 1997]. In recent years several papers have also been published on other viral forms of HIV-1 such as: HIV-1 subtypes A1, F1, G, K, some circulating recombinant viral forms (CRF02_AG) and several unique recombinant forms [Engelbrecht *et al.*, 1999; van Harmelen *et al.*, 1999; Hunt *et al.*, 2001; Engelbrecht *et al.*, 2001; Bredell *et al.*, 2002; Scriba *et al.*, 2001; Gordon *et al.*, 2003; Bessong *et al.*, 2005; Jacobs *et al.*, 2006; Bredell *et al.*, 2007; Jacobs *et al.*, 2008; Huang *et al.*, 2009; Jacobs *et al.*, 2009; Wilkinson and Engelbrecht, 2009; Papathanasopoulos *et al.*, 2010; Fish *et al.*, 2010; Iweriebor *et al.*, 2011; Wilkinson *et al.*, 2013 in press]. However, these rare viral forms of HIV-1 contribute less than 1% of the epidemic in South Africa.

HIV-1 subtype C is also responsible for the overwhelming majority of infections in other Southern African countries. In Botswana, Zimbabwe, Malawi, and Swaziland HIV-1 subtype C accounts for over 99,0% of all viral genotyped isolates [Papathanasopoulos *et al.*, 2003; Lihana *et al.*, 2012]. Even in Zambia, the Southern African country with the largest degree of genetic diversity, HIV-1 subtype C accounts for over 90% of all infections [Papathanasopoulos *et al.*, 2003; Lihana *et al.*, 2012]. The rest of the non-subtype C isolates in Southern African countries is largely made up of HIV-1 subtype A, D or complex viral recombinant forms [Novitsky *et al.*,

2000; McCormack *et al.*, 2002; Novitsky *et al.*, 2007; Gnanakaran *et al.*, 2010; Campbell *et al.*, 2011].

1.6 Phylogenetic analysis and HIV

HIV was discovered when modern molecular biology and phylogenetic methods became widely used. Therefore, the advances in molecular biology such as, DNA amplification and sequencing, as well as advances in computer technology and evolutionary biology, have revolutionized HIV based research. Since a large variety of different phylogenetic methods are widely used throughout the course of this study it is of importance to briefly introduce some of the basic concepts of modern phylogenetic practices.

1.6.1 An introduction to phylogenetics

Phylogenetics forms a small part of modern evolutionary biology. Traditionally, the evolutionary relationships between taxa or species were inferred from phenotypic differences or similarities, since the days of Charles Darwin. In the early days these trees were drawn by hand and the branching order between the different taxa was based on observed phenotypic differences or similarities. In the late 1950's and early 1960's two critical technological advances gave a new impetus to modern phylogenetics. These were the advancements in molecular biology (nucleic-acid and amino acid sequence composition) and the development of large centralized computers, which were powerful enough to handle complex computations. With the genetic information and computational power now readily available, scientists set out to develop algorithmic means of analysing the genetic data to infer evolutionary relationships.

The first major breakthrough came with the development of parsimony methods of inferring evolutionary relationships in the early 1960's [Edwards and Cavalli-Sforza, 1963]. This method is rooted in the assumption that the evolutionary tree that requires the least number of changes to explain the current set of data would be the best possible tree topology (most parsimonious). Since the development of parsimony methods, several algorithmic processes have been developed to infer evolutionary trees. These include the edition of the unweighted pair group method with arithmetic means or UPGMA [Sokal and Michener, 1958; Murtagh, 1984], the Maximum Likelihood method [Edwards and Cavalli-Sforza, 1964], the Fitch-Margoliash method [Fitch and Margoliash, 1967], the Neighbor-Joining method [Saitou and Nei, 1987], the Minimum Evolution method [Rzhetsky and Nei, 1992; Rzhetsky and Nei, 1993], and lastly the Bayesian method [Rannala and Yang, 1996; Yang and Rannala, 1997; Mau and Newton, 1997; Li

et al., 2000] of tree inference. These various techniques can broadly be divided into two main categories (Table 1.3) based on the kind of data they use to infer tree topologies: distance based methods and character based methods [Hall, 2008; Salemi and Vandamme 2003].

Table 1.3: Summary of the various methods of tree construction.

Method	Optimality search criterion	Clustering algorithm
Distance based methods	Fitch-Margoliash	UPGMA
	Minimum Evolution	Neighbor-Joining
Character based methods	Maximum Parsimony	
	Maximum Likelihood	
	Bayesian Inference	

In the following section, some of the most widely used methods of tree inference will be discussed. However, the construction of a phylogenetic tree is in some cases the final product of any phylogenetic investigation. Several important steps precede the inference of a phylogenetic tree topology (Figure 1.4).

One of the first steps in any phylogenetic investigation is the retrieval of relevant genetic information to compare against newly sequenced information. The next step involves the aligning of the different sequences with one another in order to obtain position homology. Thirdly some assumption about the evolutionary process needs to be made. For this an appropriate model of nucleotide substitution needs to be selected. Finally the sequence alignment and the inferred model of substitution can be used to infer an evolutionary relationship [Baldauf, 2003]. A schematic breakdown of the basic steps involved in any phylogenetic investigation is presented in Figure 1.4. These basic steps (steps 1 to 4) in any phylogenetic inference will be introduced briefly in the following section (section 1.6.2 – 1.6.4).

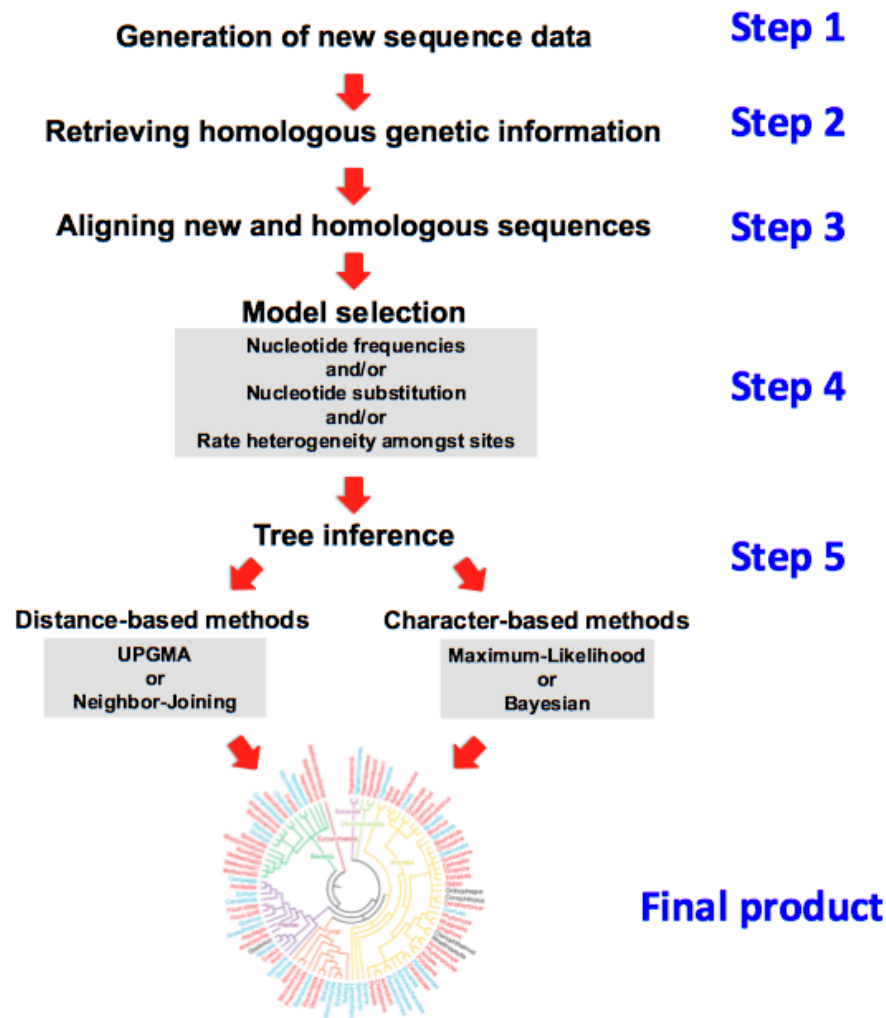


Figure 1.4: A breakdown of the basic steps involved in any phylogenetic investigation. The diagram illustrates the basic steps involved in the generation of a phylogenetic tree [Authors own artwork].

1.6.2 Retrieving of relevant genetic information

The goal of any phylogenetic analysis is to establish the evolutionary relationship between newly sequenced data and other known sequences. One of the first steps in any phylogenetic analysis is to obtain reference sequences to compare to newly sequenced data in order to establish the evolutionary relationship of the newly sequenced information.

The aim is therefore to obtain enough genetic information that shares a close genetic relationship with the new sequence(s) of interest. However, it is important to understand the difference between homology and similarity where sequence information is concerned. Genetic similarity merely reflects the proportion of sites over the length of sequences that are identical [Koonin, 2005]. Homology on the other hand implies that two taxa or sequences are descended from a

common ancestor and thus will imply that in a sequence alignment identical residues at a site are identical by descent [Koonin, 2005].

The easiest method to obtain homologous sequence information is to use the Basic Local Alignment Search Tool or BLAST method [Altschul *et al.*, 1990]. BLAST uses the input sequence as a query to search databases for any protein or nucleic acid sequence that share similarity. After the search is complete the program will produce a list of sequences that it found to be similar to the query sequence. The BLAST program also produces an *E* value for every “hit”, which indicates the level of confidence in that particular result. If a sequence *E* value is below 0.1, one can assume with high confidence that the sequence will be a homologue to your query sequence [Altschul *et al.*, 1990].

Currently, the majority of genetic information is stored in online sequence databases, either in a nucleic acid or amino acid format. There are a large number of sequence databases in existence, the most important of which are: GenBank (at NCBI), EMBL (European Molecular Biology Laboratory), and the DDBJ (DNA Data Bank of Japan) [Learn *et al.*, 1996]. Other specialized databases exist for sequences only associated with a single organism or research topic such as, the HCV database (containing Human Hepatitis C Viral sequences), and the HIV database (containing sequences of HIV and other related lentiviruses) [Rodrigo and Learn, 2001]. Most of these sequence specific online databases allow users to search the database for homologous sequences directly through the BLAST application.

1.6.3 Sequence alignments

In bioinformatics, a multiple sequence alignment is a method of arranging the different sequences of nucleic or amino acids to identify regions of similarity and form the basis of all phylogenetic analysis. Apart from its wide use in modern evolutionary biology, it is also widely used in functional and structural evaluations of protein sequences. Aligned sequences of nucleotides or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues in order to obtain position homology. Operating under the assumption that two sequences in an alignment share a common ancestor, one can interpret mismatches within the alignment as point mutations and gaps as indels (indels can be defined as insertions or deletions) which were introduced in one or both of the taxa in the time since they diverged [Abecasis *et al.*, 2007]. Sequences of a few nucleotides long that share a high degree of sequence similarity can be aligned easily by hand. Most sequences alignments however, require the alignment of large numbers of lengthy, and sometime highly variable, sequences that cannot be aligned solely by

human effort. In the modern digital age, algorithms are used for the construction of sequence alignments. Even with the development of several alignment algorithms, the quality of most of these alignments is still very poor and they require manual editing (with special alignment editing tools) in order to obtain accurate codon alignments.

Pairwise sequence alignment methods are commonly employed to find the best matching alignment of two query sequences and can therefore only be used between two sequences at a time [Abecasis *et al*, 2007]. They are however extremely easy to calculate and are therefore often used for methods that do not require extreme precision.

Multiple sequence alignment is an extension of pairwise alignment to accommodate more than two sequences at a time [Salemi and Vandamme, 2003]. This method is often used for the identification of conserved sequence regions across a group of sequences, which are related back in time (share a common ancestor). Multiple sequence alignments also form the backbone of modern phylogenetic analysis since they are used for the construction of phylogenies [Salemi and Vandamme, 2003]. The most commonly used method for the construction of multiple sequence alignment is the progressive method (also called the tree method of alignment) in which the program first draws a “guide tree” and then aligns sequences according to the tree topology [Salemi and Vandamme, 2003]. Taxa that appear within the tree to be most closely related are first aligned with one another, then successively less related sequences are added to the alignment until the entire set of sequences has been resolved [Salemi and Vandamme, 2003].

1.6.4 Nucleotide substitution models

Phylogenetic analysis makes certain assumptions about the process and rate of DNA substitutions or amino acid replacements in the model of evolution they employ. Point mutations can either be due to transversions (when a purine base is replaced by a pyrimidine base) or due to transitions (the replacement of a purine or pyrimidine base with another purine or pyrimidine respectively). Due to the chemical similarity between purine bases (Adenine or Guanine) or pyrimidine bases (Cytosine or Thymine) transitions (Ts) are more common than transversions (Tv), which would alter the chemical composition of the DNA molecule [Graur and Li, 1997; Salemi and Vandamme, 2003]. To study the dynamics of these changes in sequences, one needs to use mathematical algorithms that take into account different rates of nucleotide substitution (e.g. to allow for transitions to occur more often than transversions). To date a large number of these models have been developed, all of which allow for different assumptions and conditionalities.

The first model of nucleotide substitution developed was the Jukes and Cantor method (JC69) in 1969 [Jukes and Cantor, 1969]. This model operates under the assumption that the equilibrium base frequencies of the four nucleotides are 25% for each nucleotide ($\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1/4$). It also assumes that any nucleotide has the same probability to be replaced by any of the other three nucleotides. This means that the only variable is the overall substitution rate or μ . By taking these considerations into account one can see that, although the process can be easily mathematically applied, there are some shortcomings to this model of nucleotide substitution. Since the development of the JC69 model in the 1960's, several extensions and improvements have been made, that can allow for unequal base frequencies or allow for different rates of transitions and transversions. A full diagrammatical representation of the most important nucleotide substitution models is presented in Figure 1.5.

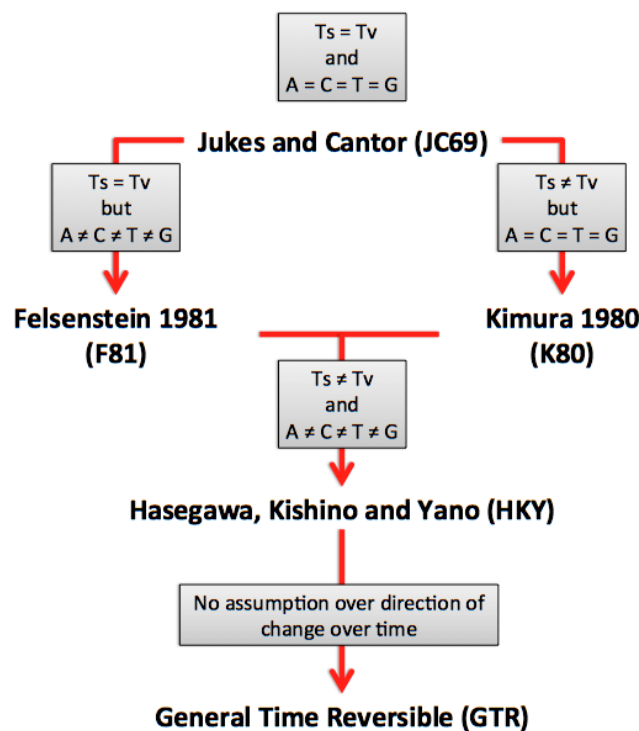


Figure 1.5: A diagrammatical breakdown of the most important nucleotide substitution models currently available. Each of the models is basically an extension of the JC69 model. The K80 model [Kimura, 1980] allows for different Transition to Transversion (Ts/Tv) ratios, but keeps the base frequencies equal, while the F81 model [Felsenstein, 1981] allows for the base frequencies to vary but keeps the Ts/Tv ratio equal. The HKY model [Hasegawa *et al*, 1985] is basically a combination of the K80 and F81 model in that it allows for unequal base frequencies and Ts/Tv ratio. The GTR model [Rodriguez *et al*, 1990; Yang *et al*, 1994] is an extension of the HKY model, but allows each of the 6 parameters to have its own probability while not assuming any direction in the change over time. [Authors own artwork].

Besides the use of a specific model of nucleotide substitution in evolutionary analysis, one also needs to account for variable substitution rates across sites. All of the model(s) that were discussed in the preceding section work under the assumption that different sites in a sequence evolve in the same way and at the same rate. Such an assumption however, may be unrealistic as some areas of a coding region may be more conserved due to their importance in determining the secondary structure of proteins. One can account for such rate variations by assuming that the rate for any site is a random variable that can be calculated from a statistical distribution.

The most commonly used distribution to accommodate for rate heterogeneity amongst sites today is the gamma distribution (Γ). A gamma distribution of 1 across sites for instance will mean that all site across the length of the alignment evolve at the same constant rate, while a gamma distribution closer to 0 ($G < 1$) will mean that different parts across the sequence length evolve at much different rates.

1.6.5 Phylogenetic inference

In a previous section (section 1.6.1) a brief introduction were given to modern phylogenetic reconstruction. In the following section the various algorithmic methods of tree inference will be discussed in more detail. However, it is of some importance to briefly review the most basic components of any phylogenetic tree.

In any phylogenetic tree, two closely related taxa (taxa are also sometime referred to as operational taxonomic units or OTU's) are connected with one another by branches. These two branches connect with one another in an internal node, which represents the hypothetical common ancestor of these two sequences. This grouping of taxa based on similarities can be expanded to include larger number of taxa. If a group of taxa are closely grouped together within a tree then they form a monophyletic group of taxa or cluster. Most phylogenies are rooted. This root represents the hypothetical common ancestor of all of the taxa within the tree topologies. If the common ancestor of the group of taxa is not known then no common ancestor can be included in the data set and the tree topology is unrooted. A diagrammatical breakdown of a basic phylogenetic tree is presented in Figure 1.6.

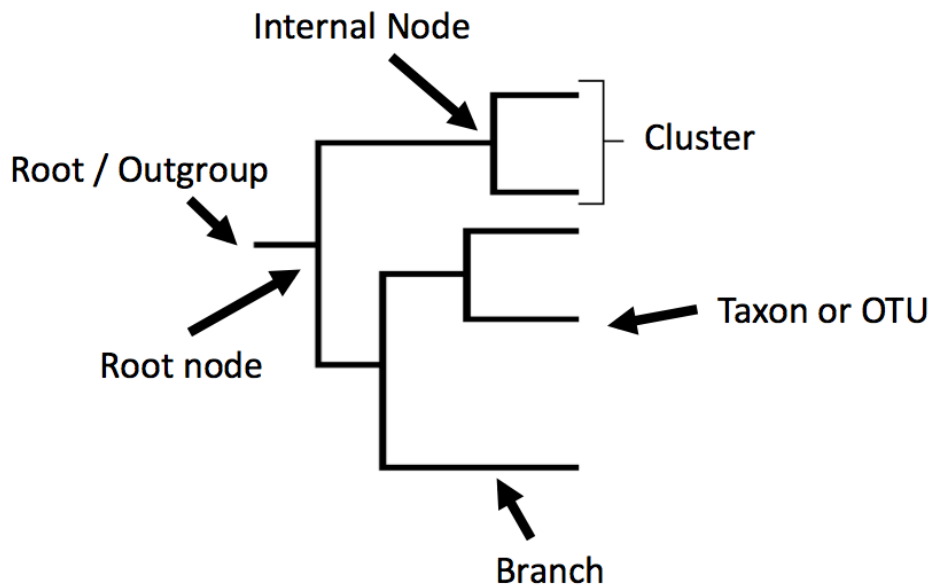


Figure 1.6: Diagrammatical representation of a phylogenetic tree. On the diagram all the important features of a phylogenetic tree has been depicted including: a branch, a taxon or OTU, the root, a clade and an internal node [Authors own artwork].

1.6.5.1 Distance based methods of tree inference

A major branch of phylogenetic methods has been the distance-based methods of tree construction. This method of tree inference was introduced by Cavalli-Sforza and Edwards [Edwards and Cavalli-Sforza, 1963; Edwards and Cavalli-Sforza, 1964] and by Fitch and Margoliash [Fitch and Margoliash, 1967] in the 1960's. The general idea is to calculate a measure of the distance between each pair of species or taxa within the dataset and then find a tree that predicts the observed set of distances as closely as possible. This leaves out all information from higher order combinations of character states and reduces the data to a matrix of numerical values. These distances are calculated using a distance matrix in which all sequences are plotted against one another. An example of a distance matrix is presented in Table 1.4.

The most simplistic way to interpret distance-based methods is to look at the calculated distance between two taxa as an estimation of the branch length between the two taxa with in the phylogeny. Several methods of distance based methods to infer phylogenies have been developed to date such as: the Fitch-Margoliash method [Fitch and Margoliash, 1967], the unweighted pair group method with arithmetic means (UPGMA) [Sokal and Michener, 1958], the Minimum Evolution (ME) method [Rzhetsky and Nei, 1992], and the Neighbour Joining (NJ) method of tree construction [Saitou and Nei, 1987].

Table 1.4: A representation of a distance matrix calculated for a data set of eight taxa or sequences. [Authors own graphic work]

Distance / Q Matrix	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5	Sequence 6	Sequence 7	Sequence 8
Sequence 1	-	2.6	3.7	5.1	4.7	9.5	1.4	7.3
Sequence 2	2.6	-	1.5	2.7	2.1	4.3	6.4	3.6
Sequence 3	3.7	1.5	-	1.9	9.7	6.9	5.7	6.2
Sequence 4	5.1	2.7	1.9	-	0.4	8.4	0.8	8.1
Sequence 5	4.7	2.1	9.7	0.4	-	7.4	4.8	5.4
Sequence 6	9.5	4.3	6.9	8.4	7.4	-	0.6	3.6
Sequence 7	1.4	6.4	5.7	0.8	4.8	0.6	-	7.4
Sequence 8	7.3	3.6	6.2	8.1	5.4	3.6	7.4	-

One limitation of this method of tree construction is that it may leave out small details of the dataset and thus could not possibly provide an accurate estimation of the phylogeny [Holder and Lewis, 2003]. Recent computer analyses on these objections have shown that the amount of information about the phylogeny that is lost in the process is remarkably small and that the estimates of the phylogeny are quite accurate [Holder and Lewis, 2003]. These methods of tree inference are particularly good for very large phylogenies as most character-based methods (e.g. Maximum Likelihood or Bayesian) are still restricted by computer power.

1.6.5.1.1 UPGMA

The unweighted pair group method with arithmetic means or UPGMA is the most simplistic method of tree construction and was first introduced by Sokal and Michener in 1958 [Sokal and Michener, 1958]. Their method was largely used for the construction of taxonomic phenograms (trees that reflect the phenotypic similarities between OTU's). However this was later adapted for the construction of phylogenetic trees from sequence data. The implementation of the UPGMA to sequence data is only possible if the rate of evolution is approximately constant amongst the different lineages in the data set [Murtagh, 1984].

The method employs a sequential clustering algorithm, in which local topological relationships are identified by similarity scores, and then the tree is constructed in a stepwise manner. The two OTU's within the dataset with the smallest observed distance in the matrix are grouped together and treated in subsequent rounds of addition as a single OTU. This process of stepwise addition

is continued until only two unique taxonomic units (UTU) remain, and these are then joined together.

The UPGMA method produces a rooted tree topology and assumes a constant rate of evolution (molecular clock hypothesized) and therefore it is generally not regarded as a good method to infer phylogenetic relationships. The UPGMA method of tree inference is also quite old, and many better inference algorithm methods have been developed to date, which reconstruct phylogenetic relationships far more accurately.

1.6.5.1.2 Fitch-Margoliash method

The Fitch-Margoliash method of tree inference can broadly be viewed as an extension of the UPGMA method of tree inference and was introduced by Fitch and Margoliash in 1967 [Fitch and Margoliash, 1967].

As with UPGMA this method of tree inference relies on the distance matrix (calculation of genetic distances) for the grouping of taxa. The algorithm also uses a sequential clustering technique (as with UPGMA) to group all of the taxa into clusters until the entire tree topology has been resolved. However, this method does not rely on the assumption of a molecular clock and produces an unrooted tree topology. As with the UPGMA method of tree inference this method is also quite old and today there are far more robust algorithmic methods of inferring evolutionary relationships.

1.6.5.1.3 Neighbor-Joining

Briefly, the Neighbor-Joining (NJ) method, as a distance based method of tree inference, uses a distance matrix (as presented in Table 1.4), which specifies the genetic distance between each pair of taxa in the given data set [Gascuel and Steel, 2006]. The algorithm starts with a completely unresolved tree (tree topology corresponds to a random star phylogeny) and repeats the following steps until all the taxa and branch lengths have been resolved:

1. The algorithm calculates the distance matrix for the entire data set.
2. The two taxa with the smallest distance in the matrix are paired together.
3. A node is introduced in the tree topology that joins these two taxa together.
4. The algorithm calculates the distance from the new node to each of the two taxa.
5. The algorithm then calculates the distance from the new node to each of the other taxa in the data set.

6. The algorithm starts from the beginning by calculating a new distance matrix, while treating the two taxa that were joined in the preceding section as a single OUT starting from their internal node.

Unlike the UPGMA method of tree inference, Neighbour Joining does not assume the implication of a molecular clock and the algorithm produces an unrooted tree. Rooted trees can be created with Neighbour Joining methods by using an appropriate outgroup.

The major advantage of the NJ-method is that it performs extremely fast in comparison to other methods of tree inference. The major disadvantage of the method is that some information may be lost by the compression of sequence data into distances and that reliable estimation of pairwise distances could be difficult to calculate for extremely divergent taxa [Gascuel and Steel, 2006].

The NJ-method of tree inference is often just treated as a simplistic starting point for an intensive search for the best phylogeny. To perform such a search a standard must be used which is called an optimality criterion. The most popular optimality criteria are the parsimony, minimum evolution and the maximum likelihood methods of tree inference [Holder and Lewis, 2003].

1.6.5.1.4 Minimum-Evolution

Rzhetsky and Nei first proposed the Minimum Evolution method of tree construction in the early 1990's [Rzhetsky and Nei, 1992; Rzhetsky and Nei, 1993]. ME method of tree inference can be viewed as an extension of the NJ-method of tree inference. As with the NJ-method, ME-tree inference requires the simplification of genetic data into a distance matrix. As with NJ the ME algorithm repeats over the following steps to find the best ME-tree topology:

1. Construct a standard NJ-tree topology.
2. Compute the total sum (S) of all the branch lengths for the NJ-tree.
3. Then all theoretically possible tree topologies, which could be an infinite number (depending on the number of taxa), that are closest to the NJ-tree are examined under certain criteria.
4. The S values for each of these trees are then computed.
5. Each of the computed S values are then compared with one another and the topology with the smallest S value will be chosen as the final "best" tree.

A major advantage of the ME-method of tree inference is that computational analyses have

shown that the ME-method is more efficient than most other distance based methods of tree construction. However, since the method relies on an exhaustive search for the best tree topology this method can be extremely computationally intensive, particularly for large data sets. This can be overcome with the use of heuristic searches, such as Nearest Neighbor Interchange (NNI) or Subtree Pruning and Regrafting (SPR) method (section 1.6.5.3), which improves the speed of tree inference for basic ME-methods.

1.6.5.2 Character based methods of tree inference

Character based methods of tree construction operate by evaluating candidate phylogenetic trees according to an explicit optimality criterion [Holder and Lewis, 2003]. The tree with the most favourable score is, regarded as the best estimation of the phylogenetic relationship of the taxa within the data set. Several character-based methods of inference have been developed to data such as: maximum parsimony, maximum likelihood, and Bayesian inference [Holder and Lewis, 2003].

1.6.5.2.1 Maximum-Parsimony

The parsimony method of tree construction is the most simplistic character based method of tree inference there is and can be easily performed in a reasonable amount of time. As was briefly introduced earlier in section 1.6.1. This method of tree inference represents one of the first tree inference methods and was introduced by Edwards and Cavalli-Sforza in the early 1960's [Edwards and Cavalli-Sforza, 1963].

Maximum parsimony (MP) aims to find the tree topology for any given data set (sequence alignment) that can be explained with the smallest number of character changes (substitutions). For a particular tree topology the MP-algorithm infers for each sequence position the minimum number of character changes required along its branches to explain the observed character states. The sum of the scores at all positions is called the parsimony length of a tree and this is computed for different tree topologies. After a reasonable number of tree topologies have been evaluated, the tree that requires the minimum number of changes is selected as the most optimal maximum parsimony tree.

It is relatively easy to find the most parsimonious tree if the data set is small, but when data sets become larger, the time of computation becomes a problem when trying to identify the most parsimonious tree. Consequently, a number of heuristic search methods for optimization have

been developed in the past to locate a highly parsimonious tree within such a large dataset. Most of these heuristic methods involve some of the mechanisms in phylogenetics, which operates on a tree rearrangement bases (section 1.6.5.3).

There are some advantages of maximum parsimony methods of inference. The method is based on shared and derived characters. It is therefore a cladistics rather than a phonetic method of tree construction. It does not reduce sequence information to a single numerical value like many of the distance-based methods and the method also tries to provide information on the ancestral sequences. Several disadvantages of maximum parsimony methods exist as well. This method is slow in comparison with distance-based methods of tree construction. The method does not utilize all the sequence information and only use the informative sites in the alignment. The method does not allow for multiple mutations and does not provide information on the branch lengths.

1.6.5.2.2 Maximum-likelihood

Maximum likelihood (ML) method of tree construction is broadly very similar to maximum parsimony in that the algorithm examines different tree topologies and evaluates the relative support for each tree. Unlike parsimony methods, maximum likelihood evaluates the support for each tree by summing over all the positions along the sequence length and not just the informative sites along the sequence length [Hall, 2008; Salemi and Vandamme 2003].

The ML algorithm search for the tree that maximizes the probability (likelihood) of observing the character states in the given data set, given a particular tree topology and an inferred model of evolution (nucleotide substitution model). Numerical optimization techniques are used to find the combination of evolutionary parameters and branch lengths that maximizes the likelihood. Depending on the search algorithm used, the likelihoods of a large number of trees topologies are searched with the method. These search algorithms include branch optimization techniques (described in detail in section 1.6.5.3) such as the nearest neighbor interchange (NNI) or the subtree pruning and regrafting (SPR) methods. The tree that yields the highest likelihood at the end is then chosen as the best possible tree to explain the given data set [Hall, 2008; Salemi and Vandamme 2003].

Unfortunately a major disadvantage to the ML-method of tree inference is that it can be extremely computationally intensive. However, the many advantages of ML-tree inference out

ways the computational intensive nature of this search algorithm. Firstly, the ML-method of tree inference relies on maximum likelihood scores, which allows for users to test trees or hypotheses and does not rely on additional statistical analysis such as Bootstrap sampling. The ML-method also utilizes all the sequence information in the data set and not only informative sites. It is also robust enough for the analysis of very short sequences [Holder and Lewis, 2003].

1.6.5.2.3 Bayesian methods of tree inference

Bayesian inference is a character-state method of tree inference that employs an optimality criterion. It was introduced into the field of molecular phylogenetics in the late 1990's [Rannala and Yang, 1996; Yang and Rannala, 1997; Mau and Newton, 1997; Li *et al.*, 2000] and represents the latest breakthrough in evolutionary biology. It is theoretically very different from other character based methods of inference such as maximum parsimony and likelihood methods, in that this method does not attempt to search only for one single best tree [Suchard *et al.*, 2001]. As with maximum likelihood, Bayesian inference also employs the concept of likelihood, but as it targets a probability distribution of trees rather than a single best tree it searches for a set of plausible trees or hypotheses for the given data set. This posterior distribution of plausible trees inherently holds a confidence estimate of any evolutionary relationship. Therefore phylogenetic statistical confidence such as inferred by bootstrapping is not necessary in Bayesian tree inference [Suchard *et al.*, 2001].

Bayesian inference requires the use of a prior, as specified by the researcher or the investigator, which is formalized as a prior distribution on the model parameters (e.g. branch lengths, tree topology, and substitution model parameters). If no biological information is available then the prior belief is preferably vague or uninformative. If biological information (e.g. mutation rate or predominant mode of nucleotide substitution) about the gene or species under study is known, then one can set a very informative prior on the analysis [Suchard *et al.*, 2001]. In Bayesian inference a uniform prior on tree topology is the foremost objective.

Posterior probabilities of trees are obtained by exploring the tree space using a sampling technique, called the Markov Chain Monte Carlo (MCMC) method [Yang, 2008]. This Bayesian MCMC algorithm can broadly be described as follows. The algorithm starts with a totally random tree, with random branch lengths and random substitution parameters, at a random spot in the tree space. Then in each of the steps in the chain the tree is rearranged, as follows:

1. The algorithm proposes a change in the tree, by using standard tree rearrangement techniques such as NNI or SPR.
2. The branch lengths of the topology are changed.
3. The algorithm proposes a change to the parameters.
4. The algorithm then calculates the likelihood and prior ratio for the new topology.
5. If the product of the likelihood and the prior ratio is better in the new state than the old one, then the new state (tree) is accepted and the old one discarded.

Following a specified number of iterations in the chain length a sample of the tree, branch lengths and parameters is taken and saved. This process of continuous tree rearrangement and sampling every few steps in the chain is continued until the end of the chain length. At the end of the chain the sampled data is summarized into a posterior distribution of trees/parameters.

The major disadvantage of Bayesian methods of tree inference is that it can be extremely computationally intensive. The major advantage of Bayesian inference is that from a single MCMC run, support values for each of the clusters in a tree can be derived from the data generated, and thus Bayesian inference provides a natural way of taking phylogenetic uncertainty into account [Hall, 2008; Lemey *et al.*, 2010].

1.6.5.3 The problem of finding the best tree topology

The problem with many of the tree inference techniques is that there is no build-in method of assessing the clustering in a tree topology. Since most character based methods such as Maximum-Likelihood and Bayesian tree construction inherently rely on likelihood scores and a posterior distribution of possible trees or parameters, this problem is largely restricted to the distance methods of tree inference (e.g. NJ- or ME-tree inference). However, additional methods have been developed to statistically assess the confidence of internal nodes in these types of tree topologies.

The most widely used method today is the Bootstrapping method of confidence testing. Bootstrap resampling as a statistical tool was invented in the late 1970's by Bradley Efron [Efron, 1979] and was introduced into the field of molecular phylogenetics by Joseph Felsenstein in the mid 1980's [Felsenstein, 1985]. Briefly, bootstrapping in modern molecular phylogenetics entails continuous resampling of taxa, over a user specified number of iterations. Following the resampling statistical confidence for branches are obtained by a single value. Therefore, a bootstrap value of 70 for a branch indicates that in 70% of the resampled cases, the taxa that are

joined by the internal node of that branch clustered together. Bootstrapping does not resolve the question of whether the tree topology that was obtained is the best possible fit for the given data set. It only provides a degree of confidence estimation for the internal branching order of the topology.

Another major problem with tree inference is to find the best tree topology that represents the given data set the best. If you were to search the entire tree space (all possible tree) you would obviously find the best possible tree. However the total number of possible trees becomes very large, even with a small number of taxa. A data set of 50 taxa contains roughly $2,75 \times 10^{76}$ possible tree topologies. Therefore, to conduct an exhaustive search through the entire tree space is usually impossible due to the obvious time constraints. Heuristic search methods have been developed in order to overcome this problem. The most widely used heuristic search algorithms today in modern phylogenetic are the Nearest Neighbor Interchange (NNI) or the Subtree Pruning and Regrafting (SPR) methods.

The NNI search algorithm allows for the swapping of two adjacent branches on the tree topology. This is done by the elimination of one of the internal branches and reconnecting the taxa or clusters by the addition of another branch in a different place. Conversely, the more widely used SPR algorithm selects and removes a small subtree from the main tree topology and reinserts it elsewhere on the tree to create a new node on the tree. These heuristic search algorithms allow for random jumps in the tree space, which prevents the tree topologies getting stuck on a local maximum (which is not the true global maximum in the tree space). Additionally heuristics, such as NNI and SPR, greatly speed up the inference of phylogenies when compared to the alternative exhaustive search algorithms.

1.6.6 The implication of a molecular clock

Since the start of modern phylogenetics in the late 1950's and early 1960's, the implication of a molecular clock has been central in the field of phylogenetic inference. The molecular clock was introduced in the early 1960's by Zuckerkandl and Pauling, who published two papers on the rate of evolution in proteins [Zuckerkandl, 1962; Zuckerkandl, 1963]. They observed that the genetic distance of two sequences coding for the same protein, but isolated from different species, increased linearly with the divergence time of the two isolates. They observed similar findings in a wide range of other proteins and hypothesized that the rate of evolution for any given protein is constant over time. These findings led to the implication of a molecular clock for different genes. This meant that if a molecular clock exists and the rate of evolution of a particular gene can be

calculated or is known, one can calculate the divergence time between the isolates by comparing their nucleic or amino acid sequences [Zuckerandl, 1962; Zuckerandl, 1963].

The branch length of a phylogenetic tree is expressed as the expected number of substitutions per site. If the implied model of evolution indicates that each site within an ancestral sequence will experience n number of substitutions then the ancestor and descendant are considered to be separated by a branch length n [Felsenstein, 2004]. A molecular clock can be applied only if the expected number of substitutions per year (indicated by the Greek letter $\mu = \mu$) is constant in all the taxa within the data set. For example when one is studying the divergence of HIV with that of other primate lentiviruses (SIV) or when studying mammalian protein families (e.g. globin sequences) [Felsenstein, 2004].

Some species (or genes) evolve at faster rates than others. The assumption of a molecular clock in these cases is unrealistic, especially across long periods of evolution. For example, primates and rodents are genetically very similar to one another, but rodents have undergone a much higher rate of substitutions in the estimated time since divergence in some areas of their genome. In cases such as this the two measures of branch lengths between the two species are not directly proportional and a molecular clock cannot be applied [Felsenstein, 2004]. However, now with the implementation of a relaxed molecular clock assumption this problem can be overcome.

The relaxed molecular clock allows for different rates of mutations in different branches within the tree topology, while the strict molecular clock assumption allows for only a constant rate in all of the branches [Lemey *et al.*, 2010]. Several modern phylogenetic software packages have been developed to implement a molecular clock within their analysis and in most analyses one has the option to employ a strict or a relaxed molecular clock.

Since the development of the molecular clock in the 1960's this method has become widely used in computational evolutionary biology. The study of the evolutionary history of HIV is no exception and the implementation of a molecular clock has been widely used. These include the pioneer work done by Betty Korber and her group at the Los Alamos National Laboratory on the origin of HIV-1 and HIV-2 [Korber *et al.*, 2000]; the work done by Michael Worobey on the origin of subtype B HIV-1 in the DRC and its subsequent spread to Haiti and the homosexual population of the USA and the rest of the industrialized world [Gilbert *et al.*, 2007; Wertheim and Worobey, 2009]; the work done by Tulio de Oliveira, Oliver Pybus, Andrew Rambaut and co-workers on the timing of the epidemic spread of HIV and HCV amongst children infected in hospitals and clinics in Libya [de Oliveira *et al.*, 2006]; and similar molecular clock work done

by Marco Salemi and his team in Florida which worked independently on the origin of the global HIV pandemic and the epidemic spread of HIV-1 in Albania [Salemi *et al.*, 2001; Ciccozzi *et al.*, 2005] and Anne-Mieke Vandamme and her team in Leuven in Belgium, which worked on the origin of HIV-2 and Group O HIV-1 [Lemey *et al.*, 2003; Lemey *et al.*, 2004].

In addition to its wide application in HIV based research the molecular clock assumption has also successfully been employed in the investigation of the origin of pandemic influenza by Rambaut and co-workers [Rambaut *et al.*, 2008; Lemey *et al.*, 2009; Nelson *et al.*, 2011].

1.6.7 The coalescent theory in modern phylodynamics

In recent years the use of genomic data for the inference of population dynamics has become ever more important in the studying of epidemics. The most widely used method in “molecular epidemiology” to elucidate the population dynamics of epidemics is based on coalescent inference. The concept of coalescent inferences is based on the knowledge that present day populations carry a “genetic signature” that is locked inside their genomic data and that their genetic information can be used to reconstruct their historical profile of population dynamics. This reconstruction of population dynamics from present day sequence data provides scientists and epidemiologists with useful insight into various evolutionary processes and has become fairly widely used in recent years (e.g. in tracing the transmission and spread of viral pathogens) [Kitchen *et al.*, 2008; Magiorkinis *et al.*, 2009].

The coalescent approach is aimed at quantifying the relationship between the genealogy of a given data set and the demographic history of the set data set and was first popularized by Kingman in the 1980’s [Kingman 1982a; Kingman 1982b]. In a coalescent framework, lineages are traced back in time from a sample of sequences at a given time point, with pairs of lineages in the genealogy coalescing with one another at random until a distant point in the past. This point where all lineages have coalesced to a single point back in time is commonly referred to as the common ancestor of all the sampled sequences. The genealogy of a sample of sequences is randomly determined by a wide range of factors (history of the population, natural selection pressures and other factors) [Donnelly and Tavaré, 1995]. The reconstruction of the demographic history thus involves the estimation of the genealogy and the inference of the effective population size at various time points along the genealogy. The effective population size directly reflects the number of individuals that contribute offspring in the subsequent generations and is almost always smaller than the actual population size.

In the past, the majority of these methods assumed that the population history of a data set could be easily described by simple parametric models (e.g. constant-, exponential-, or logistic growth). The reality however, is that most population histories are far more complex, and they cannot sufficiently be described by simplistic parametric means. This led to the development of non-parametric methods for the inference of demographic histories from sequence data [Fu, 1994; Polanski *et al.*, 1998; Pybus *et al.*, 2000]. The development of non-parametric means of inferring demographic histories led to the introduction of the “skyline plot” framework which was first introduced by Pybus and co-workers in 2000 [Pybus *et al.*, 2000]. Since the introduction of the “classic skyline plot” by Pybus and co-workers, a small “family” of skyline plot methods has been developed [Strimmer and Pybus, 2001; Opgen-Rhein *et al.*, 2005; Drummond *et al.*, 2005; Minin *et al.*, 2008; Heled and Drummond, 2008].

In viral epidemic population dynamics the coalescent framework and demographic reconstruction by means of non-parametric means has extensively been used in the recent past to investigate the epidemiology of pathogens such as; the inter-host evolutionary dynamics of longitudinally sampled HIV-1 *env* genes [Lemey *et al.*, 2006], the estimation of pandemic growth of H1N1 influenza in the USA [de Silva *et al.*, 2012], the epidemic sexual transmission and the phylodynamics of HIV-1 amongst the MSM population in the UK [Lewis *et al.*, 2008], and the epidemic reconstruction of the HCV epidemic in Egypt [Drummond *et al.*, 2005].

The wide spread use of coalescence, along with the implementation of molecular clock and Bayesian techniques has recently given rise to a new branch in phylogenetic inference termed phylodynamics. This new branch is aimed at the elucidation of population characteristics, such as the calculation of the population size (N_e) and the rate of viral genetic expansion, of epidemics through the analysis of sequence data and the use of advanced phylogenetic methods (molecular clock, Bayesian inference and the coalescence).

AIM OF THE STUDY

The aim of this study was to look at the evolutionary history of the heterosexual HIV-1 subtype C epidemic in Cape Town, South Africa. Four specific scientific questions were formulated: (1) What is the evolutionary history and dynamic aspects of the Cape Town epidemic? (2) How does the evolutionary history of the Cape Town data sets relate to the epidemic across the Southern African region? (3) What is the evolutionary relationship of the Cape Town data sets and how does it fit into the global HIV-1 subtype C pandemic? (4) What is the nature of the highly

monophyletic clustering of Cape Town sequences and what is the evolutionary history of the putative transmission events?

The strategy was to obtain a comprehensive data set of longitudinal sampled specimens from the Cape Town region of South Africa. From these longitudinal sampled specimens *gag* p24 and *pol* sequences were generated through standard molecular characterization techniques. These sequences were then used to elucidate the evolutionary history of the HIV-1 subtype C epidemic in Cape Town, South Africa with the use of advanced phylogenetic methods such as Bayesian inference and the implementation of a molecular clock and coalescence. This included the estimation of the date of origin and phylodynamic aspects of the epidemic in Cape Town.

Additionally, large numbers of homologous HIV-1 subtype C sequences were obtained from public and/or private sequence databases. A small subset of these sequences was then used to infer evolutionary histories for other areas of the Southern African region. The results of the Cape Town sequence data set were then compared to those from other areas of the Southern African region. The genetic information that is produced throughout this study represents some of the oldest HIV-1 subtype C sequences from South Africa and the Southern African region.

Specific objectives of the study included:

1. To obtain a longitudinal sampled data set of *gag* p24 and *pol* sequences from Cape Town, South Africa.
2. To investigate the evolutionary history, including the date of origin and phylodynamics aspects, of the Cape Town HIV-1 subtype C epidemic.
3. To compare the evolutionary history of the epidemic in Cape Town, South Africa with the epidemic(s) from other areas of the Southern African region.
4. To conduct a basic phylogenetic investigation to establish the evolutionary relationship of the isolates in the Cape Town data set(s) with other HIV-1 subtype C isolates from around the world.
5. To investigate the nature of highly monophyletic clustering of Cape Town isolates, which was observed during the basic phylogenetic investigation, and to establish whether these monophyletic clades represent transmission events of HIV-1.

CHAPTER TWO - TABLE OF CONTENTS

	Page
2.1 Ethical permission	65
2.2 Reagents and equipment	65
2.3 Generation of Cape Town data sets	66
2.3.1 Sample selection	67
2.3.2 Sample preparation & Nucleic acid isolation	68
2.3.3 Amplification of the viral genomic segments	68
2.3.4 Gel electrophoresis and sample clean up	71
2.3.5 Sequencing of amplified products	72
2.3.6 Sequence subtyping	73
2.3.7 Testing the molecular clock of the Cape Town data set	73
2.4 Phylogenetic investigation	74
2.4.1 Bayesian inference and epidemic reconstruction	76
2.4.1.1 Data set composition for the Bayesian analysis	76
2.4.1.2 Sequence alignment for Bayesian inference	76
2.4.1.3 Setting up of the Bayesian runs	76
2.4.1.4 Analysis of Bayesian MCMC runs	77
2.4.1.5 Epidemic reconstruction	78
2.4.2 Establishing the evolutionary relationship of Cape Town isolates	79
2.4.2.1 Data Mining	79

2.4.2.2 Sequence alignments and Editing	80
2.4.2.3 Inference of large-scale phylogenies	80
2.4.2.4 Analysis of large-scale phylogenies	81
2.4.3 Phylogenetic investigation of transmission events	82
2.4.3.1 Identifying and testing potential clusters	83
2.4.3.2 Timing the internal nodes of transmission clusters	85
2.4.3.3 Time resolved tree topologies with clustering of the various monophyletic clades	85

CHAPTER TWO

The methodology that was used during the course of this study will be discussed in this chapter. This project involved three main steps: (1) the production of 21-years of genetic information from patient samples that were stored mostly in -20°C refrigerators, (2) the analysis of the genetic information, and (3) the interpretation of the analysed data.

Briefly, the general aim was to obtain longitudinal sampled sequence data sets of patients from Cape Town, spanning roughly over a 21-year time period. Samples were selected from the patient database at the Division of Medical Virology (Tygerberg Academic Hospital). These samples, which were submitted for routine HIV diagnostic purposes, were selected from the Division's cold storage at -20°C . The *gag* p24 and a partial segment of the *pol* fragment of HIV-1 were targeted for characterization due to the highly standardised nature of these PCR and sequencing assays, as well as for the availability of primer sets.

The analysis of these sequence fragments involved the separation of sequence data into three different data sets: a *gag* p24, a *pol* and a concatenated *gag-pol* data set (containing *gag* and *pol* sequence fragments from patients who were represented in both data sets). Firstly, the estimated date of origin and phylodynamic aspects of the Cape Town epidemic was estimated through Bayesian reconstruction from the three different Cape Town data sets. Following the epidemic reconstruction from the three Cape Town data sets, similar analyses were performed on Southern African data sets (some including and others excluding sequence information from the Cape Town data sets). This was done in order to assess whether any potential similarities or differences exists between the Cape Town epidemic and the Southern African HIV-1 subtype C epidemic and to calculate the estimated date of origin of the HIV-1 subtype C epidemic in the Southern African region. Thereafter, a phylogenetic investigation was performed to establish the evolutionary relationship of the Cape Town isolates with other HIV-1 subtype C strains from around the world. This was done through the inference of large-scale phylogenetic tree topologies, which were then analysed through both manual and automated methods. Highly monophyletic clustering of Cape Town isolates, which were observed during the inference of large-scale tree topologies, was further investigated following the epidemic reconstruction. This was done in order to establish the nature of these monophyletic clusters, which may be representative of transmission events of HIV-1.

2.1 Ethical permission

The study received ethical approval from Human Research Ethics Committee (HREC) at the Faculty of Medicine and Health Sciences at Stellenbosch University (Tygerberg Campus). Ethics permission was granted in August 2009 and the project was registered under the application number N09/08/221 and renewed annually.

2.2 Reagents and equipment

All the reagents, equipment, and software applications that were used during the course of this study are listed in Table(s) 2.1 - 2.3. All chemical and biological agents or commercial kits that were used in this study are summarized in Table 2.1. Registered trade mark items and trade mark products are indicated by the symbols ® and TM respectively.

Table 2.1: List of chemicals and commercial products used in the study.

Chemical or Commercial products and kits used	Supplying Company	Catalogue number
QIAamp Viral RNA Mini Kit	QIAGEN, Dusseldorf, Germany	52 906
QIAamp DNA Blood Mini Kit	QIAGEN, Dusseldorf, Germany	51 106
Access RT-PCR System	Promega, Madison, WI, USA	A 1 260
GoTaq® DNA Polymerase	Promega, Madison, WI, USA	M 8 305
dNTP's	Promega, Madison, WI, USA	U 1 330
Nuclease free water	Promega, Madison, WI, USA	P 1 193
Agarose	Whitehead Scientific (Pty) Ltd.	#D1 - LE
Ethidium Bromide	Promega, Madison, WI, USA	H 5 041
6x Blue Orange Loading Dye	Promega, Madison, WI, USA	G 1 881
QIAquick PCR Purification Kit	QIAGEN, Dusseldorf, Germany	28 106
BigDye™ Terminator cycle sequence ready Kit	Applied BioSystems, CA, USA	4 337 035
5x Sequencing Buffer	Applied BioSystems, CA, USA	4 305 603
BigDye XTerminator Purification Kit	Applied BioSystems, CA, USA	4 374 408

A brief summary of all the equipment that was used during the course of this study is listed in Table 2.2.

Table 2.2: Equipment used to perform sample analysis.

Piece of Equipment	Supplying Company
QIAcube nucleic acid isolation system	QIAGEN, Dusseldorf, Germany
GeneAmp PCR System 9700 thermal cycler	Applied BioSystems, CA, USA
Hoefer EPS 2 A 200, Power Pack	Pharmalac Biotechnologies, CA, USA
Syngene™ GeneGenius Computer System	Synoptics Ltd., Cambridge, UK
ABI 3130xl automated DNA sequencer	Applied BioSystems, CA, USA

The various software applications, and/or, online analytical tool that were used during the phylogenetic analysis of the sequence data are listed in Table 2.3.

Table 2.3: Software programs and online analytical tool that were used in the analysis of sequence information.

Software package	References and/or licensed companies
Sequencher v 4.8	Gene Codes Corporation, Ann Arbor, MI, USA
ClustalW v 2.1	Thompson [©] <i>et al.</i> , 1997
Se-al v 2.0	Rambaut (http://tree.bio.ed.ac.uk)
MrBayes v 3.0	Huelsenbeck and Ronquist, 2001
Path-O-Gen	Rambaut (http://tree.bio.ed.ac.uk/software/pathogen)
PhyML v 3.0	Guindon <i>et al.</i> , 2010
BEAST v 1.4	Drummond and Rambaut, 2007
FigTree v 1.3.1	Rambaut (http://tree.bio.ed.ac.uk)
MEGA v 5.0	Tamura <i>et al.</i> , 2011
fastME	Desper and Gascuel, 2002
PhyloType	Chevenet <i>et al.</i> , 2013
jpHMM	Spang <i>et al.</i> , 2002
REGA v 2.0 HIV subtyping tool	de Oliveira <i>et al.</i> , 2005

2.3 Generation of Cape Town data sets

The following section contains the methodology that was used in the generation of the various Cape Town data sets (Figure 2.1). In this study the *gag* p24 and a partial segment of the *pol* genes of HIV-1 were targeted for investigation. These regions of the HIV-1 genome were specifically targeted since the molecular assays, both PCR and sequencing, for these regions of the HIV genome has been well standardized [Swanson *et al.*, 2003; Plantier *et al.*, 2005] and the various primers were readily available at our laboratory. Furthermore, these segments of the HIV-1 genome have routinely been used successfully in the past to reconstruct evolutionary relationships of HIV isolates [Dalai *et al.*, 2009; Novitsky *et al.*, 2010].

A large number of *gag* p24 and *pol* sequences that have previously been characterized within the Division of Medical Virology were chosen for inclusion in this study. A total of 163 previously characterized *gag* p24 sequences and 93 previously characterized *pol* sequences were selected for inclusion into the study.

These sequences were generated by either me or other by colleagues [Jacobs *et al.*, 2006; Wilkinson *et al.*, 2013 in press; Isaacs *et al.*, in preparation]. Additionally, a large number of HIV-positive patient samples were selected for characterization from the HIV sample database of the Division of Medical Virology (Tygerberg Academic Hospital). These samples have been stored in -20°C freezers for several years since their submission to the Division for various routine diagnostic tests.

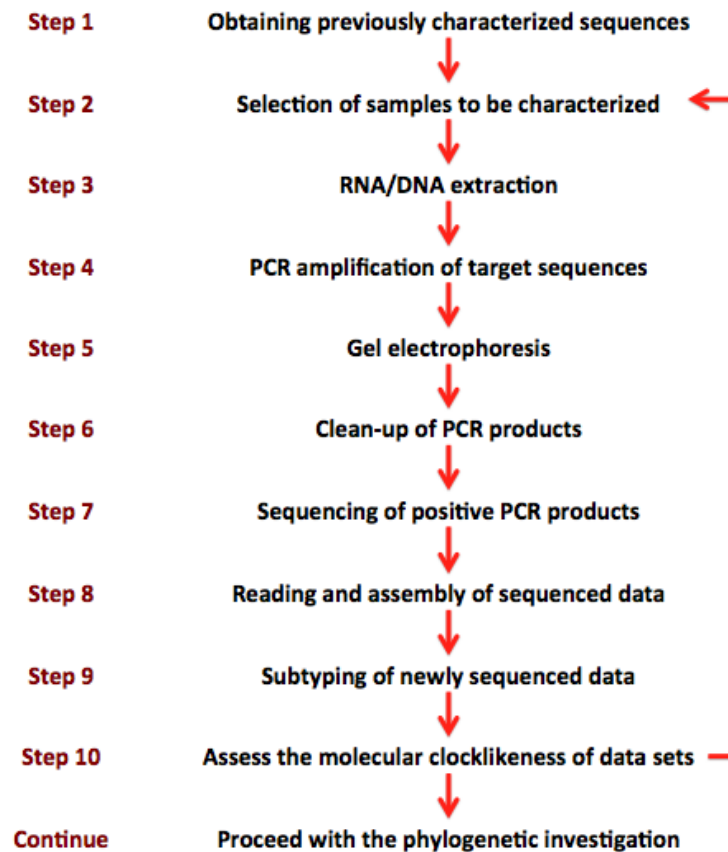


Figure 2.1: A step-by-step breakdown of the methodology that was used in the generation of the Cape Town data sets is illustrated in the different steps.

2.3.1 Sample selection

Longitudinally sampled specimens, spanning roughly over a 21-year period (1989 to 2010), were selected from the Division of Medical Virology's HIV-1 sample database. The database contains over 35,000 HIV-1 samples since the early 1980's till the present. Adequate patient demographic data was available on many of the samples included in the database. Samples were selected based on strict inclusion and exclusion criteria. A list of inclusion and exclusion criteria is presented in Table 2.4.

Table 2.4: Specific inclusion and exclusion criteria for sample selection.

Inclusion Criteria	Exclusion Criteria
1. Specimens from South African nationals	1. Specimens from non-South African citizens.
2. Patients older than 15-years of age.	2. Patients younger than 15-years of age.
3. Epidemiologically unlinked patient samples.	3. Specimens from epidemiologically linked patients.

A total of 250 patient samples were selected for genotyping, which met all of the inclusion criteria. These 250 patient samples only represent a very small number of the total number of patient samples that were available for selection ($n > 35,000$). These samples included 63 previously isolated DNA samples (originating between 1989 and 1992), 34 samples from 1996 from which new DNA or RNA were extracted (depending on the sample type), and 153 previously isolated RNA samples which were sampled between 2000 and 2004. The previously isolated DNA and RNA samples were stored at 4°C and -80°C respectively since their isolation.

2.3.2 Sample preparation & Nucleic acid isolation

Nucleic acid isolations were performed on all samples for which no nucleic acid had previously been isolated. Since the majority of the samples that were selected for characterization had nucleic acids isolated at a previous point in time, the isolation of new nucleic acids were only limited to the 34 samples from 1996. Due to the variation in the type of specimen (whole blood or blood plasma) either DNA or RNA was isolated from these samples.

For DNA isolation the QIAamp DNA Blood Mini Kit and the QIAcube automated nucleic acid isolation system (QIAGEN, Dusseldorf, Germany) was used with the proposed protocol of the manufacturer. After the DNA was isolated from whole blood specimens, the DNA samples were labelled and stored at 4°C for use at a later date.

RNA was isolated from blood plasma samples with the use of the QIAamp Viral RNA Mini Kit and the QIAcube automated nucleic acid isolation system (QIAGEN, Dusseldorf, Germany) according to the manufactures proposed protocol. After RNA was isolated from the blood plasma, the samples were labelled and stored at -80°C for further use.

2.3.3 Amplification of the viral genomic segments

The amplification of viral genomic segments involved two main steps, a reverse transcription polymerase chain reaction (RT-PCR) and a standard polymerase chain reaction, which are described in detail in the following paragraphs.

For the genotyping of RNA samples a coupled reverse transcription and PCR amplification assay were used to reverse transcribe the RNA into complimentary DNA (cDNA) and then amplify the target gene of interest within a single reaction. This was achieved with the use of the Access RT-PCR System and GoTaq DNA polymerase (Promega, Madison, WI, USA). All reverse transcription and PCR steps were performed on the GeneAmp PCR System 9700 thermal cycler

(Applied BioSystems, CA, USA). PreNested and Nested PCR's were performed on samples that were characterized.

For all RNA samples the reverse transcription reactions for the formation of cDNA were set up as follows: each reaction contained 10 µl of AMV/*Tfl* 5x Reaction Buffer (1x concentration), 1 µl of dNTP's (10mM concentration), 1 µl of each primer (40 pmol concentration), 2 µl of MgSO₄ (at a concentration of 25mM), 1µl of *Tfl* DNA Polymerase (concentration of 0.1U/µl) and 1 µl of AMV Reverse Transcriptase (concentration of 0.1U/µl). The PreNested cycling conditions were preceded with 2 cycles for the synthesis of cDNA from the RNA template copy. This was done by heating each sample to 48°C for 45 minutes and then to 94°C for 2 minutes (for the inactivation of the AMV reverse transcriptase enzyme and denaturation of nucleic acid) before the regular cycling conditions of the PreNested PCR.

The PCR methods and primers, for the amplification of target genes from DNA or cDNA templates, were adapted from Swanson and co-workers [Swanson *et al.*, 2003] and Plantier and co-workers [Plantier *et al.*, 2005] for the *gag* p24 and *pol* assays respectively. The *pol* amplification assay as described by Plantier and co-workers involves the amplification of the *pol* target region in two overlapping fragments: one spanning the protease (PR) segment and one spanning the reverse transcriptase (RT) segment of the *pol* gene. For all DNA samples, both the PreNested and Nested PCR reactions of the *gag* p24 and *pol* PCR assays contained 10 mM of each dNTP, 20 µM of each primer, 1,5 mM of MgCl₂ and 1U of *Taq* polymerase in a total reaction volume of 50 µl. The following cycling conditions were used for the *gag* p24 as well as for the protease and reverse transcriptase *pol* PCR assays: one cycle of denaturation at 94°C for 2 minutes; followed by 40 cycles of: denaturing at 94°C for 20 seconds, primer annealing for 30 seconds, and primer extension at 68°C; one final step of primer extension at 68°C for 10 minutes, after which samples were cooled down and stored at 4°C. Five µl of the PreNested product was carried over to each nested PCR reaction. Table 2.5 contains an outline of the cycling conditions of the *gag* p24 PreNested and Nested PCR assay.

Table 2.5: Cycling conditions for the PreNested and Nested *gag* 24 PCR assays. These assays were conducted on either DNA or RNA templates depending on the sample type.

Cycling conditions for the PreNested <i>gag</i> p24 PCR assay			
Step	Temperature	Duration	Cycles
Reverse Transcription	48°C	45 min	x1
Initial Denature Step	94°C	2 min	x1
Denature	94°C	20 sec	x40
Anneal	45°C	30 sec	
Extend	68°C	90 sec	
Final Extension	68°C	10 min	x1
Cycling conditions for the Nested <i>gag</i> p24 PCR assay			
Step	Temperature	Duration	Cycles
Initial Denature Step	94°C	2 min	x1
Denature	94°C	20 sec	x40
Anneal	50°C	30 sec	
Extend	68°C	60 sec	
Final Extension	68°C	10 min	x1

sec – seconds; min – minutes; °C – Degrees Celsius

Similarly, Table(s) 2.6 and 2.7 contains the cycling conditions of both the PreNested and Nested PCR assays of the PR-*pol* and RT-*pol* fragments, respectively.

Table 2.6: Cycling conditions for the PreNested and Nested protease-*pol* PCR assays. These assays were conducted on either DNA or RNA templates depending on the sample type.

Cycling conditions for the PreNested PR- <i>pol</i> PCR assay			
Step	Temperature	Duration	Cycles
Reverse Transcription	48°C	45 min	x1
Initial Denature Step	94°C	2 min	x1
Denature	94°C	20 sec	x40
Anneal	55°C	30 sec	
Extend	68°C	90 sec	
Final Extension	68°C	10 min	x1
Cycling conditions for the Nested PR- <i>pol</i> PCR assay			
Step	Temperature	Duration	Cycles
Initial Denature Step	94°C	2 min	x1
Denature	94°C	20 sec	x40
Anneal	55°C	30 sec	
Extend	68°C	60 sec	
Final Extension	68°C	10 min	x1

sec – seconds; min – minutes; °C – Degrees Celsius; RT – Reverse Transcriptase; *pol* – polymerase

Table 2.7: Cycling conditions for the PreNested and Nested reverse transcriptase-*pol* PCR assays. These assays were conducted on either DNA or RNA templates depending on the sample type.

Cycling conditions for the PreNested RT- <i>pol</i> PCR assay			
Step	Temperature	Duration	Cycles
Reverse Transcription	48°C	45 min	x1
Initial Denature Step	94°C	2 min	x1
Denature	94°C	20 sec	x40
Anneal	55°C	30 sec	
Extend	68°C	90 sec	
Final Extension	68°C	10 min	x1
Cycling conditions for the Nested RT- <i>pol</i> PCR assay			
Step	Temperature	Duration	Cycles
Initial Denature Step	94°C	2 min	x1
Denature	94°C	20 sec	x40
Anneal	55°C	30 sec	
Extend	68°C	60 sec	
Final Extension	68°C	10 min	x1

sec – seconds; min – minutes; °C – Degrees Celsius; RT – Reverse Transcriptase; *pol* – polymerase

Additionally, all PCR assays were run with a positive HIV-1 control sample that amplified well under the same conditions and primer sets used in a previous study [Wilkinson and Engelbrecht, 2009]. Due to the age of these patient samples, some degree of difficulty was experienced in the amplification of some samples. Therefore, in some cases a large amount of time and resources were spent optimizing PCR conditions such as: various primer concentrations, MgCl₂ titrations at various concentrations and adjusting annealing temperatures. The final amplification conditions are summarized in Table(s) 2.5 through to 2.7.

2.3.4 Gel electrophoresis and sample clean up

All nested PCR products were separated on 0,8%, ethidium bromide stained, agarose gels (10 cm long). Eight µl of PCR product was mixed with 3 µl of Blue Orange loading dye (Promega, Madison, WI, USA) and loaded onto the gels. The samples were then run at 50 Volts for approximately 45 minutes. After the samples migrated through the gel, each gel was exposed to UV light and photographed with the use of the SyngeneTM GeneGenius Computer System (Synoptics Ltd., Cambridge, UK). All positive PCR samples were then cleaned-up to remove all excess dNTP's and residual unbound primers, with the QIAquick PCR Purification kit (QIAGEN, Dusseldorf, Germany). The kit uses silica based spin columns to bind amplified PCR products while all residual dNTP's and primers are removed through continuous wash steps using high speed centrifugation or a vacuum. The amplified target DNA is then finally eluded from the spin column with elution buffer or nuclease free water. The molecular concentration of each

cleaned up product were then determined with the Nanodrop™ ND 1000 (Nanodrop Technologies Inc., Delaware, USA) before it was stored at 4 °C for later use.

2.3.5 Sequencing of amplified products

Amplified PCR products were all directly sequenced with the use of the primers listed in Table 2.8. The BigDye® Terminator Cycle Sequencing Kits (Applied BioSystems, CA, USA), was used for the sequencing reactions.

Table 2.8: Sequencing primers and the annealing temperatures that were used for the sequencing of amplified products.

Primers and Cycling conditions used for the sequencing of <i>gag</i> p24 PCR products			
Primer	Forward / Reverse	Annealing Temperature	Reference
<i>gag</i> p6	Reverse	50°C	Swanson <i>et al.</i> , 2003
<i>gag</i> p2	Forward	50°C	Swanson <i>et al.</i> , 2003
Primers and Cycling conditions used for the sequencing of PR- <i>pol</i> PCR products			
Primer	Forward / Reverse	Annealing Temperature	Reference
JA 217	Reverse	50°C	Plantier <i>et al.</i> , 2005
30 prot 2	Reverse	50°C	Plantier <i>et al.</i> , 2005
Pol 1D	Forward	50°C	A Loxton personal communication
Primers and Cycling conditions used for the sequencing of RT- <i>pol</i> PCR products			
Primer	Forward / Reverse	Annealing Temperature	Reference
AK 10	Forward	50°C	Plantier <i>et al.</i> , 2005
AK 11	Forward	50°C	Plantier <i>et al.</i> , 2005
<i>pol</i> 3	Forward	50°C	S Engelbrecht personal communication
NE 135	Reverse	50°C	Plantier <i>et al.</i> , 2005

°C – Degrees Celsius; PCR – Polymerase Chain Reaction; RT – *reverse transcriptase*; PR – *protease*; *pol* – polymerase; *gag* – glycoprotein

Briefly, each sequencing reaction contained approximately 50 ng of the purified PCR product, 5 pmol of sequencing primer, 1.3 µl of Big Dye terminator enzyme mix (BigDye® Terminator Cycle Sequencing Kits, Applied BioSystems, CA, USA), and 2.7 µl of 5x Sequencing Buffer (Applied BioSystems, CA, USA). Nuclease free water was added to each reaction mixture to make up a final reaction volume of 10 µl. Each sequencing reaction was performed under the following conditions on a GeneAmp PCR System 9700 thermal cycler (Applied BioSystems, CA, USA): 25 cycles of denaturation at 96 °C for 10 seconds, primer annealing for 5 seconds (Table 2.8) and an elongation step at 60 °C for 4 minutes. Afterwards the samples were cooled down to 4 °C for storage.

Each sequencing reaction was cleaned up with the BigDye XTerminator Purification Kit (Applied BioSystems, CA, USA) according to the manufactures protocol before it was run on the ABI 3130xl automated DNA sequencer (Applied BioSystems, CA, USA). The trace data files of each sequencing run was then retrieved and imported into Sequencer v 4.8 (Gene Codes

Corporation, Ann Arbor, Michigan, USA). The quality of each chromatogram fragments was assessed and manually trimmed in order to remove ambiguous sections (portions with multiple peaks). Each patient's chromatogram fragments was then assembled into a single contiguous sequence fragment and manually proofread in order to assure good sequence quality. After the fragments were proofread they were exported in a text file (.txt) format and labelled.

2.3.6 Sequence subtyping

All assembled sequenced data that was generated during the course of the study, as well as all sequences that was selected for inclusion in the study, were subtyped with the use of two different online HIV subtyping tools: the HIV-1 & -2 viral subtyping tool of REGA v 2.0 (<http://www.bioafrica.net/rega-genotype/html/subtypinghiv.htm>), which is accessible from the bioafrica.net webpage, and the jpHMM (<http://jphmm.gobics.de>) at GOBICS to ensure that only HIV-1 subtype C sequences were included into the study.

The REGA subtyping tool is an easy online method of subtyping full-length or subgenomic fragments by combining different phylogenetic approaches with bootscanning methods [de Oliveira *et al.*, 2005]. The jpHMM method of subtyping employs a jumping alignment approach, which was first proposed by Spang and colleagues [Spang *et al.*, 2002], for the subtyping of sequence fragments or the identification of recombinant viruses. Instead of a query sequence being compared with a multiple alignment, which is the standard approach, the query sequence is compared and aligned to individual sequences from the alignment. In the case of recombination events the query sequence can then jump between different sequences of the multiple alignment as a sliding window moves over the alignment. This tool also makes the identification of particular breakpoints within a recombinant isolate much easier [Schultz *et al.*, 2006; Zhang *et al.*, 2006].

2.3.7 Testing the molecular clock of the Cape Town data set

It is important to assess the clocklikeness of any given data set before any phylogenetic analysis, which relies on the assumption of a molecular clock, is conducted. The generation of *gag* p24 and *pol* sequences from Cape Town samples continued until a strong evolutionary signal was obtained in each of the data sets. The evolutionary signal was tested with Path-O-Gen v 1.3 (<http://tree.bio.ed.ac.uk/software/pathogen>), which is part of the BEAST software package [Drummond and Rambaut, 2007]. The Path-O-Gen program was designed to read phylogenies

that were not constructed under the assumption of a molecular clock to test the clocklikeness of the given data set.

Briefly, dated sequences (corresponding to the year of sampling) were assembled in a single text (.txt) file for both the *gag* p24 and *pol* sequences. Alignments were constructed for each of the data sets in ClustalW v 2.1 (<http://www.clustal.org/download/current>) with the use of a quick tree function to speed up the time of the alignment. Alignment files were then imported into Se-AL v 2.0 (<http://tree.bio.ed.ac.uk/software/seal>) and manually edited till a perfect codon alignment was obtained. Each of the alignments (*gag* p24 and *pol*) was then exported in the appropriate file format (.nex & .fasta) and labelled. Each of the file formats were then used for the inference of ML- and ME-tree topologies which was constructed with the use of the HKY+G+I model of nucleic acid substitution and the use of the Subtree Pruning and Regrafting (SPR) method in in PhyML v 3.0 (<http://www.atgc-montpellier.fr/phyml/>) [Guindon *et al.*, 2010] and fastME (<http://www.atgc-montpellier.fr/fastme/binaries.php>) [Desper *et al.*, 2002] respectively. Newly constructed topologies were then analysed manually in FigTree v 1.2.1 (<http://tree.bio.ed.ac.uk/software/figtree>) and saved in a nexus tree file format. Each of the trees were then imported in Path-O-Gen v 1.3 (<http://tree.bio.ed.ac.uk/software/pathogen>) software package and analysed to evaluate the temporal signal, general clock likeness and mutation rate (slope rate) of each of the data sets. For a diagrammatical representation of a molecular clock analysis that was performed in Path-O-Gen on the *gag* p24 ME-tree topology please refer to Figure 6.1 in Appendix B on page 236.

2.4 Phylogenetic investigation

The following section contains the methodology that was used in the analysis of the various Cape Town data sets. The different methodologies were used in order to answer the following scientific questions: (1) What is the evolutionary history and dynamic aspects of the Cape Town epidemic? (2) How does the evolutionary history of the Cape Town data sets relate to the epidemic across the Southern African region? (3) What is the evolutionary relationship of the Cape Town data sets and how does it fit into the global HIV-1 subtype C pandemic? (4) What is the nature of the highly monophyletic clustering of Cape Town sequences and what is the evolutionary history and aspects of these potential transmission events? A diagrammatical breakdown of the phylogenetic methods is presented in Figure 2.2.

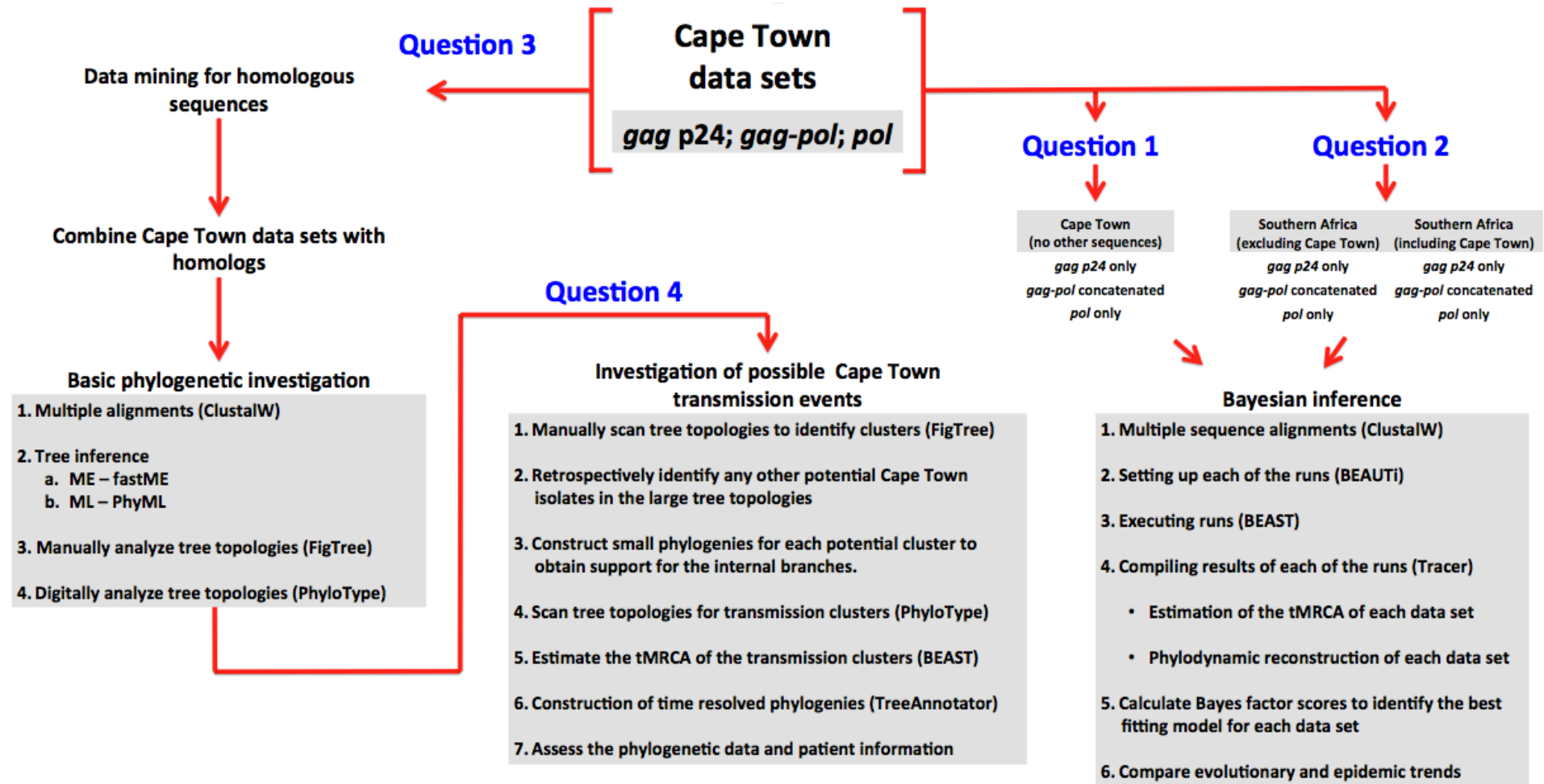


Figure 2.2: A diagrammatical breakdown of the phylogenetic methodology that was used during in order to answer the 4 main scientific questions in this study. Question 1 involves the epidemic reconstruction of the Cape Town HIV-1 subtype C epidemic from the three data sets. Question 2 involves the epidemic reconstruction of the epidemic of Southern Africa from three genomic regions of HIV. Question 3 involved the retrieval of homologous sequences and the inference of large-scale tree topologies in order to establish the evolutionary relationship of the Cape Town isolates. Question 4 involves the methodology that was used in the investigation of possible transmission events of HIV-1 in Cape Town.

2.4.1 Bayesian inference and epidemic reconstruction

2.4.1.1 Data set composition for the Bayesian analysis

For the estimation of the evolutionary histories of the entire Southern African region, as well as individual countries (e.g. Botswana or Zimbabwe) or regions (in the case of Cape Town), a large number of spatiotemporal samples were selected at random with known dates of sampling from each of the large data sets. For a full breakdown of the total number of taxa for each Southern African data set please refer to Table 6.4 in Appendix C on page 237.

All three of the data sets (*gag*, *gag-pol*, and *pol*) were analysed in three different Bayesian inference runs: a run containing only Cape Town sequence data, a run of Southern Africa sequences excluding Cape Town sequences, and a run of Southern African sequences including sequence data from Cape Town.

2.4.1.2 Sequence alignment for Bayesian inference

Each of the alignments was constructed with a quicktree method of alignment in ClustalW v 2.1 (<http://www.clustal.org/download/current>) to speed up the alignment process. Each alignment file was manually edited in Se-AL v 2.0 (<http://tree.bio.ed.ac.uk/software/seal>). Gaps were excluded from alignments if the gaps were present in more than 20% of the taxa within the data set by deleting the entire column in which these gaps appeared. Alignments were then manually edited until a perfect codon alignment was obtained. Once again a considerable amount of time was spent on the manual codon aligning of each of the alignment due to the importance of a good alignment on any subsequent analysis. Aligned files were then exported in a nexus (.nex) and fasta (.fasta) file formats and labelled.

2.4.1.3 Setting up of the Bayesian runs

Bayesian Markov Chain Monte Carlo (MCMC) runs were set up under various demographic models in the BEAUTi software application, which is part of the BEAST v 1.7.3 (http://beast.bio.ed.ac.uk/Main_Page) software package. Sequence alignments, containing longitudinal sampled sequences were imported into BEAUTi in a nexus (.nex) file format. Sampling dates was estimated from the taxa names within each of the alignments.

For each of the data sets, evolutionary histories were estimated under a parametric (Constant population size) and non-parametric (Bayesian Skyline Plot) model with the use of both a strict

and relaxed molecular clock assumption. Each run was set up with 100 million steps in the MCMC with sampling every 10,000 steps in the chain. Each of the runs was run with the implementation of the SRD06 model of nucleotide substitution [Shapiro *et al.*, 2006]. This model has two rate partitions, one for the first- and second-base and another for the third-base in a codon. This allows for a different rate for G + I than the first two coding bases in an amino acid. Fixed and estimated mutation rates were used for each of these models. Mutation rates were fixed for all of the model parameters that were executed under a fixed mutation rate according to the known mutation rates for the various regions of the HIV-1 genome: *gag* p24 [Novitsky *et al.*, 2010], concatenated *gag-pol* [Hue *et al.*, 2004], and *pol* [Hue *et al.*, 2004] that are routinely used in the literature. Each of these run parameters was carefully set up and exported and saved. Each of the files (.xml) was then executed in the BEAST v 1.7.3 software application (http://beast.bio.ed.ac.uk/Main_Page).

The average time of each of the runs varied depending on the total number of taxa contained within each data set, as well as the sequence length of each data set. The smallest data sets took on average around 24 hours, while the largest took up to 14 days. Each of the runs was executed on a standard multi processor (32-processors) Linux based computer platform with Mac OS X 10.6 (Snow Leopard).

2.4.1.4 Analysis of Bayesian MCMC runs

Convergence in each of the runs was sporadically assessed with the use of Tracer v 1.5 (<http://tree.bio.ed.ac.uk/software/tracer/>) on the basis of the effective sample size (EES) and good convergence in the trace files. It is generally regarded that an EES greater than 200 shows good sampling from the posterior distribution of parameters. If good convergence was attained before the run attained the final chain length (100 million steps) the run was terminated prematurely. Each of the corresponding log files was used to extract the estimated time to the most recent common ancestor (tMRCA), mutation rate, and the coefficient of variation. Uncertainty in the estimation in each of the runs was assessed by the 95% highest posterior density (95% HPD) intervals.

Bayesian inference relies heavily on the use of priors, for example the assumption of a relaxed or strict molecular clock or assumptions based on population sizes (constant or exponential). Therefore, one needs to test which of the chosen priors represents the given data set the best. Since the output of any Bayesian analysis is a posterior probability of possible solutions, standard probability methods can be used to address this question. For each of the various data sets the

“best-fitting model” of inference was determined through the calculation of Bayes factors [Suchard et al., 2001]. This was performed in Tracer with the use of the various log output files from the Bayesian runs in BEAST v 1.7.3 following 1000 bootstrap replicates. The best fitting model for each data set was then chosen based on the calculated Bayes factors.

Time resolved phylogenies of each of the data sets were also constructed. The tree output file from the corresponding Bayesian MCMC run was used to construct a time resolved phylogenetic tree in TreeAnnotator v 1.7.3, which is part of the BEAST (http://beast.bio.ed.ac.uk/Main_Page) software application. The “best fitting” non-parametric model that was identified through the Bayes factor model comparison was used for the inference of these time resolved phylogenies.

2.4.1.5 Epidemic reconstruction

HIV demographic histories were reconstructed from the data that were inferred during the BEAST analysis. Bayesian Skyline Plots (BSPs) were constructed from the “best fitting” non-parametric run in the *gag* and *pol* data sets for the Cape Town epidemic, the Southern African epidemic (excluding sequence data from Cape Town) and the entire Southern African epidemic. These BSPs were inferred from the corresponding log (.log) and tree (.tree) files in Tracer v 1.5 (<http://tree.bio.ed.ac.uk/software/tracer/>). The plots, as well as the raw data, were exported and saved. The raw data was then imported into Excel and the BSPs were carefully reconstructed for each plot. The median and 95% Highest Posterior Density (HPD) intervals were used for this reconstruction.

The estimated percentage lineages through time (PLTT) were also inferred from the best fitting non-parametric models for the Cape Town *gag* p24 and *pol* data sets. These, as with the BSPs, were inferred in Tracer v 1.5 (<http://tree.bio.ed.ac.uk/software/tracer/>) from the corresponding log (.log) and tree (.tree) files. These plots, along with the raw crude data, were once again exported and saved. The raw data were then imported into Excel and the percentages were calculated by standard mathematical techniques. Time resolved tree topologies were then inferred from the corresponding *gag* p24 and *pol* Bayesian data (.tree files) in TreeAnnotator v 1.7.3, which is part of the BEAST software package (http://beast.bio.ed.ac.uk/Main_Page). Each of the time resolved tree topologies along with the corresponding PLTT were combined in PowerPoint in a single diagram for ease of interpretation.

2.4.2 Establishing the evolutionary relationship of Cape Town isolates

The Cape Town data sets were compared with other HIV-1 subtype C homologous sequences. This was accomplished by the inference of large-scale phylogenies to assess the evolutionary relationship between the sequences from Cape Town with other sequences from around the world. This involved four main steps, which are described in detail in the next four sub-sections.

2.4.2.1 Data Mining

Public sequence databases were searched for all homologous HIV-1 subtype C sequences, which could be used to compare the newly sequenced data from the Cape Town cohort in a basic phylogenetic investigation. For this purpose both *gag* p24 (1246 – 1727 bp relative to HXB2), *pol* (2264 – 3321 bp relative to HXB2), and *gag-pol* (1246 – 3321 bp relative to HXB2) homologous sequences were collected. All sequences were downloaded with the appropriate patient information that was available. All sequences were then assembled into the appropriate file formats (.fasta or .txt) and labelled.

A large number of sequences in these public sequence databases have been generated from a single patient over a long period of time. Therefore, it is possible to find multiple sequences, which have been generated from a single patient. Each of the three different data sets was imported into Se-AL v 2.0 (<http://tree.bio.ed.ac.uk/software/seal>) and exported in a text file format. The files were then imported into an Excel spreadsheet and arranged alphabetically. The alphabetical arrangement of taxa based on their unique patient ID (Accession number) simply allows for easier exclusion of duplicates. This alphabetical arrangement of taxa was then used to screen for sequences with a high degree of sequence identity (>98%) as identified with the HIV-1 Sequence Quality Analysis Tool (<http://bioafrica.mrc.ac.za/tools/pppweb.html>) from the bioafrica.net webpage.

The HIV-1 Sequence Quality Analysis Tool is normally used to screen newly genotyped isolates for any potential sequence contamination. This is done by comparing the genetic diversity between strains, with two highly similar strains (> 98%) being identified as possible contaminants. However, this method can also be used to identify multiple genotypes from the same patient, as the genetic diversity of these genotypes are very low, even if these sequences were generated from specimens that were sampled months or years apart. Therefore, this method was used to manually exclude all multiple sequences (duplicates) from a single patient in the final data set. The removal of multiple sequences from the same patient is of the utmost importance.

Even though multiple sequences may not have any serious effect on the inference of standard phylogenies, they may seriously impact on advanced Bayesian methods of inference as inter and intra host evolutionary rates may vary considerably. Since South African and Southern African genotypes were selected from these large HIV-1 subtype C data sets, duplicates may therefore significantly influence any of the Bayesian results in the subsequent analyses.

2.4.2.2 Sequence alignments and Editing

The Cape Town data sets (*gag* p24, *gag-pol*, and *pol*) were combined with all the “cleaned-up” homologous reference sequences that were obtained in the data mining section. Each of these files were then saved in the appropriate file format and labelled carefully for alignment.

ClustalW v 2.1 (<http://www.clustal.org/download/current>) [Thompson *et al.*, 1997] was used for the construction of each of the alignments. A quicktree method of alignment was used to speed up alignments due to the large number of taxa contained within the data sets. Each alignment file was then imported into Se-AL v 2.0 (<http://tree.bio.ed.ac.uk/software/seal>) for manual editing. Gaps were excluded from the alignment if the gaps were not present in more than 20% of the taxa in each of the alignments. Alignments were then manually edited until a perfect codon alignment was obtained. This is an extremely time consuming process and may sometimes take days of manual alignment for some data sets (depending on the size of the data set). However, this is an extremely important process since all further analysis will rely on these alignments and the quality of these analysis ultimately depend on the quality of the alignment that is used. After manual editing aligned files were exported in a nexus (.nex) and fasta (.fasta) file format and saved for later use.

2.4.2.3 Inference of large-scale phylogenies

Two different methods of tree inference were used to infer phylogenies for each of the three data sets. This was done in order to compare clustering of samples across the two different methods. No model test was performed on any of the alignments due to their large size. For example the *gag* p24 data set contained 1895 taxa while the *pol* data set contained 2333 taxa. Therefore, all phylogenies were inferred under a single model of nucleotide substitution (HKY85), which allows for base frequencies and Ts/Tv ratios to vary.

Large-scale Minimum Evolution (ME) trees were inferred in fastME (<http://www.atgc-montpellier.fr/fastme/binaries.php>) [Desper *et al.*, 2002] for each of the data sets. Firstly, the ME-

tree topologies were inferred with the HKY85 model of nucleotide substitution, estimated gamma shape parameter, and the Nearest Neighbor Interchange (NNI) of branch optimization. This was done in order to identify any possible miss aligned sequences. Final ME-tree topologies were then inferred with the use of the HKY85+G (alpha = 0.8) model of nucleotide substitution. Each of these final ME-phylogenies was constructed with the implementation of the Subtree Pruning and Regrafting (SPR) method of tree search optimization technique. Bootstrap resampling for all ME-SPR tree topologies were also performed with a total of 100 bootstrap replicates for each data set.

Maximum Likelihood (ML) tree topologies were also inferred in phyML v 3.0 [Guindon *et al.*, 2010] from the three large HIV-1 subtype C data sets in addition to the ME-tree topologies. These ML-tree topologies were inferred with an HKY85 model of nucleotide substitution, a proportion of invariant sites and estimation of a Gamma shape parameter. These ML-tree topologies were inferred with an approximate likelihood ratio test (aLRT) in order to obtain some support for the internal branching order of these phylogenies.

These ME- and ML-tree topologies were constructed using a MacPro computer with 32 processors and 64 GB of memory. Some of the ML-tree topologies took more than 40-days of computing time.

2.4.2.4 Analysis of large-scale phylogenies

Each of the newly constructed phylogenies was imported into FigTree v 1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree>) and manually investigated. The clustering of Cape Town sequences with other sequences was manually assessed through careful assessment of the branching patterns in each of the phylogenies.

ME- and ML-phylogenies were also examined with the use of the PhyloType (<http://www.phylotype.org>) online software application. The PhyloType application is a recently published application [Chevenet *et al.*, 2013], that allows for the quick, easy and unbiased analysis of large phylogenies that would normally have to be done manually, which is an extremely time consuming method. Briefly, all sequences were classified into three main categories based on their dates of sampling. All sequences in the large data sets that were generated from patient samples prior to the year 2000 (≤ 1999) were designated as “oldest” samples. Sequences from patient samples dating 2000 – 2005 were designated as “middle”, and sequences from patient samples dating from 2006 onwards were designated as “youngest”.

Sequences from patients with no known date of sampling were designated as “none”. The separation of taxa based on temporal classes, above their geographical classification, allows for a more in-depth look at the clustering of taxa in these large phylogenies. Taxa were carefully arranged, according to these dated criteria, in a PhyloType text file annotation. An illustration of a PhyloType file annotation is presented in Figure 2.3. These PhyloType annotated files along with the corresponding tree files (in Newick specific file formats) were used to assess sequence-clustering patterns based on geographical and temporal classification in the PhyloType application with a total of 1000 shuffling iterations in order to calculate p-values for each of the identified clades in the tree topologies.

```
Sequences , Country , Date , Country_Date
'C.AR.2001.ARG4006.AY563170','AR','Middle','AR_Middle'
'C.BR.2002.02BR2022.JN692434','BR','Middle','BR_Middle'
'C.BR.2004.04BR013.AY727522','BR','Middle','BR_Middle'
'C.BR.2004.04BR021.AY727523','BR','Middle','BR_Middle'
'C.BR.2004.04BR038.AY727524','BR','Middle','BR_Middle'
'C.BR.2004.04BR073.AY727525','BR','Middle','BR_Middle'
'C.BW.2000.00BW07621.AF443088','BW','Middle','BW_Middle'
'C.BW.2000.00BW076820.AF443089','BW','Middle','BW_Middle'
'C.BW.2000.00BW087421.AF443090','BW','Middle','BW_Middle'
'C.BW.2000.00BW147127.AF443091','BW','Middle','BW_Middle'
'C.BW.2000.00BW16162.AF443092','BW','Middle','BW_Middle'
'C.BW.2000.00BW1686.AF443093','BW','Middle','BW_Middle'
'C.BW.2000.00BW17593.AF443094','BW','Middle','BW_Middle'
'C.BW.2000.00BW17732.AF443095','BW','Middle','BW_Middle'
'C.BW.2000.00BW17835.AF443096','BW','Middle','BW_Middle'
```

Figure 2.3: An illustration of a PhyloType file annotation. This particular file annotation contains the taxa (sequence) of each isolate, the country code, the date category of the particular taxa and a country date combination. This is the query file annotation that PhyloType use to search the tree topology. This particular file annotation was generated in TextWrangler v 4.0.1.

Analyses of these large phylogenies based only on their geographical classification were performed at a later stage (section 2.4.4.1).

2.4.3 Phylogenetic investigation of transmission events

During the course of the basic phylogenetic investigation of sequence data a large number of sequences from Cape Town clustered closely together. Such monophyletic clustering patterns are very unusual. Particularly in the context of a generalized HIV-1 subtype C epidemic in a country such as South Africa. As such, these are the first large monophyletic clusters detected of subtype C isolates in South Africa.

These sequences that clustered in these monophyletic clusters however are not identical. Patient samples clustered in these monophyletic clusters in both the *gag* p24, and *pol* in tree topologies as well as the *gag-pol* (for those select few patient samples that were represented in both the *gag*

and *pol* data sets) phylogenies. Similarly, these patient samples were sampled at different time points and their genotypes have been generated in different studies [Jacobs *et al.*, 2006; Wilkinson *et al.*, in press; Isaacs *et al.*, submitted], which dispel possible sequence contamination. Additionally, the previous mentioned analysis (in section 2.4.2.1) is normally used to exclude sequence contamination and therefore any potential sequence contamination would have been detected in these analyses. Furthermore, other South African sequences that cluster closely with the Cape Town isolates in these phylogenies were identified through retrospective analysis to also have originated from patients in the greater Cape Metropolitan area. Lastly, the majority of the patient samples were obtained from a few hospitals or clinics in the Cape Metropolitan area, which was then sent to the Tygerberg Academic Hospital for diagnostic testing, and therefore did not originate from a single source. Therefore it is highly possible that these monophyletic clusters may represent transmission events of HIV-1 amongst local communities in the Cape Town region.

2.4.3.1 Identifying and testing potential clusters

Each of the large-scale *gag* p24 and *pol* phylogenies that were constructed in the basic phylogenetic investigation in section 2.4.2 was imported into FigTree v 1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree/>) and manually investigated for potential transmission clusters of Cape Town sequences. Traditional knowledge regarding transmission events, based on findings in the established scientific literature (e.g. the length of internal and external branch lengths) was used to assess potential clusters. This method of analysing evolutionary trees is still regarded as the best way of identifying large clusters of similarity.

Each of the inferred large-scale phylogenies was also assessed with the use of the PhyloType application (<http://lamarck.lirmm.fr/phyloptype/>) to identify any potential clusters within the phylogenies. A total of 1000 shuffling iterations were used to provide statistical confidence in the assessment of the transmission clusters.

During the manual and automatic scanning of tree topologies several Cape Town clusters were observed, however no adequate bootstrap support was obtained for the internal branches of these clusters. Similarly, adequate support values for more established HIV-1 subtype C clusters (e.g. the Indian and Brazilian clades) were also very poor. Due to the lack of adequate support for the internal branches of the putative transmission clusters, additional analyses were performed in order to investigate whether these clusters could be trusted.

Therefore, additional clustering analyses were set up in order to assess the validity of these putative clusters and investigate whether they hold up against a different reference set. Each potential cluster was analysed against a reference set of HIV-1 sequences in five different phylogenies. For this the full full-length HIV-1 subtype C reference set was obtained from the Los Alamos National Laboratory (LANL) HIV-1 sequence database (<http://www.hiv.lanl.gov/content/index>) along with the HXB2 reference strain of HIV-1, which were used to root phylogenies in the subsequent analyses. Multiple sequence alignments were constructed for each of the putative transmission clusters, along with homologous reference strains in ClustalW (<http://www.clustal.org/clustal2/>) [Thompson *et al.*, 1997]. Each alignment was then manually edited in Se-Al v 2.0 (<http://www.tree.bio.ed.ac.uk/software/seal/>). Gaps were excluded from the alignment if the gaps were present in more than 20% of the taxa in each of the alignments. Sequences were manually edited until a perfect codon alignment was achieved.

Five different tree topologies were inferred for each of these alignments. This included a Neighbor-Joining (NJ) tree topology and a Minimum Evolution (ME) tree topology, which were both inferred with the K2P model of nucleotide substitution, an estimated gamma shape parameter, and 1000 bootstrap replicates. The NJ-tree topologies were inferred in MEGA v 5.0 [Tamura *et al.*, 2011] while the ME-tree topologies were inferred in fastME [Desper and Gascuel, 2002]. Furthermore, two Maximum Likelihood (ML) tree topologies were inferred for each data set in phyML v 3.0 [Guindon *et al.*, 2010], with the use of the HKY85+G (G = 0.8) model of nucleotide substitution. One of the ML-tree topologies was inferred with aLRT while the other was inferred with bootstrap resampling (n = 1000). Finally, a Bayesian tree topology was also inferred for each of the alignments in MrBayes [Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003] with the GTR+G+I model of nucleotide substitution. Posterior tree files were summarized in TreeAnnotator v 1.7.4, which is part of the BEAST software package (Drummond and Rambaut, 2007).

All trees were inspected in FigTree v 1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree/>) to assess the support for internal branches of the clusters. Additionally, all of these phylogenies were also analysed with the PhyloType application (<http://lamarck.lirmm.fr/phyloptype/>) to identify clusters, and to assess support for each of these clusters. Only Cape Town isolates that consistently clustered within the monophyletic clades were accepted as “true” transmission clusters, while any taxa that cluster outside of these clades were excluded from any of the subsequent analyses.

2.4.3.2 Timing the internal nodes of monophyletic clades

The estimated root height (date of origin or tMRCA) of the internal nodes of each of the identified clades was calculated with the use of standard Bayesian inference methods.

Briefly, each of the taxa in the clusters was annotated according to the specific year, month and day of sampling (represented as a fraction of a sampling year e.g. 1990.67), which was manually calculated in Excel. Sequence alignments of each of the clusters were performed in ClustalW v 2.1 with a quicktree method of alignment (<http://www.clustal.org/download/current>) and manually edited in Se-AI v 2.0 (<http://tree.bio.ed.ac.uk/software/seal>) to obtain a perfect codon alignment. Each of the alignments was then exported and saved in the appropriate file format (.nex). Dated tree topologies, evolutionary rates and root heights (tMRCA of each cluster) were co-estimated with the implementation of a standard Bayesian MCMC approach, which was executed in BEAST v 1.7.3 (http://beast.bio.ed.ac.uk/Main_Page) [Drummond and Rambaut, 2007] with a the use of the SRD06 model of nucleotide substitution [Shapiro *et al.*, 2006]. Two parametric (Constant Population Size and Exponential Growth) and one non-parametric (BSP) tree prior, was compared under a strict and relaxed molecular clock conditions. Each of the run parameters was analysed under both a fixed and estimated mutation rates. For the *gag* p24 data sets the mutation rate were fixed at $3,0 \times 10^{-3}$ mutations/site/year [Novitsky *et al.*, 2010] and for the *pol* analyses the mutation rate were fixed at $2,55 \times 10^{-3}$ mutations/site/year [Hue *et al.*, 2004]. Chains in the MCMC were run for 30 million generations and sampled every 3000 steps in the chain. Convergence in each of the runs was assessed with the use of Tracer v 1.5 (<http://tree.bio.ed.ac.uk/software/tracer/>) on the basis of the effective sample size (EES) after a 10% burn-in. Uncertainty in the estimation was indicated by 95% highest posterior density (95% HPD) intervals. Bayes factor comparisons were conducted for all runs for each of the various data sets in Tracer v 1.5 [Suchard *et al.*, 2001].

2.4.3.3 Time resolved tree topologies with clustering of the various monophyletic clades

Time resolved phylogenies of the entire *gag* p24 and *pol* Cape Town data sets were also constructed.

Briefly, sequence alignments of the various (Cape Town only) data sets were constructed in ClustalW v 2.1 (<http://www.clustal.org/download/current/>) and manually edited in Se-AI v 2.0 (<http://tree.bio.ed.ac.uk/software/seal/>) until a perfect codon alignment was obtained. Each of the

alignments was then exported and saved for later use. Each of the sequences within the clusters was forced into monophyletic clades in BEAUTi (http://beast.bio.ed.ac.uk/Main_Page).

A Bayesian Markov Chain Monte Carlo (MCMC) approach, executed in BEAST v 1.7.3 (http://beast.bio.ed.ac.uk/Main_Page), was used for the estimation of dated tree topologies, evolutionary rates and population growth rates. Each of the runs was conducted with the use of the SPR06 model of nucleotide substitution. Two parametric models (constant population size and exponential growth) and one non-parametric model (Bayesian skyline plot) were compared under strict and relaxed molecular clock conditions. Each of the run parameters was analysed with the use of both a fixed and estimated mutation rate. For the *gag* p24 analysis the mutation rate was fixed at $3,00 \times 10^{-3}$ mutations/site/year [Novitsky *et al.*, 2010] and for the *pol* analysis the mutation rate was fixed at $2,55 \times 10^{-3}$ mutations/site/year [Hue *et al.*, 2004]. Two independent runs were performed for each model tree prior. A total of 50 million steps was run in the MCMC and sampled every 5000 steps. Convergence in each of the runs was assessed in Tracer v 1.5 (<http://tree.bio.ed.ac.uk/software/tracer/>) on the basis of the effective sample size (ESS). Uncertainty in the estimation was indicated by 95% highest posterior density (95% HPD) intervals.

The posterior distribution of trees were summarized in a target tree with the use of TreeAnnotator v 1.7.3 program, included in the BEAST software package (http://beast.bio.ed.ac.uk/Main_Page), by choosing the tree with the maximum product of posterior probabilities after a 10% burn-in. Each of the time resolved tree topologies were examined in FigTree v 1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree/>) and manually edited for better interpretation and visualization.

CHAPTER THREE - TABLE OF CONTENTS

	Page
3.1 Results of the selection of patient samples	89
3.2 Results of the genotyping of patient samples	90
3.2.1 Amplification of <i>gag</i> p24 and <i>pol</i> fragments from Cape Town	90
3.2.2 Sequencing and the composition of the final sequence data sets from Cape Town	90
3.2.3 Results of the subtyping of various sequence data sets	90
3.2.4 The results of the regression analyses of root-to-tip divergence of the Cape Town data sets	91
3.3 Results of the HIV-1 subtype C epidemic reconstruction	92
3.3.1 Outcome of the epidemic reconstruction of the HIV-1 subtype C epidemic in Cape Town	92
3.3.1.1 Estimated tMRCA of the Cape Town epidemic	92
3.3.1.2 Phylodynamic aspects of the Cape Town data sets	96
3.3.1.3 Estimates of the percentage lineages through time in Cape Town	98
3.3.2 Outcome of the epidemic reconstruction of the Southern African epidemic (excluding sequence information from the Cape Town data sets)	102
3.3.2.1 The estimated tMRCA of the Southern African epidemic (excluding sequence information from the Cape Town data sets)	102
3.3.2.2 Results of the dynamic epidemic reconstruction from Southern Africa sequences (excluding Cape Town)	110

3.3.3 Southern African data sets (including Cape Town)	112
3.3.3.1 tMRCA of the Southern African epidemic (including Cape Town)	112
3.3.3.2 Outcome of the dynamic reconstruction of the Southern Africa epidemic (including sequence information from Cape Town)	119
3.4 Results of the basic phylogenetic investigation into the evolutionary relationship of the Cape Town isolates	121
3.4.1 Outcome of the data mining	121
3.4.2 Large-scale phylogenetic inference	123
3.4.3 Outcome of the PhyloType analysis of large tree topologies	133
3.5 Large monophyletic clades of HIV-1 subtype C in Cape Town	144
3.5.1 Identification of monophyletic clusters	145
3.5.2 Testing putative transmission clusters	150
3.5.3 Timing the root height of the internal nodes of each cluster	154
3.5.4 Time resolved tree topologies with transmission clusters	158

CHAPTER THREE

The following chapter contains the results of the study and are organised in several parts: (1) Results on sample selection and the genotyping of patient samples. This part (section 3.1) will cover the selection of samples from cold storage for genotyping. (2) Results on the genotyping of patient samples. This part (section 3.2) will cover the genotyping of the selected patient samples for the generation of longitudinal *gag* p24 and *pol* sequence data sets from Cape Town. (3) Results of the dating analyses. This part (section 3.3) will cover the results of the viral subtyping, results of the molecular clock analyses of the Cape Town sequence data sets, the results of the evolutionary reconstruction of the Cape Town HIV-1 subtype C epidemic and the results of the evolutionary reconstruction of the Southern African HIV-1 subtype C epidemic. (4) This will be followed by results from the basic phylogenetic investigation (section 3.4), which looked at the evolutionary relationship of the isolates from the Cape Town data sets with other HIV-1 subtype C isolates from around the world. (5) Results of the transmission cluster analyses. This part (section 3.5) will contain results on the identification, verification and analyses of putative transmission clusters of HIV-1 subtype C in Cape Town.

These results that are presented in this chapter will then be discussed and compared with one another, as well as with the findings from other studies in the established scientific literature, in the following chapter.

3.1 Results of the selection of patient samples

Several HIV-1 subtype C *gag* p24 and *pol* sequences have been characterized in the past from patients in the Cape Town area. These sequences, which were generated within the Division of Medical Virology at the Tygerberg Academic Hospital, were selected for inclusion into the study. A total of 168 previously characterized *gag* p24 sequences (2002 – 2010), and 92 *pol* sequences (2008 – 2010), were obtained by this method. These patient sequences were characterized by either myself or by other students within the Division of Medical Virology [Jacobs *et al.*, 2006; Wilkinson *et al.*, 2013 in press; Isaacs *et al.*, in preparation].

Additionally, 250 patient samples were selected from the -20°C freezers for the generation of new sequences, particularly from the very early years of the HIV-1 subtype C epidemic in Cape Town. This was done in order to achieve comprehensive longitudinally sampled sequence data sets that would span over roughly a 21-year time period.

3.2 Results of the genotyping of patient samples

3.2.1 Amplification of *gag* p24 and *pol* fragments from Cape Town

The amplification of *gag* p24 and *pol* genomic fragments from isolated nucleic acid (either RNA & DNA) produced a total of 49 PCR positive *gag* p24 fragments (spanning between 1989 and 1993) and 110 PCR positive *pol* fragments (spanning between 1989 and 2004). No positive PCR products were obtained for the 34 RNA samples from 1996, following extensive methods to optimize the PCR conditions (e.g. changing the concentrations of MgCl₂ and/or slightly adjusting the annealing temperatures of primers). The 110 successfully amplified *pol* fragments only refer to isolates for which both the PR- and RT PCR-*pol* amplification assays were positive. This small number of characterized patient samples (particularly from the very oldest samples) represents only a small number of the patient samples that were selected for characterization (n = 250). The characterization of such old patient samples are extremely difficult and time consuming since the storage of these samples, over long periods of time, may have led to a considerable degree of sample degradation.

3.2.2 Sequencing and the composition of the final sequence data sets from Cape Town

Sequencing of amplified *gag* p24 and *pol* fragments produced 25 new *gag* p24 sequences (1989 - 1992) and 74 new *pol* sequences (1989 - 2004). The newly characterized *gag* p24 and *pol* sequences were combined with the previously characterized sequences from Cape Town, that were genotyped as part of other studies within the Division of Medical Virology (Tygerberg Hospital). This produced a final *gag* p24 data set, containing a total of a 193 taxa, and a final *pol* data set of 166 sequences. Only 52 patient samples were represented in both the *gag* p24 and *pol* data set. These 52 sequences were also arranged into a different data set to produce a concatenated *gag-pol* data set. A full list of *gag* p24, concatenated *gag-pol* and *pol* sequences that were generated from Cape Town patients, along with relevant demographic and health information are presented in Tables 6.1 to 6.3 in Appendix A (page 223 - 235).

3.2.3 Results of the subtyping of various sequence data sets

All of the *gag* p24 and *pol* sequences from Cape Town were subtyped with two online HIV-1 subtyping tools. All of the *gag* p24 and *pol* sequences contained in the two data sets were classified with high bootstrap support (> 70%) as HIV-1 subtype C isolates with the REGA v 2.0

subtyping tool (<http://www.bioafrica.net/rega-genotype/html/subtypinghiv.htm>). Similarly, all isolates were also classified with high posterior probability values (> 0.9) as HIV-1 subtype C isolates with the jumping profile Hidden Markov Model (jpHMM) method from GOBICS (<http://jphmm.gobics.de>).

3.2.4 The results of the regression analyses of root-to-tip divergence of the Cape Town data sets

The *gag* p24 and *pol* data sets were analysed with the Path-O-Gen application (<http://tree.bio.ed.ac.uk/software/pathogen/>) in order to assess the degree of temporal signal and molecular clock likeness of each of the data sets.

The Path-O-Gen analysis of the final Cape Town *gag* p24 data set (data shown in Figure 6.1 on page 242) placed the estimated tMRCA (based on the crude root-to-tips regression) around 1953.50 with an R squared (R^2) value of 0.213. The estimated mutation rate of the *gag* p24 Cape Town data set (calculated from the crude root-to-tip regression in Path-O-Gen) was estimated at around 2.79×10^{-3} mutations/site/year.

Similarly, the Path-O-Gen analysis of the final Cape Town *pol* data set (Figure 3.1) placed the estimated tMRCA for the data set (based on the crude root-to-tips regression) around 1952.77 with a calculated R^2 value of 0.294. The mutation rates of the *pol* Cape Town data set were estimated at around 2.23×10^{-3} mutations/site/year based on the root-to-tip regression in Path-O-Gen. The preliminary molecular clock analysis of both the *gag* p24 and *pol* data sets, suggest that both data sets holds enough genetic information and does evolve in a clock like manner.

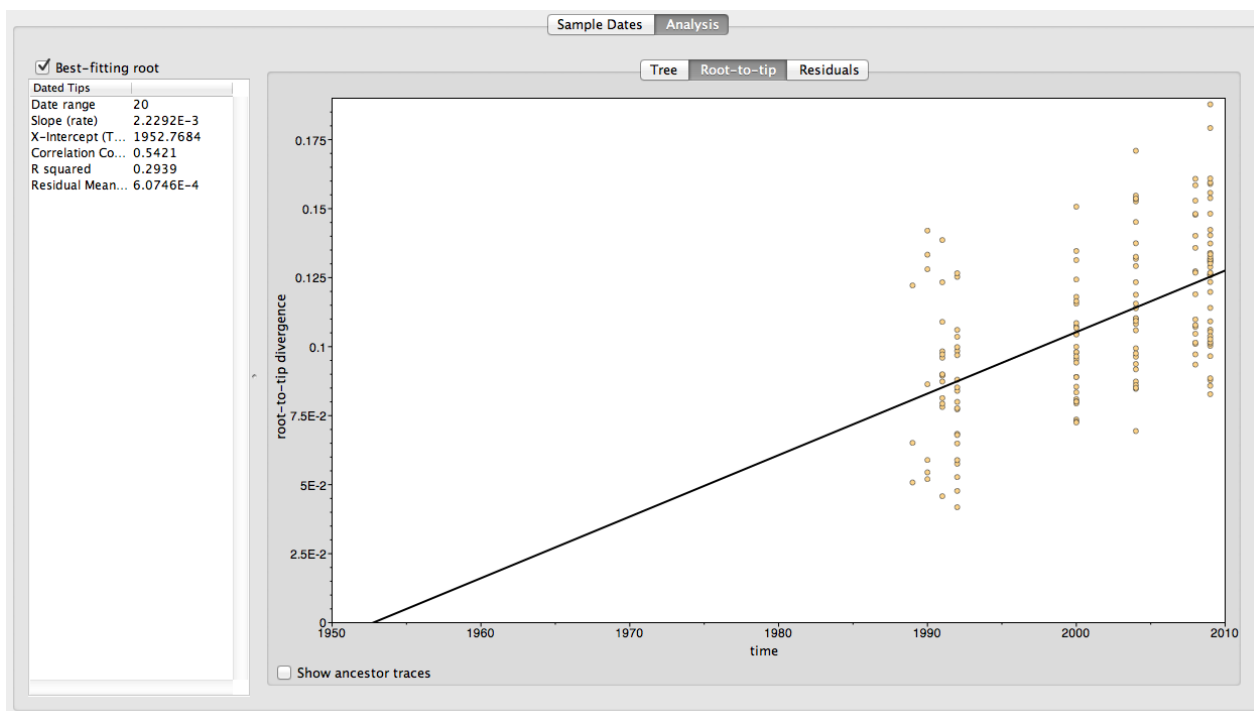


Figure 3.1: The root-to-tip regression analysis of the *pol* Cape Town data set. The 166 *pol* sequences from Cape Town were analysed in Path-O-Gen v 1.3. The ML-tree that was used for the analysis was constructed in PhyML with the HKY+G+I model of nucleotide substitution and the SPR method of branch swapping was utilized. The average R^2 of the data set was 0.29, the estimated tMRCA calculated from the crude root to tip regression was 1952.77 and the mutation rate (slope rate) is 2.23×10^{-3} mutations/site/year.

3.3 Results of the HIV-1 subtype C epidemic reconstruction

3.3.1 Outcome of the epidemic reconstruction of the HIV-1 subtype C epidemic in Cape Town

HIV epidemic reconstruction was undertaken to investigate the evolutionary history of the Cape Town subtype C HIV-1 epidemic. Firstly, the estimated tMRCA was inferred from the three different Cape Town data sets. Secondly, the phylodynamic aspect of the Cape Town epidemic was investigated through dynamic construction from selected *gag* p24 and *pol* data runs.

3.3.1.1 Estimated tMRCA of the Cape Town epidemic

Three different genomic regions (*gag* p24, *gag-pol* concatenated, and *pol*) were used to infer the evolutionary history of the Cape Town epidemic. Manual inspection of each of the log files was assessed in Tracer and good convergence in the Markov Chains were observed for all the runs. The estimated tMRCA of the Cape Town epidemic that was inferred from the three different

genomic regions (*gag* p24, partial *pol* and a concatenated *gag-pol* data set) gave very similar results (Tables 3.1 – 3.3).

The inferred tMRCA of the Cape Town epidemic from the *gag* p24 data set (Table 3.1) ranged between 1952,5 (95% HPD 1935,9 – 1967,0) and 1973,7 (95% HPD 1969,2 – 1977,7). The calculation of Bayes factors identified **Const.relax.est.2** as the best fitting tree prior used for the *gag* p24 Cape Town data set (data shown in Table 6.5 in the Appendix D on page 238). The estimated mean tMRCA of this model was 1966,6 with the 95% HPD ranging between 1956,4 and 1975,8 while the mean estimated mutation rate for this model tree prior was $3,1 \times 10^{-3}$ mutations/site/year with the 95% HPD interval ranging between $2,3 \times 10^{-3}$ and $3,8 \times 10^{-3}$ mutations/site/year. The average coefficient of variation for the *gag* p24 Cape Town relaxed runs was 0,33 for the BSP model and 0,47 for the constant population size tree prior. These coefficients of variation indicate that there is a very small variation in the evolutionary rate amongst the different branches in the tree topology irrespective of the evolutionary model that was employed.

Table 3.1: The estimated tMRCA and mutation rates for the Cape Town *gag* p24 data set. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Estimated tMRCA					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	ESS
BSP.relaxed.est.1	1952,6	1953,9	1967,8	1935,2	669,6
BSP.relaxed.est.2	1954,2	1955,5	1968,7	1937,6	533,3
BSP.relaxed.fix.1	1973,4	1973,7	1977,8	1968,7	1030,5
BSP.relaxed.fix.2	1973,4	1973,6	1977,7	1968,7	400,2
BSP.strict.est.1	1952,8	1953,9	1966,5	1936,8	808,7
BSP.strict.est.2	1952,5	1953,6	1967,0	1935,9	827,7
BSP.strict.fix.1	1973,7	1973,9	1977,6	1969,4	1164
BSP.strict.fix.2	1973,7	1973,9	1977,7	1969,2	1262,9
Const.relax.est.1	1967,7	1968,5	1976,3	1958,2	539,0
Const.relax.est.2	1966,6	1967,4	1975,8	1956,4	351,4
Const.relax.fix.1	1968,0	1968,5	1973,6	1961,3	819,2
Const.relax.fix.2	1967,3	1967,9	1973,5	1959,5	914,2
Average	1964,7	1965,4	1973,3	1954,7	776,7
Estimated mutation rates					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	ESS
BSP.relaxed.est.1	1,30E-03	1,30E-03	9,50E-04	1,70E-03	598,9
BSP.relaxed.est.2	1,40E-03	1,40E-03	1,00E-03	1,90E-03	117,6
BSP.strict.est.1	1,30E-03	1,30E-03	9,70E-04	1,70E-03	55,1
BSP.strict.est.2	1,30E-03	1,30E-03	9,70E-04	1,70E-03	663,5
Const.relax.est.1	3,10E-03	3,00E-03	2,30E-03	3,80E-03	253,6
Const.relax.est.2	2,90E-03	2,90E-03	2,20E-03	3,70E-03	136,8

tMRCA – time to the Most Recent Common Ancestor; HPD – Highest Posterior Density; ESS – Effective Sample Size; BSP – Bayesian Skyline Plot, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Constant – Constant Population Size

Similarly, the inferred tMRCA of the Cape Town epidemic from the concatenated *gag-pol* data set (Table 3.2) ranged between 1941,4 (95% HPD 1917,1 – 1960,2) and 1973,4 (95% HPD 1970,0 – 1976,6). The calculation of Bayes factors identified **BSP.relax.est.1** as the best fitting tree prior used for the *gag-pol* Cape Town data set (data shown in Table 6.6 in the Appendix D on page 238). The estimated mean tMRCA of this model was 1969,9 with the 95% HPD ranging between 1961,0 and 1977,9. The mean estimated mutation rate for this model tree prior was $2,5 \times 10^{-3}$ mutations/site/year with the 95% HPD interval ranging between $1,6 \times 10^{-3}$ and $3,3 \times 10^{-3}$. The average coefficient of variation for the *gag-pol* concatenated Cape Town relaxed runs was 0,32 for the BSP model and 0,49 for the constant population size tree prior, which suggest that there is a very small variation in the evolutionary rates amongst the different branches in the tree topology irrespective of the various models that were employed.

Table 3.2: The estimated tMRCA and mutation rates for the Cape Town *gag-pol* concatenated data set. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Estimated tMRCA					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	ESS
BSP.relax.fix.1	1940,9	1943,7	1962,0	1914,4	1166,1
BSP.relax.fix.2	1941,3	1944,3	1962,4	1915,1	1175,7
BSP.relax.est.1	1969,9	1970,6	1977,9	1961,0	840,1
BSP.relax.est.2	1968,4	1969,8	1978,1	1956,6	128,8
BSP.strict.fix.1	1972,4	1972,5	1975,2	1969,3	4429,3
BSP.strict.fix.2	1972,3	1972,4	1975,0	1969,1	5403,4
BSP.strict.est.1	1941,7	1944,1	1961,0	1918,1	2024,2
BSP.strict.est.2	1941,4	1943,7	1960,2	1917,1	1940,9
Const.relax.fix.1	1971,3	1971,7	1975,7	1965,9	2747,1
Const.relax.fix.2	1970,9	1971,3	1975,6	1965,8	3110,8
Const.relax.est.1	1973,3	1973,4	1976,5	1969,9	4742,8
Const.relax.est.2	1973,4	1973,5	1976,6	1970,0	6003,6
Average	1961,4	1962,6	1971,4	1949,4	2809,4
Estimated mutation rates					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	ESS
BSP.relax.est.1	2,50E-03	2,40E-03	1,60E-03	3,30E-03	887,2
BSP.relax.est.2	2,40E-03	2,40E-03	1,40E-03	3,30E-03	162,7
BSP.strict.est.1	9,90E-04	1,00E-03	6,30E-04	1,30E-03	1462,8
BSP.strict.est.2	9,90E-04	9,90E-04	6,10E-04	1,30E-03	1221
Const.relax.est.1	9,90E-04	1,00E-03	6,00E-04	1,40E-03	1004,9
Const.relax.est.2	1,00E-03	1,00E-03	6,10E-04	1,40E-03	823,6

tMRCA – time to the Most Recent Common Ancestor; HPD – Highest Posterior Density; ESS – Effective Sample Size; BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

From the *pol* data set the inferred tMRCA of the Cape Town epidemic (Table 3.3) was estimated between 1951,4 (with the 95% HPD ranging between 1937,0 and 1964,7) and 1969,9 (with the 95% HPD ranging between 1965,1 and 1974,3). Bayes factor comparison identified **Const.relax.est.1** as the best fitting model tree prior used for the *pol* Cape Town data set (data presented in Table 6.7 in the Appendix D on page 239). For this model the mean tMRCA was 1964,8 with the 95% HPD ranging between 1955,3 and 1973,5. The mean inferred mutation rate for this model was estimated at $2,1 \times 10^{-3}$ mutation/site/year (with the 95% HPD ranging between $1,8 \times 10^{-3}$ and $2,5 \times 10^{-3}$). The median coefficients of variation were 0,34 and 0,48 for the relaxed BSP and constant population size tree priors. This suggests that there is a very small variation in the evolutionary rates amongst the different branches in the tree topology irrespective of the various models that were employed.

Table 3.3: The estimated tMRCA and mutation rates for the Cape Town *pol* data set. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Estimated tMRCA					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	ESS
BSP.relax.est.1	1951,4	1952,4	1964,7	1937,0	772,4
BSP.relax.est.2	1951,6	1952,6	1963,7	1937,0	1118,4
BSP.relax.fix.1	1969,9	1970,1	1974,3	1965,1	362,6
BSP.relax.fix.2	1969,6	1969,9	1974,2	1964,7	171,9
BSP.strict.est.1	1951,6	1952,1	1961,2	1940,6	725,1
BSP.strict.est.2	1951,3	1951,9	1961,0	1940,8	582,0
BSP.strict.fix.1	1966,4	1966,6	1971,3	1961,2	1363,7
BSP.strict.fix.2	1966,2	1966,4	1971,1	1961,2	2324,3
Const.relax.est.1	1964,8	1965,5	1973,5	1955,3	380,0
Const.relax.est.2	1964,9	1965,6	1973,7	1955,1	365,2
Const.relax.fix.1	1960,5	1961,1	1968,4	1951,4	1591,3
Const.relax.fix.2	1960,7	1961,3	1968,5	1952,6	1020,4
Const.strict.est.1	1960,2	1960,5	1967,5	1952,5	939,4
Const.strict.est.2	1960,0	1960,3	1967,3	1952,2	638,4
Const.strict.fix.1	1965,1	1965,3	1969,4	1960,4	5117,8
Const.strict.fix.2	1964,9	1965,0	1969,4	1960,1	3732,4
Average	1961,2	1961,7	1968,7	1953,0	1325,3
Estimated mutation rates					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	ESS
BSP.relax.est.1	1,40E-03	1,40E-03	9,90E-04	1,70E-03	404,8
BSP.relax.est.2	1,40E-03	1,40E-03	9,90E-04	1,70E-03	872,3
BSP.strict.est.1	3,00E-03	3,00E-03	2,30E-03	3,60E-03	149,2
BSP.strict.est.2	3,00E-03	3,00E-03	2,40E-03	3,70E-03	172,6
Const.relax.est.1	2,10E-03	2,10E-03	1,80E-03	2,50E-03	456,1
Const.relax.est.2	2,10E-03	2,10E-03	1,80E-03	2,50E-03	395,2
Const.strict.est.1	1,40E-03	1,40E-03	1,10E-03	1,70E-03	332,3
Const.strict.est.2	1,40E-03	1,40E-03	1,10E-03	1,70E-03	370,9

tMRCA – time to the Most Recent Common Ancestor; HPD – Highest Posterior Density; ESS – Effective Sample Size; BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

In summary, the dating analyses from the three different Cape Town data sets placed the estimated date of origin for the Cape Town/South African HIV-1 subtype C epidemic around the mid 1960's with the 95% HPD interval ranging from 1955 to 1977.

3.3.1.2 Phylodynamic aspects of the Cape Town data sets

Phylodynamic reconstruction of the demographic history of the Cape Town HIV-1 subtype C epidemic was achieved through non-parametric remodelling from various *gag* p24 and *pol* runs.

Close examination of the reconstructed BSP under a relaxed molecular clock assumption from the *gag* p24 data set (Figure 3.2) revealed a relatively smooth BSP with a linear growth over

roughly a 30-year period (1980 – 2010). A brief period of exponential growth in the EPS is visible in the mid 1980's (1984 – 1987). Another possible brief period of exponential growth can be seen in the mid 1990's (1994 – 1997), particularly when looking at the growth trajectory of the 95% lower HPD.

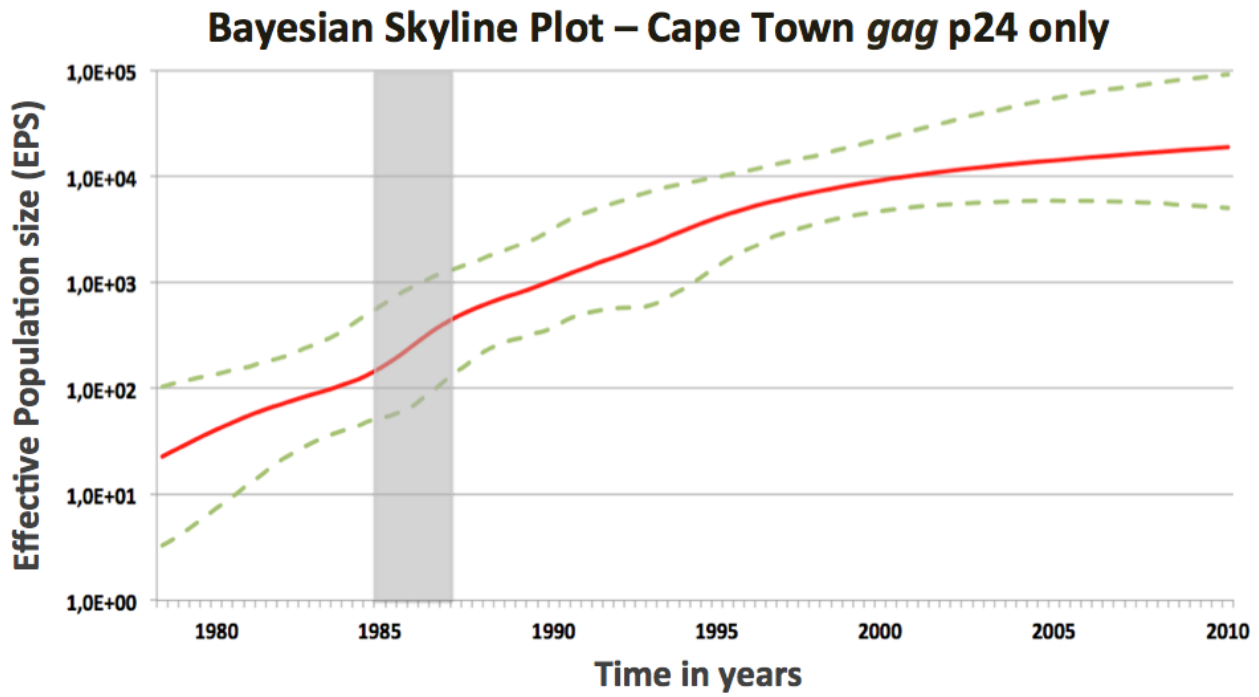


Figure 3.2: Bayesian skyline plot of Cape Town HIV-1 subtype C *gag* p24 data sets. This non-parametric estimate of demographic history was reconstructed in Excel from raw data. The solid red lines are the “traced” median effective population size with the 95% upper and lower highest posterior density intervals indicated in the green dashed lines. This plot starts in the late 1970's and stretches over roughly a 30-year period (1980 - 2010). The period marked in grey indicates period(s) of exponential growth in the effective population size.

Similarly, the BSP from the *pol* data set (Figure 3.3) revealed a strong linear growth in the HIV-1 subtype C epidemic in Cape Town from the middle of the 1970's till the late 1980's, with a period of exponential growth in the late 1980's and early 1990's. A plateau, or even a slight decline (as seen in the lower 95% HPD), in the rate of epidemic growth, was observed from the mid 1990's till the present.

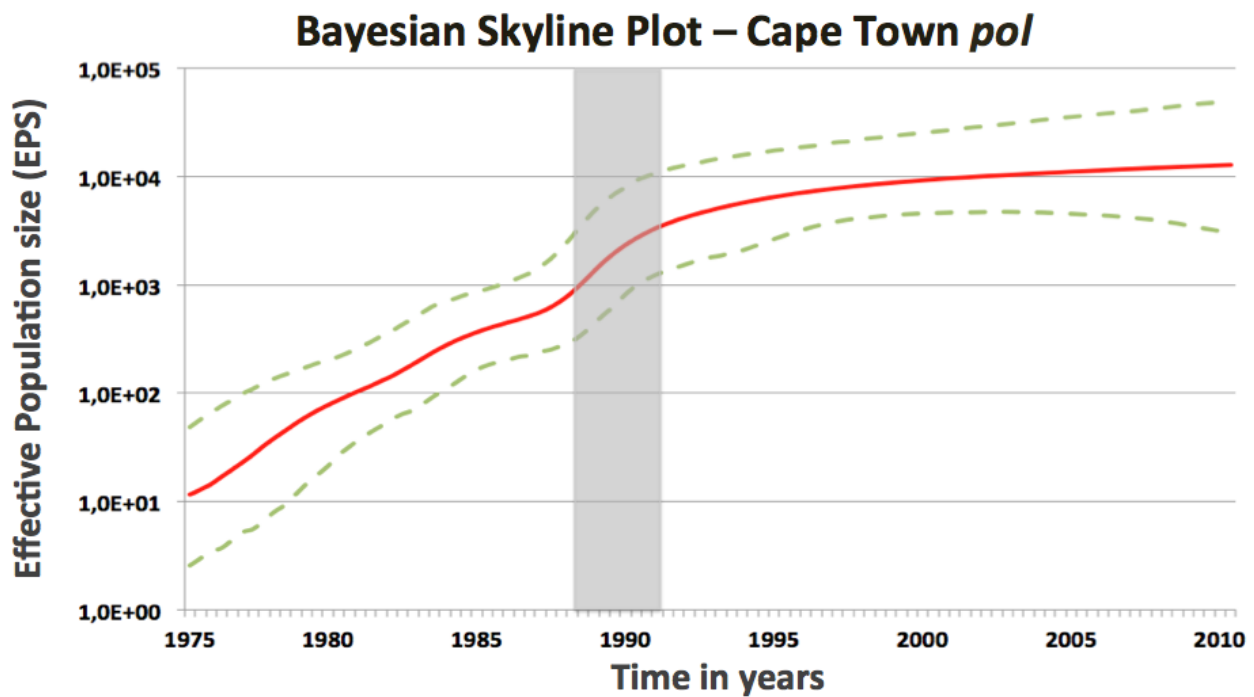


Figure 3.3: Bayesian skyline plot of Cape Town HIV-1 subtype C *pol* data sets. This non-parametric estimate of demographic history was reconstructed in Excel from raw data. The solid red lines are the “traced” median effective population size with the 95% upper and lower highest posterior density intervals indicated in the green dashed lines. This plot begins in the mid 1970’s and stretches over roughly a 35-year period (1975 – 2010). The period marked in grey indicates period(s) of exponential growth in the effective population size.

In summary, the dynamic reconstruction of the HIV-1 subtype C epidemic from the two Cape Town data sets (*gag* p24 and the partial *pol*) suggests a slow linear growth in the epidemic in Cape Town/South Africa since the start of the epidemic till the mid to late 1990’s. Small periods of exponential growth in the effective population size were observed during the mid to late 1980’s. Since the turn of the century it would appear that the growth of the epidemic has stabilized.

3.3.1.3 Estimates of the percentage lineages through time in Cape Town

The estimated percentage lineages through time, as well as time resolved tree topologies, were also inferred from selected runs for both the *gag* p24 and *pol* data sets. Each of the inferred time resolved tree topologies were combined with the median estimated percentage lineages through time.

The estimated percentage lineages through time (PLTT) of the *gag* p24 inferred data (Figure 3.4) reveal a slow increase in the genetic diversity in the period before 1980. This was followed by a massive increase in genetic variation during the course of the 1980’s. By the start of the 1990’s

an estimated 79,2% of current genetic isolates were already present within the Cape Town region or the surrounding environment. Since then the genetic diversity has continued to expand till the present day.

The estimated percentage lineages through time (PLTT) that were inferred from the *pol* Cape Town sequence data (Figure 3.5) reveals a slow increase in the genetic diversity in the early years of the epidemic in Cape Town (1965 - 1980). This period of slow viral genetic expansion was followed by a massive increase in genetic variation during the 1980's. By the end of the 1980's more than 80% of the present day genetic variants were already circulating amongst the infected population of Cape Town. Since the start of the 1990's the genetic diversity of HIV-1 subtype C has continued to increase till the present.

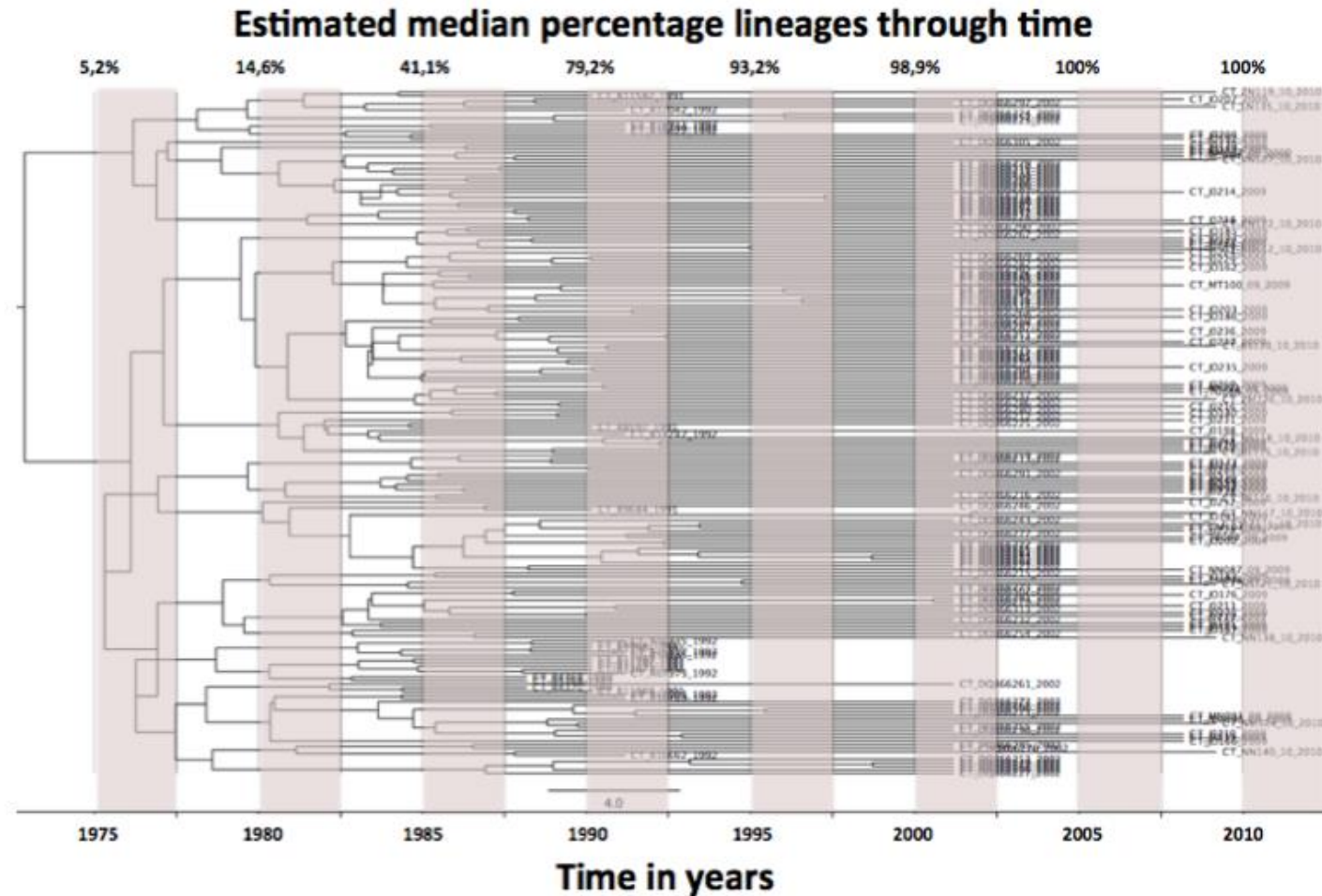


Figure 3.4: Estimated percentage lineages through time for the *gag* p24 Cape Town data set. The tree topology and the corresponding percentage lineages through time were reconstructed from data that was inferred in BEAST following 100 million iterations in the MCMC. This MCMC was conducted under a non-parametric relaxed molecular clock assumption with a fixed mutation rate of 3.0×10^{-3} mutations/site/year. The estimated percentage lineages through time were calculated in Tracer with the use of the corresponding log and tree files. The raw data were exported into Excel and converted into percentages. The time resolved tree topology was constructed from the corresponding tree file in TreeAnnotator following the discarding of a 10% burn-in.

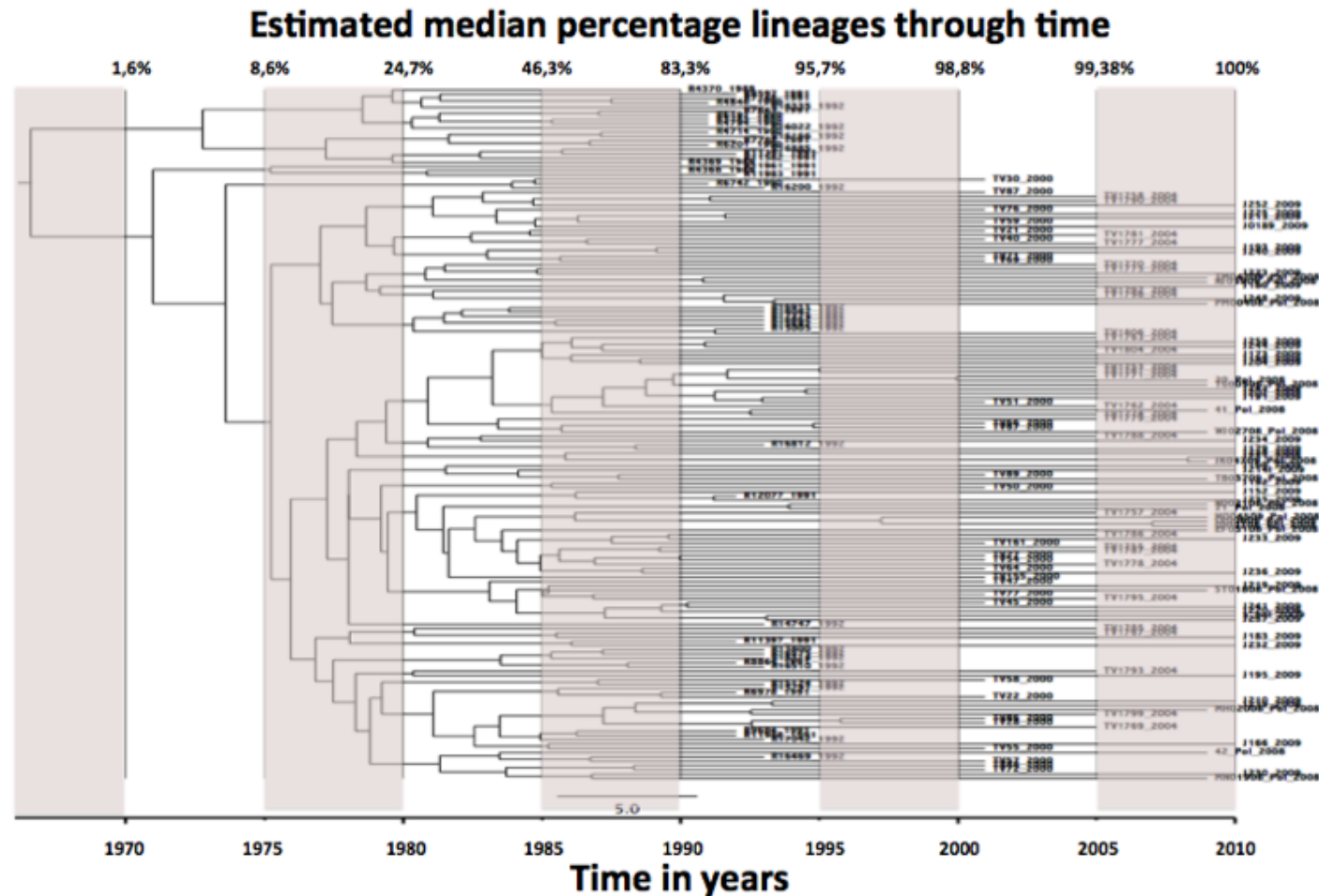


Figure 3.5: Estimated percentage lineages through time for the *pol* Cape Town data set. The tree topology and the corresponding percentage lineages through time were reconstructed from data that was inferred in BEAST following 100 million iterations in the MCMC. This MCMC was conducted under a non-parametric relaxed molecular clock assumption with a fixed mutation rate of 2.5×10^{-3} mutations/site/year. The estimated percentage lineages through time were calculated in Tracer with the use of the corresponding log and tree files. The raw data were exported into Excel and converted into percentages. The time resolved tree topology was constructed from the corresponding tree file in TreeAnnotator following the discarding of a 10% burn-in.

3.3.2 Outcome of the epidemic reconstruction of the Southern African epidemic (excluding sequence information from the Cape Town data sets)

Evolutionary histories were also inferred from sequence data from the Southern African region in order to compare the evolutionary trends of the Cape Town HIV-1 subtype C epidemic with that of the Southern African region. Once again, evolutionary histories were inferred from three data sets spanning over different genomic regions: a *gag* 24 data set, a *gag-pol* concatenated data set, and a *pol* data set. These data sets contained no sequence information from the Cape Town data sets.

3.3.2.1 The estimated tMRCA of the Southern African epidemic (excluding sequence information from the Cape Town data sets)

The inferred tMRCA of the Southern African epidemic from the *gag* p24 data set (Table 3.13) ranged between 1927,6 (with the 95% HPD ranging between 1871,0 and 1964,6) and 1970,9 (with the 95% HPD ranging between 1963,0 and 1977,1). Low effective sample sizes (ESS < 200) were obtained for these runs, however good convergence in the trace files was observed. An example of this can be seen in Figure 6.2 in Appendix E on page 243.

The estimated inferred mean tMRCA of the Southern African (excluding Cape Town sequence data) epidemic, from the *gag* p24 data set (Table 3.4) ranged between 1927 (95% HPD 1871,0 – 1964,6) and 1970,9 (95% HPD 1963,0 – 1977,1). Bayes factor comparison between the various tree priors (Table 6.8 in Appendix D on page 239) identified **BSP.relax.est.2** as the best fitting prior used for this data set. The mean tMRCA of this run was 1955,9 with the 95% HPD ranging between 1936,5 and 1972,0. The mean mutation rate for the best fitting model was $2,4 \times 10^{-3}$ mutations/site/year with the 95% HPD ranging between $1,7 \times 10^{-3}$ and $3,2 \times 10^{-3}$.

Table 3.4: Estimated tMRCA and mutation rates for the Southern African *gag* p24 data set. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Estimated tMRCA					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relax.est.1	1963,5	1964,6	1978,1	1947,5	36,2
BSP.relax.est.2	1955,9	1957,3	1972,0	1936,5	31,6
BSP.strict.fix.1	1970,9	1971,8	1977,1	1963,0	15,8
BSP.strict.fix.2	1965,3	1965,6	1970,1	1960,4	191,8
BSP.strict.est.1	1937,6	1941,9	1968,6	1898,6	191,2
BSP.strict.est.2	1927,6	1937,0	1964,6	1871,0	566,4
Const.relax.est.1	1967,6	1968,7	1979,5	1952,8	52,0
Const.relax.est.2	1970,6	1972,1	1981,6	1956,3	56,5
Const.strict.fix.1	1966,2	1966,6	1972,7	1958,2	19,5
Const.strict.fix.2	1967,2	1967,5	1972,5	1960,6	197,6
Const.strict.est.1	1935,6	1945,4	1972,3	1894,9	252,2
Const.strict.est.2	1929,3	1940,7	1970,8	1863,0	333,8
Average	1954,7	1958,2	1973,3	1930,2	162,0
Estimated mutation rate					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relax.est.1	2,40E-03	2,40E-03	1,70E-03	3,20E-03	50
BSP.relax.est.2	2,80E-03	2,70E-03	1,80E-03	3,90E-03	24,3
BSP.strict.est.1	1,20E-03	1,20E-03	6,10E-04	1,80E-03	147,9
BSP.strict.est.2	1,10E-03	1,10E-03	4,80E-04	1,70E-03	462,3
Const.relax.est.1	3,60E-03	3,50E-03	2,60E-03	4,70E-03	22,9
Const.relax.est.2	3,70E-03	3,70E-03	2,60E-03	4,80E-03	18,5
Const.strict.est.1	1,30E-03	1,30E-03	6,10E-04	2,00E-03	241,2
Const.strict.est.2	1,20E-03	1,20E-03	4,30E-04	2,00E-03	275,8

tMRCA – time to the Most Recent Common Ancestor; HPD – Highest Posterior Density; ESS – Effective Sample Size; BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Constant – Const Population Size tree prior

The estimated tMRCA of the various countries of origin in the Southern African data sets (Table 3.5); Botswana, South Africa (excluding Cape Town), Zambia, Zimbabwe, and Malawi, were also compiled from the log files. The estimated mean tMRCA for the model with the best fitting tree prior (**BSP.relax.est.2**), as was determined through Bayes factor comparison, for the *gag* p24 data set for each of these countries were as follows: Botswana 1957,3 (1939,5 – 1972,6); Malawi 1960,0 (1944,6 – 1974,5), South Africa excluding Cape Town sequence data 1958,7 (1940,4 – 1973,7), Zimbabwe 1960,8 (1945,4 – 1973,6), and Zambia 1957,5 (1938,8 – 1973,7).

Table 3.5: Inferred tMRCA of the various countries in the Southern African only *gag* p24 data set. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Model parameter	Botswana			Malawi			South Africa (excluding Cape Town)			Zimbabwe			Zambia		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
BSP.relax.est.1	1964	1978	1948	1967	1981	1953	1965	1978	1950	1966	1979	1952	1964	1979	1949
BSP.relax.est.2	1957	1972	1939	1960	1974	1944	1958	1973	1940	1960	1973	1945	1957	1973	1938
BSP.strict.fix.1	1971	1977	1963	1971	1977	1964	1971	1977	1963	1972	1977	1964	1971	1977	1963
BSP.strict.fix.2	1966	1970	1961	1966	1970	1962	1965	1970	1961	1966	1971	1962	1965	1970	1961
BSP.strict.est.1	1938	1968	1899	1939	1968	1900	1938	1968	1899	1938	1969	1900	1938	1968	1899
BSP.strict.est.2	1928	1967	1873	1929	1967	1875	1928	1966	1873	1929	1967	1875	1928	1966	1873
Const.relax.est.1	1969	1980	1955	1972	1981	1962	1970	1980	1959	1971	1981	1962	1969	1980	1957
Const.relax.est.2	1972	1982	1958	1975	1982	1965	1973	1982	1962	1974	1982	1964	1972	1982	1961
Const.strict.fix.1	1967	1973	1960	1969	1974	1964	1967	1973	1961	1969	1974	1964	1967	1973	1960
Const.strict.fix.2	1968	1973	1962	1970	1975	1965	1968	1974	1963	1970	1975	1964	1968	1973	1962
Const.strict.est.1	1936	1972	1895	1937	1972	1896	1936	1972	1895	1936	1972	1896	1936	1972	1895
Const.strict.est.2	1930	1972	1865	1931	1973	1868	1930	1971	1863	1930	1973	1867	1930	1973	1865
Average	1956	1974	1932	1957	1975	1935	1956	1974	1932	1957	1974	1935	1955	1974	1932

HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot tree prior, fix – Fixed mutation rate; est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Const – Const population size tree prior

The estimated inferred mean tMRCA of the Southern African (excluding Cape Town), epidemic from the concatenated *gag-pol* data set (Table 3.6) ranged between 1941,5 and 1975,5. Once again good convergence in the trace files were obtained even though the majority of the runs reported low (>200) effective sample sizes. The best fitting model, as was determined through Bayes factor comparison analysis (Table 6.9 in Appendix D on page 240), was identified as **Const.relax.est.2**. The mean tMRCA for this model was 1946,0 (95% HPD ranging between 1926,7 and 1972,1) while the estimated mutation rate for the model was $2,7 \times 10^{-3}$ mutations/site/year ($2,3 \times 10^{-3} - 3,6 \times 10^{-3}$ mutations/site/year).

Table 3.6: The estimated tMRCA and mutation rates for the Southern African only *gag-pol* concatenated data set. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Estimated tMRCA					
Model parameters	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relax.fix.1	1953,7	1954,1	1957,6	1949,0	47,7
BSP.relax.fix.2	1951,6	1952,1	1956,9	1945,9	34,0
BSP.relax.est.1	1948,5	1948,9	1960,7	1933,8	11,9
BSP.relax.est.2	1948,6	1949,4	1961,9	1933,7	36,1
BSP.strict.fix.1	1951,3	1951,3	1956,1	1946,2	77,7
BSP.strict.fix.2	1949,9	1950,0	1955,8	1944,4	39,9
BSP.strict.est.1	1949,3	1949,9	1959,1	1938,7	16,4
BSP.strict.est.2	1946,1	1946,4	1956,2	1936,4	82,3
Const.relax.fix.1	1941,5	1942,0	1948,4	1934,1	42,1
Const.relax.fix.2	1975,5	1975,7	1980,5	1970,4	13,6
Const.relax.est.1	1964,7	1966,0	1972,6	1955,2	14,3
Const.relax.est.2	1946,0	1965,1	1972,1	1926,7	8,8
Average	1952,2	1954,2	1961,5	1942,9	35,4
Estimated mutation rate					
Model parameters	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relax.est.1	2,40E-03	2,40E-03	2,00E-03	3,00E-03	30,1
BSP.relax.est.2	2,40E-03	2,40E-03	1,70E-03	2,90E-03	30,9
BSP.strict.est.1	2,50E-03	2,50E-03	2,10E-03	2,80E-03	51,9
BSP.strict.est.2	1,70E-03	1,70E-03	1,40E-03	1,90E-03	71,5
Const.relax.est.1	2,90E-03	2,90E-03	2,30E-03	3,60E-03	9,7
Const.relax.est.2	2,70E-03	3,00E-03	5,30E-04	3,40E-03	4,2

tMRCA – time to the Most Recent Common Ancestor; ESS – Effective Sample Size; HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot tree prior, fix – Fixed mutation rate; est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Constant – Constant Population Size tree prior

The mean estimated tMRCA from the best fitting model in this *gag-pol* data set for each of the various countries (Table 3.7) were as follows: Botswana 1946 (95% HPD 1927 – 1972); Swaziland 1948 (95% HPD 1931 – 1972), South Africa (excluding Cape Town) 1945 (95% HPD

1927 – 1970), Zimbabwe 1949 (95% HPD 1930 – 1971), and Zambia 1945 (95% HPD 1926 – 1971).

Table 3.7: Inferred tMRCA of the various countries in the Southern African only *gag-pol* concatenated data set. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Model parameter	Botswana			Swaziland			South Africa (excluding Cape Town)			Zambia			Zimbabwe		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
BSP.relax.fix.1	1952	1958	1949	1955	1959	1949	1953	1959	1950	1952	1959	1949	1955	1959	1949
BSP.relax.fix.2	1951	1957	1946	1952	1959	1947	1951	1956	1945	1950	1957	1945	1952	1958	1948
BSP.relax.est.1	1948	1961	1934	1950	1962	1935	1949	1961	1935	1948	1961	1935	1950	1961	1935
BSP.relax.est.2	1948	1962	1934	1950	1961	1935	1948	1962	1934	1949	1961	1934	1949	1962	1934
BSP.strict.fix.1	1951	1956	1946	1952	1958	1948	1951	1957	1945	1951	1956	1945	1952	1958	1948
BSP.strict.fix.2	1949	1956	1944	1950	1961	1940	1949	1961	1939	1949	1960	1939	1950	1939	1960
BSP.strict.est.1	1949	1959	1939	1950	1960	1941	1949	1960	1940	1949	1960	1938	1950	1939	1961
BSP.strict.est.2	1946	1956	1936	1948	1957	1938	1946	1955	1936	1945	1955	1937	1947	1956	1937
Const.relax.fix.1	1941	1948	1934	1943	1948	1935	1942	1948	1935	1941	1948	1934	1943	1949	1936
Const.relax.fix.2	1975	1981	1970	1976	1980	1971	1975	1981	1970	1975	1981	1970	1976	1980	1971
Const.relax.est.1	1964	1973	1955	1965	1973	1956	1964	1974	1956	1963	1973	1954	1965	1972	1956
Const.relax.est.2	1946	1972	1927	1948	1972	1931	1945	1970	1927	1945	1971	1926	1949	1971	1930
Average	1952	1962	1943	1953	1963	1944	1952	1962	1943	1951	1962	1942	1953	1959	1947

HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot, fix – Fixed mutation rate; est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Const – Constant population size

The estimated inferred mean tMRCA of the Southern African (excluding Cape Town), epidemic from the *pol* data set (Table 3.8) ranged between 1958,1 (with the 95% HPD ranging between 1946,2 and 1969,2) and 1965,6 (with the 95% HPD ranging between 1954,0 and 1976,5). The best fitting model, as was determined through Bayes factor comparison analysis (Table 6.10 in Appendix D on page 240), was identified as **BSP.relax.est.1**. The mean tMRCA for this model was 1960,6 (1948,0 – 1972,4) while the estimated mutation rate for the model was $2,3 \times 10^{-3}$ mutations/site/year ($1,8 \times 10^{-3} - 2,8 \times 10^{-3}$).

Table 3.8: The estimated tMRCA and mutation rates for the Southern African *pol* data set (excluding Cape Town). The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Estimated tMRCA					
Model parameter	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relax.est.1	1960,6	1961,4	1972,4	1948,0	1258,5
BSP.relax.est.2	1960,6	1961,4	1972,1	1946,9	970,1
BSP.strict.est.1	1958,3	1958,9	1968,9	1945,9	726,1
BSP.strict.est.2	1958,1	1958,8	1969,2	1946,2	456,3
Const.relax.est.1	1965,5	1966,2	1976,1	1953,8	425,1
Const.relax.est.2	1965,6	1966,2	1976,5	1954,0	353,3
Const.strict.est.1	1964,4	1964,8	1972,2	1955,0	537,2
Const.strict.est.2	1964,6	1964,9	1972,4	1955,8	756,6
Average	1962,2	1962,8	1972,5	1950,7	685,4
Estimated Mutation Rates					
Model parameter	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relax.est.1	2,30E-03	2,30E-03	1,80E-03	2,80E-03	964,7
BSP.relax.est.2	2,30E-03	2,30E-03	1,80E-03	2,90E-03	654,8
BSP.strict.est.1	2,20E-03	2,20E-03	1,70E-03	2,70E-03	635,4
BSP.strict.est.2	2,20E-03	2,20E-03	1,70E-03	2,70E-03	413,2
Const.relax.est.1	3,60E-03	3,60E-03	2,80E-03	4,40E-03	163,5
Const.relax.est.2	3,60E-03	3,60E-03	2,80E-03	4,40E-03	92,3
Const.strict.est.1	3,00E-03	3,00E-03	2,40E-03	3,60E-03	250,0
Const.strict.est.2	3,00E-03	3,00E-03	2,50E-03	3,60E-03	508,7

tMRCA – time to the Most Recent Common Ancestor; ESS – Effective Sample Size; HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot, est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Constant – Constant Population size

The mean estimated tMRCA for the various countries in the *pol* data set (Table 3.9) for the best fitting model in the Bayesian MCMC were as follows: Botswana 1961,0 (1948,0 – 1972,4); Swaziland 1969,7 (1960,3 – 1977,9), South Africa excluding Cape Town sequence data 1961,1 (1948,9 – 1973,8), Zimbabwe 1969,6 (1960,2 – 1977,8), and Zambia 1961,1 (1948,9 – 1972,5).

Table 3.9: Inferred tMRCA of the various countries in the *pol* Southern African data set. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Model parameter	Botswana			South Africa (excluding Cape Town)			Zambia			Zimbabwe			Swaziland		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
BSP.relax.est.1	1961	1972	1948	1961	1974	1949	1961	1973	1949	1970	1978	1960	1969	1977	1960
BSP.relax.est.2	1961	1972	1947	1961	1972	1947	1961	1972	1947	1969	1977	1960	1969	1977	1960
BSP.strict.est.1	1958	1969	1946	1958	1969	1946	1958	1969	1946	1968	1976	1959	1968	1976	1959
BSP.strict.est.2	1958	1969	1946	1958	1970	1947	1958	1969	1946	1968	1975	1959	1968	1975	1959
Const.relax.est.1	1966	1976	1955	1966	1976	1954	1966	1976	1955	1975	1981	1968	1975	1981	1968
Const.relax.est.2	1966	1976	1955	1966	1977	1954	1966	1977	1955	1975	1981	1968	1975	1981	1969
Const.strict.est.1	1964	1972	1955	1965	1973	1955	1964	1972	1955	1972	1978	1966	1972	1978	1966
Const.strict.est.2	1965	1972	1956	1965	1973	1956	1965	1972	1956	1972	1978	1966	1972	1978	1966
Average	1962	1972	1951	1963	1973	1951	1962	1973	1951	1971	1978	1963	1971	1978	1963

tMRCA – time to the Most Recent Common Ancestor; ESS – Effective Sample Size; HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot, est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Const – Constant Population Size

In summary, the estimated date of origin of the Southern African HIV-1 subtype C epidemic, as was inferred from Southern African sequence data (excluding sequences from the original Cape Town data sets), placed the date of origin for the epidemic around the mid 1950's with the 95% HPD confidence interval stretching from 1926 to 1972.

3.3.2.2 Results of the dynamic epidemic reconstruction from Southern Africa sequences (excluding Cape Town)

Close inspection of the BSP plot in Figure 3.6, which was reconstructed from the Southern African *gag* p24 data sets (excluding Cape Town), revealed exponential growth in the effective population size in the late 1970's (1974 - 1980) and late 1980's (1985 - 1990). A possible reduction in the rate of epidemic expansion can be seen on the graph from the early 1990's until the present. This BSP was inferred from the "best fitting" non-parametric model, as determined through Bayes factor comparison, which was the **BSP.Relax.est.1** run.

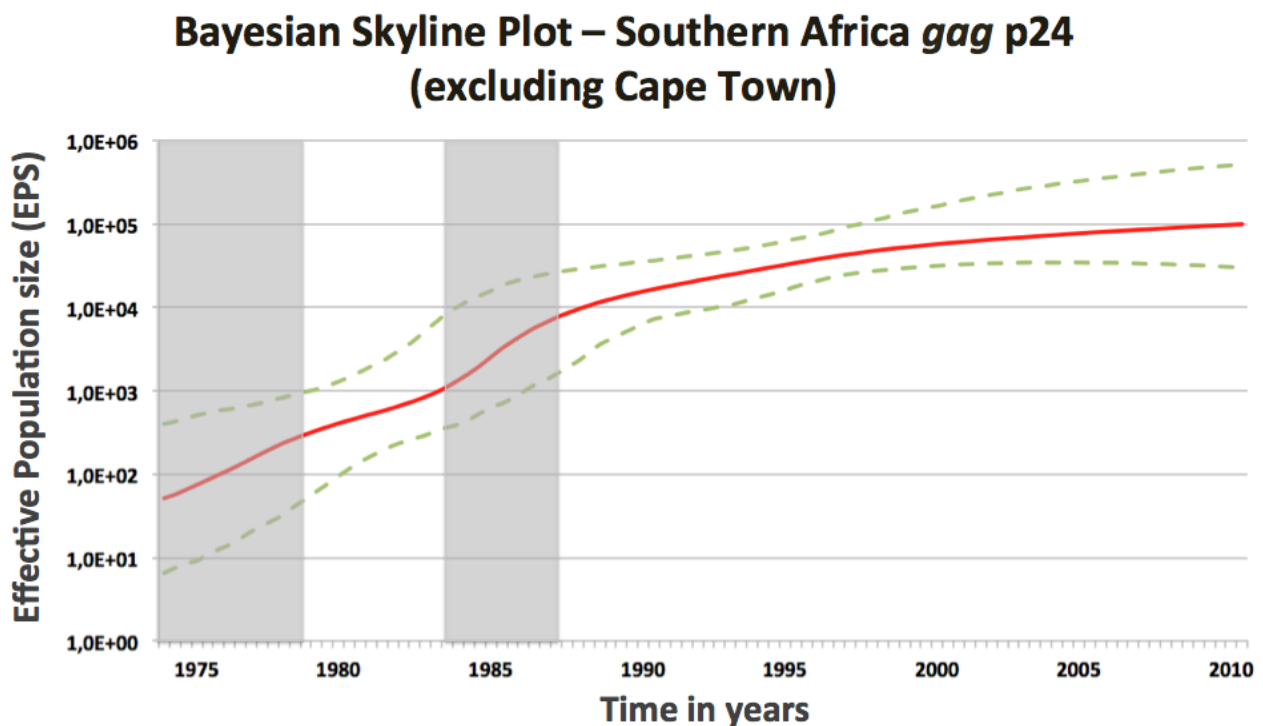


Figure 3.6: BSP of the Southern African *gag* p24 data set (excluding Cape Town). This non-parametric estimate of demographic history was reconstructed in Excel from raw data. The solid red lines are the "traced" median effective population size with the 95% upper and lower highest posterior density intervals indicated in the green dashed lines. This plot begins in the mid 1970's and stretches over roughly a 35-year period (1975 – 2010). The period(s) marked in grey represent periods of exponential growth in the effective population size of the epidemic.

Close inspection of the reconstructed BSP plot in Figure 3.7, which was inferred from the Southern African *pol* data sets (excluding Cape Town), revealed exponential growth in the effective population size in the early 1980's (1982 - 1986) and the mid 1990's (1993 - 1997). A possible reduction in the rate of epidemic expansion can be seen on the graph from the early 1990's till the present. This BSP was inferred from the “best fitting” non-parametric model, as determined through Bayes factor comparison, which was the **BSP.Relax.est.1** run.

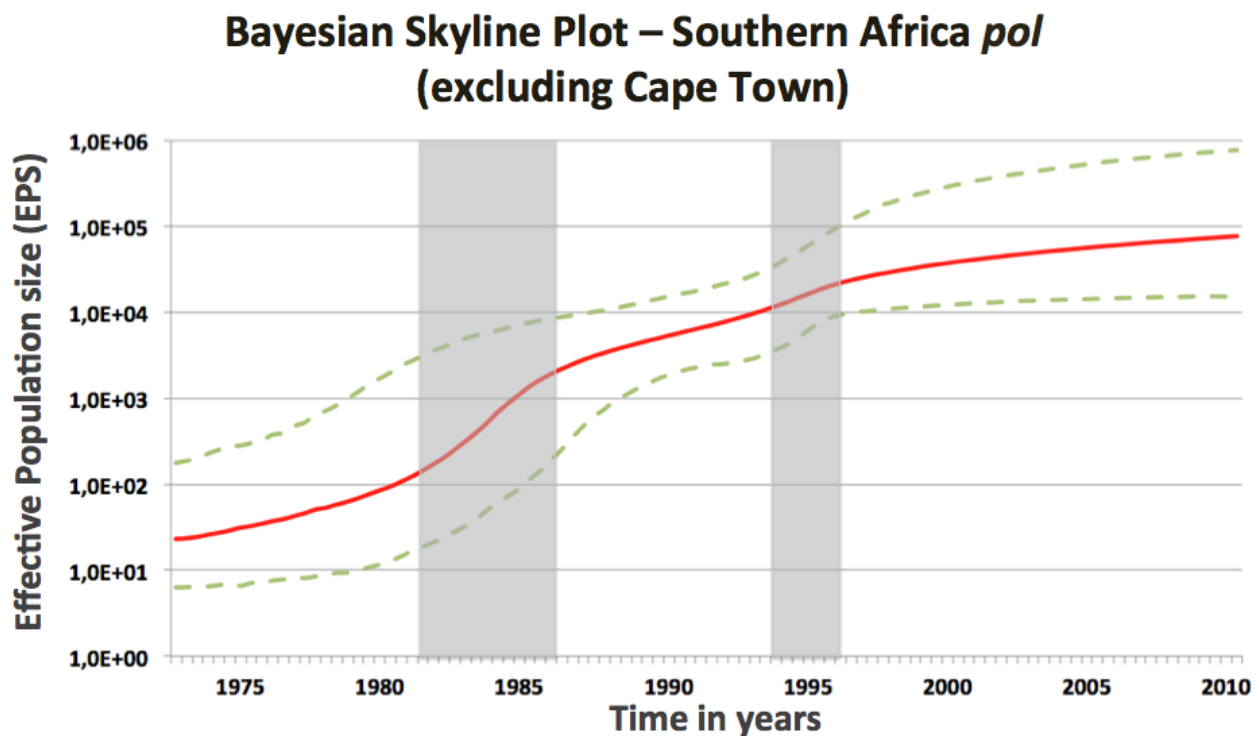


Figure 3.7: BSP of the Southern African *pol* data set (excluding Cape Town). This non-parametric estimate of demographic history was reconstructed in Excel from raw data. The solid red lines are the “traced” median effective population size with the 95% upper and lower highest posterior density intervals indicated in the green dashed lines. This plot begins in the mid 1970's and stretches over roughly a 35-year period (1975 – 2010). The period(s) marked in grey represent periods of exponential growth in the effective population size of the epidemic.

In summary, the epidemic reconstruction of Southern African HIV-1 subtype C (excluding sequence data from Cape Town) from the two data sets (*gag* p24 and partial *pol*) suggests a linearly increasing in the effective population size over the course of the epidemic, with small short periods of exponential growth during the mid 1980's and mid 1990's.

3.3.3 Southern African data sets (including Cape Town)

3.3.3.1 tMRCA of the Southern African epidemic (including Cape Town)

Root height or tMRCA estimations were also made from a Southern African data set, which contained sequence information from the Cape Town data sets. These evolutionary histories, as in the previous two sections, were reconstructed from three different genomic regions (*gag* 24, *gag-pol*, and *pol*). Convergence in the log files was manually assessed in Tracer for each of the model parameters. Good convergence in the Markov Chain was observed for all of the runs, even though in some cases the estimated sample size (ESS) remained very low for some of the runs. These low ESS values are due to the large data sets that were used ($n > 500$). With such extremely large data sets, spanning over a large time frame and comprising genetically diverse isolates from a large geographical region, it is extremely difficult for the Bayesian MCMC to obtain a good posterior distribution of the various model parameters. However, given the good convergence in the trace files (Figures 6.3 – 6.5 in Appendix E on pages 243 and 244) one can interpret the following results with confidence.

The mean inferred tMRCA of the Southern African *gag* p24 data set, with sequence data from Cape Town (Table 3.10), ranged between 1946,3 (1928,8 – 1960,8) and 1974,3 (1962,8 – 1988,0). The best fitting tree prior used (Table 6.11 in Appendix D on page 241), as was determined through Bayes factor comparison, was **BSP.relax.est.1**. This model had a mean tMRCA of 1950,3 with the 95% HPD intervals ranging between 1933,5 and 1964,0. The mean estimated mutation rate for this particular tree prior was 2.1×10^{-3} mutations/site/year with the 95% HPD intervals ranging between 1.4×10^{-3} and 3.0×10^{-3} mutations/site/year.

Table 3.10: The estimated tMRCA and mutation rates for the entire Southern Africa *gag* p24 data set. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Estimated tMRCA					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relax.fix.1	1974,3	1971,4	1988,0	1962,8	11,3
BSP.relax.fix.2	1967,8	1968,1	1973,7	1960,9	125,8
BSP.relax.est.1	1950,3	1951,4	1964,0	1933,5	198,7
BSP.relax.est.2	1950,3	1951,3	1964,4	1933,9	46,3
BSP.strict.est.1	1947,1	1948,2	1961,5	1931,4	192,5
BSP.strict.est.2	1946,3	1947,5	1960,8	1928,8	124,4
Const.relax.est.1	1968,8	1969,5	1976,3	1960,0	125,7
Const.relax.est.2	1968,5	1969,4	1975,7	1959,2	220,2
Const.strict.est.1	1967,1	1967,5	1972,8	1960,3	50,1
Const.strict.est.2	1967,0	1967,6	1973,2	1960,2	85,7
Average	1960,7	1961,2	1971,0	1949,1	118,07
Estimated Mutation Rates					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relax.est.1	2,10E-03	2,10E-03	1,40E-03	3,00E-03	6,0
BSP.relax.est.2	2,10E-03	2,10E-03	1,30E-03	2,90E-03	6,6
BSP.strict.est.1	2,10E-03	2,00E-03	1,50E-03	2,70E-03	31
BSP.strict.est.2	2,10E-03	2,10E-03	1,30E-03	2,70E-03	19,7
Const.relax.est.1	5,10E-03	5,00E-03	3,40E-03	7,40E-03	9,2
Const.relax.est.2	4,50E-03	4,20E-03	3,30E-03	6,50E-03	6,2
Const.strict.est.1	4,10E-03	3,60E-03	2,90E-03	5,70E-03	3,2
Const.strict.est.2	5,10E-03	5,10E-03	4,40E-03	6,00E-03	19,1

tMRCA – time to the Most Recent Common Ancestor; ACT – Auto-Correlation Time; ESS – Effective Sample Size; HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot, fix – Fixed mutation rate; est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Constant – Constant Population Size

The estimated tMRCA of the various countries of origin in the entire Southern African *gag* p24 data set; Botswana, Cape Town, South Africa (not including Cape Town sequences), Zambia, Zimbabwe and Malawi, were also compiled from the log files in Tracer. The mean estimated tMRCA for each of the various countries/regions from this *gag* p24 data set, as determined through Bayes factor model comparison were as follows: Botswana 1951 (1933 - 1966), Cape Town 1950 (1932 - 1966), South Africa 1950 (1932 - 1966), Zambia 1950 (1932 - 1966), Zimbabwe 1952 (1935 - 1966), and Malawi 1951 (1935 - 1966). A summary of the estimated tMRCA's of the various countries and regions contained within the Southern African *gag* p24 data set are listed in Table 3.11.

Table 3.11: Inferred tMRCA of the various countries in the entire Southern African *gag* p24 data set (including Cape Town). The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Model parameters	Botswana			Cape Town			Rest of South Africa		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
BSP.relax.fix.1	1976	1991	1963	1974	1988	1963	1976	1994	1963
BSP.relax.fix.2	1968	1974	1962	1968	1974	1961	1968	1974	1962
BSP.relax.est.1	1951	1966	1933	1950	1966	1932	1950	1966	1932
BSP.relax.est.2	1951	1965	1933	1951	1965	1932	1951	1965	1932
BSP.strict.est.1	1947	1961	1929	1947	1960	1929	1947	1960	1929
BSP.strict.est.2	1946	1961	1929	1946	1960	1929	1946	1960	1929
Const.relax.est.1	1967	1976	1952	1967	1975	1952	1968	1976	1952
Const.relax.est.2	1964	1975	1939	1965	1975	1946	1965	1975	1946
Const.strict.est.1	1967	1973	1961	1968	1973	1962	1968	1973	1963
Const.strict.est.2	1967	1973	1961	1968	1973	1962	1968	1973	1962
Average	1960	1972	1946	1960	1971	1947	1961	1972	1947
Model parameters	Zambia			Zimbabwe			Malawi		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
BSP.relax.fix.1	1976	1989	1963	1977	1995	1964	1976	1991	1963
BSP.relax.fix.2	1968	1974	1961	1968	1974	1962	1968	1974	1962
BSP.relax.est.1	1950	1966	1932	1952	1966	1935	1951	1966	1935
BSP.relax.est.2	1951	1965	1932	1953	1966	1934	1952	1967	1934
BSP.strict.est.1	1947	1960	1929	1948	1962	1932	1947	1961	1931
BSP.strict.est.2	1946	1960	1929	1948	1962	1931	1947	1962	1930
Const.relax.est.1	1967	1975	1953	1969	1978	1955	1969	1979	1955
Const.relax.est.2	1964	1975	1939	1967	1977	1947	1966	1975	1947
Const.strict.est.1	1968	1973	1962	1969	1975	1963	1969	1974	1963
Const.strict.est.2	1968	1973	1962	1969	1974	1963	1969	1974	1963
Average	1961	1971	1946	1962	1973	1949	1961	1972	1948

HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot tree prior, fix – Fixed mutation rate; est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

The mean inferred tMRCA of the Southern African *gag-pol* concatenated data set, with sequence data from Cape Town (Table 3.12), ranged between 1927,5 (1881,0 – 1974,2) and 1970,1 (1965,4 – 1974,5). Bayesian factor model comparison identified **BSP.relaxed.fix.1** as the best fitting model tree prior used for the data set (Table 6.12 in Appendix D on page 241). The mean tMRCA for this run was 1963,3 with the 95% HPD ranging between 1952,9 and 1972,3. The mutation rate for this run was fixed at 2.5×10^{-3} mutations/site/year.

Table 3.12: The estimated tMRCA and mutation rates for the entire Southern Africa *gag-pol* concatenated data set (including Cape Town). The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Estimated tMRCA					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relaxed.fix.1	1963,3	1963,4	1972,3	1952,9	153,9
BSP.relaxed.fix.2	1970,1	1970,2	1974,5	1965,4	61,9
BSP.relaxed.est.1	1936,3	1937,4	1950,5	1921,2	125,4
BSP.relaxed.est.2	1937,9	1938,6	1952,4	1922,1	187,2
BSP.strict.est.1	1938,1	1938,8	1950,1	1925,0	241,0
BSP.strict.est.2	1934,2	1934,8	1947,6	1919,6	64,2
Const.relaxed.fix.1	1964,0	1964,6	1969,5	1958,9	46,3
Const.relaxed.fix.2	1964,0	1964,2	1968,9	1958,6	116,4
Const.relaxed.est.1	1927,5	1963,1	1974,2	1881,0	8,0
Const.relaxed.est.2	1966,7	1967,0	1972,7	1959,9	18,3
Average	1950,2	1954,2	1963,2	1936,4	102,26
Estimated Mutation Rates					
Model Parameters	Mean	Median	95% HPD lower	95% HPD upper	EES
BSP.relaxed.est.1	1,50E-03	1,50E-03	1,20E-03	1,80E-03	77,5
BSP.relaxed.est.2	1,50E-03	1,50E-03	1,20E-03	1,80E-03	148,6
BSP.strict.est.1	1,50E-03	1,50E-03	1,30E-03	1,80E-03	111,8
BSP.strict.est.2	1,50E-03	1,50E-03	1,20E-03	1,70E-03	194,1
Const.relaxed.est.1	3,40E-03	3,10E-03	2,30E-03	4,90E-03	4
Const.relaxed.est.2	3,20E-03	3,20E-03	2,70E-03	3,70E-03	20,4

tMRCA – time to the Most Recent Common Ancestor; ACT – Auto-Correlation Time; ESS – Effective Sample Size; HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot, fix – Fixed mutation rate; est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Constant – Constant Population Size

The estimated tMRCA of the various countries of origin in the Southern African *gag-pol* concatenated data set; Botswana, Cape Town, South Africa (not including Cape Town sequences), Zambia, and Tanzania, were also compiled from the log files of the best fitting model as was determined in the Bayes factor model comparison. The average estimated tMRCA from this *gag-pol* concatenated data set for each of the regions and countries were as follows: Botswana 1964 (1952 - 1972), Cape Town 1968 (1965 - 1971), South Africa 1963 (1952 - 1972), Zambia 1963 (1951 - 1975), and Tanzania 1964 (1959 - 1969). A summary of the estimated tMRCA under the various model parameters for each of the countries or regions for this data set is listed in Table 3.13.

Table 3.13: Inferred tMRCA of the various countries in the Southern African *gag-pol* concatenated data set, including sequence information from Cape Town. The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Model parameter	Botswana			Cape Town			Rest of South Africa			Zambia			Tanzania		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
BSP.relax.fix.1	1964	1972	1952	1968	1971	1965	1963	1972	1952	1963	1975	1951	1964	1969	1959
BSP.relax.fix.2	1970	1975	1966	1973	1975	1970	1970	1975	1966	1970	1975	1967	1971	1975	1967
BSP.relax.est.1	1937	1951	1922	1942	1955	1928	1937	1951	1922	1937	1950	1921	1939	1953	1923
BSP.relax.est.2	1938	1952	1922	1943	1956	1929	1938	1952	1922	1938	1951	1922	1940	1954	1925
BSP.strict.est.1	1938	1950	1925	1945	1956	1934	1938	1950	1925	1938	1950	1925	1949	1959	1938
BSP.strict.est.2	1934	1948	1920	1941	1952	1927	1934	1948	1920	1934	1948	1920	1946	1957	1931
Const.relax.fix.1	1965	1970	1960	1969	1973	1966	1964	1969	1959	1965	1970	1960	1966	1971	1960
Const.relax.fix.2	1964	1969	1959	1969	1972	1967	1964	1969	1960	1964	1969	1959	1965	1970	1960
Const.relax.est.1	1956	1973	1903	1934	1975	1895	1956	1974	1902	1931	1974	1896	1932	1974	1896
Const.relax.est.2	1967	1973	1960	1971	1976	1966	1967	1973	1961	1967	1973	1961	1968	1973	1961
Average	1953	1963	1939	1956	1966	1945	1953	1963	1939	1951	1964	1938	1954	1966	1942

HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot tree prior, fix – Fixed mutation rate; est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Constant – Constant Population Size tree prior

The mean inferred tMRCA of the Southern African *pol* data set, with sequence data from Cape Town (Table 3.14), ranged between 1909,6 (1877,7 – 1937,0) and 1956,9 (1935,5 – 1972,3). The best fitting model tree prior used, as was determined through Bayes factor comparison, for this *pol* data set was **Const.relax.est.1** (Table 6.13 in Appendix D on page 242). The mean estimated tMRCA for this model was 1951,9 with the 95% HPD ranging between 1927,2 and 1971,0. The mean mutation rate for this model was $2,6 \times 10^{-3}$ mutations/site/year with the 95% HPD ranging between $2,1 \times 10^{-3}$ and $3,0 \times 10^{-3}$ mutations/site/year.

Table 3.14: The estimated tMRCA and mutation rates for the Southern Africa *pol* data set (including Cape Town). The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Estimated tMRCA					
Model parameters	Mean	Median	95% HPD lower	95% HPD upper	ESS
BSP.relax.fix.1	1947,8	1950,0	1966,0	1923,6	12
BSP.relax.fix.2	1950,2	1953,0	1966,9	1925,1	13,8
BSP.relax.est.1	1937,2	1939,3	1955,7	1912,1	74,6
BSP.relax.est.2	1938,9	1940,6	1957,6	1917,9	242,7
BSP.strict.est.1	1911,0	1911,9	1937,2	1882,6	848,3
BSP.strict.est.2	1909,6	1910,6	1937,0	1877,7	543,2
BSP.strict.fix.1	1927,1	1928,0	1949,9	1904,2	19,9
BSP.strict.fix.2	1930,1	1930,6	1949,8	1909,9	162
Const.relax.est.1	1951,9	1954,4	1971,0	1927,2	42,8
Const.relax.est.2	1956,9	1959,3	1972,3	1935,5	35,9
Const.strict.est.1	1939,3	1940,2	1956,3	1920,6	181,7
Const.strict.est.1	1941,7	1942,4	1957,9	1924,3	319,2
Average	1936,8	1938,4	1956,5	1913,4	208,0
Estimated Mutation Rates					
Model parameters	Mean	Median	95% HPD lower	95% HPD upper	ESS
BSP.relax.est.1	1,60E-03	1,60E-03	1,20E-03	1,90E-03	122,8
BSP.relax.est.2	1,70E-03	1,70E-03	1,30E-03	2,00E-03	113,6
BSP.strict.est.1	1,60E-03	1,60E-03	1,30E-03	2,00E-03	21,5
BSP.strict.est.2	1,50E-03	1,50E-03	1,20E-03	1,80E-03	134,6
Const.relax.est.1	2,60E-03	2,60E-03	2,10E-03	3,00E-03	15,5
Const.relax.est.2	2,60E-03	2,60E-03	2,10E-03	3,00E-03	51,8
Const.strict.est.1	3,50E-03	3,50E-03	2,90E-03	4,20E-03	15,5
Const.strict.est.1	3,60E-03	3,60E-03	3,00E-03	4,30E-03	51,8

tMRCA – time to the Most Recent Common Ancestor; ACT – Auto-Correlation Time; ESS – Effective Sample Size; HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot tree prior, fix – Fixed mutation rate; est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption, Constant – Constant Population Size tree prior

The estimated tMRCA of the various countries and regions of origin in the entire Southern African *pol* data set (including sequence information from the Cape Town data sets) was also compiled from the log files in Tracer. The mean estimated tMRCA from this *pol* data set for each of these countries and regions were as follows: Botswana 1956 (1940 – 1969), Cape Town 1954

(1927 – 1972), South Africa excluding Cape Town sequence data 1957 (1944 – 1970), Zambia 1957 (1944 – 1970), Zimbabwe 1968 (1962 – 1974), and Swaziland 1968 (1962 – 1974). A summary of the estimated tMRCA's of the various countries or regions of the Southern African *pol* data set are presented in Table 3.15.

Table 3.15: Inferred tMRCA of the various countries in the Southern African *pol* data set (including Cape Town). The best fitting model tree prior used for this data set, as was determined by Bayes factor calculation, are marked in bold. The average mean, median, 95% HPD lower and upper estimates, as well as the average ESS's are indicated at the bottom.

Model parameters	Botswana			Cape Town			Rest of South Africa		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
BSP.relax.fix.1	1952	1965	1938	1950	1969	1923	1952	1965	1938
BSP.relax.fix.2	1954	1965	1941	1954	1971	1926	1954	1965	1941
BSP.relax.est.1	1939	1956	1921	1937	1956	1913	1939	1956	1922
BSP.relax.est.2	1940	1956	1921	1939	1957	1918	1940	1956	1921
BSP.strict.est.1	1934	1948	1918	1910	1938	1882	1934	1948	1918
BSP.strict.est.2	1934	1948	1918	1911	1938	1883	1934	1948	1918
BSP.strict.fix.1	1947	1958	1936	1927	1950	1904	1947	1958	1936
BSP.strict.fix.2	1949	1958	1940	1930	1950	1910	1949	1958	1940
Const.relax.est.1	1956	1969	1940	1954	1972	1927	1957	1970	1944
Const.relax.est.2	1959	1972	1943	1961	1974	1938	1963	1974	1948
Const.strict.est.1	1950	1960	1940	1939	1957	1921	1951	1960	1941
Const.strict.est.1	1952	1961	1941	1942	1958	1924	1952	1961	1942
Average	1947	1960	1933	1938	1958	1914	1948	1960	1934
Model parameters	Zambia			Zimbabwe			Swaziland		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
BSP.relax.fix.1	1953	1967	1939	1962	1969	1954	1962	1969	1954
BSP.relax.fix.2	1955	1967	1943	1964	1969	1957	1964	1969	1957
BSP.relax.est.1	1939	1956	1922	1953	1964	1942	1953	1964	1942
BSP.relax.est.2	1940	1958	1922	1953	1964	1941	1954	1964	1941
BSP.strict.est.1	1934	1948	1918	1951	1961	1942	1951	1961	1942
BSP.strict.est.2	1934	1948	1918	1952	1961	1941	1952	1961	1941
BSP.strict.fix.1	1947	1958	1936	1961	1967	1953	1961	1968	1953
BSP.strict.fix.2	1949	1958	1940	1962	1967	1956	1962	1967	1956
Const.relax.est.1	1957	1970	1944	1968	1974	1962	1968	1974	1962
Const.relax.est.2	1960	1972	1945	1967	1974	1960	1967	1974	1960
Const.strict.est.1	1951	1960	1941	1962	1968	1955	1962	1968	1955
Const.strict.est.1	1952	1961	1942	1964	1969	1957	1964	1969	1957
Average	1948	1960	1934	1960	1967	1952	1960	1967	1952

HPD – Highest Posterior Density; BSP – Bayesian Skyline Plot tree prior, fix – Fixed mutation rate; est – Estimated mutation rate; relax – Relaxed Molecular Clock assumption; strict – Strict Molecular Clock assumption; Constant – Constant Population Size tree prior

In summary, the estimated tMRCA of the Southern African epidemic that was inferred from three different data sets and included sequence data from the Cape Town data sets, places the date of origin of the epidemic around the early to mid 1950's. Some small variation in the estimates was

observed with the estimated tMRCA inferred from the concatenated *gag-pol* data set suggesting a slightly younger date of origin around 1960.

3.3.3.2 Outcome of the dynamic reconstruction of the Southern Africa epidemic (including sequence information from Cape Town)

Phylogenetic reconstruction of the demographic history of the entire Southern African HIV-1 subtype C epidemic was performed from *gag* p24 and *pol* sequence data sets. Close examination of the reconstructed BSP of the Southern African *gag* p24 data set (Figure 3.8) reveals a slow linear increase in the EPS from the early 1960's till the mid 1970's, which was followed by a brief period of exponential growth (1975 – 1980). Since then the growth has returned to a linear trajectory again.

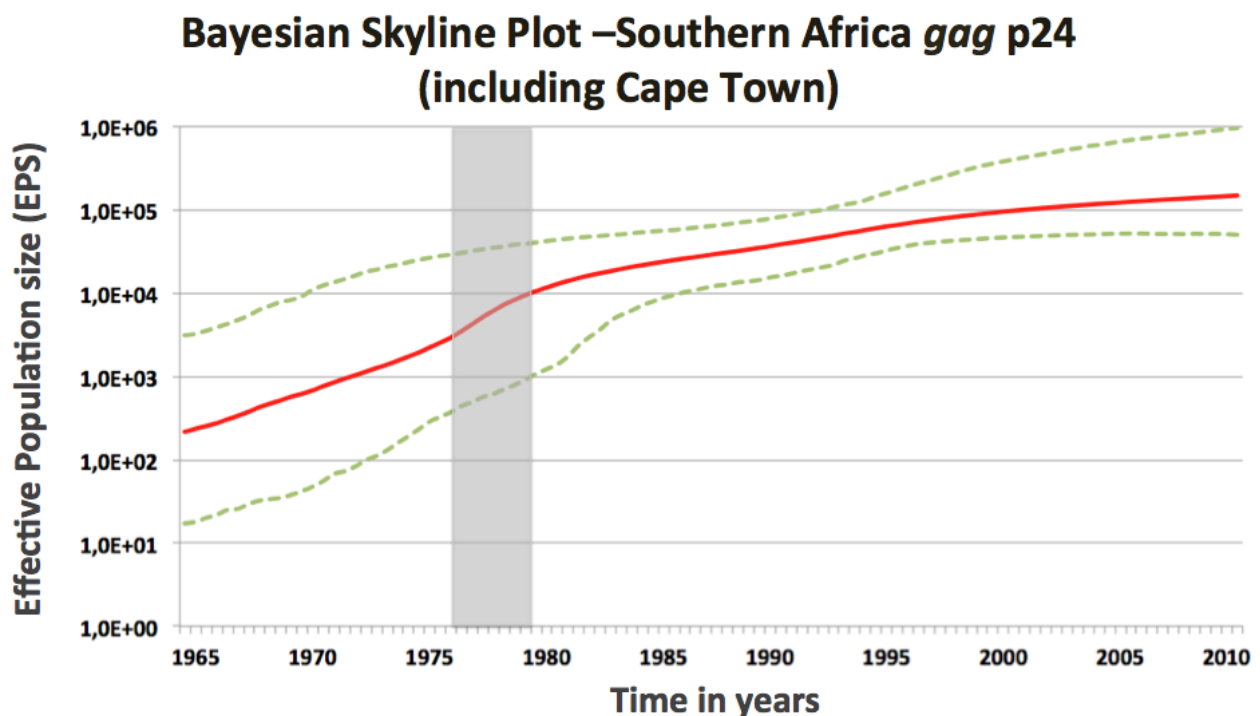


Figure 3.8: BSP of the Southern African *gag* p24 data set (including Cape Town). This non-parametric estimate of demographic history was reconstructed in Excel from raw data. The solid red lines are the “traced” median effective population size with the 95% upper and lower highest posterior density intervals indicated in the green dashed lines. This plot begins in the mid 1960's and stretches over roughly a 45-year period (1965 – 2010).

Close examination of the BSP of the entire Southern African *pol* data set (Figure 3.9) reveals a strikingly similar course in epidemic growth when compared to the data that was inferred from the entire Southern African *gag* p24 data set. As with the *gag* p24 data, the *pol* BSP indicates a period of slow linear epidemic growth from the mid 1960's till the late 1970's. This period then

followed by a brief period (1977 - 1983) of exponential epidemic growth. After the period of exponential growth the epidemic growth returns to a more linear phase till the mid 1990's. Once again a possible stabilization in the epidemic growth can be observed in the period from the mid 1990's till the present.

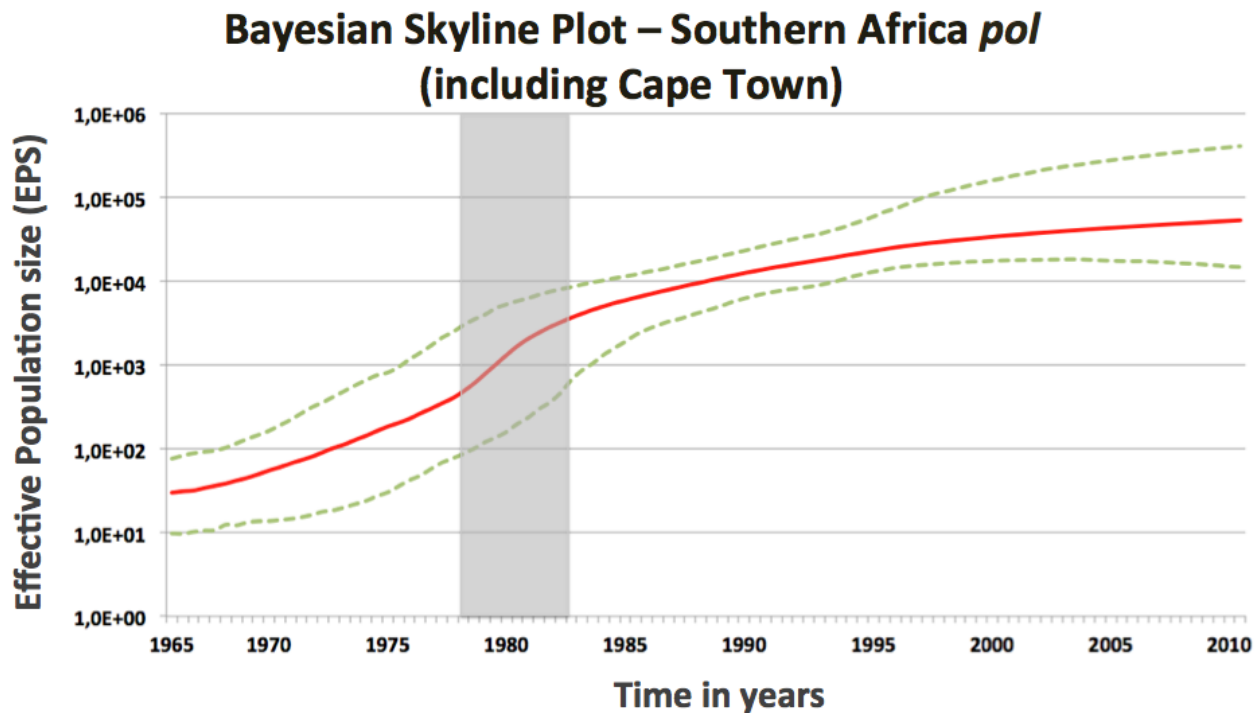


Figure 3.9: BSP of the Southern African *pol* data set (including Cape Town). This non-parametric estimate of demographic history was reconstructed in Excel from raw data. The solid red lines are the “traced” median effective population size with the 95% upper and lower highest posterior density intervals indicated in the green dashed lines. This plot begins in the mid 1960's and stretches over roughly a 45-year period (1965 – 2010).

In summary, the reconstruction of the Southern African epidemic, including sequence data from the Cape Town data sets, suggests a slow linear increase in the effective population size (EPS) from the mid 1960's till the mid 1970's. This was then followed by a brief period of epidemic expansion with exponential increases in the EPS in the mid to late 1970's. During the 1980's the growth in the epidemic returned to a linear trajectory. When looking at the 95% HPD intervals it would seem that another period of epidemic expansion occurred during the late 1990's (particularly when looking at the upper 95% confidence interval).

3.4 Results of the basic phylogenetic investigation into the evolutionary relationship of the Cape Town isolates

3.4.1 Outcome of the data mining

The data mining of homologous sequences from the HIV sequence database at the Los Alamos National Laboratory (<http://www.hiv.lanl.gov>) produced 2347 *gag* p24 and 5377 *pol* homologous sequences respectively. The screening of sequences with the HIV-1 Sequence Quality Analysis Tool (<http://bioafrica.mrc.ac.za>) revealed a total of 645 duplicate *gag* p24 and 398 duplicate *pol* sequences. No duplicates were found in the concatenated *gag-pol* data set. These duplicates were removed manually while the remaining sequences were combined with the Cape Town sequences to form three large data sets. Due to the large size of the *pol* data set (initially > 5,000) phylogenies were difficult to compute. Therefore this data set was reduced by roughly half.

The large *gag* p24 data set contained a total of 1895 sequences, including 193 sequences from Cape Town. A full break down of the large *gag* p24 data set is presented in Table 3.16.

Table 3.16: The complete *gag* p24 data set for the large-scale phylogenetic inference.

Final <i>gag</i> p24 data set for large-scale phylogenetic inference					
Code	Name of country	Number of taxa	Code	Name of country	Number of taxa
AR	Argentina	1	JP	Japan	1
AU	Australia	3	KE	Kenya	12
BR	Brazil	17	MW	Malawi	9
BW	Botswana	224	NG	Nigeria	1
CA	Canada	7	SA	Saudi Arabia	16
CD	Democratic Republic of the Congo	6	SN	Senegal	5
CM	Cameroon	2	SO	Somalia	1
CPT	Cape Town	193	TZ	Tanzania	74
CY	Cyprus	10	UG	Uganda	3
DK	Denmark	1	US	United States of America	3
ES	Spain	3	UY	Uruguay	1
ET	Ethiopia	13	YE	Yemen	1
FR	France	18	ZA	South Africa	935
GB	United Kingdom	4	ZM	Zambia	285
IL	Israel	18	ZW	Zimbabwe	5
IN	India	23	Total		1895

Due to the lack of large sequence fragments available in public sequence databanks the reference sequences for the *gag-pol* concatenated data set was largely restricted to all available full (FLG) or near full-length genomes (NFLG) of HIV-1 subtype C. The final *gag-pol* concatenated data set contained a total of 507 sequences including the 52 patient samples from Cape Town for whom

both *gag* p24 and *pol* sequence data was available. Once again a full break down of the large concatenated *gag-pol* data set is presented in Table 3.17.

Table 3.17: The complete *gag-pol* concatenated data set for the large-scale phylogenetic inference.

Final <i>gag-pol</i> concatenated data set for large-scale phylogenetic inference					
Code	Name of country	Number of taxa	Code	Name of country	Number of taxa
AR	Argentina	1	MM	Myanmar	1
BR	Brazil	8	MW	Malawi	1
BW	Botswana	50	SN	Senegal	1
CPT	Cape Town	52	SO	Somalia	1
CN	China	2	TZ	Tanzania	21
CY	Cyprus	11	US	USA	2
DK	Denmark	1	UY	Uruguay	1
ES	Spain	4	YE	Yemen	1
ET	Ethiopia	7	ZA	South Africa	304
IL	Israel	5	ZM	Zambia	16
IN	India	14	Total		507
KE	Kenya	3			

Similarly, the final *pol* data set contained a total of 2333 sequences including 166 sequences from Cape Town. A full break down of the large *pol* data set is presented in Table 3.18.

Table 3.18: The complete *pol* data set for the large-scale phylogenetic inference.

Final <i>pol</i> data set for large-scale phylogenetic inference					
Code	Name of country	Number of taxa	Code	Name of country	Number of taxa
AR	Argentina	9	KR	South Korea	2
AT	Austria	3	LC	Saint Lucia	1
BE	Belgium	20	LU	Luxembourg	4
BI	Burundi	30	ML	Mali	2
BR	Brazil	178	MM	Myanmar	1
BW	Botswana	119	MW	Malawi	87
BY	Belarus	1	MZ	Mozambique	79
BZ	Belize	2	NG	Nigeria	6
CD	Democratic Republic of the Congo	23	NL	Netherlands	9
CH	Switzerland	13	NP	Nepal	2
CM	Cameroon	1	PH	Philippines	1
CN	China	10	PK	Pakistan	1
CPT	Cape Town	166	PL	Poland	2
CS	Serbia and Montenegro	1	PT	Portugal	23
CU	Cuba	15	RO	Romania	23
CY	Cyprus	11	RU	Russia	5
CZ	Czech Republic	27	SD	Sudan	8
DE	Germany	7	SE	Sweden	26
DK	Denmark	21	SN	Senegal	23
ES	Spain	20	SZ	Swaziland	32
ET	Ethiopia	40	TH	Thailand	2
FI	Finland	4	TW	Taiwan	1
FR	France	8	TZ	Tanzania	48
GA	Gabon	1	UA	Ukraine	2
GB	United Kingdom	9	UG	Uganda	28
GE	Georgia	1	US	the United States of America	25
GQ	Equatorial Guinea	1	VE	Venezuela	1
HN	Honduras	1	YE	Yemen	5
IL	Israel	5	ZA	South Africa	635
IN	India	183	ZM	Zambia	119
IT	Italy	22	ZW	Zimbabwe	125
JP	Japan	7	Total		2333
KE	Kenya	42			

3.4.2 Large-scale phylogenetic inference

Large-scale tree topologies were inferred from the three large data sets that were obtained in the previous section. For the inference of ME-tree topologies exploratory runs were performed with the ME-NNI method of tree inference in order to check for any possible miss aligned sequences (e.g. taxa with extremely long branch lengths). Following this basic exploratory investigation these alignments were then used for the construction of large-scale ME-SPR tree topologies, which took a considerable amount of time due to the exhaustive nature of the tree inference method that was used. The ME tree inference with bootstrap resampling (n = 100 replicates) took three day for the *gag* p24 data set, one day for the concatenated *gag-pol* data set, and six days for the *pol* data set. Furthermore, Maximum Likelihood (ML) tree topologies were also inferred from

the three different data sets. These ML-tree topologies were inferred with approximate likelihood ratio test (aLRT) and took 16 days for the *gag* p24 data set, one day for the concatenated *gag-pol* data set and 46 days for the *pol* data set. Each of the inferred phylogenies was inspected in FigTree v 1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree>) and manually edited for better visual interpretation.

In the manual investigation of the *gag* p24 ME-SPR tree topology (Figure 3.10) a close clustering of Cape Town isolates with other Cape Town or South African isolates was observed. A total of 80 Cape Town isolates (40,6%) clustered with other Cape Town isolates, while 81 (41,1%) of the sequences clustered with isolates from other areas of South Africa. Five Cape Town sequences clustered with East African isolates (e.g. Kenya, Tanzania) and another five sequences cluster with isolates from other areas of the world (e.g. Europe and the Middle East). The remaining 26 Cape Town sequences (13,2%) in the *gag* p24 data set clustered with isolates from other Southern African countries. Several large monophyletic clades of Cape Town isolates were observed in this ME-SPR tree topology. The largest of which contained 15 Cape Town isolates. However, the bootstrap support for this internal branch of the cluster was 2,0%. In addition to the large monophyletic cluster of Cape Town isolates, highly monophyletic clades were also observed for the Indian (bootstrap = 0,0%) and South American (bootstrap = 0,0%) HIV-1 subtype C clusters.

The manual inspection of the ML.aLRT tree topology of the large *gag* p24 data set (Figure 3.11) a similar close relationship between the Cape Town isolates and other South African isolates was observed. A total of 85 Cape Town isolates (43,1%) clustered with other Cape Town isolates in the tree topology, while 81 of the Cape Town isolates (41,1%) clustered with other South African isolates. Six of the Cape Town isolates (3,0%) clustered with HIV-1 subtype C sequences that originated in other areas of the world, such as East Africa, India and Europe. The remaining 25 Cape Town isolates (12,7%) clustered with other isolates from the Southern African region.

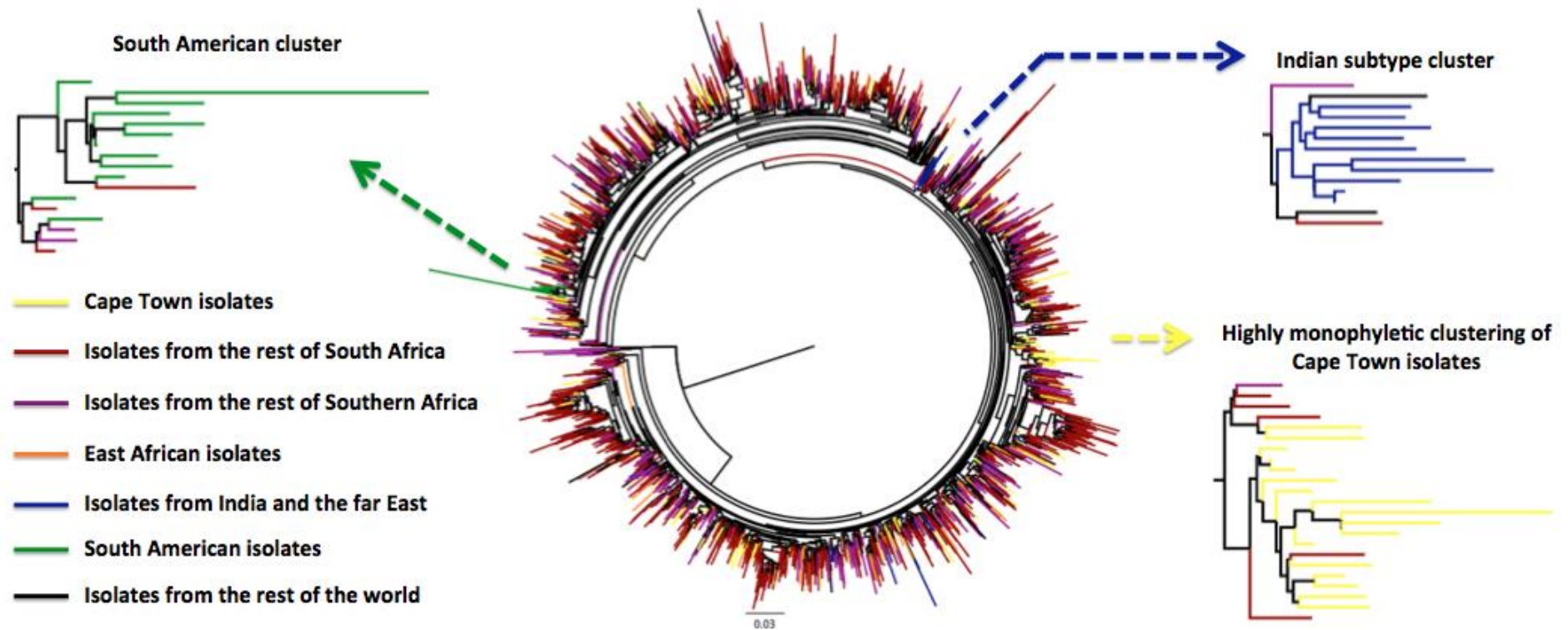


Figure 3.10: Large-scale ME-SPR tree of the *gag* p24 data set. The tree contains a total of 1985 taxa and was constructed in fastME, with the use of the HKY85+G ($\alpha = 0.8$) method of nucleotide substitution. The genetic distance is shown in the bottom line and corresponds to the length of the branches. Each of the branches has been colour coded to correspond to the place of origin of each of the taxa in the tree topology. A distinct South American, Asian (excluding the Middle East and the former Soviet Union) and one highly monophyletic Cape Town cluster have been highlighted. Due to the sheer size of the tree topology the bootstrap support values for each of the branches have been removed.

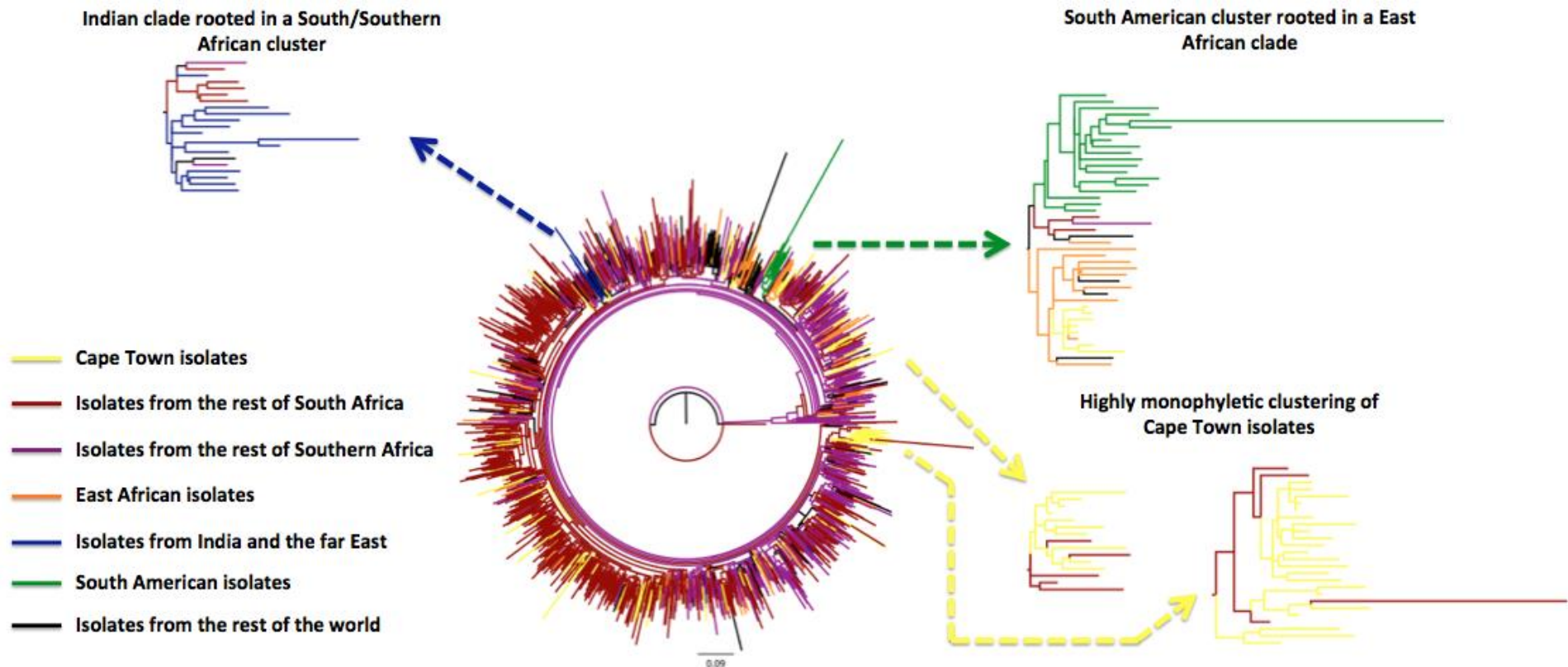


Figure 3.11: Large-scale ML tree with aLRT of the *gag* p24 data set. The tree contains a total of 1985 taxa and was constructed in phyML v 3.0, with the use of the HKY85+G ($\alpha = 0.8$) method of nucleotide substitution. The genetic distance is shown in the bottom line and corresponds to the length of the branches. Each of the branches has been colour coded to correspond to the place of origin of each of the taxa in the tree topology. A distinct South American, Asian (excluding the Middle East and the former Soviet Union) and two highly monophyletic Cape Town cluster have been highlighted. Due to the sheer size of the tree topology the aLRT support values for each of the branches have been removed.

The manual inspection of the ME-SPR *gag-pol* concatenated tree topology (Figure 3.12) showed a close relationship amongst Cape Town isolates. Furthermore, a large number of Cape Town isolates also showed a close relationship with other South African isolates. A total of 27 Cape Town isolates (51,9%) cluster with other Cape Town isolates, while 15 Cape Town isolates (28,8%) clustered with South African sequences. Seven of the Cape Town sequences clustered with isolates from other Southern African nations, while the remaining three Cape Town isolates in the data set clustered with subtype C isolates from other areas of the world (e.g. East Africa or Europe). Two highly monophyletic clusters were observed in the tree topology. The first was a pure monophyletic cluster of 11 Cape Town sequences, with a bootstrap support of 4% for the internal branch of this cluster. The second clade contained 10 Cape Town isolates broken eight times by isolates from Zimbabwe, Botswana, Malawi, South Africa and Tanzania. The bootstrap support for this internal branch was 0,0%. Furthermore, the other two monophyletic clusters that were observed were the Indian and Brazilian HIV-1 subtype C clades, with bootstrap support of 9,0% and 95,0% respectively. The Indian clade was rooted in a cluster of South/Southern African isolates, while the Brazilian clade were rooted in a clade of East African isolates (Figure 3.12).

Close examination of the ML.aLRT tree topology that was inferred from the concatenated *gag-pol* data set (Figure 3.13) showed a similar close relationship of Cape Town isolates with one another. A total of 32 of the Cape Town isolates in this data set (61,5%) clustered with other Cape Town isolates, while 12 of the Cape Town isolates (23,1%) clustered with other South African reference strains in the data set. Of the remaining seven isolates, 6 all clustered with isolates from other Southern African countries (11,5%), while one clustered with an East African isolates (1,9%). Two Cape Town clusters were observed in the tree topology. One was a pure monophyletic cluster contained 11 Cape Town isolates with an aLRT support of 0,98. The other cluster contained 6 Cape Town patients and was broken three times, once by an isolate from Tanzania, and twice by South African isolates. The aLRT support for the internal branch of this cluster was 0,63. A similar clustering pattern was observed in the ML-tree topology for Indian and Brazilian HIV-1 subtype C isolates as in the corresponding ME-tree topology.

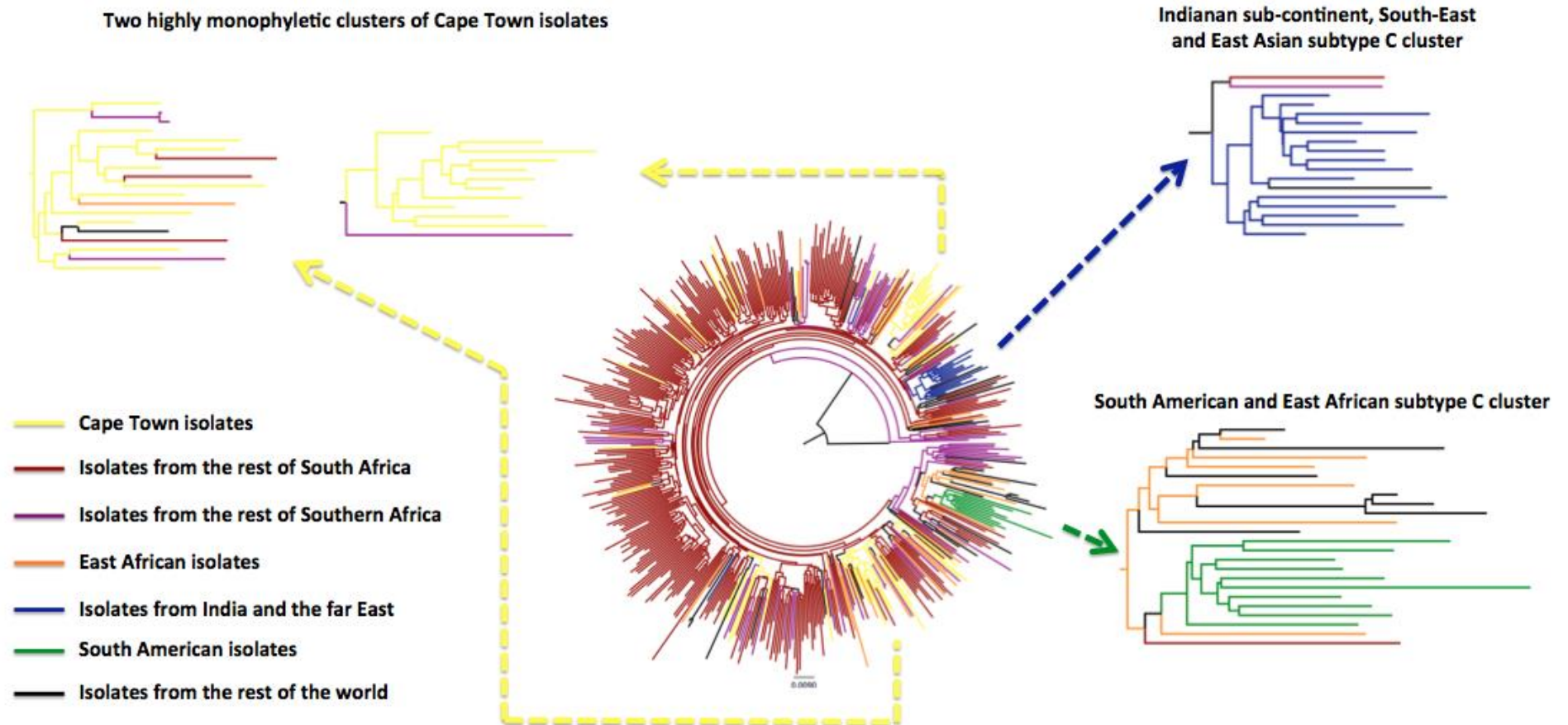


Figure 3.12: Large-scale ME-SPR tree of the *gag-pol* concatenated data set. The tree contains a total of 507 taxa and was constructed in fastME, with the use of the HKY85+G ($\alpha = 0.8$) method of nucleotide substitution. The genetic distance is shown in the bottom line and corresponds to the length of the branches. Each of the branches has been colour coded to correspond to the place of origin of each of the taxa in the tree topology. A distinct South American, Asian (excluding the Middle East and the former Soviet Union) and two monophyletic Cape Town cluster have been highlighted. Due to the sheer size of the tree topology the bootstrap support values for each of the branches have been removed.

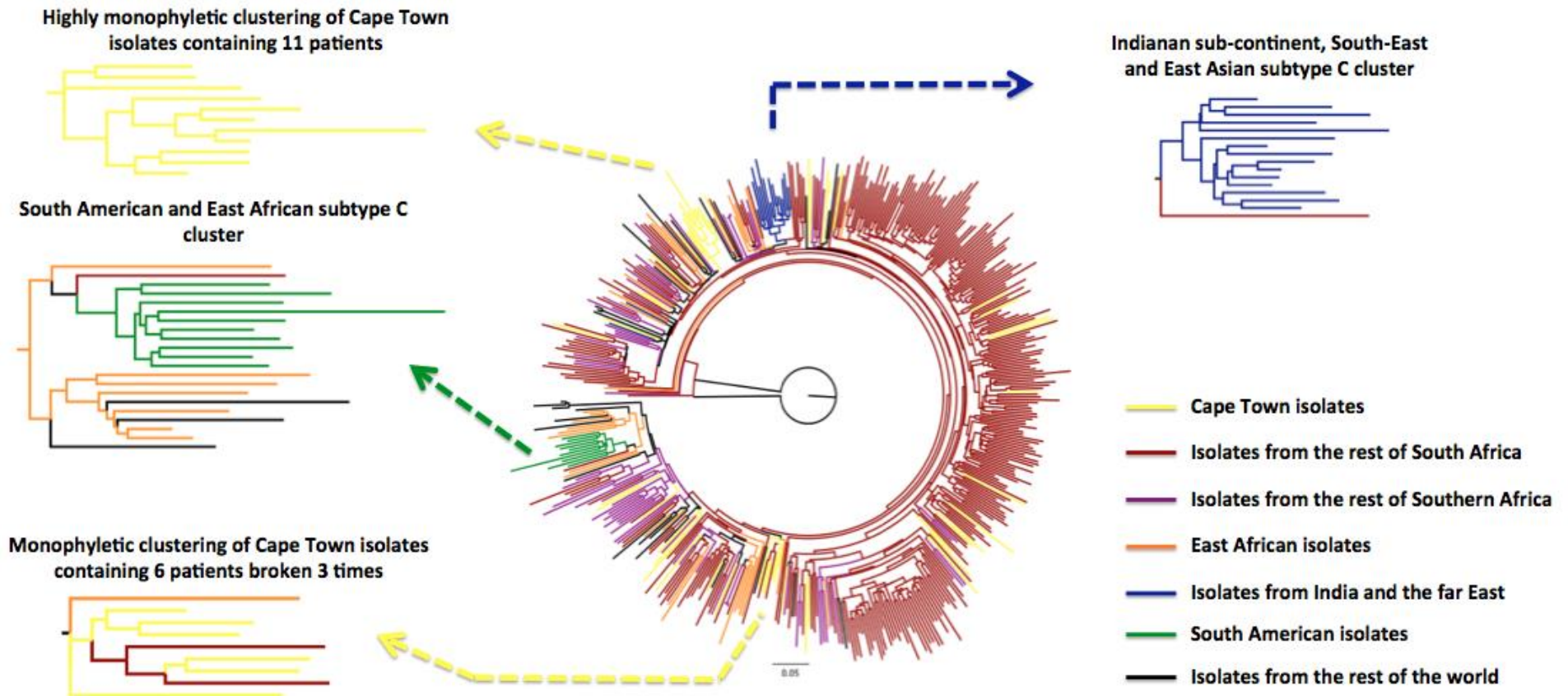


Figure 3.13: Large-scale ML tree with aLRT of the *gag-pol* concatenated data set. The tree contains a total of 507 taxa and was constructed in phyML v 3.0, with the use of the HKY ($\alpha = 0.8$) method of nucleotide substitution. The genetic distance is shown in the bottom line and corresponds to the length of the branches. Each of the branches has been colour coded to correspond to the place of origin of each of the taxa in the tree topology. A distinct South American, Asian (excluding the Middle East and the former Soviet Union) and two monophyletic Cape Town cluster have been highlighted. Due to the sheer size of the tree topology the aLRT support values for each of the branches have been removed.

The manual inspection of the ME-SPR *pol* tree topology (Figure 3.14) also indicated a close relationship between Cape Town sequences with themselves and other subtype C isolates from South Africa. A total of 69 (41,6%) sequences out of the entire Cape Town *pol* data set clustered with other Cape Town sequences. A total of 61 (37,0%) of the Cape Town *pol* sequences in the data set clustered with other South African isolates. Four Cape Town *pol* sequences clustered with European isolates, one with an Indian isolate, two with isolates from the US, while the remaining 27 (16,4%) sequences in the *pol* data set clustered with other subtype C isolates from other Southern African countries (e.g. Zimbabwe, Botswana, Swaziland and Mozambique). Furthermore, three highly monophyletic clades of Cape Town sequences were observed. The first contained 10 Cape Town isolates. The other two clades contained 12 and 21 Cape Town isolates respectively. No bootstrap support (0,0%) was obtained for any of these clades. Similarly, highly distinct clustering of Indian and Brazilian taxa was observed. The Indian clade was rooted in a small cluster of Southern Africa sequences (bootstrap = 0,0%), while the Brazilian clade was rooted in a larger East African cluster (bootstrap = 0,0%).

Lastly, the manual inspection of the ML-aLRT tree topology that was inferred from the large-scale *pol* data set (Figure 3.15) also revealed a close clustering of Cape Town isolates with one another. In total 83 Cape Town isolates (50,2%) clustered with other Cape Town *pol* sequences in the data set, while 59 of the Cape Town isolates (35,8%) clustered with other South African sequences. Of the remaining Cape Town isolates, 16 clustered with other isolates from Southern Africa (9,7%), while 7 isolates (4,3%) clustered with other subtype C isolates from around the world. Furthermore, two highly monophyletic clades of Cape Town sequences were observed. The first contained 10 Cape Town isolates clustering with 5 other South African isolates. The aLRT support for the internal branch of this cluster was 0,899. The second cluster contained 20 Cape Town isolates clustering along with two isolates from Tanzania. The aLRT support for the internal branch of this cluster was 0,886. Highly monophyletic clades were also observed for the Indian and Brazilian taxa in the ML.aLRT tree topology. The aLRT support values for the Indian and Brazilian clades were 0,918 and 0,982 respectively.

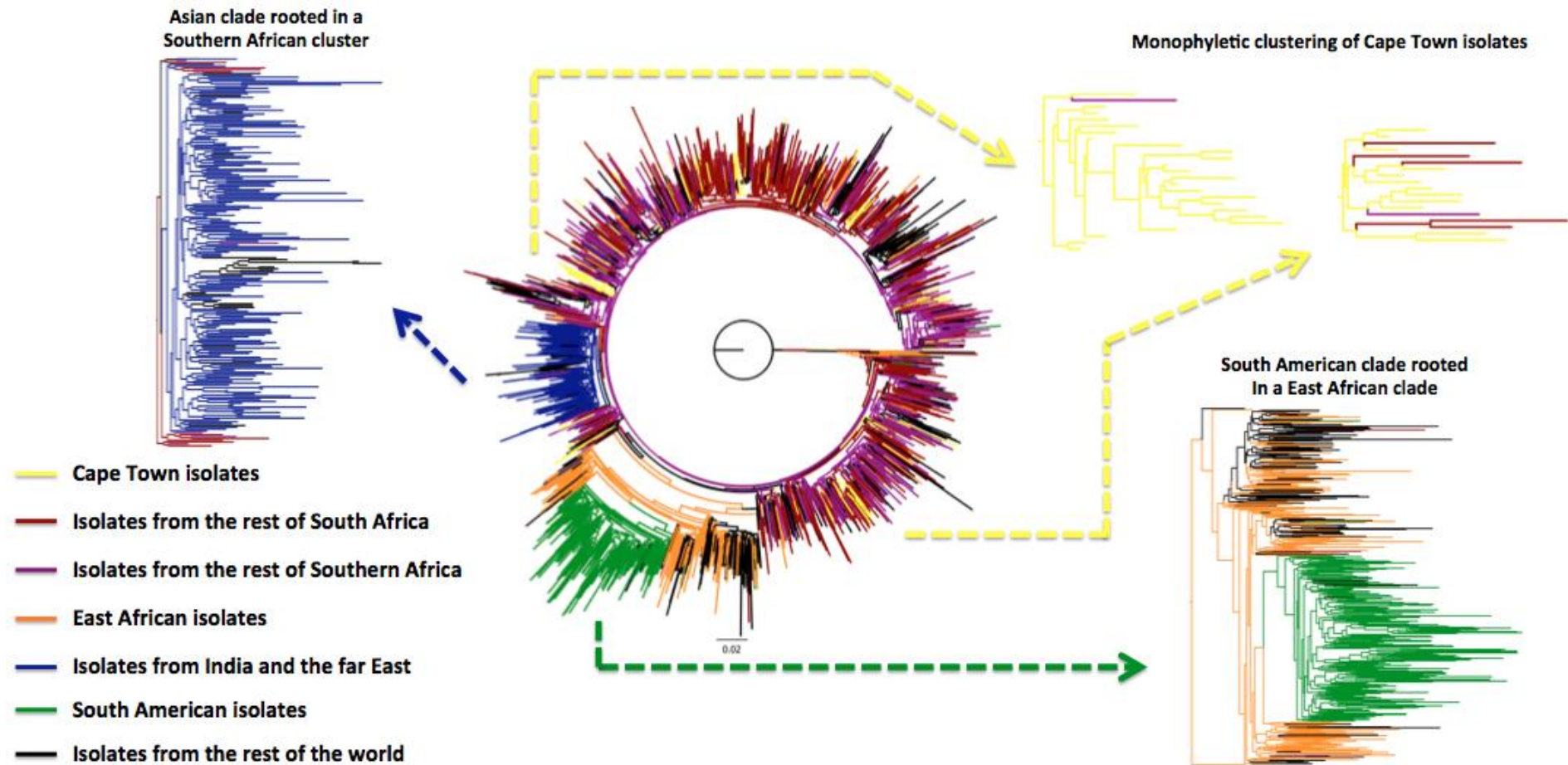


Figure 3.14: Large-scale ME-SPR tree of the *pol* data set. The tree contains a total of 2334 taxa and was constructed in fastME, with the use of the HKY85+G (alpha = 0.8) method of nucleotide substitution. The genetic distance is shown in the bottom line and corresponds to the length of the branches. Each of the branches has been colour coded corresponding to the place of origin of each of the taxa. Large monophyletic clusters of Cape Town isolates have also been highlighted. Due to the sheer size of the tree topology the bootstrap support values for each of the branches have been removed.

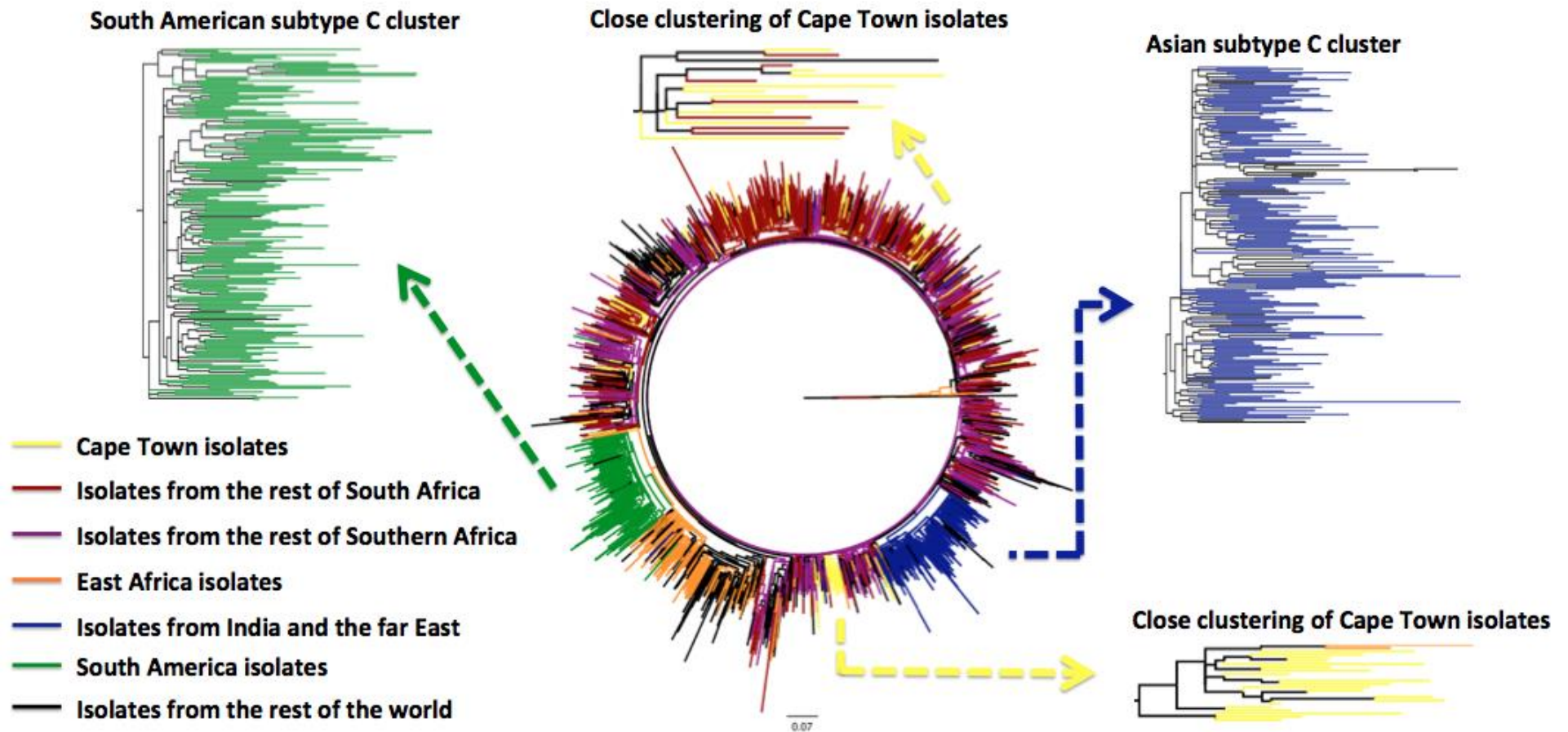


Figure 3.15: Large-scale ML tree with aLRT of the *pol* data set. The tree contains a total of 2334 taxa and was constructed in phyML, with the use of the HKY85+G (alpha = 0.8) method of nucleotide substitution. The genetic distance is shown in the bottom line and corresponds to the length of the branches. Each of the branches has been colour coded corresponding to the place of origin of each of the taxa. A clear clustering of Asian, East African and South American samples are visible and have been highlighted. Similarly large monophyletic clusters from the Cape Town data set have also been highlighted. Due to the sheer size of the tree topology the aLRT support values for each of the branches have been removed.

3.4.3 Outcome of the PhyloType analysis of large tree topologies

Taxa in these large trees were classified into different temporal classes and analysed with the PhyloType application (<http://lamarck.lirmm.fr/phyloptype/>) to assess the clustering of samples based on temporal and geographical criteria.

The PhyloType analysis of the large-scale *gag* p24 ME-SPR tree topology revealed a strong clustering of Cape Town sequences with one another (Table 3.19). Nearly three quarters (72,0%) of the oldest Cape Town sequences (1989 - 1992) contained in the data set clustered with other old Cape Town sequences. This is remarkable since these sequences make up only 1,3% of the total *gag* p24 data set. Similarly, 33,0% of the Cape Town isolates that was sampled between 2000 and 2005 clustered with one another, considering these Cape Town isolates only makes up 4,8% of the entire *gag* p24 data set. Lastly, the most recently sampled Cape Town isolates in the data set also showed a fairly high clustering with one another (21,0%).

Similarly, the large-scale *gag* p24 ML-tree topology with aLRT was also investigated with the PhyloType software application (Table 3.20). A very large proportion (80,0%) of the oldest Cape Town sequences in the data set, which were sampled between 1989 and 1992, clustered with other Cape Town isolates that were sampled in the same period. This is highly significant since these Cape Town isolates only comprise 1,3% of the entire *gag* p24 data set. Furthermore, 40,0% of the Cape Town isolates that were sampled between 2000 and 2005 also clustered with one another, even though they only comprise 4,8% of the entire data set. Lastly, 34,0% of the more recently samples Cape Town isolates (2006 till the present) clustered with one another, while they only comprise 4,1% of the full *gag* p24 data set.

Table 3.19: Results of the PhyloType analysis of the *gag* p24 ME-SPR tree topology.

Category or criteria	Number of clusters	Average size of cluster	Total number of taxa in cluster(s)	Total number of taxa that fits this criteria	Percentage of this criteria that cluster with taxa of the same criteria	Number of taxa with criteria in total data set
BR_Middle	1	4,00	4	6	67,00%	0,30%
BR_Youngest	2	2,00	4	9	44,00%	0,50%
BW_Middle	1	2,00	2	25	8,00%	1,30%
BW_Youngest	17	3,12	53	133	40,00%	7,00%
CPT_Middle	10	3,00	30	91	33,00%	4,80%
CPT_Oldest	4	4,50	18	25	72,00%	1,30%
CPT_Youngest	6	2,67	16	78	21,00%	4,10%
ET_Oldest	2	2,00	4	11	36,00%	0,60%
FR_Oldest	3	2,00	6	19	32,00%	1,00%
IL_Middle	1	2,00	2	13	15,00%	0,70%
IN_Middle	1	2,00	2	6	33,00%	0,30%
IN_Oldest	1	2,00	2	13	14,00%	0,70%
KE_Oldest	1	2,00	2	9	20,00%	0,50%
SA_Middle	2	2,00	4	15	25,00%	0,80%
TZ_Middle	4	2,50	10	66	15,00%	3,50%
ZA_Middle	60	5,70	342	464	74,00%	24,50%
ZA_None	11	2,18	24	203	12,00%	10,70%
ZA_Oldest	1	2,00	2	61	30,00%	3,20%
ZA_Youngest	24	2,25	54	208	26,00%	11,00%
ZM_Middle	44	2,52	111	224	50,00%	11,80%
ZM_Youngest	2	2,00	4	32	13,00%	1,70%

Table 3.20: Results of the PhyloType analysis of the *gag* p24 ML.aLRT tree topology.

Category or criteria	Number of clusters	Average size of cluster	Total number of taxa in cluster(s)	Total number of taxa that fits this criteria	Percentage of this criteria that cluster with taxa of the same criteria	Number of taxa with criteria in total data set
BR_Oldest	1	2,00	2	2	100,00%	0,10%
BR_Youngest	1	6,00	6	9	67,00%	0,50%
BW_None	3	2,00	6	46	13,00%	2,50%
BW_Youngest	19	3,16	60	133	45,00%	7,00%
CPT_Middle	13	2,77	36	90	40,00%	4,80%
CPT_Oldest	6	3,33	20	25	80,00%	1,30%
CPT_Youngest	10	2,60	26	76	34,00%	4,10%
ET_Oldest	3	2,33	7	11	64,00%	0,10%
FR_Oldest	3	3,67	11	19	58,00%	1,00%
IL_Middle	1	2,00	2	13	15,00%	0,70%
IN_Middle	1	3,00	3	6	50,00%	0,30%
IN_None	1	2,00	2	3	67,00%	0,20%
IN_Oldest	2	3,50	7	14	50,00%	0,70%
KE_Oldest	1	6,00	6	10	60,00%	0,50%
SA_Middle	2	2,50	5	16	31,00%	0,80%
SN_Oldest	1	2,00	2	5	40,00%	0,30%
TZ_Middle	6	2,67	16	67	24,00%	3,50%
TZ_Oldest	1	2,00	2	4	50,00%	0,20%
ZA_Middle	40	8,88	355	461	77,00%	24,50%
ZA_None	25	2,44	61	203	30,00%	10,70%
ZA_Oldest	7	2,14	15	60	25,00%	3,20%
ZA_Youngest	24	2,46	59	211	28,00%	11,00%
ZM_Middle	43	3,26	140	222	63,00%	11,80%
ZM_None	2	2,00	4	20	20,00%	1,10%
ZM_Youngest	1	2,00	2	33	6,00%	1,70%

The PhyloType analysis of the *gag-pol* ME-SPR tree topology (Table 3.21) revealed also a high degree of clustering of Cape Town isolates. A total of 92,0% of the oldest Cape Town isolates in the data set clustered with one another, while they only comprise 5,70% of the entire data set. Similarly, 34,0% of the most recently sampled Cape Town isolates in the data set also clustered with one another even though these Cape Town isolates only comprise a total of 2,2% of the entire *gag-pol* concatenated data set.

The analysis of the ML.aLRT tree topology that was constructed from the concatenated *gag-pol* data set with the PhyloType application also revealed very similar results (Table 3.22). A total of 79,0% of the oldest Cape Town isolates in the data set clustered with one another, while they only comprise 5,70% of the entire data set. Similarly, 48,0% of the most recently sampled Cape Town isolates in the data set also clustered with one another even though these Cape Town isolates only comprise a total of 2,2% of the entire *gag-pol* concatenated data set.

Table 3.21: PhyloType analysis of *gag-pol* ME-SPR phylogeny based on temporal and geographical criteria.

Category or criteria	Number of clusters	Average size of cluster	Total number of taxa in cluster(s)	Total number of taxa that fits this criteria	Percentage of this criteria that cluster with taxa of the same criteria	Number of taxa with criteria in total data set
BR_Middle	1	5,0	5	5	100,00%	1,00%
BR_Oldest	1	2,0	2	2	100,00%	5,30%
BW_Middle	3	2,0	6	27	22,00%	4,50%
BW_Oldest	1	3,0	3	23	13,00%	4,70%
CPT_Oldest	3	7,3	22	24	92,00%	5,70%
CPT_Youngest	5	2,0	10	29	34,00%	2,20%
CY_Youngest	1	3,0	3	9	33,00%	0,80%
ET_Oldest	1	3,0	3	6	50,00%	1,20%
IN_Oldest	1	9,0	9	11	82,00%	2,20%
TZ_Middle	3	2,0	6	17	35,00%	3,40%
TZ_Oldest	1	2,0	2	4	50,00%	0,80%
ZA_Middle	2	128,0	256	261	98,00%	46,80%
ZA_Oldest	2	2,0	4	24	17,00%	4,50%
ZM_Middle	2	2,5	5	12	42,00%	7,90%

Table 3.22: PhyloType analysis of *gag-pol* ML.aLRT phylogeny based on temporal and geographical criteria.

Category or criteria	Number of clusters	Average size of cluster	Total number of taxa in cluster(s)	Total number of taxa that fits this criteria	Percentage of this criteria that cluster with taxa of the same criteria	Number of taxa with criteria in total data set
BR_Middle	1	5,0	5	5	100,00%	1,00%
BR_Oldest	1	2,0	2	2	100,00%	5,30%
BW_Middle	3	2,0	6	27	22,00%	4,50%
BW_Oldest	3	2,7	8	23	35,00%	4,70%
CPT_Oldest	4	4,8	19	24	79,00%	5,70%
CPT_Youngest	7	2,0	14	29	48,00%	2,20%
CY_Youngest	1	3,0	3	9	33,00%	0,80%
ET_Oldest	2	3,0	6	6	100,00%	1,20%
IN_Oldest	1	9,0	9	11	82,00%	2,20%
TZ_Middle	3	2,3	7	17	41,00%	3,40%
TZ_Oldest	1	2,0	2	4	50,00%	0,80%
ZA_Middle	1	260,0	260	260	100,00%	46,80%
ZA_Oldest	1	2,0	2	22	9,00%	4,50%
ZM_Middle	2	2,5	5	12	42,00%	7,90%

The analysis of the ME-SPR tree topology that was inferred from the large *pol* data set revealed similarly high clustering of Cape Town isolates with one another (Table 3.23). A total of 73% of the oldest sequences in the Cape Town *pol* data set clustered with one another, while these sequences only comprise 1,9% of the entire *pol* data set. Similarly, 16,0% of the sequences that were sampled between 2000 and 2005 also clustered with one another. A total of 18,0% of the most recently sampled isolates in the Cape Town *pol* data set also clustered with one another. Similar, high clustering was observed in the Brazilian and Indian HIV-1 subtype C clades. A total of 92,0% of the Brazilian isolates that were sampled between 2000 and 2005 clustered with one another, while they only comprise 4,7% of the entire *pol* data set.

Lastly, the analysis of the ML.aLRT tree topology that was inferred from the large *pol* data set revealed a similarly high clustering of Cape Town isolates with one another (Table 3.24). In total 80,0% of the oldest Cape Town isolates clustered with one another, while these isolates only comprised 1,9% of the entire data set. A total of 21,0% of the Cape Town isolates that were sampled between 2000 and 2005 also clustered with one another. These Cape Town isolates only comprised 2,6% of the entire data set. Lastly, 28,0% of the more recently sampled Cape Town isolates also clustered with other Cape Town isolates from the same time period, while they only made up 2,6% of the final data set. As with the analysis of the ME-SPR tree topology similarly high clustering was observed for the Brazilian and Indian isolates. A total of 85,0% of the Brazilian isolates that were sampled between 2000 and 2005 clustered with one another, while 70,0% of the Brazilian sequences that were derived from more recently sampled specimens also clustered with one another. For the more recently sampled Indian samples a total of 92,0% of samples clustered with one another, while they only comprised 5,4% of the entire data set. Similarly, 39,0% and 43,0% of the Indian samples that were sampled during the temporal periods “Middle” and “Oldest” also clustered with one another.

Table 3.23: PhyloType analysis of *pol* ME-SPR phylogeny based on spatiotemporal criteria.

Category or criteria	Number of clusters	Average size of cluster	Total number of taxa in cluster(s)	Total number of taxa that fits this criteria	Percentage of this criteria that cluster with taxa of the same criteria	Number of taxa with criteria in total data set
AR_Middle	1	2,00	2	8	25,00%	0,30%
BE_Middle	3	2,00	6	11	55,00%	0,50%
BE_Oldest	2	3,50	7	9	78,00%	0,40%
BI_Middle	4	4,00	16	30	53,00%	1,30%
BR_Middle	3	33,67	101	110	92,00%	4,70%
BR_Youngest	5	5,80	29	66	44,00%	2,80%
BW_Middle	12	2,50	30	88	34,00%	3,80%
BW_Oldest	1	3,00	3	21	14,00%	0,90%
BZ_Middle	1	2,00	2	2	100,00%	0,10%
CN_Youngest	3	2,00	6	8	75,00%	0,30%
CPT_Middle	4	2,50	10	63	16,00%	2,60%
CPT_Oldest	4	8,00	32	44	73,00%	1,90%
CPT_Youngest	5	2,20	11	61	18,00%	2,60%
CU_Middle	3	3,00	9	14	64,00%	0,60%
CZ_Middle	1	3,00	3	10	30,00%	0,40%
CZ_Noneee	1	5,00	5	17	29,00%	0,70%
ES_Youngest	3	2,33	7	17	41,00%	0,70%
ET_Middle	7	2,29	16	36	44,00%	1,50%
ET_Oldest	1	2,00	2	4	50,00%	0,20%
GQ_Youngest	1	2,00	2	2	100,00%	0,10%
IN_Middle	1	4,00	4	44	9,00%	1,90%
IN_Youngest	2	58,00	116	125	93,00%	5,40%
IT_Middle	1	2,00	2	13	15,00%	0,60%
JP_Middle	1	2,00	2	7	29,00%	0,30%
KE_Youngest	5	4,40	22	38	58,00%	1,60%
MW_Middle	2	2,00	4	19	21,00%	0,80%
MW_Youngest	8	2,00	16	67	24,00%	2,90%
MZ_Middle	10	2,50	25	78	32,00%	3,40%
NL_Middle	3	2,00	6	9	67,00%	0,40%
PT_Middle	2	2,00	4	18	22,00%	0,80%
PT_Youngest	1	2,00	2	5	40,00%	0,20%
RO_Middle	2	3,00	6	12	50,00%	0,50%
RO_Youngest	1	3,00	3	11	27,00%	0,50%
SE_Middle	3	2,33	7	18	39,00%	0,80%

SN_Middle	2	6,00	12	14	86,00%	0,60%
SN_Oldest	1	2,00	2	5	40,00%	0,20%
SZ_Middle	3	4,67	14	32	44,00%	1,40%
TZ_Middle	1	6,00	6	22	27,00%	0,90%
TZ_Oldest	1	2,00	2	4	50,00%	0,20%
TZ_Youngest	3	2,33	7	22	32,00%	0,90%
UA_Middle	1	2,00	2	2	100,00%	0,10%
UG_Youngest	1	3,00	3	14	21,00%	0,60%
ZA_Middle	15	2,60	39	150	26,00%	6,50%
ZA_Oldest	1	2,00	2	20	10,00%	0,90%
ZA_Youngest	58	6,00	348	464	75,00%	19,90%
ZM_Middle	2	2,00	4	44	9,00%	2,00%
ZM_Oldest	1	2,00	2	25	8,00%	1,00%
ZM_Youngest	8	2,00	16	47	34,00%	2,00%
ZW_Youngest	10	2,20	22	100	22,00%	4,40%

Table 3.24: PhyloType analysis of *pol* ML.aLRT phylogeny based on spatiotemporal criteria.

Category or criteria	Number of clusters	Average size of cluster	Total number of taxa in cluster(s)	Total number of taxa that fits this criteria	Percentage of this criteria that cluster with taxa of the same criteria	Number of taxa with criteria in total data set
AR_Middle	2	2,00	4	8	50,00%	0,30%
BE_Middle	3	2,00	6	11	55,00%	0,50%
BE_Oldest	2	3,50	7	9	78,00%	0,40%
BI_Middle	5	3,80	19	30	63,00%	1,30%
BR_Middle	8	11,63	93	109	85,00%	4,70%
BR_Youngest	6	7,67	46	66	70,00%	2,80%
BW_Middle	13	3,23	42	88	48,00%	3,80%
BW_Youngest	2	2,00	4	10	40,00%	0,40%
BZ_Middle	1	2,00	2	2	100,00%	0,10%
CN_Youngest	3	2,00	6	8	75,00%	0,30%
CPT_Middle	6	2,17	13	62	21,00%	2,60%
CPT_Oldest	3	11,67	35	44	80,00%	1,90%
CPT_Youngest	7	2,43	17	61	28,00%	2,60%
CU_Middle	2	5,50	11	14	79,00%	0,60%
CZ_Middle	1	3,00	3	10	30,00%	0,40%
CZ_None	1	6,00	6	17	35,00%	0,70%
ER_Middle	1	2,00	2	2	100,00%	0,10%
ES_Youngest	3	2,33	7	17	41,00%	0,70%
ET_Middle	6	2,83	17	36	47,00%	1,50%
GQ_Youngest	1	2,00	2	2	100,00%	0,10%
IN_Middle	4	4,25	17	44	39,00%	1,90%
IN_Oldest	3	2,00	6	14	43,00%	0,60%
IN_Youngest	4	28,75	115	125	92,00%	5,40%
IT_Middle	1	2,00	2	13	15,00%	0,60%
IT_Youngest	1	2,00	2	8	25,00%	0,30%
JP_Middle	1	2,00	2	7	29,00%	0,30%
KE_Youngest	5	5,00	25	38	66,00%	1,60%
MW_Youngest	11	2,46	27	68	40,00%	2,90%
MZ_Middle	9	3,22	29	78	37,00%	3,40%
NL_Middle	3	2,00	6	9	67,00%	0,40%
PT_Middle	2	2,00	4	18	22,00%	0,80%
PT_Youngest	1	2,00	2	5	40,00%	0,20%
RO_Middle	2	3,00	6	12	50,00%	0,50%
RO_Youngest	1	3,00	3	11	27,00%	0,50%

SE_Middle	3	2,00	6	18	33,00%	0,80%
SN_Middle	2	6,00	12	14	86,00%	0,60%
SZ_Middle	4	3,50	14	32	44,00%	1,40%
TZ_Middle	1	5,00	5	22	23,00%	0,90%
TZ_Oldest	1	2,00	2	4	50,00%	0,20%
TZ_Youngest	3	5,00	15	22	68,00%	0,90%
UA_Middle	1	2,00	2	2	100,00%	0,10%
UG_Middle	1	2,00	2	5	40,00%	0,20%
UG_Youngest	2	2,00	4	14	29,00%	0,60%
ZA_Middle	17	2,65	45	150	30,00%	6,50%
ZA_Youngest	35	10,37	363	465	78,00%	19,90%
ZM_Middle	4	2,00	8	47	17,00%	2,00%
ZM_Oldest	1	2,00	2	25	8,00%	1,00%
ZM_Youngest	6	2,00	12	46	26,00%	2,00%
ZW_Middle	3	2,33	7	23	30,00%	1,00%
ZW_Youngest	17	2,59	44	19	43,00%	4,40%

This PhyloType method is a new and easy way of analysing the clustering of taxa in tree topologies, particularly large trees, which may not always be possible to do manually. As the manual interpretation of clustering may be subject to some bias, this method also provides an unbiased assessment of the clustering relationship in any tree topology. In Table 3.6 the clustering of various taxa based on spatiotemporal classifications were assessed for the concatenated *gag-pol* ME-SPR tree topology. For the Cape Town sequences contained in this data set there were 2 clusters containing old sequences (sampled between 1989 and 1992) with the average size of these clusters measuring 10 taxa per cluster. Additionally, there were 4 clusters, where Cape Town sequences which belonged to the most recently sampled (2006 - 2010), clustered with one another with the average size measuring 2,5 taxa per cluster. Manual inspection of the tree topology that was used for the PhyloType analysis can confirm these findings (Figure 3.16).

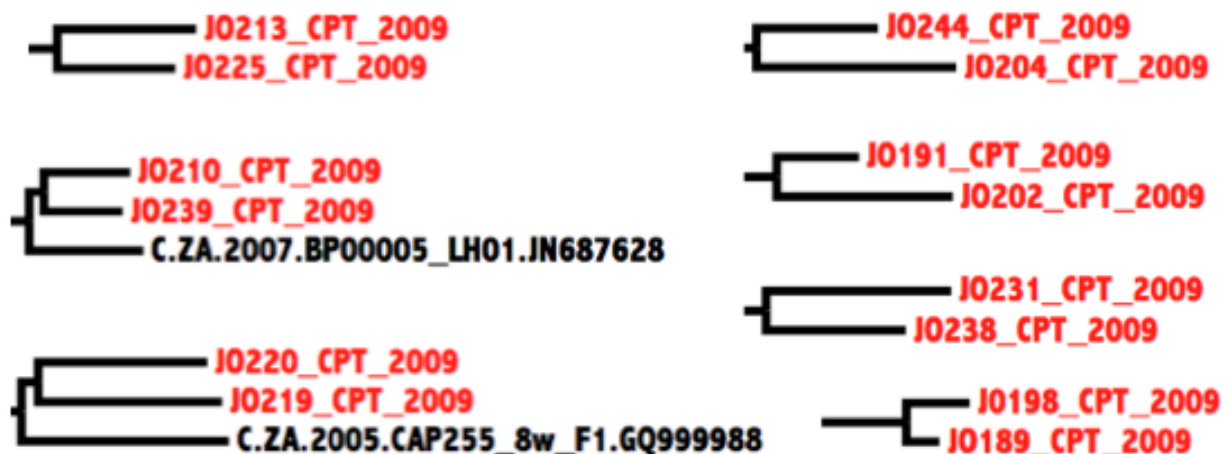


Figure 3.16: Observed clustering of Cape Town isolates in the concatenated *gag-pol* ML.aLRT tree topology as was observed through the PhyloType analysis. The 7 different clusters with the youngest Cape Town isolates are shown. Each cluster contains on average 2,0 taxa from Cape Town corresponding to the average that was obtained in PhyloType. There are 29 Cape Town isolates in the concatenated *gag-pol* data set that satisfy this criteria which means that on average roughly 48,3% of the isolates clustered with an isolate in the same spatiotemporal class (in correlation with data presented in Table 3.21).

3.5 Large monophyletic clades of HIV-1 subtype C in Cape Town

During the examination of the large-scale *gag* p24, *gag-pol*, and *pol* tree topologies highly monophyletic clusters of Cape Town isolates were observed, which are very unique for the clustering of patient samples in a generalized HIV epidemic. Additionally, several of the South African taxa that clustered with these taxa in the large phylogenies were identified through retrospective investigation to also have originated from the Cape Metropolitan area. These

clusters along with the retrospectively identified Cape Town isolates were further investigated to establish their true nature.

3.5.1 Identification of monophyletic clusters

Due to small variations in the clustering patterns between the ME and ML tree topologies only the ML.aLRT tree topology were used for any of the subsequent analyses. However, the putative clusters that were observed in the ME-SPR tree topologies, as well as the isolates contained within them, were also present in the ML.aLRT tree topologies, with only small minor variations.

The manual assessment of the *gag* p24 and *pol* ML.aLRT tree topologies revealed five monophyletic clusters of Cape Town isolates in the *gag* p24 tree topology, while three putative clusters were observed in the *pol* phylogeny. These clusters were also present in the corresponding Minimum-Evolution tree topologies that were inferred with small variations in the clustering in each cluster due to the varying degree of tree inference algorithms used. The five clusters in the ML.aLRT *gag* p24 tree topology ranged from five taxa in size in the smallest cluster to 20 taxa in size in the biggest cluster. Similarly, the size of the three clusters in the ML.aLRT *pol* tree topology ranged between 22 taxa in the smallest cluster to 32 in the largest of the clusters. PhyloType analyses of the tree topologies also supported these findings (data not shown).

Five putative transmission clusters of Cape Town sequences were observed in the *gag* p24 ML.aLRT tree topology (Table 3.25 and Figure 3.17). The first cluster (*gag*.cluster.1) contained 25 Cape Town taxa (1991 - 2010), with an additional two South African taxa intermingled in the cluster. The second cluster (*gag*.cluster.2) contained seven Cape Town taxa (1989 - 1991) clustering with an isolate from Kenya, while the third cluster (*gag*.cluster.3) contained ten sequences from Cape Town (2002 - 2010) and two additional isolates from South Africa. The fourth cluster (*gag*.cluster.4) contained a total of five taxa from Cape Town (1989 - 1992), while the fifth cluster (*gag*.cluster.5) contained five taxa (all from the year 2002). The internal branch support for each of the five putative clusters in the ML.aLRT *gag* p24 tree topology was all above 0,700 with the highest 0,904 (*gag*.cluster.2) while the lowest support value was 0,739 (*gag*.cluster.1). The taxa in each of the clusters are tabulated in Table 3.25, while an illustration of the various *gag* p24 clusters can be seen in Figure 3.17.

Table 3.25: The five putative transmission clusters of Cape Town sequences that were found through manual assessment of the *gag* p24 ML.aLRT tree topology. Sequences and their respective years of sampling are listed in the right hand column, while sequences marked in bold were identified through the retrospective identification of additional Cape Town sequences.

Cluster	Taxa and Sampling Year
<i>gag</i> .cluster.1	AF543976_1999 , DQ866235_2002, DQ866208_2002, DQ866265_2002, DQ866283_2002, DQ866244_2002, DQ866243_2002, AZ111_2010, CM103_2009, JO228_2009, DQ866221_2002, DQ866222_2002, JO202_2009, NN087_2009, DQ866277_2002, TB089_2009, NN117_2010, JO191_2009, GU201717_2007 , DQ866246_2002, R9684_1991, JO166_2009, JO232_2009, JO244_2009
<i>gag</i> .cluster.2	R16812_1992, R11961_1991, R11397_1991, R11983_1991, R11391_1991, R4369_1989, R4368_1989
<i>gag</i> .cluster.3	DQ866296_2000, JO239_2009, JO210_2009, MN091_2009, DQ866300_2000, DQ866271_2000, DQ866262_2000, DQ866255_2000, NM114_2010, NM090_2009
<i>gag</i> .cluster.4	R4370_1989, R11988_1991, R16335_1992, R8864_1991, R16885_1992
<i>gag</i> .cluster.5	DQ866227_2002, DQ866206_2002, DQ866249_2002, DQ866253_2002, DQ866248_2002

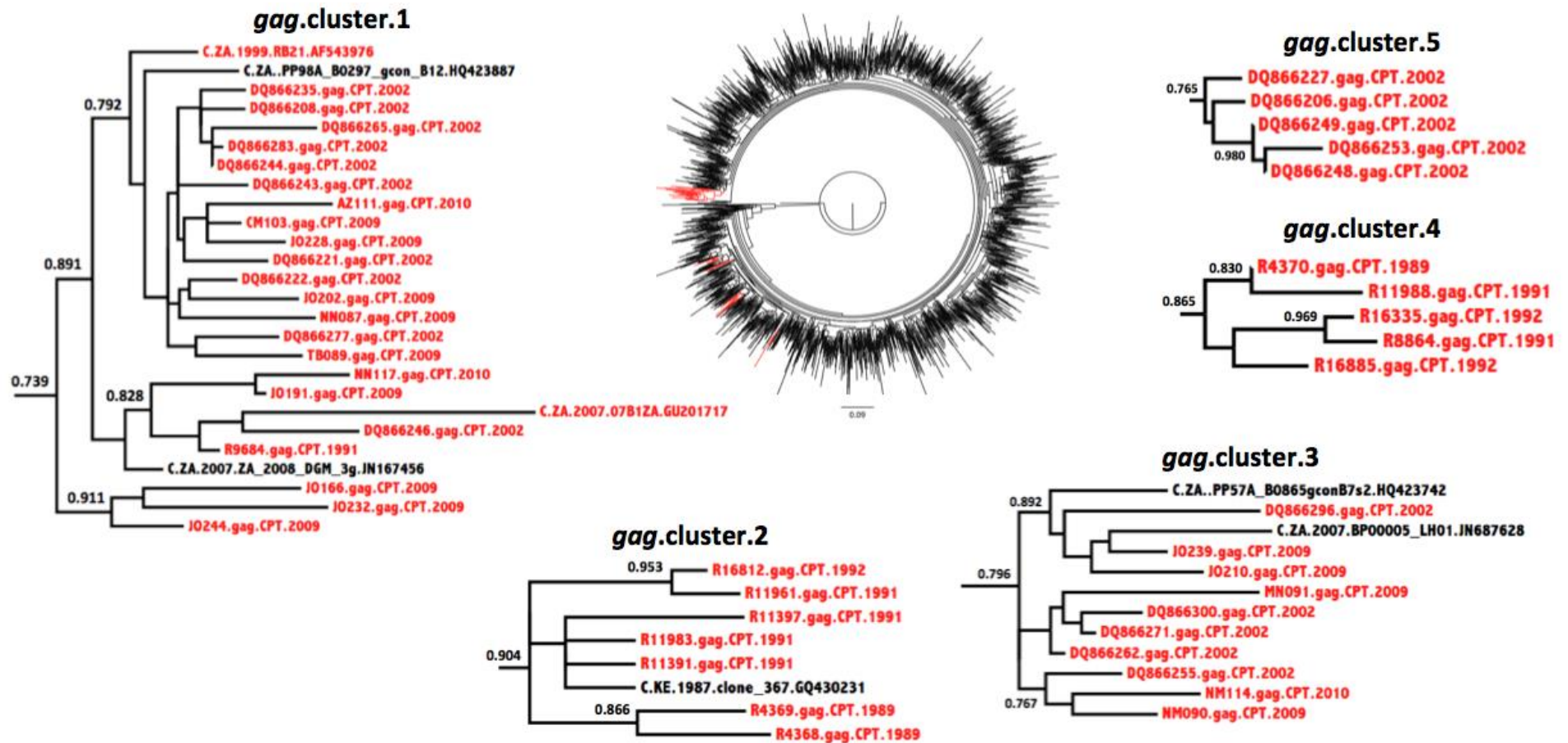


Figure 3.17: The five putative transmission clusters of Cape Town sequences that were observed through manual inspection of the *gag* p24 ML.aLRT tree topology. This ML.aLRT tree topology contains 1895 taxa and is rooted by the subtype B reference strain HXB2. It was constructed in phyML v 3.0, with the HKY85 model of nucleotide substitution and an estimated Gamma shape parameter (HKY+G). The scale bar at the bottom of the large circular tree topology correspond the branch lengths in that phylogeny. All of the isolates from Cape Town in the putative clusters are marked in red, while sequences from other locations are marked in black. The branch support for some of the internal branches of each putative cluster is also indicated.

Similarly, the analysis of the *pol* ML.aLRT tree topology revealed three clusters of Cape Town sequences (Table 3.26). The first cluster (*pol.cluster.1*) contained 22 isolates, mostly from the very early years of the HIV-1 subtype C epidemic (1989 - 2000), clustering with two isolates from Tanzania. The second cluster, *pol.cluster.2* contained 32 Cape Town isolates (1992 - 2009). This cluster contained 16 *pol* sequences from the original Cape Town data set, while another 16 sequences were retrospectively identified as being from Cape Town. The third cluster contained 33 isolates, originating from more recently sampled patients (2000 - 2009). This cluster contains 16 isolates from the original Cape Town *pol* data set, while another 17 isolates were retrospectively identified as originating from Cape Town. Once again the support values for the internal branches of the three putative clusters in the ML.aLRT *pol* tree topology were all large than 0,700 with the highest support 0,920 (*pol.cluster.3*) while the lowest support value was 0,740 (*pol.cluster.2*). The taxa in each of the *pol* clusters are tabulated in Table 3.26, while an illustration of the various clusters can be seen in Figure 3.18.

Table 3.26: The three putative transmission clusters of Cape Town sequences that were found through manual assessment of the *pol* ML.aLRT tree topology. Sequences and their respective years of sampling are listed in the right hand column, while sequences marked in bold were identified through the retrospective identification of additional Cape Town sequences.

Cluster	Taxa and Sampling Year
<i>pol.cluster.1</i>	R4370_1989, R16022_1992, R4714_1990, R16166_1992, R11961_1991, R4368_1989, R11983_1991, R4369_1989, R16885_1992, R11582_1991, R11391_1991, R8597_1991, R7663_1991, R6191_1990, R7148_1991, R4846_1990, R16335_1992, R7788_1991, R6201_1990, TV30_2000, R6742_1990, R16200_1992
<i>pol.cluster.2</i>	EF602189_2002 , R15791_1992, R6984_1991, R11988_1991, JN638177_2009 , JN638094_2009 , JN638160_2009 , EU854512_2006 , JO210_2009, JN638079_2009 , TV86_2000, TV28_2000, AX455917_1998 , TV1769_2004, EF602191_2002 , JN638170_2009 , JN638230_2009 , JN700932_2009 , JN638218_2009 , TV1799_2004, JN638115_2009 , JN638101_2009 , JN638110_2009 , MH020_2008, JN638095_2009 , JN638102_2009 , JN638213_2009 , JO239_2009, R17042_1992, TV22_2000, R6978_1991, JO166_2009, TV55_2000
<i>pol.cluster.3</i>	JO258_2009, HQ994437_2007 , TV1761_2004, EF602205_2002 , EF602211_2002 , TV1771_2004, 20_pol_2008, EF602186_2002 , JO247_2009, TV1753_2004, JO191_2009, JN638165_2009 , JO202_2009, TV51_2000, JN638119_2009 , TG005_2008, JN638192_2009 , EF602245_2007 , TV1762_2004, JO204_2009, JO244_2009, JO173_2009, HQ994450_2007 , HQ994455_2007 , HQ994569_2007 , HQ994591_2007 , JN638153_2009 , HQ994470_2007 , HQ994600_2007 , JN638120_2009 , HQ994418_2007 , 41_pol_2008, TV1774_2004

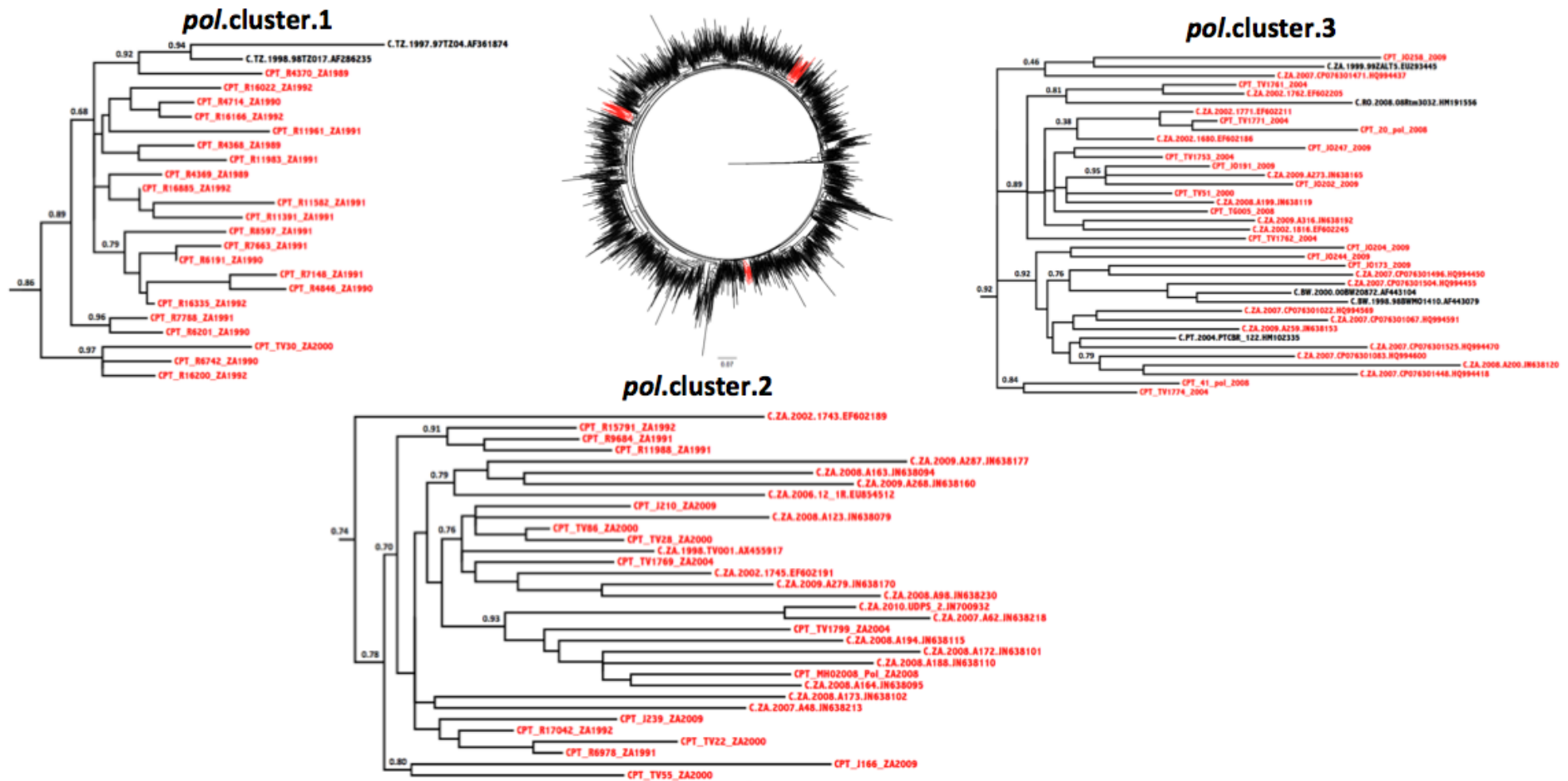


Figure 3.18: The three putative transmission clusters of Cape Town sequences that were observed through manual inspection of the *pol* ML.aLRT tree topology. This ML.aLRT tree topology contains 2333 taxa and is rooted by the subtype B reference strain HXB2. It was constructed in phyML v 3.0, with the HKY85 model of nucleotide substitution and an estimated Gamma shape parameter (HKY+G). The scale bar at the bottom of the large circular tree topology correspond the branch lengths in that phylogeny. All of the isolates from Cape Town in the putative clusters are marked in red, while sequences from other locations are marked in black. The branch support for some of the internal branches of each putative cluster is also indicated.

3.5.2 Testing putative transmission clusters

Following the identification of these putative transmission clusters, additional clustering analyses were performed in order to dispel any doubt over their validity. This included the construction of five phylogenies (NJ.bootstrap, ME.bootstrap, ML.aLRT, ML.bootstrap and Bayesian) for each of the observed putative clusters to assess support values for internal branches and consistency in clustering across methods. Due to the large size of these results (totalling 40 tree topologies) they have not been included in the main results section, but can be found in Appendix F (Figures 6.6 – 6.45). A summary of these tree topologies are provided in Table(s) 3.27 and 3.28, for the putative *gag* p24 and *pol* transmission clusters respectively.

Careful examination of the results of the clustering analyses for the five putative transmission clusters that was observed in the *gag* p24 tree topology provided some support for each cluster. For the first cluster, *gag*.cluster.1, only 20 out of the 25 Cape Town isolates consistently clustered with one another. For *gag*.cluster.2, *gag*.cluster.3 and *gag*.cluster.4 all of the Cape Town isolates clustered consistently across all five methods of tree inference with one another. Only four out of the five Cape Town isolates in the *gag*.cluster.5 data set clustered consistently with one another. The support values for these clusters were fairly low, in particularly the bootstrap support. However, the posterior support and aLRT support values were much higher. The support for these putative clusters were compared against support values for the Indian and Brazilian clades (Table 3.27), which were fairly similar to those found for the Cape Town clusters. Support values for putative clusters are normally set above 70,0% with 90% being optimal. However, branch support values are influence by the size of the alignment as well as the degree of genetic variation within the region under study. Therefore branch support values will perform poorly with such small variable alignments. The analysis of these putative clusters would have benefited from the analysis of complete genomes, however the generation of complete genomes would be difficult from such old samples. However, given the consistent clustering of the Cape Town isolates with one another across the five different methods of tree inference one can assume that they are truly epidemiologically linked. However, given the lack of confidence in the one cannot interpret as true transmission events of HIV-1. Furthermore, the p-values for each of the clusters that were inferred during the PhyloType analyses (Table 3.27), all indicate that the clustering of the Cape Town sequences were highly significant.

Table 3.27: Summary of the clustering analyses for the five putative *gag* p24 clusters of Cape Town sequences, as well as the results of the clustering analyses for the Indian and Brazilian clades. In the table the total number of Cape Town or Indian or Brazilian taxa for each cluster is indicated at the top of each column. The clustering refers to the number of isolates (either Cape Town, Brazilian and/or Indian) that clustered together out of the total number of isolates in each of the data sets. The Phylotype support corresponds to the support values for each of the clusters given as a fraction. The p values indicate the likelihood of those isolate clustering together given the data set as was determined by following 1000 shuffling iterations in Phylotype. These results correspond to the phylogenies presented in Figures 6.6 – 6.30.

Method of tree inference	Clustering / Support	<i>gag</i> .cluster.1 (25 taxa)	<i>gag</i> .cluster.2 (7 taxa)	<i>gag</i> .cluster.3 (10 taxa)	<i>gag</i> .cluster.4 (5 taxa)	<i>gag</i> .cluster.5 (5 taxa)	IN.clusters (16 taxa)	BR.clusters (10 taxa)
NJ.bootstrap	Clustering	22/25	7/7	10/10	5/5	5/5	5/16	5/10
	Support	0.000	91.000	0.000	0.000	93.000	20.000	28.000
	Phylotype	0.000	0.910	0.000	0.000	0.930	0.200	0.276
	p value	0.004	0.000	0.001	0.002	0.000	0.000	0.000
ME.bootstrap	Clustering	21/25	7/7	10/10	5/5	5/5	5/16	5/10
	Support	0.000	94.000	0.000	0.000	93.000	20.000	26.000
	Phylotype	0.000	0.940	0.000	0.000	0.930	0.200	0.256
	p value	0.000	0.000	0.001	0.002	0.000	0.000	0.000
ML.aLRT	Clustering	21/25	7/7	10/10	5/5	5/5	10/16	6/10
	Support	83.200	95.200	90.400	91.900	98.000	78.000	92.000
	Phylotype	0.832	0.952	0.904	0.919	0.981	0.780	0.919
	p value	0.001	0.001	0.000	0.000	0.001	0.001	0.001
ML.bootstrap	Clustering	21/25	7/7	10/10	5/5	5/5	10/16	6/10
	Support	8.000	69.000	1.000	23.000	93.000	0.000	18.000
	Phylotype	0.080	0.690	0.010	0.230	0.930	0.000	0.180
	p value	0.002	0.001	0.000	0.000	0.001	0.001	0.001
Bayesian	Clustering	21/25	7/7	10/10	5/5	4/5	10/16	9/10
	Support	62.570	94.310	21.750	13.090	76.000	91.000	93.000
	Phylotype	0.626	0.943	0.218	0.131	0.760	0.910	0.930
	p value	0.001	0.002	0.001	0.001	0.000	0.000	0.000

Careful examination of the results of the clustering analyses for the three putative transmission clusters that were observed in the *pol* ME-SPR tree topology provided some support for each cluster. For the first cluster, *pol*.cluster.1, 21 out of the 22 Cape Town isolates consistently clustered with one another. For *pol*.cluster.2 all 32 of the Cape Town isolates consistently clustered with one another while for *pol*.cluster.3 only 32 out of the 33 Cape Town isolates consistently clustered together across all five methods of tree inference.

The support values for these putative clusters were fairly low, in particularly the bootstrap support. However, the posterior support and aLRT support values were a bit higher. The support for these putative clusters were compared against support values for Indian and Brazilian clades (Table 3.28), which were fairly similar to those found for the Cape Town clusters. As mentioned earlier support values for putative clusters are normally set above 90,0% (95,0% - 98,0%). However, given the consistent clustering of the Cape Town isolates with one another across the five different methods of tree inference one can conclude that they represent monophyletic clusters of Cape Townian sequence as they held true during the additional clustering analyses that was performed. As with the analyses of the five putative *gag* p24 clusters, the p-values that were inferred during the PhyloType analyses (Table 3.28), suggest that the clustering of the Cape Town isolates were highly significant.

Table 3.28: Summary of the clustering analyses for the five putative *pol* clusters of Cape Town sequences, as well as the results of the clustering analyses for the Indian and Brazilian clades. In the table the total number of Cape Town or Indian or Brazilian taxa for each cluster is indicated at the top of each column. The clustering refers to the number of isolates (either Cape Town, Brazilian and/or Indian) that clustered together out of the total number of isolates in each of the data sets. The Phylotype support corresponds to the support values for each of the clusters given as a fraction. The p values indicate the likelihood of those isolate clustering together given the data set as was determined by following 1000 shuffling iterations in Phylotype. These results correspond to the phylogenies presented in Figures 6.31 – 6.45.

Method of tree inference	Clustering / Support	<i>pol</i> .cluster.1 (22 taxa)	<i>pol</i> .cluster.2 (32 taxa)	<i>pol</i> .cluster.3 (33 taxa)	IN.clusters (16 taxa)	BR.clusters (10 taxa)
NJ.bootstrap	Clustering	21/22	32/32	32/33	15/16	10/10
	Support	0.000	19.000	0.100	2.000	28.000
	Phylotype	0.000	0.189	0.001	0.020	0.281
	p value	0.002	0.003	0.000	0.001	0.001
ME.bootstrap	Clustering	21/22	32/32	32/33	15/16	10/10
	Support	0.000	18.900	0.100	0.94	2.000
	Phylotype	0.000	0.189	0.001	0.009	0.020
	p value	0.002	0.001	0.000	0.001	0.001
ML.aLRT	Clustering	22/22	32/32	32/33	14/16	10/10
	Support	80.20	34.800	43.900	94.000	40.000
	Phylotype	0.802	0.348	0.439	0.940	0.401
	p value	0.004	0.002	0.001	0.000	0.000
ML.bootstrap	Clustering	22/22	32/32	32/33	14/16	10/10
	Support	1.000	1.000	0.000	64.000	59.000
	Phylotype	0.010	0.010	0.000	0.639	0.591
	p value	0.004	0.002	0.001	0.000	0.000
Bayesian	Clustering	22/22	32/32	30/33	14/16	10/10
	Support	80.100	48.89	43.890	42.000	38.000
	Phylotype	0.801	0.489	0.439	0.420	0.380
	p value	0.003	0.000	0.004	0.001	0.001

3.6.3 Timing the root height of the internal nodes of each cluster

Bayesian inference through the use of various model parameters was employed to time the estimated date of origin (age of the internal node) of the various clusters. Manual inspections of the log trace files for each of the various runs showed good convergence in the trace files, following a Bayesian MCMC totalling 30 million steps. Extremely high ESS's (between 200 - 9000) were also recorded for each of the runs, which is a good "diagnostic indicator" of the quality of the various runs. For some of the transmission clusters selected non-parametric runs could not be inferred in BEAST due to the extremely small size of the data sets (< 10 taxa), in which case the root-height estimation was restricted to the use of only a parametric tree prior.

The inferred tMRCA of *gag*.cluster.1 (Table 3.29) was estimated around 1986,4 (with the 95% HPD ranging between 1981,1 and 1990,8) according to the "best-fitting model" as was determined through Bayes factor comparison. Similarly, the estimated inferred tMRCA of *gag*.cluster.2 and *gag*.cluster.3 were placed around 1986,7 (95% HPD 1983,3 – 1989,3) and 1993,6 (95% HPD 1989,4 – 1997,4) respectively. Similarly, the root height of *gag*.cluster.4 was placed around 1980,8 (95% HPD 1972,9 – 1988,1), while the tMRCA of *gag*.cluster.5 was estimated around 1995,1 (95% HPD 1991,2 – 1998,3).

Table 3.29: The estimated tMRCAs of the various *gag* p24 transmission clusters of Cape Town sequences. The “best fitting” run as was identified through Bayesian model comparison has been highlighted in bold and the average estimated date of origin is indicated at the bottom of each column. Due to the small number of taxa contained within some clusters only parametric tree priors could be used to infer the dates of origin in some clusters.

Model parameters	<i>gag</i> .cluster.1			<i>gag</i> .cluster.2			<i>gag</i> .cluster.3			<i>gag</i> .cluster.4			<i>gag</i> .cluster.5		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
Relax.bsp.est.1	1985,8	1991,5	1976,1	-	-	-	-	-	-	-	-	-	-	-	-
Relax.bsp.est.2	1985,6	1991,5	1975,7	-	-	-	-	-	-	-	-	-	-	-	-
Relax.bsp.fix.1	1985,6	1991,5	1975,8	-	-	-	-	-	-	-	-	-	-	-	-
Relax.bsp.fix.2	1985,9	1991,5	1976,4	-	-	-	-	-	-	-	-	-	-	-	-
Relax.const.est.1	1985,9	1991,5	1976,5	1986,0	1989,2	1981,9	1984,8	1998,3	1963,6	1981,9	1988,2	1972,4	1993,0	1998,6	1984,9
Relax.const.est.2	1968,7	1990,3	1933,8	1985,9	1989,3	1981,9	1984,6	1998,2	1962,6	1981,8	1988,4	1972,1	1992,5	1998,6	1984,4
Relax.const.fix.1	1969,3	1989,5	1933,4	1985,8	1989,2	1981,7	1984,5	1998,2	1962,6	1980,2	1988,2	1972,6	1992,9	1998,7	1984,9
Relax.const.fix.2	1969,7	1990,1	1934,9	1986,1	1989,2	1982,1	1984,9	1998,2	1963,6	1980,1	1988,0	1972,5	1992,9	1998,6	1985,5
Relax.expo.est.1	1969,8	1990,1	1937,3	1986,0	1989,1	1981,9	1985,1	1997,8	1964,7	1979,2	1988,0	1972,5	1994,7	1998,7	1990,0
Relax.expo.est.2	1982,5	1990,7	1971,9	1986,5	1989,4	1983,1	1990,7	1997,8	1981,7	1979,8	1988,1	1972,5	1995,0	1998,5	1990,5
Relax.expo.fix.1	1986,4	1990,8	1981,1	1986,5	1989,4	1983,0	1994,5	1997,7	1990,9	1980,8	1987,9	1972,6	1995,1	1998,3	1991,2
Relax.expo.fix.2	1982,8	1991,1	1972,7	1986,6	1989,3	1983,2	1990,8	1997,8	1981,7	1980,8	1988,1	1972,9	1994,5	1998,6	1989,1
Strict.bsp.est.1	1982,9	1991,1	1972,7	-	-	-	-	-	-	-	-	-	-	-	-
Strict.bsp.est.2	1985,6	1991,5	1978,7	-	-	-	-	-	-	-	-	-	-	-	-
Strict.bsp.fix.1	1985,7	1991,5	1978,9	-	-	-	-	-	-	-	-	-	-	-	-
Strict.bsp.fix.2	1986,8	1991,5	1981,1	-	-	-	-	-	-	-	-	-	-	-	-
Strict.const.est.1	1987,0	1991,5	1981,1	1986,7	1989,3	1983,3	1990,8	1997,6	1981,3	1985,6	1988,3	1982,6	1995,2	1998,4	1991,4
Strict.const.est.2	1983,3	1990,0	1975,4	1987,1	1989,4	1984,7	1993,1	1997,0	1988,4	1985,7	1988,2	1983,1	1995,2	1998,5	1991,5
Strict.const.fix.1	1983,4	1989,5	1975,5	1987,1	1989,5	1984,5	1993,1	1997,3	1988,5	1985,6	1988,1	1982,7	1995,3	1998,3	1991,7
Strict.const.fix.2	1984,4	1990,0	1977,4	1986,7	1989,1	1983,7	1993,5	1997,3	1989,2	1985,6	1988,1	1982,8	1995,2	1998,5	1992,0
Strict.expo.est.1	1984,4	1990,1	1977,9	1987,2	1989,4	1984,7	1993,6	1997,4	1989,4	1985,7	1988,2	1982,6	1996,0	1998,6	1992,9
Strict.expo.est.2	1985,3	1990,6	1979,5	1986,8	1989,2	1984,1	1994,0	1997,7	1990,1	1985,7	1988,2	1982,6	1995,8	1998,5	1992,7
Strict.expo.fix.1	1985,2	1990,3	1979,2	1986,8	1989,1	1984,1	1994,0	1997,6	1990,1	1985,7	1988,2	1983,0	1995,9	1998,5	1993,0
Strict.expo.fix.2	1986,4	1990,8	1981,1	1987,1	1989,4	1984,5	1994,4	1997,8	1990,9	1985,6	1988,3	1982,6	1995,9	1998,6	1992,9
Average	1982,4	1990,8	1970,2	1986,5	1989,3	1983,3	1990,4	1997,7	1980,0	1983,1	1988,1	1977,6	1994,7	1998,5	1989,9

HPD – Highest Posterior Density, BSP – Bayesian Skyline Plot tree prior, relax – relaxed molecular clock assumption, strict – strict molecular clock assumption, Const – Constant population size tree prior, expo – exponential growth tree prior, fix – fixed mutation rate, est – estimated mutation rate

Similarly, the inferred tMRCA of the “best fitting model” as was determined in a Bayesian model test comparison, for the *pol.cluster.1* data set was 1980,8 with the 95% HPD ranging between 1977,4 – 1984,0. Similarly, the inferred tMRCA of the *pol.cluster.2* data set was estimated at 1985,9 with the 95% HPD ranging between 1983,5 and 1987,9 (1976,8 – 1984,7). Lastly, the inferred tMRCA’s of the *pol.cluster.3* data set was estimated at around 1992,2 with the 95% HPD ranging between 1990,1 and 1994,2. The full results of the estimated tMRCA for each of the three different *pol* transmission clusters are tabulated in Table 3.30.

Table 3.30: The estimated tMRCA's of the various *pol* transmission clusters of Cape Town sequences. The “best fitting” run as was identified through Bayesian model comparison has been highlighted in bold for each data set and the average estimated date of origin is indicated at the bottom of each column.

Model parameters	<i>pol.cluster.1</i>			<i>pol.cluster.2</i>			<i>pol.cluster.3</i>		
	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper	Mean	95% HPD lower	95% HPD upper
Relax.bsp.est.1	1979,7	1985,1	1973,1	1985,7	1987,5	1983,8	1992,3	1994,4	1990,1
Relax.bsp.est.2	1979,8	1985,4	1973,4	1985,0	1987,3	1982,5	1990,2	1996,1	1983,2
Relax.bsp.fix.1	1980,1	1985,0	1974,4	1985,9	1988,0	1983,6	1990,8	1996,0	1984,8
Relax.bsp.fix.2	1980,1	1984,9	1974,2	1985,9	1987,9	1983,5	1990,8	1995,7	1984,9
Relax.const.est.1	1965,9	1983,3	1937,5	1982,8	1986,8	1978,0	1966,5	1993,1	1917,2
Relax.const.est.2	1966,1	1983,4	1937,3	1982,8	1986,7	1978,2	1966,9	1993,6	1921,6
Relax.const.fix.1	1966,9	1982,9	1940,3	1983,5	1987,1	1978,7	1965,8	1992,5	1918,4
Relax.const.fix.2	1966,3	1982,9	1938,1	1983,3	1987,2	1977,9	1966,3	1992,3	1919,5
Relax.expo.est.1	1976,5	1982,8	1969,7	1984,4	1986,8	1981,8	1987,7	1993,9	1981,2
Relax.expo.est.2	1976,3	1982,6	1969,0	1984,4	1986,7	1981,7	1987,9	1993,7	1981,6
Relax.expo.fix.1	1977,3	1982,5	1970,5	1985,6	1987,6	1983,3	1989,2	1994,1	1983,6
Relax.expo.fix.2	1977,4	1982,8	1971,2	1985,6	1987,6	1983,3	1989,2	1994,1	1983,7
Strict.bsp.est.1	1982,0	1984,6	1979,2	1985,0	1987,1	1982,6	1991,2	1994,3	1988,2
Strict.bsp.est.2	1982,0	1984,6	1979,1	1985,0	1987,1	1982,8	1991,2	1994,3	1988,1
Strict.bsp.fix.1	1982,6	1984,6	1980,5	1986,0	1987,8	1984,1	1993,1	1995,0	1990,9
Strict.bsp.fix.2	1982,6	1984,4	1980,5	1986,0	1987,7	1984,0	1993,1	1995,1	1990,9
Strict.const.est.1	1980,8	1984,0	1977,4	1983,9	1986,4	1981,5	1989,2	1993,3	1985,6
Strict.const.est.2	1980,8	1984,0	1977,3	1983,9	1986,3	1981,3	1989,2	1993,0	1985,3
Strict.const.fix.1	1981,3	1983,7	1978,5	1985,2	1987,1	1983,2	1991,4	1993,8	1988,7
Strict.const.fix.2	1981,3	1983,8	1978,7	1985,2	1987,0	1983,1	1991,4	1993,8	1988,6
Strict.expo.est.1	1980,3	1983,3	1977,3	1984,4	1986,7	1982,1	1990,1	1993,1	1987,1
Strict.expo.est.2	1980,3	1983,3	1977,3	1984,4	1986,5	1982,0	1990,1	1993,2	1987,2
Strict.expo.fix.1	1981,5	1983,6	1979,2	1985,7	1987,5	1983,8	1992,2	1994,2	1990,1
Strict.expo.fix.2	1981,5	1983,5	1979,1	1985,0	1987,3	1982,5	1990,2	1996,4	1983,2
Average	1977,9	1983,8	1969,7	1984,8	1987,2	1982,1	1986,5	1994,1	1975,2

HPD – Highest Posterior Density, BSP – Bayesian Skyline Plot tree prior, relax – relaxed molecular clock assumption, strict – strict molecular clock assumption, Const – Constant population size tree prior, expo – exponential growth tree prior, fix – fixed mutation rate, est – estimated mutation rate

3.6.4 Time resolved tree topologies with transmission clusters

Time resolved tree topologies were inferred from the entire Cape Town *gag* p24 and *pol* data sets. These were inferred with the use of a BSP tree prior and a relaxed molecular clock assumption. Time resolved tree topologies were inferred in TreeAnnotator and imported into FigTree for visual interpretation and manual adjustment. The time resolved tree topology of the *gag* p24 and *pol*, along with the different clusters is presented below in Figures 3.19 and 3.20, respectively.

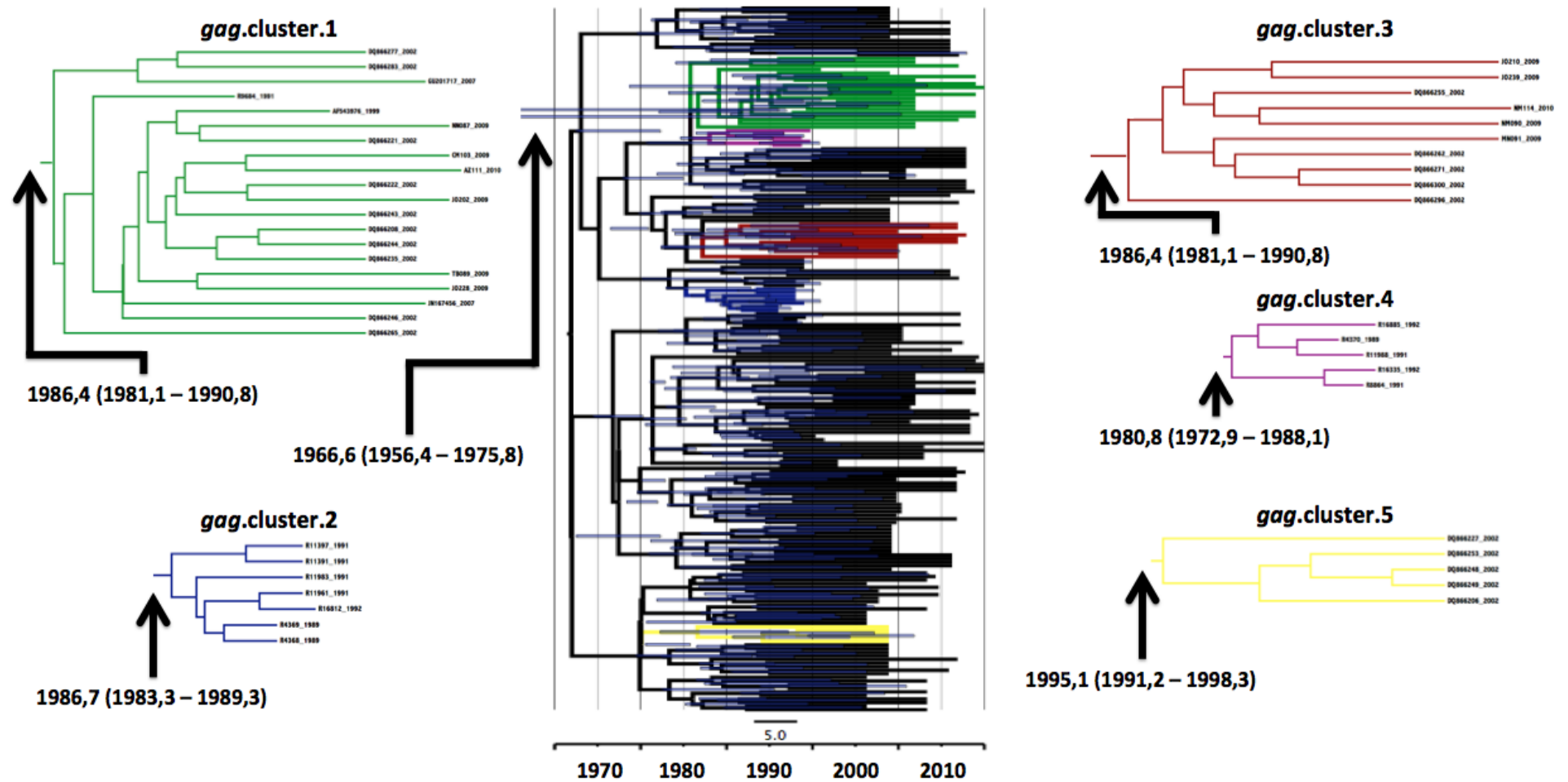


Figure 3.19: A time resolved tree topology of the *gag* p24 Cape Town data set with the 5 different clusters. The tree was reconstructed in TreeAnnotator from a non-parametric BSP tree prior that was executed in BEAST. The five different transmission clusters has been highlighted in red. The estimated root height as well as the 95% HPD intervals of each cluster as well as the full tree is indicated. The blue bars represent the 95% HPD intervals of each of the internal nodes in the tree topology.

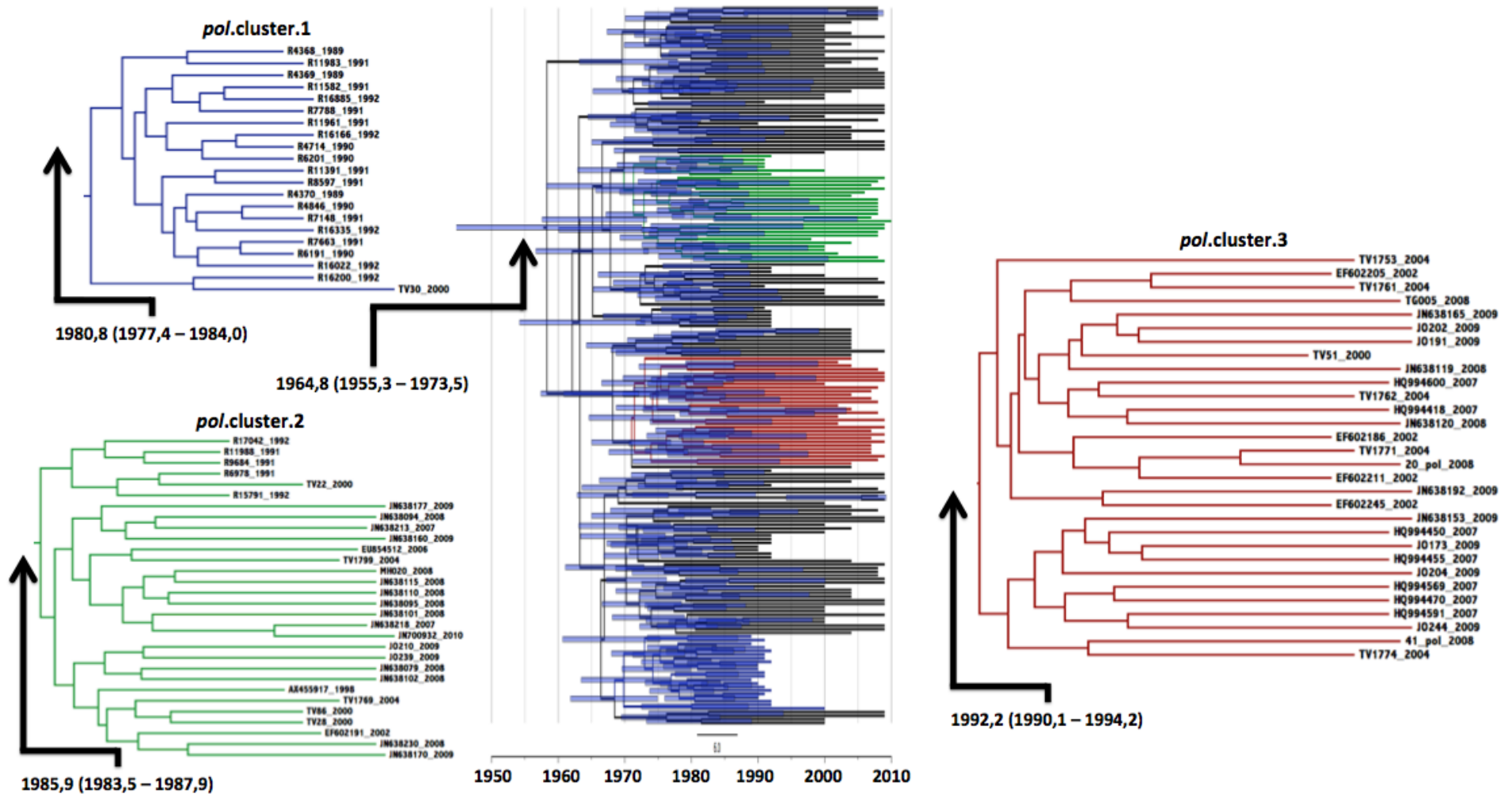


Figure 3.20: Time resolved tree topology of the *pol* Cape Town data sets with the 3 different clusters. The tree was reconstructed in TreeAnnotator from a non-parametric BSP tree prior that was executed in BEAST. The three different transmission clusters has been highlighted in red. The estimated root height as well as the 95% HPD intervals of each cluster as well as the full tree is indicated. The blue bars represent the 95% HPD intervals of each of the internal nodes in the tree topology

CHAPTER FOUR - TABLE OF CONTENTS

	Page
DISCUSSION	163
4.1 Sample selection and genomic characterization	163
4.2 Subtyping of samples	165
4.3 Preliminary analyses of the molecular clock signal of the Cape Town data sets	166
4.4 Evolutionary inference through Bayesian means	167
4.4.1 Inference of the tMRCA of the Cape Town and Southern African HIV-1 subtype C epidemics.	167
4.4.1.1 Estimated date of origin of Cape Town epidemic	168
4.4.1.2 tMRCA of the Southern African epidemic (excluding Cape Town)	169
4.4.1.3 tMRCA of the entire Southern African epidemic (including Cape Town)	171
4.4.2 Epidemic reconstruction of the Cape Town and Southern African HIV-1 subtype C epidemics.	172
4.4.2.1 Phylodynamic aspects of the Cape Town epidemic	172
4.4.2.2 Reconstruction of the Southern African epidemic (excluding Cape Town)	173
4.4.2.3 Reconstruction of the Southern African epidemic (including Cape Town)	175
4.5 Factors that may influence HIV-1 epidemic reconstruction	175
4.6 Basic phylogenetic investigation to establish the evolutionary relationship of the HIV-1 subtype C isolates from Cape Town	178
4.6.1 Data mining	178

4.6.2 Large-scale phylogenetic inference	179
4.7 Monophyletic clades of HIV circulating in the Cape Metropolitan area	181
4.8 Global HIV-1 subtype C perspective	184
4.9 Southern African subtype C epidemic	186
4.9.1 The role of migration on the epidemic	186
4.9.2 The two stage HIV-1 epidemic in Southern Africa	188
CONCLUSION	190

CHAPTER FOUR

DISCUSSION

4.1 Sample selection and genomic characterization

One of the primary objectives of this study was to generate a comprehensive longitudinal sequence data set of Cape Town isolates that span over several years or decades. In order to achieve this objective a large number of *gag* p24 and *pol* sequence fragments that were previously generated within the Division of Medical Virology (Faculty of Medicine and Health Sciences of Stellenbosch University) were selected for inclusion into the study [Jacobs *et al.*, 2006; Wilkinson *et al.*, 2013 in press; Isaacs *et al.*, in preparation]. In addition to these genotypes, 250 patient samples were selected (particularly from the earlier years of the epidemic) in order to obtain a comprehensive longitudinally sampled sequence data sets.

Patient samples were selected based on stringent inclusion and exclusion criteria (Table 2.4), from the HIV-1 sample database at the Virology laboratory of the Tygerberg Academic Hospital. This sample database contains more than 35,000 patient samples, which were sent in over the past 30-years for routine HIV diagnostic testing. For a large number of these patients, particularly from the very early years of the epidemic (1986 - 1993), nucleic acid was isolated from patient samples at previous occasions. Therefore, patient samples for which previously isolated nucleic acid was available were preferred above other patient samples for inclusion in order to reduce cost and to increase amplification success. A total of 250 patient samples were selected for genotyping. These 250 patient samples included 63 previously isolated DNA samples (originating between 1989 and 1992), 34 patient samples from 1996 from which DNA were extracted, and 153 previously isolated RNA samples which originated between 2000 and 2004.

These patient samples were used for the generation of *gag* p24 and *pol* sequence fragments through standard molecular techniques (PCR and Sanger sequencing). As mentioned in section 3.2 in the previous chapter the amplification of target gene fragments were very troublesome. The poor amplification could be described to several factors such as problems with primer annealing, primer extension, nucleic acid degradation, and suboptimal buffer and/or salt concentrations. Therefore a considerable amount of time was taken to try and improve the amplification of target DNA. This included the changing of primer annealing temperatures and magnesium titrations. Success for PCR assays was also measured with the use of a positive control. The amplification

of the positive control worked consistently across the various methods that were tried. However, only 49 of the 250 patient samples produced a strong positive PCR product in the *gag* p24 amplification assay. Similarly, only 110 positive PCR products could be obtained from the *pol* PCR assay. The primer sets that were used during the course of this study for the amplification of target DNA were designed to work against a large range of different subtypes of HIV-1 including HIV-1 subtypes B and C [Swanson *et al.*, 2003; Plantier *et al.*, 2005]. Given that the positive control worked consistently across the various methods that were tried and the fact that the changing of the PCR conditions did not perform any better than the standard amplification conditions that are listed in Table(s) 2.5 – 2.7 one can conclude that the problem with the amplification does not lie with the PCR conditions but with the samples itself. These samples and/or nucleic acid have been stored over several years or decades at -20°C. Non optimal storage and several power failures coupled with the breaking down of refrigeration equipment may therefore have led to significant degradation in the nucleic acid.

The sequencing of these PCR products produced 25 new *gag* p24 sequences (1989 - 1992) and 74 new partial *pol* sequences fragments (1989 - 2004) while 24 *gag* p24 and 36 *pol* PCR products could not be sequenced due to sequencing problems. These problems could be due to primer annealing or primer extension problems in the sequencing assays. In addition to the problems with primer annealing and primer extension, another possible problem may be due to sample degradation due to prolonged storage at 4°C or -80°C for DNA and RNA specimens, as well as the effect that repeated freezing and thawing cycles may have on the integrity of the specimens. These two factors in HIV sample degradation have been well characterized previously [Gessoni *et al.*, 2004; Baleriola *et al.*, 2010] and can seriously affect the outcome of PCR amplification assays.

However the primary aim of this study was to obtain a longitudinal sample of *gag* p24 and *pol* sequences from Cape Town spanning over several years or decades. Therefore, the number of characterized specimens combined with the sequence data that were previously characterized was sufficient enough for this study. In total our *gag* p24 data set contained 193 patient sequences spanning from 1989 till 2010. Similarly, the final *pol* data set of Cape Town patients contained 166 sequences sampled between 1989 and 2010. The concatenated *gag-pol* data set of Cape Town patients, which were represented in both the *gag* p24 and the *pol* data sets, contained 52 sequences (1989 - 2010).

In previous studies, which reconstructed HIV-1 subtype C epidemic in other areas of the world, the total number of taxa that were used differs significantly. In the estimation of the date of origin of the HIV-1 subtype C epidemic from Malawian sequence data two different genomic regions (*gag* p17-p24 and *env* C2-V3) were used for their analyses. These data sets differed in size from as little as 59 taxa while the largest data sets contained up to 376 taxa [Travers *et al.*, 2004]. In the reconstruction of HIV-1 epidemics in Zimbabwe [Dalai *et al.*, 2009] and Ethiopia [Tully and Wood, 2010] a total of 177 (1991 - 2006), and 165 (1988 - 1998) taxa were used respectively. In another study a total of 82 (V3), 40 (partial *env* C2 – C5), and 72 partial *pol* fragments, spanning over a 15-year period (1991 - 2006), were used for the epidemic reconstruction of the Brazilian HIV-1 subtype C epidemic [Bello *et al.*, 2008]. More recently, the demographic reconstruction of the HIV-1 subtype C epidemics in Senegal and Angola used even smaller sequence data sets. The Angolan epidemic [Afonso *et al.*, 2012] was inferred from only 31 partial *pol* sequences (spanning over a 9-year period), while the epidemic in Senegal [Jung *et al.*, 2012] was reconstructed from 56 partial *pol* sequences (1990 - 2008). Given these numbers the 193 *gag* p24, 166 partial *pol*, and 52 concatenated *gag-pol* sequences from Cape Town, are large enough for epidemic reconstruction.

Given the large variation in the number of taxa that is used for the inference of demographic histories it becomes clear that several factors may influence the size of the data set such as, availability of patient samples and the genomic region of interest. The most important factor however is the molecular clocklikeness of any data set and therefore the size of the final data set will ultimately be determined by the degree of clocklikeness that it carries [Brown, 2002].

4.2 Subtyping of samples

The subtyping of newly acquired HIV sequence information is one of the most basic principles of modern HIV-1 phylogenetic analysis. The subtyping of new HIV sequence information is essential in the study design of any further phylogenetic analysis that any researcher may want to conduct. For instance, in phylogenetic methods that employ the coalescent theory, HIV isolates from different subtypes can have a major effect on the outcome of the results.

The city of Cape Town is generally regarded by many as one of the most cosmopolitan on the African continent and is home to thousands of domestic and international migrants, as well as a substantial population of MSM. This large demographic variation has resulted in the regular identification and characterization of non-subtype C isolates in and around Cape Town [Engelbrecht *et al.*, 1999; Engelbrecht *et al.*, 2001; Jacobs *et al.*, 2006; Jacobs *et al.*, 2008; Jacobs

et al., 2009; Wilkinson and Engelbrecht, 2009; Jacobs *et al.*, 2013 in press; Wilkinson *et al.*, 2013 in press]. However, even though these studies have identified a large number of non-subtype C isolates, HIV-1 subtype C still accounts for the overwhelming number of infections (> 90%) in the Cape Metropolitan area. However, the circulation of such strains of HIV within the area under study does warrant the need to screen all new isolates before their inclusion in the analysis of this study, particularly for those phylogenetic analyses that relies heavily on the assumption of a molecular clock and the coalescence theory.

Several methods of subtyping HIV isolates have been developed. The most prominent and widely used of which are; the REGA viral subtyping tool (<http://www.bioafrica.net/rega-genotype/html/subtyping.html>) and the jumping profile Hidden Markov Model methods (<http://jphmm.gobics.de>). Since the main objective of this study was to investigate the evolutionary history of HIV-1 subtype C isolates, all newly characterized isolates were carefully screened with both these two methods as well as through manual phylogenetic methods, before they could be included in any further analysis. All of the newly sequenced isolates, as well as those sequences that were selected from previous studies for inclusion into this project, were characterized with high confidence as HIV-1 subtype C isolates. Therefore, intrasubtype diversity did not have any effect on the outcomes of the epidemic reconstruction that was performed in this study.

4.3 Preliminary analyses of the molecular clock signal of the Cape Town data sets

The molecular clock analysis that was performed in this study, prior to the start of the Bayesian epidemic inference strongly supports the assumption of a molecular clock for the *gag* p24 and *pol* Cape Town data sets. The crude root-to-tip regression roughly placed the tMRCA of the *gag* p24 Cape Town data set at 1953,50 with an estimated R^2 value of 0,213. The mutation rate (slope rate) for the *gag* p24 Cape Town data set was roughly estimated to be around $2,79 \times 10^{-3}$ mutations/site/year.

Similarly, the estimated tMRCA of the *pol* data set was placed at 1952,77 with an estimated R^2 value of 0,290. The estimated mutation rate for the *pol* Cape Town data set was calculated to be around $2,23 \times 10^{-3}$ mutations/site/year. The estimated tMRCA fits well into the estimated date of origin of the global HIV-1 subtype C lineage that was estimated by Novitsky and co-workers [Novitsky *et al.*, 2010], but is slightly older than the estimated date of origin that was inferred by Travers and co-workers [Travers *et al.*, 2004]. Similarly, the mutation rates for both the *gag* p24 and *pol* data sets are in line with findings from other studies found in the scientific literature,

which were reconstructed under a comprehensive number of models in BEAST [Hue *et al.*, 2004; Novitsky *et al.*, 2010].

It is generally regarded that a coefficient of variation (R^2) for an HIV data set greater than 0,2 does support the assumption of a molecular clock. For instance, in the Ethiopian study, which characterized the demographic history of the HIV-1 subtype C epidemic in the East African nation, the coefficient of variation was 0.20 [Tully and Wood, 2010]. Similarly, the Zimbabwean study by Dalai and co-workers [Dalai *et al.*, 2009] the coefficients of variation was 0.43, 0.24 and 0.25 for the constant, exponential and Bayesian Skyline Plot tree priors respectively. In the Brazilian study by Bello and co-workers [Bello *et al.*, 2008] the median coefficient of variation was 0.24 (HPD 0.13 – 0.34) for the *pol* data set and 0.32 (HPD 0.20 – 0.44) for the *env* data set. Finally, the median inferred coefficient of variation in the Angolan study by Afonso and co-workers [Afonso *et al.*, 2012] was 0.32 (HPD 0.28 – 0.36), while the Indian study by Shen and co-worker [Shen *et al.*, 2011] the median R^2 was 0.384 (HPD 0.373 – 0.596). Given these figures, ranging between 0.20 and 0.40, it would be reasonable to assume that the *gag* p24 and *pol* Cape Town data sets that were generated during the course of this study do hold enough genetic information to infer phylogenies with the implementation of a molecular clock. Moreover, these estimations were consistent within and amongst runs and different models for both data sets.

4.4 Evolutionary inference through Bayesian means

Bayesian means were used to reconstruct aspects of the HIV-1 subtype C epidemic(s) in Cape Town/South African, as well as the greater Southern African regions. This included the estimation of the date of origin and the reconstruction of demographic histories of these epidemics.

4.4.1 Inference of the tMRCA of the Cape Town and Southern African HIV-1 subtype C epidemics.

The estimated date of origin of the Cape Town/South African HIV-1 subtype C epidemic was reconstructed from three different longitudinal data sets. Similarly the estimated date of origin of the HIV-1 subtype C epidemic in the greater Southern African region was also inferred from three data sets. This was done both with and without sequence information from the original Cape Town data sets

4.4.1.1 Estimated date of origin of Cape Town epidemic

The estimated date of origin from the three different Cape Town data sets, as was inferred in BEAST under various different model parameters, produced very similar results (Tables 3.1 – 3.3).

The inference of the tMRCA from the *gag* p24 sequence data from Cape Town (Table 3.1) suggests a date of origin around the mid 1960's. The estimated tMRCA for the “best fitting model” as was compared under the Bayes factor comparison placed the date of origin around 1966,6, with the 95% HPD confidence intervals stretching from 1956,4 to 1975,3. The mean estimated mutation rate for this model tree prior was $2,9 \times 10^{-3}$ mutations/site/year with the 95% confidence intervals stretching between $2,2 \times 10^{-3}$ and $3,7 \times 10^{-3}$ mutations/site/year. This mutation rate compares well against the estimated mutation rate that was obtained from the epidemic reconstruction of the global HIV-1 subtype C epidemic from *gag* sequence data by Novitsky and co-workers [Novitsky *et al.*, 2010]. Additionally, four of the run parameters estimated a tMRCA around the early 1950's (1952 – 1954). These runs were inferred under slower estimated mutation rates, which can account for their older tMRCA's when compared to the runs (either fixed or estimated) that were inferred under a more realistic mutation rate as was identified through the Bayes factor comparison.

Similarly, the estimation of the tMRCA from the concatenated *gag-pol* Cape Town sequence data set (Table 3.2) also suggests a date of origin around the mid to late 1960's. The estimated tMRCA for the “best fitting model” as was compared under the Bayes factor comparison placed the date of origin around 1969,9 with the 95% HPD confidence intervals stretching from 1961,0 to 1977,9. The mean estimated mutation rate for this model tree prior was $2,5 \times 10^{-3}$ mutations/site/year with the 95% confidence intervals stretching between $1,6 \times 10^{-3}$ and $3,3 \times 10^{-3}$ mutations/site/year. This mutation rate compares very well with the rate that was used for the inference of the models under the assumption of a fixed mutation rate. The fixed mutation rate that was used for the concatenated *gag-pol* region was based on the findings from the study of Hue and co-workers [Hue *et al.*, 2004]. However, once again the mutation rate of some of the other runs played a major role in the estimation of their tMRCA's. Four of the model parameters that were ran under an estimated mutation rate produced extremely slow mutation rates ranging between $9,9 \times 10^{-3}$ and $1,00 \times 10^{-3}$ mutations/site/year. These slow mutation rates pushed back the estimated date of origin of these runs to the early 1940's.

Lastly, the estimation of the tMRCA from the Cape Town *pol* sequence data set (Table 3.3) also suggests a date of origin around the mid 1960's for the subtype C epidemic in Cape Town. The estimated tMRCA for the "best fitting model" as was compared under the Bayes factor comparison placed the date of origin around 1964,8 with the 95% HPD confidence intervals stretching from 1955,3 to 1973,5. The mean estimated mutation rate for this model tree prior was $2,1 \times 10^{-3}$ mutations/site/year with the 95% confidence intervals stretching between $1,8 \times 10^{-3}$ and $2,5 \times 10^{-3}$ mutations/site/year. This rate is slightly lower than the fixed mutation rate that was used, which were based on the findings of Hue and co-workers [Hue *et al.*, 2004].

It is therefore clear, based on the data that were inferred during the course of this study, that the estimated date of origin of the HIV-1 subtype C epidemic in Cape Town can be traced back to some point in the mid 1960's. The estimated date of origin for the Cape Town epidemic (mid 1960's) fits in well within any of the estimated tMRCA of HIV-1 group M [Korber *et al.*, 2000; Salemi *et al.*, 2001; Wertheim and Worobey, 2009]. Similarly, the estimated tMRCA of the Cape Town epidemic also fits in well when compared to the estimated tMRCA of the global HIV-1 subtype C lineage of 1950 (under a constant population size tree prior) and 1948 (under a BSP tree prior), that was inferred by Novitsky and co-workers from *gag* sequence data [Novitsky *et al.*, 2010]. Similarly, the estimated date of origin currently pre-seeds any of the previously inferred dates of origin, with the exception of Malawi (mid to late 1960's) [Travers *et al.*, 2004], for some of the other countries for whom the demographic histories of their HIV-1 subtype C epidemic have been characterized: Ethiopia (early 1980's) [Abebe *et al.*, 2001], Zimbabwe (early 1970's) [Dalai *et al.*, 2009], Malawi (mid 1960's) [Travers *et al.*, 2004], India (mid 1970's) [Shen *et al.*, 2011], Brazil (early 1980's) [Bello *et al.*, 2008], Angola (multiple introductions between the late 1970's and early 2000's) [Afonso *et al.*, 2012], Senegal (early 1980's) [Jung *et al.*, 2012], and the United Kingdom (early 1990's) [de Oliveira *et al.*, 2010].

4.4.1.2 tMRCA of the Southern African epidemic (excluding Cape Town)

Once again the estimated date of origin from the three different Southern African data sets (excluding Cape Town sequence information from Cape Town), as was inferred in BEAST under various different model parameters, produced very similar results (Tables 3.4 – 3.9).

The estimation of the date of origin from the Southern Africa *gag* p24 sequence data (Table 3.4) suggests a date of origin around the late 1950's when comparing the results of all the date that was inferred under the various model tree priors. However, the estimated tMRCA from the "best-fitting" model as was identified through Bayes factor comparison identified the date of origin at

1955,9 (1936,5 – 1972,0). The mean mutation rate for this model that was inferred under a variable mutation rate was $2,8 \times 10^{-3}$ mutations/site/year (95% HPD ranging between $1,8 \times 10^{-3}$ and $3,9 \times 10^{-3}$ mutations/site/year). Once again the mutation rate played a crucial role in the outcome of the analysis. Four of the run parameters produced an estimated tMRCA around the late 1920's or 1930's (1927 – 1937) due their slower estimated mutation rates. Additionally two of the other runs were inferred from mutation rates that were slightly faster than the mutation rate for the “best-fitting model” for this *gag* p24 data set. These two runs slightly underestimated the average root height or the tMRCA (1967 – 1970).

The estimated tMRCA of the concatenated *gag-pol*, Southern African only data set (Table 3.6), as was identified through Bayes factor comparison placed the date of origin around 1946,0. However, the 95% confidence intervals for this particular runs are quite large, stretching from 1926,7 to 1972,1. This estimated tMRCA was inferred under a mean mutation rate of $2,7 \times 10^{-3}$ mutations/site/year with the 95% HPD ranging between $5,4 \times 10^{-4}$ and $3,4 \times 10^{-3}$ mutations/site/year.

Lastly, the estimated inferred tMRCA of the Southern African *pol* data set (Table 3.8) from various runs parameters placed the date of origin around the late 1950's. However, the tMRCA from the “best-fitting model” as was identified through Bayes factor comparison was 1960,6 (1948,0 – 1972,4). This tMRCA was inferred under an estimated mutation rate of around $2,3 \times 10^{-3}$ mutations/site/year with the 95% HPD ranging between $1,8 \times 10^{-3}$ and $2,8 \times 10^{-3}$ mutations/site/year, which is slightly slower than the mutation rate of $2,55 \times 10^{-3}$ that was used by Hue and co-workers [Hue *et al.*, 2004].

It is therefore clear, based on the data that was inferred during the course of this study, that the estimated date of origin of the HIV-1 subtype C epidemic, from Southern African sequence information (excluding Cape Town), can be traced back to some point in the mid to late 1950's, with the 95% HPD ranging between 1929 and 1971. This estimated date of origin fits in well with the findings of the previous studies by Novitsky and co-workers [Novitsky *et al.*, 2010] that estimated the date of origin for the global HIV-1 subtype C epidemic around mid 1950's from *gag* sequence data. The estimated date however, is slightly older than that found by Travers and co-workers [Travers *et al.*, 2004] that placed the date of origin for the global HIV-1 subtype C around the mid 1960's. However, an estimated date of origin somewhere in the late 1950's still falls within the lower 95% confidence interval for the estimated tMRCA's that was found in the study of Travers and co-workers.

4.4.1.3 tMRCA of the entire Southern African epidemic (including Cape Town)

The estimated inferred tMRCA of the entire Southern African HIV-1 subtype C epidemic (including Cape Town sequence data) from the *gag* p24 genomic region (Table 3.10), placed the date of origin around the 1950's. The date of origin of the “best-fitting run” for this *gag* p24 data set, as was determined through Bayes factor model comparison, was 1950,3 (1933,5 – 1964,0). The estimated mutation rate for this run was $2,1 \times 10^{-3}$ mutations/site/year with the 95% HPD ranging between $1,5 \times 10^{-3}$ and $2,7 \times 10^{-3}$ mutations/site/year. This rate is slower than the rate that was estimated for the other two *gag* p24 data sets and the mutation rate that was obtained by Novitsky and co-workers [Novitsky *et al.*, 2010]. However, the upper limit of the 95% confidence interval still falls within the range for those mutation rates that were found in the other *gag* p24 data sets.

The concatenated *gag-pol* and *pol* data sets of the entire Southern African region were all inferred with an extremely low effective sample size (ESS), which is normally an indication of poor performance in the MCMC. However as was explained previously, the reason for these low ESS values is due to the large size of these data sets ($n > 500$). With such large data sets, spanning over a large time frame and comprising genetically diverse isolates from a large geographical region, it is extremely difficult for the Bayesian MCMC to obtain a good posterior distribution of the various model parameters. However, given the good convergence in the trace files (Figures 6.3 – 6.6) one can interpret these results with confidence. Careful, consideration of the mutation rate and the ESS does however indicate that the estimated tMRCA of these data sets can still be traced back to around the some point in the 1950's or early 1960's. The estimated date of origin for the concatenated *gag-pol* data set, as determined through Bayes factor calculation, places the date of origin around 1963,3 (1952,9 – 1972,3). This run was inferred under a fixed mutation rate of $2,5 \times 10^{-3}$ mutations/site/year (Table 3.12).

Similarly, the estimated tMRCA of the Southern African *pol* data set, with sequence information from the Cape Town data sets (Table 3.14), for the “best-fitting model” was placed around 1951,9 with the 95% HPD ranging between 1927,2 and 1971,0. This estimated tMRCA was inferred under an estimated mutation rate of $2,6 \times 10^{-3}$ mutations/site/year with the 95% HPD ranging between $2,1 \times 10^{-3}$ and $3,0 \times 10^{-3}$ mutations/site/year.

As with the inferred estimated tMRCA of the Southern African epidemic (excluding Cape Town sequence data) this inferred tMRCA fits well with the previously inferred dates of origin for the global subtype C pandemic [Novitsky *et al.*, 2010] from *gag* p24 sequence data. The estimated

date of origin for this data set is slightly older than that found for the global HIV-1 subtype C pandemic and Malawian epidemic [Travers *et al.*, 2004] that was inferred under various conditions from *gag* p24 and *env* sequence data.

4.4.2 Epidemic reconstruction of the Cape Town and Southern African HIV-1 subtype C epidemics.

The growth in the HIV-1 subtype C epidemic of Cape Town/South African, as well as the greater Southern African region were reconstructed from selected non-parametric tree priors that were used during the dating of the epidemic. Furthermore, the estimated percentage lineages through time were also estimated for the Cape Town/South African epidemic.

4.4.2.1 Phylodynamic aspects of the Cape Town epidemic

The demographic history of the HIV-1 subtype C epidemic in Cape Town was reconstructed from the “best-fitting” non-parametric runs for both the *gag* p24 and *pol* data sets. The *gag* p24 BSP was reconstructed under an estimated mutation rate, while the BSP for the *pol* data set was reconstructed under a fixed mutation rate. The inferred BSPs is only a plot of the Effective Population Size (EPS), which is not a direct reflection of the actual size of the epidemic, but rather a representation of the number of infected individuals who actively contribute to the next generation.

Close examination of the two plots reveals a linear growth in the epidemic over time with a brief period of strong exponential growth. This period of epidemic growth for the *gag* p24 BSP (Figure 3.2) appear around the mid 1980's (1984 - 1987), while the period of strong epidemic growth in the *pol* BSP (Figure 3.3) appears to be around the late 1980's (1988 - 1992). The *gag* p24 BSP plot is possibly not a good reflection of the demographic history of the epidemic in Cape Town, given the fixed mutation rate, the short time span of the inferred *gag* p24 BSP plot (1975 till the present), and the short fragment length of the *gag* p24 data sets. However, both plots suggest a 3-fold log increase in the EPS. This implies an increase in the total number of infected from 250 around the mid 1970's to 250,000 in the year 2010, given the modest assumption of a 5% population level HIV prevalence in the city of Cape Town in 2010.

When compared to the reconstructed BSP's from earlier studies, the Cape Town epidemic appears to have a far more linear increase in the EPS. This contrasts with the reconstructed EPS

from the Zimbabwean study [Dalai *et al.*, 2009] and the Indian study [Shen *et al.*, 2011] showing a very clear increase (roughly 4-fold log increase) during the 1980's.

Apart from the reconstruction of the demographic history from sequence data, the estimated percentage lineages through time (PLTT) were also inferred from the "best-fitting" *gag* p24 and *pol* runs. The inferred PLTT from the *gag* p24 data set (Figure 3.4) suggest a slow increase in viral genetic diversity leading up to the start of the 1980's. During the course of the 1980's it would appear that a massive increase in the genetic diversity (increase from 14,9% to 79,2%) of the viral variants occurred. By the end of the 1980's almost 80% of all of the current genetic variants were already present. The estimated PLTT from the *pol* Cape Town data set (Figure 3.5) suggests a similar slow increase in viral diversity leading up to the start of the 1980's, after which a massive increase in viral genetic diversity was observed (from 24,7% to 83,3%). By the end of the 1980's more than 80% of the current genetic variants were already present. Since then the process of viral expansion has continued until the present day.

Therefore, the inferred PLTT data suggest an initial slow increase in viral variance, followed by a massive increase during the 1980's. Since then the genetic diversity has continued to increase at a slow but steady rate. This compares well with the reconstructed Zimbabwean data that was inferred by Dalai and co-workers that also seen a similar explosion in viral genetic diversity during the same time period [Dalai *et al.*, 2009].

The use of different *gag* p24 and *pol* data runs in the epidemic reconstruction, both for the construction of BSPs as well as the calculation of PLTT, does give very similar results with only small variations in the estimated parameters. Such small variations however, are to be expected when comparing two different gene fragments and inferring demographic histories under variable model parameters. This suggests that these methods are quite robust in their performance and can be used readily for the inference of demographic histories from sequence data.

4.4.2.2 Reconstruction of the Southern African epidemic (excluding Cape Town)

Epidemic reconstruction of the Southern African epidemic (excluding sequence data from Cape Town) was achieved through standard non-parametric means (Bayesian Skyline Plot) under a relaxed molecular clock assumption from selected *gag* p24 (Figure 3.6) and *pol* (Figure 3.7) runs. Close inspection of the two inferred epidemic plots reveals a distinct pattern of constant linear

growth in the epidemic (based on the EPS) with two brief periods of exponential growth in the epidemic. This is suggestive of periods of strong epidemic growth in Southern Africa.

For the *pol* BSP plot that was inferred from sequence data from Southern Africa, excluding isolates from Cape Town (Figure 3.7), it would appear that there was a major growth in the HIV-1 subtype C epidemic in the early 1980's (1981 - 1985), which was followed by a more linear increase (1986 - 1992). This was then followed by another brief period of epidemic growth between 1993 and 1997. This brief period of epidemic growth was once again followed by a more linear increase over time (1998 - 2010).

This growth is broadly reflective of the known epidemiological data and prevalence trends of HIV that is available for the Southern African region. The first reported cases of heterosexually acquired HIV in Southern Africa were first reported in the Northern countries of the region (e.g. Zambia, Zimbabwe, and Malawi) [Hira *et al.*, 1989; Reeve, 1989; Ingstad, 1990]. These countries experienced massive increases in HIV prevalence during the 1980's and by the end of the decade they recorded the highest number of documented AIDS cases in the Southern African region (Table 1.1). Considering the basic assumption of a 10-year latency period, from infection till the progression to AIDS, it would appear that HIV may have increased rapidly during the late 1970's and early 1980's before health officials in these countries became aware of the epidemic. The initial exponential growth phase in the EPS in the *pol* BSP Southern African plot (Figure 3.7) could therefore represent this epidemic explosion in these countries in the region. Furthermore, the second brief period of epidemic growth in the *pol* BSP Southern African plot could be representative of the massive increases in HIV prevalence in the rest of the Southern African region (e.g. Mozambique, South Africa, and Swaziland) in the period between 1993 and 1997.

The more linear nature of the reconstructed BSP of the Southern African epidemic (excluding sequence information from Cape Town) can possibly be contributed to the averaging of epidemic reconstruction over the entire Southern African region, with different countries experiencing epidemic growth at different time points. This contrasts sharply with the reconstructed population dynamics of HIV-1 subtype C epidemics in countries such as Zimbabwe [Dalai *et al.*, 2009] and India [Shen *et al.*, 2011] that experience a distinct prolonged period of strong epidemic growth. The more linear growth of Southern African epidemic could be attributed to the fact that various countries in the region experienced exponential growth in their respective epidemics at different time points which can change the plot to a more linear trajectory.

4.4.2.3 Reconstruction of the Southern African epidemic (including Cape Town)

From the dynamic epidemic reconstruction of the entire Southern African HIV-1 subtype C epidemic from selected *gag* p24 (Figure 3.8) and *pol* (Figure 3.9) data sets, revealed a strikingly similar pattern of epidemic growth. Both the inferred plots show a slow linear growth in the effective population size of the epidemic from the start of the epidemic till the present. During this period a similar period of exponential growth in the EPS was observed in the early 1980's (Figure 3.9) just as with the reconstructed BSP from the Southern African *pol* data set (Figure 3.7) containing no Cape Town sequence information. This first exponential growth phase in the BSP as was mentioned in the previous section may be representative of epidemic growth in HIV in the northern most countries of Southern Africa (e.g. Zambia, Malawi and Zimbabwe). Such an exponential increase in the EPS in the Zimbabwean analysis was also observed during the 1980's [Dalai *et al.*, 2009].

This first exponential phase in the BSP was followed by a linear increase in the EPS. No clear second exponential growth phase could be observed in *gag* p24 (Figure 3.8) and *pol* (Figure 3.9) BSP plots when looking at the median growth curves. However, the 95% HPD of these two plots does reveal a small increase which may be suggestive of the second Southern African growth phase in the epidemic in the southern most countries (e.g. Swaziland, Lesotho, South Africa).

Once again, the strikingly similar growth phases from two different regions of the HIV-1 genome and two very different data sets, one containing Cape Town information (Figures 3.6 and 3.7) and the other excluding Cape Town sequence information (Figures 3.8 and 3.9), suggests that these methods are fairly robust in their attempts to reconstruct demographic histories from sequence data.

4.5 Factors that may influence HIV-1 epidemic reconstruction

In phylogenetics there are a wide range of factors that may influence the results or outcomes of any investigation, and the inference of demographic HIV epidemic histories through Bayesian methods is no exception. Since the start of HIV epidemic reconstruction in the 1990's several concerns have been raised that may influence the validity of the results of such endeavours. These include: (1) the effect of viral genetic diversity between strains or subtypes, (2) the effect of viral recombination, (3) the number of the taxa and fragment size of the data set, (4) the specific model parameters that was used, and (5) the effect of mutation rates.

HIV-1 subtypes have a large effect on the reconstruction of HIV demographics. Since the zoonosis of HIV from non-human primates to humans a large degree of genetic variation has accumulated amongst HIV-1 isolates [Sharp and Hahn, 2011]. These genetic variations have led to the rise of distinct HIV-1 strains or subtypes [Hemelaar, 2011]. The reconstruction of demographic histories from sequence data relies heavily on the assumption of a molecular clock and the coalescent theory. The coalescent theory is broadly based on the tracing of isolates back in time until all isolates, and their genetic information, has coalesced to a single point back in the distant past [Kingman 1982a; Kingman 1982b]. The inclusion of isolates from multiple subtypes of HIV will therefore inherently have a major effect on the reconstruction of past demographics [Lemey *et al.*, 2006]. In order to control for this all isolates were carefully subtyped in this study to insure that only HIV-1 subtype C sequences were included in the analysis.

Secondly, the effect of viral recombination on the dynamic reconstruction of HIV histories has attracted a significant amount of attention in the past. Viral recombination of HIV occurs in dually infected cells. This can occur between isolates from a single subtype (inter-subtype recombinants) or within patients who has been infected with two different subtypes, which then give rise to URFs or CRFs. The influence of HIV recombination on the reconstruction of demographic histories has been one of the main objections to the validity of these phylogenetic methods [Lemey *et al.*, 2006]. There are a couple of factors that needs to be considered when dealing with the problem of recombination on the reconstruction of demographic histories of HIV.

Firstly the demographic reconstruction would only be seriously hampered by recombinants between distinct HIV variants, which require co-infection or super-infection. Although a large number of such cases and their corresponding mosaic genomic structures have been documented the estimated rate of their prevalence across the entire infected populace remains fairly low. Secondly, recombination events between more divergent parental lineages (e.g. subtype B and C) will have a more profound impact on the results when compared with inter-subtype recombination. The intra-subtype recombinants can be manually excluded from the analysis. This method may seem like a round about way of dealing with the problem of recombination, but if the main study objective were to investigate the demographic history of a single subtype from an area, such as in this study, this method would be an appropriate way of dealing with the problem. Finally, recombination events between lineages in a single infected individual (inter-subtype recombinants) will not bias the inference process on a population level [Lemey *et al.*, 2006], even though they may lead to variation in the evolutionary rate [Rousseau *et al.*, 2007]. This problem

can now easily be overcome with the implementation of a relaxed molecular clock, which allows for variation in the rate in different branches of the topology.

Another concern regarding demographic reconstruction has been the total size of the number of taxa included in the analysis. Too many isolates would unnecessarily slow down to analysis, while too few isolates would leave out too much of the genetic information that is needed to infer epidemic histories [Philippe *et al.*, 2011]. It has been shown that by increasing the number of taxa in any given phylogenetic investigation may greatly reduce the degree of phylogenetic error [Zwickl and Hillis, 2002]. Similarly, the total size of the nucleic fragments, as with any phylogenetic investigation, plays another important role. The size of the nucleic acid fragments not only determines the speed of the analysis, but larger fragments carry more genetic information than smaller fragments. It is generally regarded that a fragment length of 500 bp or more for HIV carriers enough genetic information for reasonable phylogenetic inference. This however, depends on the type of investigation. Therefore, careful consideration needs to be taken regarding the sample size and fragment size in the study design of any phylogenetic investigation.

Lastly, the effect of mutation rates may have a profound impact on the outcome of any phylodynamic reconstruction. Once again the coalescent theory and the implication of a molecular clock apply here. The mutation rate is the rate at which mutations arise within a given lineage. If the mutation rate is too slow or too fast the estimated tMRCA of this lineage will be either too far back in the distant past or too recent. Great care therefore, also needs to be taken with the setting up of analysis. It would be advantageous, if time and computational power allows it, to conduct “exploratory” runs of a new data set, in order to establish the mutation rate before conducting a full in-depth analysis. In this study, such exploratory runs were conducted on several occasions and with the use of several different model parameters as well as the use of randomly sampled sequences from the Southern African data sets, before the final analysis were set up. These exploratory runs consistently produced similar results regarding the estimated data of origin and mutation rates (data not shown).

Furthermore, the mutation rate within host and at an epidemic level vary considerably and may therefore play a crucial role in the reconstruction of epidemic histories as was suggested by Lythgoe and Fraser [Lythgoe and Fraser, 2012]. It is generally accepted that the mutation rate of HIV-1 are significantly higher within patients when compared to the epidemic rate of mutations. In the study by Lythgoe and Fraser, the authors proposed three main mechanisms to explain this higher rate of mutation within patients. However, the mutation rates for only the *env* portion of

the HIV-1 genome was compared in this study. The *env* portion of the HIV-1 genome is highly variable and is under more selective pressure when compared to the *gag* or the *pol* regions of the HIV-1. This is largely due to strong immunological pressure from the host immune response. However, the problem of the inter- and intra-host level of mutation and their effect on the molecular clock is still a major issue. Currently, Bayesian models are not sophisticated enough to incorporate for such variations in the mutation rate and therefore the models for the calculation of tMRCA's or the reconstruction of the dynamic histories of epidemics should be calibrated to incorporate for these variations in the mutation rate.

4.6 Basic phylogenetic investigation to establish the evolutionary relationship of the HIV-1 subtype C isolates from Cape Town

A basic phylogenetic investigation was performed in order to determine the evolutionary relationship of the HIV-1 subtype C isolates contained within the Cape Town data sets. This was done through the inference of large-scale Minimum Evolution and Maximum Likelihood tree topologies, which were analysed with both manually as well as with new software called Phylotype.

4.6.1 Data mining

From the data mining section it is clear that a large amount of sequence data can be retrieved from public sequence databases. The most recent assessment of the HIV specific database at the Los Alamos National Laboratory (LANL) contains more than 400 000 HIV-1 sequences of which, roughly 57 000 are HIV-1 subtype C isolates (<http://www.hiv.lanl.gov>).

The data mining process in this study produced 2 347 homologous *gag* p24 sequences (1246 – 1727 bp relative to HXB2), 5 377 homologous *pol* sequences (2264 – 3321 bp relative to HXB2), and 455 homologous *gag-pol* sequences (1246 – 3321 bp relative to HXB2). The overwhelming number of *pol* sequences in the database can largely be attributed to the fact that this part of the HIV-1 genome is regularly characterized for the screening of drug resistance mutations. The smaller number of *gag-pol* concatenated sequences is largely due to the absence of long sequence fragments within the HIV sequence database and therefore this data set were largely restricted to full-length or near full-length HIV-1 subtype C sequences.

However, some of the homologous sequences that were retrieved were characterized from the same patient. The characterization of genomic data from the same patient is normally done to

study the *in-vivo* evolutionary process of HIV or to track the emergence of HIV drug resistance mutations in patients. It is important to remove multiple sequences from the same patient as they may seriously compound the results of phylogenetic inference methods [Lemey *et al.*, 2006]. Since one of the aims of this study was to look at the origin and evolutionary dynamic aspects of the HIV-1 subtype C epidemic in Cape Town and the Southern African region at large, multiple sequences from the same patient would not only distort the phylogenetic inference process, but the inclusion of inter-host genetic information would be completely distort the evolutionary and historical inference of results.

It has been found that the evolutionary rate of HIV on a population level (intra-host) to be much slower than when compared to inter-host HIV-1 mutation rates [Lemey *et al.*, 2006]. This is due to the fact that neutral selection pressures largely govern HIV evolution on a population level, while the evolution of HIV within hosts may be influenced by a multitude of selection pressures (e.g. host immune escape, drug selection pressure) [Pybus, 2000]. It is thus of utmost importance to remove multiple sequences from data set, if there is no need for such genetic information, as they could seriously distort the inference of evolutionary processes. This may often, especially when one is working with extremely large data sets, prove to be an extremely laborious process, but may be of paramount importance.

In total, 645 duplicate *gag* p24 sequences and 398 duplicate *pol* sequences were removed from each data set through careful screening. No duplicate sequences were found in the concatenated *gag-pol* data set. However, the large size of the original *pol* data set ($n = 5\ 145$) was too big to compute phylogenies in a reasonable time frame. Therefore this data set was reduced to around 2 334 taxa.

4.6.2 Large-scale phylogenetic inference

The inference of large-scale ME-SPR and ML.aLRT tree topologies, from three different genomic regions of the HIV genome, revealed information about the genetic relationship of the Cape Town isolates and the relationship to other subtype C isolate from around the world.

A close genetic relationship between the sequences in the Cape Town data sets and with other isolates from the rest of South Africa was observed. In the *gag* p24 and *pol* ME-SPR tree topologies, roughly 40% of the Cape Town isolates clustered with other sequences contained within the Cape Town data sets, while roughly another 40% clustered with other isolates from South Africa (including Cape Town sequences that were characterized in other studies). This

close clustering was also observed in the concatenated *gag-pol* ME-SPR tree were 51,9% of the Cape Town isolates clustered with one another, while roughly 28,0% clustered with other South African isolates. Examination of the corresponding ML tree topology with aLRT support also revealed broadly similar trends. In the *gag* and *pol* ML.aLRT tree topologies 43,1% and 50,2% clustered with other Cape Town isolates, while 41,1% and 35,8% clustered with other South African sequences. This trend was also noticeable in the concatenated *gag-pol* ML.aLRT tree topology where 23,1% of the Cape Town isolates clustered with other South Africa taxa. This is slightly less than for the *gag* and *pol* tree topologies. However, a large number of Cape Town isolates clustered with one another (61,5%) in this ML.aLRT tree topology when compared with the corresponding *gag* and *pol* tree topologies.

Following the close genetic relationship between the Cape Town isolates with other isolates from Cape Town and the rest of South Africa, the third most important genetic relationship was those of isolates from other areas of the Southern African region. Close inspection of the tree topologies also revealed a close clustering between 9,7% and 16,4% of the Cape Town isolates with those from other Southern African countries, most notably isolates from Botswana, Malawi, Mozambique and Zambia. Only a small number of the Cape Town isolates clustered with other subtype C sequences from outside the Southern African region. These sequences were largely of an East African, Middle Eastern or European origin. Only a single isolate in the Cape Town data set shared a close genetic relationship with an isolate from India, while none clustered with any of the South American isolates.

This kind of clustering may be an indication of some degree of a more isolated heterosexual HIV-1 subtype C epidemic occurring in the Cape Metropolitan area than in other areas of South Africa. However, the largest number of Cape Town isolates that clustered with other Cape Town isolates was amongst the oldest samples in the data sets (1989 - 1992). These old samples represent some of the earliest sequence information from HIV-1 subtype C isolates from the Southern African region. In the PhyloType analysis of the ME-SPR tree topologies, based on temporal-geographical criteria, roughly 72% of the *gag* p24, 92% of the concatenated *gag-pol*, and 73% of the oldest *pol* Cape Town isolates clustered with one another. This trend was also observed in the corresponding ML.aLRT tree topologies where 80% of the *gag* p24, 79% of the concatenated *gag-pol*, and 80% of the oldest *pol* isolates from Cape Town also clustered with one another. As the epidemic progressed, and the numbers of HIV-1 sequences available from other areas of South Africa and Southern Africa increased over time, this high degree of clustering of Cape Town isolates with one another decreased.

In addition to the high degree of clustering of Cape Town isolates based on their temporal classification, as was observed through the PhyloType analyses, highly monophyletic clusters of Cape Town isolates were observed in all of the inferred tree topologies. In the *gag* p24 tree topologies one monophyletic cluster of Cape Town sequences were observed in both of the two different tree topologies. Similarly, in the *gag-pol* tree topologies two monophyletic cluster of Cape Town sequences were also observed. The biggest of which contained 11 taxa while the smallest contained 6 Cape Town isolates. For the *pol* ML.aLRT tree topology three large monophyletic clusters of Cape Town sequences were observed. The biggest of which contained 33 Cape Town isolates, while the smallest contained only 22 Cape Town isolates. However, in the ME-SPR tree topologies these clusters were not supported by the bootstrap values for their internal branches. However, in the cluster in the ML.aLRT tree topologies each cluster in both the *gag* p24 and *pol* tree topology, were supported by their likelihood scores for their internal branches.

Such highly monophyletic clustering of HIV-1 subtype C sequences is highly unusual and is normally an indication of some transmission network in such a small geographic area. Several HIV-1 subtype C transmission events in other areas of the world have been uncovered. For instance 5 transmission clusters of HIV-1 subtype C were uncovered in Senegal, all circulating amongst the local MSM population in Dakar [Ndiaye *et al.*, 2009]. These samples were then later used for the inference of the demographic history of the Senegalese HIV-1 subtype C epidemic [Jung *et al.*, 2012].

4.7 Monophyletic clades of HIV circulating in the Cape Metropolitan area

With the inference of the large-scale tree topologies a large degree of monophyletic clustering of several of the Cape Town isolates was observed. The manual analysis of the *gag* p24 and *pol* ML.aLRT tree topologies identified 5 putative clusters in the *gag* p24 tree topology and 3 putative clusters in the *pol* tree topology (Figures 3.17 and 3.18). Such monophyletic clustering is unusual for the country with a generalized HIV-1 epidemic. As such I initially thought that they were representative of HIV-1 transmission clusters as this was very strange particularly for sequence that were sampled from several different studies within the same geographical region (e.g. Cape Metropolitan region) over several years. This suggested that the sequences could not be contaminates of one another. However, HIV-1 transmission clusters in currently published accounts uses very high bootstrap support values. The monophyletic clades that were observed in this study had very low bootstrap support (<70%) and in many cases the support for these clusters

where zero. However, the support for these putative transmission clusters in the Maximum Likelihood tree topology much higher (>80%). This is large due to the fact that the ML-tree topologies were inferred with an approximate likelihood ratio test method of branch testing. In large phylogenies of HIV-1 bootstrap values tend to decline due to the nature of the resampling strategy of bootstrap resampling as well as the highly variable nature of HIV-1 sequences.

However, due to the lack of adequate bootstrap support for these monophyletic clades of HIV-1 subtype C sequences there was no certainty about their clustering. As such, I tested whether these clusters held up in additional clustering analyses. For this I tested each cluster against the HIV-1 subtype C reference data set from the Los Alamos National Laboratory's HIV-1 database. For each putative cluster 5 different phylogenies were inferred. One NJ-tree topology with bootstrap resampling (1,000 bootstrap replicates). One ME-tree topology also with bootstrap resampling (1,000 bootstrap replicates). Two ML-tree topologies, one of which were inferred with the approximate likelihood ratio test of branch support and one with bootstrap resampling. Lastly, a Bayesian tree topology for each of the data sets were also inferred. Only samples that clustered consistently with one another were considered to be true.

Bayesian inferences on each of the identified clusters were performed in order to calculate the date of origin of each of the clusters. The average estimated root height of *gag*.cluster.1 (Table 3.29) could be traced back to around 1986,4 (95% HPD 1981,1 – 1990,8). Similarly the estimated dates of origin of *gag*.cluster.2 and *gag*.cluster.3 were placed around 1986,7 (95% HPD 1983,3 – 1989,3) and 1993,6 (95% HPD 1989,4 – 1997,4) respectively, while the estimated root height of *gag*.cluster.4 was placed around 1980,8 (95% HPD 1972,9 – 1988,1). Lastly, the estimated tMRCA of the *gag*.cluster.5 was traced back to around 1995,1 (95% HPD 1991,2 – 1998,3). Likewise, the average inferred tMRCA of the *pol*.cluster.1 and *pol*.cluster.2 (Table 3.30) was estimated around 1980,8 (95% HPD 1977,4 – 1984,0) and 1985,9 (95% HPD 1983,5 – 1987,9) respectively. The estimated root height of the *pol*.cluster.3 was placed around 1992,2 (95% HPD 1990,1 – 1994,2). Due to the small number of taxa in some of the clusters only a parametric tree prior (e.g. constant population size or exponential growth) could be used for the inference of their respective tMRCA's.

Due to the sampling and characterization of patient samples, the patients that are presented in the *gag* p24 Cape Town data set does not reflect the same way in the *pol* Cape Town data set, and only a small number of the very oldest (1989 - 1992) and very youngest samples (2008 - 2010) are found in both data sets. However, the transmission clusters *gag*.cluster.2 and *gag*.cluster.4

almost contains the same patient samples contained in *pol.cluster.1*. All of these clusters contain some of the oldest patient samples from the Cape Town data sets and their estimated tMRCA's can be traced back to the early to mid 1980's. It would therefore appear through the analysis of these longitudinally sampled data sets that two major clusters of HIV-1 subtype C occurred within the Cape Metropolitan area in South Africa. The first appears early in the heterosexual HIV-1 subtype C epidemic during the late 1970's and 1980's. This period was followed by massive increases in HIV transmission as is evident in the growth of the epidemic during the 1990's. With the massive increases in HIV-1 sequence data due to continued drug resistance testing newer HIV transmission events in recent years have also emerged from the analyses that was performed. However given the relative small size of the two data sets of Cape Town sequences, these identified clusters may only represent the "tip of the iceberg". Even though these patients contained in these clusters are indirectly linked through transmission networks the identification and characterization of the transmission clusters that are presented here do not fully describe the complex dynamics of HIV-1 transmission in these networks due to the small sample sizes of the data sets.

Close examination of data for each of the patients contained within these transmission clusters, revealed that the majority of the patients became infected through heterosexual contact while the mode of transmission for ten patients were not known and for four of the patients were via the MSM route. Four of the patients were Caucasian (2,89%), 19 were of Mixed Race (13,77%), 81 of the patients were of African origin (58,69%), while the racial background of 34 of the patients are unknown (24,65%). The male to female sex ratio of patients contained within these transmission clusters are broadly reflective of the national trends in HIV prevalence with 47 male (34,06%) patients and 76 female (55,08%) patients. For 15 patients the sex was unknown (10,86%). The average age of the patients was 33 years, 10 months and 13 days. The median age of the men were 36 years, 9 months and 15 days, while for the women the median age were 32 years, 2 months and 6 days. The small variation in age between the male and female patients within the transmission clusters is consistent with the national variation in age difference in HIV positive young adults within the country. The majority of the patients contained in these clusters (65,94%) were treatment naïve, while 19 patients (13,77%) were on antiretroviral therapy at the time of their sample collection. The treatment status of 28 patients (20,29%) was unknown. The average CD4 cell count for patients whom data was available for was 248 cells/ml, while the lowest cell count was 0 cells/ml of blood and the highest was 810 cells/ml.

These monophyletic clusters represent possible patients that are epidemiologically linked in some way. However, due to the weak branch support for these large clusters one cannot accept them as transmission clusters of HIV-1. Their consistency across different methods however, suggests that they are epidemiologically linked. Therefore, they may represent a small subset of a large sexual network within the Cape Metropolitan area. The large demographic background and dynamic profile suggest that HIV-1 sexual networks are quite complex and unpredictable. It is therefore important that HIV prevention and care not just target the general heterosexual population but to target all people from different demographic, sexual, racial and socio-economic backgrounds.

The methodology that was used for the identification and characterization of the monophyletic clades in this study provides an easy and effective method of identifying and testing potential transmission events in the future, which can be applied to analyse larger numbers of patient samples. This will provide us with valuable insight into factors that may be driving the epidemic on a micro-level. Therefore, the continued monitoring and testing of the HIV epidemic, including the investigation of the nature of transmission of the virus, is of critical importance in the fight against the HIV/AIDS epidemic not only within South Africa but also in other sub-Saharan African countries and the world at large.

4.8 Global HIV-1 subtype C perspective

Since the ME-SPR and ML.aLTR tree topologies that were inferred in this study contains almost all homologous HIV-1 subtype C isolates that is currently available in sequence databases, these tree topologies also provides a good perspective on the genetic relationship of the global HIV-1 subtype C pandemic. To date, these tree topologies (particularly the ME-SPR *pol* tree topology), to the best of my knowledge, are amongst the largest reconstructions of HIV-1 subtype C genetic relationships done globally, representing over 2,300 taxa.

A closer examination of the South American clusters in the three different tree topologies, particularly the *pol* ME-SPR tree topology (due to its larger size), suggest a common ancestor of the South American HIV-1 subtype C epidemic which originated from the East African region. This indication confirms the findings of two independent studies that were conducted by Bello and co-workers [Bello *et al.*, 2008] and Fontella and co-workers [Fontella *et al.*, 2008] on the evolutionary history of the HIV-1 subtype C epidemic in the South American region. The study that was conducted by Bello and co-workers [Bello *et al.*, 2008] suggested a single introduction of a HIV-1 subtype C strain from the East African country of Burundi. In the large-scale ME-

SPR tree topology this close relationship between HIV-1 subtype C isolates from Burundi and those from South America, particularly those from Brazil, is clearly visible. In addition to the evolutionary origin of the South American HIV-1 subtype C epidemic, another study [de Oliveira *et al.*, 2010] have found a possible epidemiological link between the subtype C epidemic in Southern Brazil and that in the United Kingdom (UK). This close relationship of subtype C isolates from Brazil and the UK is also visibly evident within the ME-SPR *pol* tree topology that was inferred during the course of this study.

Additionally, closer inspection of the South American subtype C cluster also revealed possible single introductions of HIV-1 subtype C from Brazil into Portugal and Japan. Since Brazil and Portugal share a common colonial history and Brazil is also home to the largest Japanese population outside of Japan, this suggests that a very clear epidemiological route of transmission may be present. However, the various studies in which these Portuguese and Japanese HIV-1 subtype C isolates were characterized, contains very little patient information on mode and suspected country of infection of these patients when compared with other published data [Esteves *et al.*, 2002; Palma *et al.*, 2007; Gatanga *et al.*, 2007; Ibe *et al.*, 2008].

Additionally, closer inspection of the Asian subtype C clusters in the ME-SPR tree topologies revealed a close genetic relationship between HIV-1 subtype C isolates from the Indian sub-continent (India, Pakistan and Nepal), as well as subtype C isolates from South-East Asia (Thailand and Myanmar) and East Asia (e.g. China, Taiwan, but with the exception of Japan and South Korea). This close genetic relationship between Indian subtype C isolates and other isolates from China, Taiwan, Thailand and Myanmar were first documented by Breyer and co-workers [Breyer *et al.*, 2000] that identified the overland heroin trafficking routes from Afghanistan through the northern Indian sub-continent into South East Asian and Southern China as the main mode of spread of the HIV-1 subtype C strain into these countries. This suggests a common ancestry in the HIV-1 subtype C epidemics in these countries, while subtype C isolates from other areas of the Asian continent such as, countries from the former Soviet Union [Zarandia *et al.*, 2006; Thomson *et al.*, 2007], the Middle East [Galai *et al.*, 1997; Saad *et al.*, 2005] may have originated from other sources.

Outside of Southern Africa, the HIV-1 subtype C epidemic in India ranks as one of the worst in the global pandemic. Several studies have been conducted on the origin and phylodynamic aspects of the HIV-1 subtype C epidemic in the Indian sub-continent [Shen *et al.*, 2011; Neogi *et al.*, 2012]. In the study of Shen and co-workers [Shen *et al.*, 2011], the authors identified a South

African origin for the HIV-1 subtype C epidemic in India and the corresponding countries in South-East and East Asia, while the study of Neogi and co-workers [Neogi *et al.*, 2012] identified a possible Southern African origin. Close examination of the ME-SPR *pol* tree topology that was inferred in this study also suggests a common Southern African origin. The country of origin however, is unclear since the Indian HIV-1 subtype C cluster were rooted in a common Southern African cluster containing isolates from Botswana, South Africa, Zimbabwe, Malawi and Zambia.

Furthermore, another clear epidemiological link between the HIV-1 subtype C isolates from East Africa and those isolated from Middle Eastern countries, particularly Saudi-Arabia, Yemen and Israel could be observed. This confirms the findings of previous studies from several Middle Eastern countries on the origin of the HIV-1 subtype C epidemic in these countries [Galai *et al.*, 1997; Saad *et al.*, 2005].

4.9 Southern African subtype C epidemic

4.9.1 The role of migration on the epidemic

The massive increases in HIV prevalence rates in the Southern African region during the 1980's and 1990's have been a topic of fierce debate in the past. Several theories have been proposed that can account for the large increases in HIV prevalence rates in the Southern African region in comparison with the rest of the sub-Saharan African region [Kreiss *et al.*, 1986; Plummer *et al.*, 1991; Allen *et al.*, 1993; Yamaguchi, *et al.*, 1994; Decosas *et al.*, 1995; Wollants *et al.*, 1995; Janssens *et al.*, 1997; Williams *et al.*, 2000; Glynn *et al.*, 2001; Buvé *et al.*, 2002; Quinn and Overbaugh, 2005; Baggaley *et al.*, 2010; Paxton, 2010; Fenwick, 2012]. The most important of these were briefly discussed in the first chapter. As was mentioned the topics of male circumcision and sexual concurrency and the respective roles that they have played in the Southern African HIV epidemic have extensively been investigated [Halperin and Bailey, 1999; Drain *et al.*, 2006; Baeten *et al.*, 2009; Dinh *et al.*, 2011; Morris *et al.*, 2009]. However, the latest scientific data still cannot reach a decisive conclusion on their respective roles in the massive increases in HIV prevalence across the Southern African region [Lurie and Rosenthal 2010; Sawers and Stillwaggon, 2010; Knopf and Morris, 2012].

Another major factor that has been raised is the effect of migration on the spread of the epidemic in the Southern African region. Migration and its relationship to the spread of the epidemic are better understood than that of male circumcision and sexual concurrency [Abdool Karim *et al.*,

1992; Decosas *et al.*, 1995; Lurie *et al.*, 1997; Wiseman, 1998; Coffee *et al.*, 2007; Newman *et al.*, 2011]. The practice of seasonal labour migration was well established in the first half of the 20th century by British colonial officials. This system was continued in several of the Southern African countries following the independence movement during the 1960's, and was expanded on by the Apartheid government of South Africa in order to ensure an adequate supply of unskilled labour for factories and mining companies in the large urban areas. This was achieved under the infamous "pass law" which restricted the number of Africans who could permanently settle in the large urban areas.

In addition to millions of African migrants from the rural South African countryside a large number of African migrants were also brought in from other Southern African nations (e.g. Botswana, Malawi, Zambia, Swaziland, Mozambique, and Lesotho). Official figures regarding the numbers of migrants are limited, but one study estimated a total of half a million foreign migrants were employed in the gold mining industry alone in the mid 1980's [Wiseman, 1998].

However, most of the studies have primarily focused on the role of labour related seasonal migration and does not include the role of politically motivated migratory patterns. In the latter part of the 20th century the Southern African region have seen a large degree of politically motivated migration. This includes the formation of guerrilla camps by paramilitary groups such as the Zimbabwean African National Union (ZANU), the Zimbabwean African Peoples Union (ZAPU), and the South West Africa Peoples Organization (SWAPO) during the wars of independence in Rhodesia and Namibia, the effects of the civil wars in Angola and Mozambique, and the activities of Umkhonto we Sizwe (MK) and other paramilitary groups during the Apartheid struggle in South Africa. The civil wars in Angola and Mozambique led to a large number of internally displaced refugees, while the Rhodesian and Namibian wars of independence led to a large number of paramilitary forces fleeing their respective countries to neighbouring countries in order to conduct frequent skirmishes across national borders. The MK of the African National Congress (ANC) in South Africa, following the Rivonia trial in the early 1960's, went into exile in neighbouring countries that were not friendly towards the Apartheid government in South Africa (Figure 4.1).

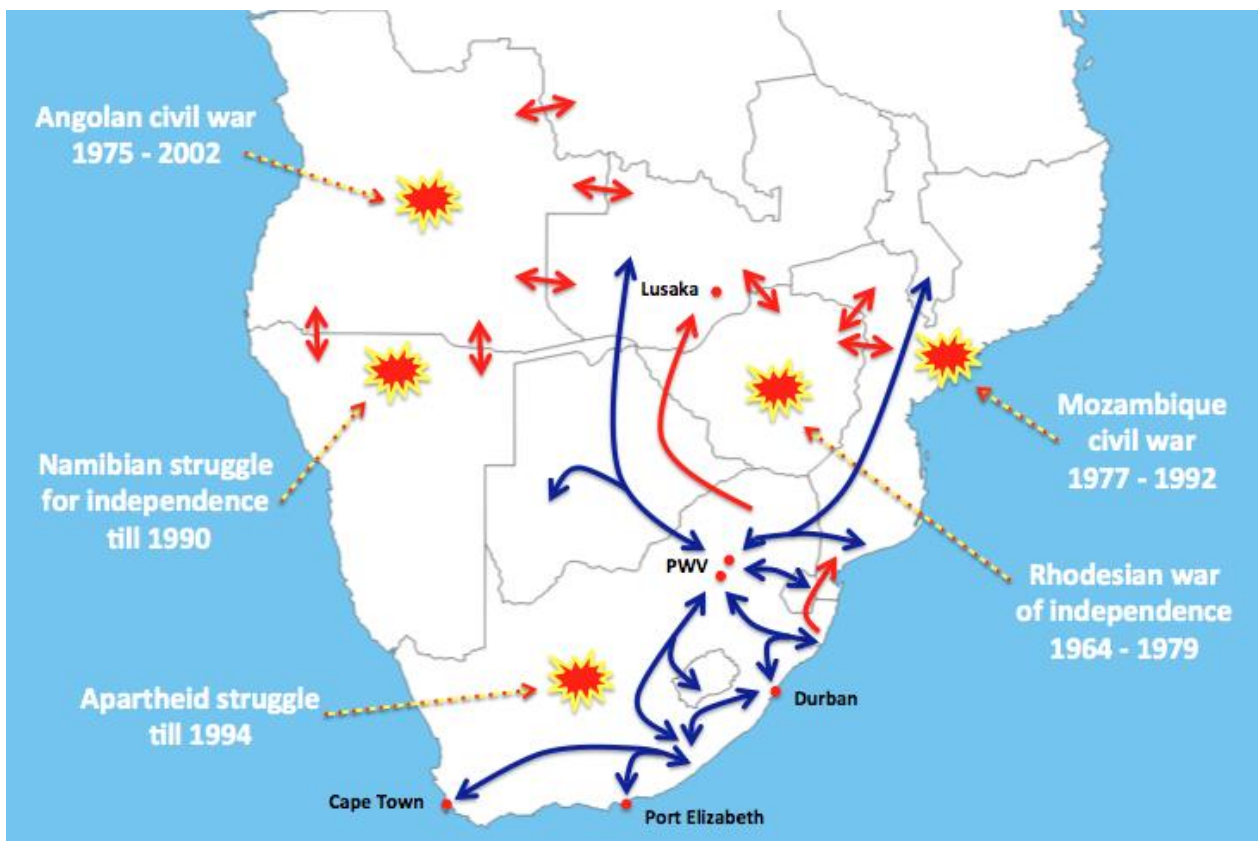


Figure 4.1: Major migratory routes across Southern African during the second half of the 20th century. The figure highlight's the most prominent foreign and domestic sources of migrant labour to South Africa (blue arrows), the biggest political struggles during the second half of the 20th century and the most prominent movements of paramilitary groups across the region from 1960 till 1994 [Authors own artwork].

Almost all of the politically motivated forms of migration were undertaken by young African men (aged 18 - 40). Due to their extreme isolation from the regular public, they may represent an extremely favourable environment for the spread of any sexually transmitted disease due to casual sexual relations and the use of commercial sex workers. Current support for the infection of a large number of freedom fighters is still lacking however one study has raised the importance of returning freedom fighters (over 40 000 returning expatriates), particularly in South African where the MK where integrated into the formal armed forces following the transition towards democracy, as an important source in the rapid spread of HIV-1 amongst local communities [Shell, 2001].

4.9.2 The two stage HIV-1 epidemic in Southern Africa

Conventional HIV prevalence estimates [UNAIDS, 2012], as well as the results contained within this and other studies, suggests that the HIV-1 subtype C epidemic in South Africa and the greater Southern African region has experienced two stages of growth. The first stage in the

Southern African HIV epidemic took place in the late 1970's and early 1980's in the more northern countries of the Southern African region (Zambia, Malawi and Zimbabwe). This is evident in the HIV prevalence estimates presented in Table 1.1 as well as in the Bayesian Skyline Plot reconstructions that was done in the epidemic reconstruction section (sections 3.3.2.2 and 3.3.3.2). In these Bayesian Skyline Plots, particularly the plots in Figure(s) 3.6 and 3.7 clearly show two phases of exponential growth in the effective population size(s). The first period of epidemic growth as observed in these Bayesian Skyline Plots possibly correlate to the growth in the HIV-1 subtype C epidemics in these countries. Furthermore, the findings from Dalai and co-workers also support this observation [Dalai *et al.*, 2009]. In their reconstruction of the Zimbabwean HIV-1 subtype C epidemic they observed massive increases in the effective population size during the 1980's, starting in the early years of the decade following the independence of Zimbabwe.

This initial stage of epidemic growth was followed by a more linear growth in the HIV epidemic in the Southern African region for a couple of years, with a second phase of exponential growth in the epidemic during the mid to late 1990's. Once again HIV prevalence estimates as well as the findings from this study are supportive of such a second growth phase. Close examination of the Bayesian Skyline Plot reconstruction from the Southern African data sets (Figures 3.6 and 3.7) clearly show a brief period of exponential growth in the effective population sizes. This growth period may correspond to the massive increases in HIV prevalence rates in the other countries of the Southern African region (Botswana, South Africa, Swaziland and Lesotho). HIV prevalence estimates for these countries by UNAIDS and government health departments [UNAIDS, 2012] support these assumptions of massive increases in HIV prevalence trends in these countries during the 1990's.

Furthermore, closer examination of the results of the Cape Town epidemic also suggests a two-staged HIV epidemic, similar to the findings in the Bayesian Skyline Plot reconstruction of the Southern African HIV epidemic. The Bayesian Skyline Plot reconstruction from Cape Town sequence data suggests two phases of exponential growth. One growth phase occurred during the mid to late 1980's (Figure 3.2), while a second growth phase occurred during the mid 1990's (Figure 3.2). These periods of epidemic growth coincide with major political changes that occurred during the time within South Africa. The first coincided with the abolishment of the infamous pass laws during the Apartheid years. The second period in the growth coincided with the end of Apartheid in the country and the transition to a fully democratic society. These events led to more freedom and an increase in the mobility of the general public. Greater mobility of

individuals ultimately leads to a greater chance of spreading any infectious disease, particularly one such as HIV with long asymptomatic phases.

It is therefore clear based on this two-stages HIV pandemic that the growth in HIV prevalence rates across the region may have been influenced largely by political factors and vacillated through regional migration.

CONCLUSION

One of the aims of this study was to generate longitudinal sampled sequence data sets from Cape Town patient samples, which were then used for the investigation of the evolutionary history of the HIV-1 subtype C epidemic in Cape Town, South Africa. This was achieved through the selection of stored patient samples from -20°C freezers and standard molecular techniques, which produced two sequence data sets: a *gag* p24 data set containing 193 sequences and a partial *pol* data set containing 166 sequences spanning over a 21-year period (1989 - 2010). The generation of a number of sequences (24 *gag* p24 and 74 partial *pol* sequence fragments) proves that sample databanks contained in freezers can be a valuable source of patient samples for the characterization of genetic data. This is particularly true for old samples, since a lot of genetic information from the early years of the HIV-1 epidemic are lacking.

The longitudinal sequence data sets were first used to investigate the evolutionary history of the Cape Town epidemic, which was then compared to epidemic trends from Southern Africa that was inferred from similar longitudinal sequence samples. The estimated date of origin for the Cape Town epidemic was found to be around the mid-1960's, which is on average about 10-years later than for the epidemic in the entire Southern African region. Additionally, epidemic reconstruction revealed a small steady increase in the growth of the epidemic for both Cape Town and the Southern African region, with small periods of exponential growth in the 1980's and 1990's. These periods of epidemic growth follows major political events in several countries and therefore it is highly plausible that major political events in the Southern African region may have played a crucial role in the spread of the epidemic in the region. The analysis of two different regions of the HIV-1 genome (*gag* p24 and a partial *pol* sequence fragment) gave very similar results even with the use of various different model parameters, which is indicative of the robustness of the results that was obtained.

The longitudinal sampled sequences were then used in a basic phylogenetic investigation to establish their evolutionary relation to other HIV-1 subtype C isolates from around the world. It

was established that the Cape Town isolates were highly associated with other isolates within the Cape Town data set, as well as from other isolates from South Africa and isolates from other Southern Africa nations. The tree topologies, to the best of our knowledge, represents some of the largest that have ever been inferred for HIV-1 subtype C sequences and provides an valuable insight into the evolutionary relationship of HIV-1 subtype C isolates from around the world.

Furthermore, highly monophyletic clusters of Cape Town sequences were further investigated. Through these analyses a small number of transmission events of HIV-1 were identified and further characterized. Molecular epidemiological reconstruction revealed that distinct transmission clusters have been spreading amongst isolated communities through-out the Cape Metropolitan region for several years or decades. It is clear that by increasing sample sizes one can start to uncover transmission events of HIV, which can provide valuable insight into factors that are driving the HIV epidemic on a micro-level.

This study is very unique as it entails the characterization of longitudinal sampled sequences derived from patient samples, which were stored for prolonged periods of time in freezers. Sequences that were generated from these patient samples represent some of the oldest available genetic data from HIV-1 subtype C isolates from South Africa and the Southern African region. The basic phylogenetic investigation that was conducted not only established the evolutionary relation of these Cape Town isolates, but the clustering pattern of global reference sequences in these large-scale phylogenies also confirms the findings of several other studies (as was discussed in section 4.8). The reconstruction of these epidemics is also the first study of its kind that has been preformed within South Africa and for the entire Southern African region as a whole. This study, along with previously mentioned studies also proves that the molecular clock can be used successfully to identify the origin of epidemics as well as to uncover epidemic growth trends. The transmission events of that were uncovered in this study also represents, to the best of my knowledge, the first identified and characterized transmission events of HIV-1 subtype C in South Africa amongst a local community. Finally, the new epidemiological model that was developed, even though still lacking empirical proof, provides a new perspective on the growth of the HIV-1 epidemic in the Southern African region.

This study however, is limited by the analysis of only Cape Town sequence data and thus the evolutionary history that was inferred from these sequences may not be representative of the entire country of South Africa. Therefore, further research will be required to validate the evolutionary history of the South African HIV-1 subtype C epidemic through the analysis of

heterochronous data sets from other areas of South Africa such as, Kwa-Zulu Natal and/or Johannesburg/Pretoria or a combination of samples from across the country. Furthermore, the new epidemiological model will need to be validated through additional phylogenetic models such as phylogeographics as well as standard epidemiological modelling.

CHAPTER FIVE

References

Abdool Karim Q, Abdool Karim SS, Singh B, Short R, Ngxongo S. Seroprevalence of HIV infection in rural South Africa. *AIDS*. 1992. 6: 1535 – 1539.

Abebe A, Lukashov VV, Pollakis G, Kliphuis A, Fontanet AL, Goudsmit J, de Wit TFR. Timing of the HIV-1 subtype C epidemic in Ethiopia based on early virus strains and subsequent virus diversification. *AIDS*. 2001. 15(12): 1555 - 1561.

Abecasis A, Vandamme A-M, Lemey P. Sequence Alignment in HIV Computational Analysis pp. 2-16 in *HIV Sequence Compendium 2006/2007*. Edited by: Thomas Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, Korber B. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. LA-UR 07-4826. 2007

Abu-Raddad LJ, Patnaik P, Kublin JG. Dual Infection with HIV and Malaria Fuels the Spread of Both Diseases in Sub-Saharan Africa. *Science*. 2006. 314: 1603 – 1606.

Afonso JM, Morgado MG, Bello G. Evidence of multiple introductions of HIV-1 subtype C in Angola. *Infection Genetics and Evolution*. 2012. 7: 1458 – 1465.

Ajoge HO, Gordon ML, de Oliveira T, Green TN, Ibrahim S, Shittu OS, Olonitola SO, Ahmad AA, Ndungú T. Genetic Characteristics, Coreceptor usage Potential and Evolution of Nigerian HIV-1 Subtype G and CRF02_AG Isolates. *PLoS ONE*. 2011. 6(3): e17865.

Allen S, Serufulira A, Bruber V, Kegeles S, Van de Perre P, Carael M, Coates TJ. Pregnancy and contraception use among urban Rwandan women after HIV testing and counseling. *American Journal of Public Health*. 1993. 83(5): 705-710.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology*. 1990. 215: 403-410.

Apetrei C, Kaur A, Lerche NW, Metzger M, Pandrea I, Hardcastle J, Falkenstein S, Bohm R, Koehler J, Traina-Dorge V. Molecular epidemiology of simian immunodeficiency virus SIVsm

in U.S. primate centers unravels the origin of SIVmac and SIVstm. *Journal of Virology*. 2005. 79: 8991 – 9005.

Apetrei C, Marx PA. African lentiviruses related to HIV. *Journal of Neurovirology*. 2005. 11(1): 33-49.

Asenkeyne OM, Lomogin J, Otim C. Struggling for survival in the era of HIV/AIDS in refugee camps in Kotido district (Uganda) experiences. In *Proceedings of the XIII International Conference on AIDS: July 2002*. Durban. Abstract MoPeD2643.

ASSA. 2003. The ASSA2003 demographic model produced by the Actuarial Society of South Africa. The model, and all relevant documentation explaining how the model was constructed and how it works can be accessed online at the following web address (<http://www.assa.org.za/default.asp?id=1000000050>).

Baeten JM, Celum C, Coates TJ. Male circumcision and HIV risks and benefits for women. *The Lancet*. 2009. 374(9685): 182 – 184.

Baggaley RF, White RG, Boily M-C. HIV transmission risk through anal intercourse: systematic review, meta-analysis and implications for HIV prevention. *International Journal of Epidemiology*. 2010. 39: 1048 – 1063.

Baldauf SL. Phylogeny for the faint of heart: a tutorial. *TRENDS in Genetics*. 2003. 19: 345 – 351.

Baleriola C, Johal H, Jacka B, Chaverot S, Bowden S, Lacey S, Rawlinson W. Stability of Hepatitis C Virus, HIV, and Hepatitis B Virus Nucleic Acids in Plasma Samples after Long-Term Storage at -20°C and -70°C. *Journal of Clinical Microbiology*. 2011. 49(9): 3163 – 3167.

Barre-Sinoussi F, Chermann J-C, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Brun-Vezinet F, Rouzioux C, Rozenbaum W, Montagnier L. Isolation of a T-Lymphotropic retrovirus from a patient at risk for Acquired Immune Deficiency Syndrome (AIDS). *Science*. 1983. 220(4599): 868 – 871.

Becker ML, Spracklen F, Becker WB. Isolation of a lymphadenopathy-associated virus from a patient with the acquired immune deficiency syndrome. *South African Medical Journal*. 1985. 68(3): 144 – 147.

Becker MLB, de Jager G, and Becker WB. MIC Analysis of Partial gag and env Gene Sequences of HIV type 1 Strains from Southern Africa. *AIDS Research and Human Retroviruses*. 1995. 11(10): 1265 – 1267.

Bello G, Passaes CPB, Guimaraes ML, Lorete RS, Almeida SEM, Medeiros RM, Alencastro PR, Morgado MG. Origin and evolutionary history of HIV-1 subtype C in Brazil. *AIDS*. 2008. 22: 1993 – 2000.

Bessong PO, Obi CL, Cilliers T, Choge I, Phoswa M, Pillay C, Papathanasopoulos M, Morris L. Characterization of Human Immunodeficiency Virus Type 1 from a Previously Unexplored Region of South Africa with a High HIV Prevalence. *AIDS Research and Human Retroviruses*. 2005. 21(1): 103 – 109.

Beyrer C, Trapence G, Motimedi F, Umar E, Lipinge S, Dausab F, Baral S. Bisexual concurrency, bisexual partnerships, and HIV among Southern African men who have sex with men. *Sexually Transmitted Infections*. 2010. 86: 323 – 327.

Bongaarts J, Reining P, Way P, Conant F. The relationship between male circumcision and HIV infection in African populations. *AIDS*. 1989. 3(6): 373 – 377.

Bredell H, Hunt G, Casteling A, Cilliers T, Rademeyer C, Coetzer M, Miller S, Johnson D, Tiemessen CT, Martin DJ, Williamson C, Morris, L. HIV-1 Subtype A, D, G, AG and Unclassified Sequences Identified in South Africa. *AIDS Research and Human Retroviruses*. 2002. 18(9): 681 – 683.

Bredell H, Martin DP, Van Harmelen J, Varsani A, Sheppard HW, Donovan R, Gray CM, HIVNET028 Study Team, Williamson C. HIV Type 1 Subtype C gag and nef Diversity in Southern Africa. *AIDS Research and Human Retroviruses*. 2007. 23: 477 – 481.

Brennan RO and Durack DT. Gay compromise syndrome. *Lancet*. 1981. 2(8259): 1338 - 1339.

Breyer C, Razak MH, Lisam K, Chen J, Lui W, Yu X-F. Overland heroin trafficking routes and HIV-1 spread in south and south-east Asia. *AIDS*. 2000. 14: 75 – 83.

Brown TA. 2002. *Genomes*. 2nd Edition. Oxford Wiley-Liss (ISBN-10: 0-471-25046-5).

Butler IF, Pandrea I, Marx PA, Apetrei C. HIV Genetic Diversity: Biological and Public Health Consequences. *Current HIV Research*. 2007. 5: 23 – 45.

Buvé A, Bishikwabo-Nsarhaza K, Mutangadura G. The spread and effect of HIV-1 infection in sub-Saharan Africa. *The Lancet*. 2002. 359: 2011 – 2017.

Campbell MS, Mullins JI, Hughes JP, Celum C, Wong KG, Raugi DN, Sorensen S, Stoddard JN, Zhao H, Deng W, Kahle E, Panteleeff D, Baeten JM, McCutchan FE, Albert J, Leitner T, Wald A, Corey L, Lingappa JR, Partners in Prevention HSV/HIV Transmission Study Team. Viral linkage in HIV-1 seroconverters and their partners in an HIV-1 prevention clinical trial. *PLoS ONE*. 2011. 6(3): E16986.

Carr JK, Foley BT, Leitner T, Salminen M, Korber B, McCutchan FE. Reference sequences representing the principle genetic diversity of HIV-1 in the Pandemic. 1998. *Human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid sequences*, Los Alamos National Laboratory, Los Alamos, New Mexico.

Carr JK, Salminen MO, Koch C, Gotte D, Artenstein AW, Hegerich PA, St Louis D, Burke DS, McCutchan FE. Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *J Virol*. 1996. 70: 5935 – 5943.

Chevenet F, Jung M, Peeters M, de Oliveira T, Gascuel O. Searching for Virus Phylotypes. *Bioinformatics*. Published online January 17, 2013

Ciccozzi M, Gori C, Boros S, Ruiz-Alvarez MJ, Harxhi A, Dervishi M, Qyra S, Schinaia N, D'Arrigo R, Ceccherini-Silberstein F, Bino S, Perno CF, Rezza G. Molecular diversity of HIV in Albania. *Journal of Infectious Diseases*. 2005. 192(3): 475 – 479.

Clumeck N, Sonnet J, Taelman H, Mascart-Lemone F, de Bruyere D, Vandepierre P, Dasnoy J, Marcelis L, Lamy M, Jonas C, Eyckmans L, Noel H, Vanhaeverbeek M, Butzler JP. Acquired immunodeficiency syndrome in African patients. *New England Journal of Medicine*. 1984. 310: 492 – 497.

Coffee M, Lurie MN, Garnett GP. Modelling the impact of migration on the HIV epidemic in South Africa. *AIDS*. 2007. 21: 343 – 350.

Coffin JM, Hughes SH, Varmus HE. *Retroviruses*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. 1997.

Cullen BR. HIV-1 auxiliary proteins: making connections in a dying cell. *Cell*. 1998. 93: 685 – 92.

Dalai SC, de Oliveira T, Harkins GW, Kassaye SG, Lint J, Manasa J, Johnston E, Katzenstein D. Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe. *AIDS*. 2009. 23(18): 2523 – 2532.

Daniel MD, Letvin NL, King NW, Kannagi M, Sehgal PK, Hunt RD, Kanki PJ, Essex M, Desrosiers RC. Isolation of a T-cell tropic HTLV-III-Like retrovirus from macaques. *Science*. 1985. 228: 1201-1204.

de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, Snoeck J, Janse van Rensburg E, Wensing AMJ, van de Vijver DA, Boucher CA, Camacho R, and Vandamme A-M. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics Applications Note*. 2005. 21: 3797 – 3800.

de Oliveira T, Pillay D, Gifford RJ, for the UK Collaborative Group on HIV Drug Resistance. The HIV-1 Subtype C Epidemic in South America Is Linked to the United Kingdom. *PLoS ONE*. 2010. 5(2): e9311.

de Oliveira T, Pybus OG, Rambaut A, Salemi M, Cassol S, Ciccozzi M, Rezza G, Gattinara GC, D'Árrigo R, Amicosante M, Perrin L, Colizzi V, Perno CF, and the Benghazi Study Group. Molecular Epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature. Brief Communications*. 2006. 444: 836-837.

de Silva E, Ferguson NM, Fraser C. Inferring pandemic growth rates from sequence data. *Journal of Royal Society Interface*. 2012. 9(73): 1797 – 1808.

Decosas J, Kane F, Anarfi JK, Sodji KDR, Wagner HU. Migration and AIDS. *The Lancet*. 1995. 346: 826 – 828.

Delatorre EO and Bello G. Phylodynamics of HIV-1 Subtype C Epidemic in East Africa. *PLoS ONE*. 2012. 7(7): e41904.

Desper R, and Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle." *Journal of Computational Biology*. 2002. 9(5): 687 - 705.

Desrosiers RC, Daniel MD, Li Y. HIV-related Lentiviruses of non-Human Primates. *AIDS Research and Human Retroviruses*. 1989. 5(5): 465 – 473.

Dinh MH, Fahrback KM, Hope TJ. The Role of the Foreskin in Male Circumcision: An Evidence-Based Review. *American Journal of Reproductive Immunology*. 2011. 65: 279 – 283.

Donnelly P, Tavaré S. Coalescents and genealogical structure under neutrality. *Annual Review in Genetics*. 1995. 29: 401 – 421.

Drain PK, Halperin DT, Hughes JP, Klausner JD, Bailey RC. Male circumcision, religion, and infectious diseases: An ecologic analysis of 118 developing countries. *BMC Infectious Diseases*. 2006. 6: 172-182.

Drummond A and Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*. 2007. 7(214): 1 – 8.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian Coalescent inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*. 2005. 22(5): 1185 – 1192.

Edwards AWF, and Cavalli-Sforza LL. Reconstruction of evolution. *Annals of Human Genetics*. 1963. 27: 105 – 106.

Edwards AWF, and Cavalli-Sforza LL. Reconstruction of evolutionary trees. *Phenetic and Phylogenetic Classification*. 1964. ed. Heywood VH and McNeil J. Systematics Association Publ. No.6. London.

Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. 1979. 7(1): 1 – 26.

Engelbrecht S, de Jager J, and van Rensburg EJ. Evaluation of commercially available assays for antibodies to HIV-1 in serum obtained from South African patients infected with HIV-1 subtypes B, C, and D. *Journal of Medical Virology*. 1994. 44(3): 223 – 228.

Engelbrecht S, De Villiers T, Sampson CC, Zur Megede J, Barnett SW, Van Rensburg E. Genetic Analysis of the Complete gag and env Genes of HIV Type 1 Subtype C Primary Isolates from South Africa. *AIDS Research and Human Retroviruses*. 2001. 17(16): 1533 – 1547.

Engelbrecht S, Smith TL, Kasper P, Faatz E, Zeier M, Moodley D, Clay CG, Van Rensburg EJ. HIV Type 1 V3 Domain Serotyping and Genotyping in Gauteng, Mpumalanga, KwaZulu-Natal, and Western Cape Provinces of South Africa. *AIDS Research and Human Retroviruses*. 1999. 15(4): 325 – 328.

Esteves A, Parreira R, Venenno T, Franco M, Piedade J, de Sousa JG, Canas-Ferreira WF. Molecular Epidemiology of HIV Type 1 Infection in Portugal: High Prevalence of Non-B Subtypes. *AIDS Research and Human Retroviruses*. 2002. 18(5): 313 – 325.

Fanning LJ. Retroviruses: Molecular Biology, Genomics, and Pathogenesis. *Clinical Infectious Disease*. 2011. 52:2 280.

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*. 1981. 17: 368 – 376.

Felsenstein J. *Inferring Phylogenies*. 2004. Sinauer Associates: Sunderland, MA, USA.

Felsenstein J. Phylogenies and the comparative method. *American Naturalist*. 1985. 125: 1 – 15.

Fenwick A. The global burden of neglected tropical diseases. *Public Health*. 2012. 126(3): 233 – 236.

Fish MQ, Hewer R, Wallis CL, Venter WDF, Stevens WS, Papathanasopoulos MA. Natural polymorphisms of integrase among HIV type 1-infected South African patients. *AIDS Research and Human Retroviruses*. 2010. 26: 498 – 493.

Fitch WM and Margoliash E. Construction of phylogenetic trees. *Science*. 1967. 155(3760): 279 – 284.

Fontella R, Soares MA, Schrago CG. On the origin of HIV-1 subtype C in South America. *AIDS*. 2008. 22(15): 2001 – 2011.

Friedman-Kien AE. Disseminated Kaposi's sarcoma syndrome in young homosexual men. *Journal of American Academy of Dermatology*. 1981. 5: 468 – 471.

Frøslash SS, Jennum P, Lindboe CF, Wefring KW, Linnestad PJ, Böhmer T. HIV-1 infection in Norwegian family before 1970. *The Lancet*. 1988. 1(8598): 1344 – 1345.

Fu YX. A phylogenetic estimator of effective population size or mutation rate. *Genetics*. 1994. 136: 685 – 692.

Galai N, Kalinkovich A, Burstein R, Vlahov D, Bentwich Z. African HIV-1 subtype C and rate of progression among Ethiopian immigrants in Israel. *The Lancet*. 1997. 349(9046): 180 – 181.

Gallo RC, Salahuddin SZ, Popovic M, Shearer GM, Kaplan M, Haynes BF, Palker TJ, Redfield R, Oleske J, Safai B, et al. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*. 1984. 224(4648): 500 – 503.

García-Calleja JM, Gouws E, Ghys PD. National population based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimation. *Sexually Transmitted Infections*. 2006. 82(3): 64 – 70.

Garrido C, Geretti AM, Zahonero N, Booth C, Strang A, Soriano V, De Mendoza C. Integrase variability and susceptibility to HIV integrase inhibitors: impact of subtypes, antiretroviral experience and duration of HIV infection. *Journal of Antimicrobial Chemotherapy*. 2010. 65(2): 320 – 326.

Garry RF, Witte MH, Gottlieb AA, Elvin-Lewis M, Gottlieb MS, Witte CL, Alexander SS, Cole WR, Drake WL Jr. Documentation of an AIDS Virus Infection in the United States in 1968. *Journal of the American Medical Association*. 1988. 260(1988): 2085 – 2087.

Gascuel, O and Steel M. Neighbor-joining revealed. *Molecular Biology and Evolution*. 2006. 23(11): 1997 – 2000.

Gatanga H, Ibe S, Matsuda M, Yoshida S, Asagi T, Kondo M, Sadamasu K, Masakane A, Mori H, Takata N, Minami R, Tateyama M, Koike T, Itoch T, Imai M, Nagashima M, Gejyo F, Ueda M, Hamaguchi M, Kojima Y, Shirsaka T, Kimura A, Yamamoto M, Fujita J, Oka S, Suiura W. Drug-resistant HIV-1 prevalence in patients newly diagnosed with HIV/AIDS in Japan. *Antiviral Research*. 2007. 75: 75 – 82.

Gelderblom H. Fine Structure of HIV and SIV. pp. IV-37-50 in *HIV Molecular Immunology Database*. 1997. Edited by: Korber B, Brander C, Haynes B, Koup R, Moore J, Walker B. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Gessoni G, Barin P, Valverde S, Giacomini A, Di Natale C, Orlandini E, Arreghini N, De Fusco G, Frigato A, Fezzi M, Antico F, Marchiori G. Biological qualification of blood Units: Considerations about the effects of sample's handling and storage on stability of nucleic acids. *Transfusion and Apheresis Science*. 2004. 30(3): 197 – 203.

Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. The emergence of HIV/AIDS in the Americas and beyond. *Proceedings of the National Academy of Science*. 2007. 104(47): 18566 – 18570.

Glynn JR, Caraël M, Auvert B, Kahindo M, Chege J, Musonda R, Kanoa F, Buvé A, and the Study Group on the Heterogeneity of HIV Epidemics in African Cities. Why do young women have a much higher prevalence of HIV than young men? A study in Kisumu, Kenya and Ndola, Zambia. *AIDS*. 2001. 15(4): 51-60.

Gnanakaran S, Daniels MG, Bhattacharya T, Lapedes AS, Sethi A, Li M, Tang H, Greene K, Gao, H, Haynes, BF, Cohen MS, Shaw GM, Seaman MS, Kumar A, Gao F, Montefiori DC, Korber B. Genetic signatures in the envelope glycoproteins of HIV-1 that associate with broadly neutralizing antibodies. *PLoS Computational Biology*. 2010. 6(10): e1000955.

Gordon M, de Oliveira T, Bishop K, Coovadia HM, Madurai L, Engelbrecht S, Janse van Rensburg E, Mosam A, Smith A, Cassol S. *Journal of Virology*. 2003. 77(4): 2587 – 2599.

Göttlinger HG. HIV-1 Gag: a Molecular Machine Driving Viral Particle Assembly and Release. pp. 2-28 in *HIV Sequence Compendium 2001*. Edited by: Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors JW, Wolinsky S, Korber B. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 02-2877.

Gouws E, Staneyki KA, Lyerla R, Ghys PD. The epidemiology of HIV infection among young people aged 15-24 years in southern Africa. *AIDS*. 2008. 22: 5 – 16.

Graur D and Li WH. *Fundamentals of molecular evolution*. Sinauer Associates, Sunderland, MA, USA. 1999.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*. 2010. 59(3): 307 - 321.

Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: Scientific and public health implications. *Science*. 2000. 287: 607–614.

Hall BG. *Phylogenetic Trees Made Easy: A How-To Manual*. 3de Edition. Sunderland Associates, Inc. Massachusetts, United States of America. 2008

Hall HI, Espinoza L, Benbow N, Hu YW, for the Urban Areas HIV surveillance Workgroup. Epidemiology of HIV Infection in Large Urban Areas in the United States. *PLoS ONE*. 2010. 5(9): e12756.

Halperin DT, Bailey RC. Male circumcision and HIV infection: 10 years and counting. *The Lancet*. 1999. 354(9192): 1813 – 1815.

Halperin DT, Mugurungi OM, Hallett TB, Muchini B, Campbell B, Magure T, Benedikt C, Gregson S. A Surprising Prevention Success: Why Did the HIV Epidemic Decline in Zimbabwe? *PLoS Medicine*. 2011. 8(2): e1000414.

Hasegawa M, Kishino K, and Yano T. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*. 1985. 22: 160 – 174.

Heled J, Drummond AJ. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*. 2008. 8(289): 1 – 15.

Hemelaar J, Gouws E, Ghys PD, Osmanov S, and WHO-UNAIDS Network for HIV Isolation and Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000 - 2007. *AIDS*. 2011. 25: 679 – 689.

Henderson LE, Bowers MA, Sowder RC 2nd, Serabyn SA, Johnson DG, Bess JW Jr, Arthur LO, Bryant DK Fenselau C. Gag proteins of the highly replicative MN strain of human immunodeficiency virus type 1: posttranslational modifications, proteolytic processings, and complete amino acid sequences. *Journal of Virology*. 1992. 66: 1856 – 1865.

Hira SK, Kamanga J, Bhat GJ, Mwale C, Tembo G, Luo N, Perine PL. Perinatal transmission of HIV-1 in Zambia. *British Medical Journal*. 1989. 299(6710): 1250.

Hirsch VM, Olmsted R A, Murphy-Corb M, Purcell RH, Johnson PR. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature*. 1989. 339(6223): 389 – 392.

Holder M and Lewis PO. Phylogeny Estimation: Traditional and Bayesian approaches. *Nature Reviews: Genetics*. 2003. 4: 275 – 284.

Holland J, Ramazanoglu C, Scott S, Sharpe S, Thomson R. Sex, gender and power: young women's sexuality in the shadow of AIDS. *Sociology of Health & Illness*. 1990. 12(3): 336 - 350.

Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Leigh Brown AJ. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proceedings of the National Academy of Science*. 1992. 89: 4835 – 4839.

Huang KH, Goedhals D, Fryer H, van Vuuren C, Katzourakis A, De Oliveira T, Brown H, Cassol S, Seebregts C, McLean A, Klenerman P, Phillips R, Frater J; Bloemfontein-Oxford Collaborative Group. Prevalence of HIV type 1 drug-associated mutations in pre-therapy patients in the Free State, South Africa. *Antiviral Therapy*. 2009. 14(7): 975 – 984.

Huelsenbeck JP and Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 2001. 17(8): 754 - 755.

Hue S, Pillay D, Clewley JP, Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proceedings of the National Academy of Science*. 2004. 102(12): 4425 – 4429.

Huet T, Cheynier R, Meyerhans A, Roelants G, Wain- Hobson S. Genetic organization of a chimpanzee lentivirus related to HIV-1. *Nature*. 1990. 345(6273): 356 – 359.

Hunt GM, Johnson D, Tiemessen C. Characterisation of the Long Terminal Repeat Regions of South African Human Immunodeficiency Virus Type 1 Isolates. *Virus Genes*. 2001. 23(1):27-34.

Hunter E. 1997.gp41, A Multifunctional Protein Involved in HIV Entry and Pathogenesis. pp. III-55-73 in *Human Retroviruses and AIDS 1997*. Edited by: Korber B, Hahn B, Foley B, Mellors JW, Leitner T, Myers G, McCutchan F and Kuiken CL. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Ibe S, Hattori J, Fujisaki S, Shigemi U, Fujisaki S, Shimizu K, Nakamura K, Kazumi T, Yokomaku Y, Mamiya N, Hamaguchi M, Kaneda. Trend of Drug-Resistant HIV Type 1

Emergence among Therapy-Naïve Patients in Nogoya, Japan: An 8-year Surveillance from 1999 to 2006. *AIDS Research and Human Retroviruses*. 2008. 24(1): 313 – 325.

Imamichi H, Koita O, Dabita D, Dao S, Ibrah M, Sogoba D, Dewar DL, Berg SC, Jiang M-K, Parta M, Washington JA, Polis MA, Lane HC, Tounkara A. Identification and Characterization of CRF02_AG, CRF06_cpx, and CRF09_cpx Recombinant Subtypes in Mali, West Africa. *AIDS Research and Human Retroviruses*. 2009. 25(1): 45 – 55.

Ingram M. Russia is on brink of AIDS epidemic. *British Medical Journal*. 1996. 313(7052): 252

Ingstad B. The Cultural Construction of AIDS and Its Consequences for Prevention in Botswana. *Medical Anthropology Quarterly*. 1990. 4(1): 28 – 40.

Iweriebor BC, Bessong PO, Mavhandu LG, Masebe TM, Nwobegahay J, Moyo SR, Mphahlele JM. Genetic Analysis of the Near Full-Length Genome of an HIV Type 1 A1/C Unique Recombinant Form from Northern South Africa. *AIDS Res Hum Retroviruses*. 2011. 27(8): 911 – 915.

Jacobs GB, de Beer C, Fincham JE, Adams V, Dhansay MA, Janse van Rensburg E, Engelbrecht S. Serotyping and Genotyping of HIV-1 Infection in Residents of Khayelitsha, Cape Town, South Africa. *Journal of Medical Virology*. 2006. 78: 1529 – 1536.

Jacobs GB, Laten AD, van Rensburg EJ, Bodem J, Weissbrich B, Rethwilm A, Preiser W, and Engelbrecht S. Phylogenetic Diversity and Low Level Antiretroviral Resistance Mutations in HIV Type 1 Treatment-Naïve Patients from Cape Town, South Africa. *AIDS Research and Human Retroviruses* 2008. 24: 1009 – 1012.

Jacobs GB, Loxton AG, Laten A, Robson B, van Rensburg EJ, Engelbrecht S. Emergence and diversity of different HIV-1 subtypes in South Africa, 2000-2001. *Journal of Medical Virology* 2009. 81: 1852 – 1859.

Janssens W, Buvé A, Nkengasong JN. The puzzle of HIV-1 subtypes in Africa. *AIDS*. 1997. 11: 705 – 712.

Jukes T, and Cantor CR. Evolution of protein molecules. In: *Mammalian protein Metabolism*, edited by Munro HN, pp 21-132. New York, Academic Press (1969).

Jung M, Leye N, Vidal N, Fargette D, Diop H Kane CT, Gascuel O, Peeters M. The Origin and Evolutionary History of HIV-1 Subtype C in Senegal. *PLoS ONE*. 2012. 7(3): e33579.

Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, et al. Chimpanzee reservoirs of pandemic and non-pandemic HIV-1. *Science*. 2006. 313: 523–526.

Kim PS, Malashkevich VN, Chan DC, Chutkowski CT. Crystal structure of the simian immunodeficiency virus (SIV) gp41 core: conserved helical interactions underlie the broad inhibitory activity of gp41 peptides. *Proceedings of the National Academy of Science USA*. 1998. 95(16): 9134–9139.

Kimura M. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press, UK. 1980

Kingman JFC. On the genealogy of large populations. *Journal of Applied Probability*. 1982a. 19: 27 – 43.

Kingman JFC. The coalescent. *Stochastic Processes and their Applications*. 1982b. 13: 235 – 248.

Kitchen A, Miyamoto MM, Mulligan CJ. Utility of DNA viruses for studying human host history: case study of JC virus. *Molecular Phylogenetics and Evolution*. 2008. 46: 673 – 682.

Knopf A and Morris M. Lack of Association Between Concurrency and HIV Infection: An Artefact of Study Design. *Journal of Acquired Immune Deficiency Syndromes*: 2012. 60(1): 20 – 21.

Kohlstaedt LA, Wang J, Rice PA, Friedman JM, and Steitz TA. The structure of HIV-1 reverse transcriptase, p. 223–249. In A. M. Skalka and S. P. Goff (ed.), *Reverse transcriptase*. 1993. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*. 2005. 39: 309 – 338.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Battacharya T. Timing the Ancestor of the HIV-1 Pandemic Strains. *Science*. 2000. 288(5472): 1789 – 1796.

Krebs FC, Hogan TH, Quiterio S, Gartner S, Wigdahl B. Lentiviral LTR-directed Expression, Sequence Variation, and disease Pathogenesis. 2001. Edited by: Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors JW, Wolinsky S, Korber B. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 02-2877.

Kreiss JK, Koech D, Plummer FA, Holmes KK, Lightfoote M, Pior P, Ronald AR, Ndinya-Achola JO, D'costa LJ, Roberts P, et al. AIDS virus infection in Nairobi prostitutes. Spread of the epidemic to East Africa. *New England Journal of Medicine*. 1986. 314: 414 – 418.

Learn GH, Korber BTM, Foley B, Hahn BH, Wolinsky SM, and Mullins JI. Maintaining the integrity of human immunodeficiency virus database. *Journal of Virology*. 1996. 70: 5720 – 5730.

Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, Roques P, Worobey M, Vandamme AM. The Molecular Population Genetics of HIV-1 Group O. *Genetics*. 2004. 167:3 1059 – 1068.

Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, and Vandamme A-M. Tracing the origin and history of the HIV-2 epidemic. *Proceedings of the National Academy of Science*. 2003. 100(11): 6588 – 6952.

Lemey P, Rambaut A, Pybus OG. HIV Evolutionary Dynamics Within and Among Hosts. *AIDS Reviews*. 2006. 8: 125 – 140.

Lemey P, Salemi M, and Vandamme AM. *The Phylogenetic Handbook. A practical approach to phylogenetic analysis and hypothesis testing*. 2nd Edition. 2010. Cambridge University Press. Cambridge, UK.

Lemey P, Suchard M, Rambaut A. reconstructing the initial global spread of human influenza pandemic. *PLOS Currents*. 2009. 2: RRN1031.

Levy JA, Hoffman AD, Kramer SM, Landis JA, Shimabukuro JM, and Oshiro LS. Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. *Science*. 1984. 225: 840 – 842.

Lewis F, Hughes GJ, Rambaut A, Pozniak A, Brown AJL. Episodic Sexual Transmission of HIV Revealed by Molecular Phylodynamics. *PLOS Medicine*. 2008. 5(3): e50.

Li S, Pearl D, and Doss H. Phylogenetic tree reconstruction using Markov chain Monte Carlo. *Journal of American Statistical Association*. 2000. 95: 493 – 508.

Li WH. *Molecular evolution*. Sinauer Associates, Sunderland, Massachusetts (MA), USA. 1997.

Lihana RW, Ssemawanga D, Abimiku A, Ndembi N. Update on HIV-1 Diversity in Africa: A Decade in Review. *AIDS Rev*. 2012. 14: 83 - 100.

Liitsola K, Tashkinova I, Laukkanen T, Korovina G, Smolskaja T, Momot O, Mashkilleysen N, Chaplinskias S, Brummer-Korvenkontio H, Vanhatalo J, Leinikki P, Salminen MO. HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. *AIDS*. 1998. 12(14): 1907 - 1919.

Lodi PJ, Ernst JA, Kuszewski J, Hickman AB, Engelman A, Craigie R, Clore GM, Gronenborn AM. Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry*. 1995. 34(31): 9826–33.

Loimera N, Presslich O, Hollerera E, Pakescha G, Pfersmana V, Wenera E. Monitoring HIV-1 infection prevalence amongst intravenous drug users in Vienna 1986 - 1990. *AIDS Care: Psychological and Socio-Medical Aspects of AIDS/HIV*. 1990. 2(3): 281 - 286.

Loxton AG, Treurnicht F, Laten A, Van Rensburg EJ, and Engelbrecht S. Sequence Analysis of Near Full-Length HIV Type 1 Subtype D Primary Strains Isolated in Cape Town, South Africa, from 1984 to 1986. *AIDS Research and Human Retroviruses*. 2005. 21(5): 410-413.

Lukashov VV, Huismans R, Rakhmanova AG, Lisitsina ZN, Akhtyrskaya NA, Vlasov NN, Melnick OB, Goudsmit J. Circulation of subtype A and gagA/envB recombinant HIV type 1 strains among injecting drug users in St. Petersburg, Russia, correlates with geographical origin of infections. *AIDS Research and Human Retroviruses*. 1999. 15(17): 1577 - 1583.

Lurie M, Harrison A, Wilkinson D, Abdool Karim SS. Circular migration and sexual networking in rural KwaZulu-Natal: implications for the spread of HIV and other sexually transmitted diseases. *Health Transition Review*. 1997. 7(3): 15 – 24.

Lurie M. Migration and AIDS in Southern Africa: a review. *South African Journal of Science*. 2000. 96: 343 - 347.

Lurie MN and Rosenthal S. Concurrent Partnerships as a Driver of the HIV Epidemic in Sub-Saharan Africa? The Evidence is Limited. *AIDS and Behaviour*. 2010. 14(19): 17-24.

Lwembe R, Lihana RW, Ochieng W, Panikulam A, Mongoina CO, Palakudy T, de Koning H, Ishizaki A, Kageyama S, Musoke R, Owens M, Songok EM, Okoth FA, Ichimura H. Changes in the HIV type 1 envelope gene from non-subtype B HIV type 1-infected children in Kenya. *AIDS Research and Human Retroviruses*. 2009. 25(2) 141-7.

Lythgoe KA and Fraser C. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proceedings of the Royal Society B*. 2012. 279: 3367-3375.

Magiorkinis G, Magiorkinis E, Paraskevis D et al. The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. *PLoS Medicine*. 2009. 6: e1000198.

Mah TL and Helperin DT. Concurrent Sexual partnerships and the HIV Epidemics in Africa: Evidence to Move Forward. *AIDS and Behaviour*. 2010. 14(1): 11-16.

Masur H, Michelis MA, Greene JB, Onorato I, Stouwe RA, Holzman RS, Wormser G, Brettman L, Lange M, Murray HW, Cunningham-Rundles S. An Outbreak of community acquired *Pneumocystis carinii* pneumonia: initial manifestation of cellular immune dysfunction. *The New England Journal of Medicine*. 1981. 305: 1431-1438.

Mau B and Newton MA. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *Journal of Computation Graphics and Statistics*. 1997. 6: 122 – 131.

McCormack GP, Glynn JR, Crampin AC, Sibande F, Mulawa D, Bliss L, Broadbent P, Abarca K, Ponnighaus JM, Fine PEM, Clewley JP. Early Evolution of the Human Immunodeficiency Virus Type 1 Subtype C Epidemic in Rural Malawi. *Journal of Virology*. 2002. 76(24) 12890-12899.

Mendelson JM, Carballo M. A policy critique of HIV/AIDS and demobilization. *Conflict, Security, and Development*. 2001. 1(2): 73-92.

Merson MH, O'Malley J, Serwadda D, Apisuk C. The history and challenge of HIV prevention. *The Lancet*. 2008. 372: 475 – 488.

Mills G. AIDS and the South African Military: Timeworn Cliché or Time bomb? Published in *HIV/AIDS: a Threat to the African Renaissance?* 2001. Part of the Konrad Adenauer Foundation.

Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*. 2008. 25: 1459–1471.

Mittal S, Cai Y, Nalam MNL, Bolon DNA, Schiffer CA. Hydrophobic Core Flexibility Modulates Enzyme Activity in HIV-1 Protease. *Journal of American Chemotherapy Society*. 2012. 134: 4163-4168.

MMWR - February 2002. Centers for Disease Control (CDC). Heterosexual Transmission of HIV --- 29 States, 1999 - 2002. 53(06): 125 - 129.

MMWR – December 1982a. Centers for Disease Control (CDC). ‘Epidemiologic Notes and Reports Possible Transfusion-Associated Acquired Immune Deficiency Syndrome, AIDS-California’. *MMWR – Morbidity and Mortality Weekly Report*. 31:48 652-654.

MMWR – December 1982b. Centers for Disease Control (CDC). ‘Unexplained Immunodeficiency and Opportunistic Infections in Infants-New York, New Jersey, California’. *MMWR – Morbidity and Mortality Weekly Report*. 31:49 665-667.

MMWR – July 1982. Centers for Disease Control (CDC). ‘Opportunistic infections and Kaposi’s sarcoma among Haitians in the United States’. *MMWR – Morbidity and Mortality Weekly Report*. 31:26 353-354

MMWR – June 1981. Centers for Disease Control (CDC). ‘Pneumocystis pneumonia – Los Angeles’. *MMWR – Morbidity and Mortality Weekly Report*. 30:21 250-252.

MMWR – September 1982. Centers for Disease Control (CDC). ‘Current Trends Update on Acquired Immune Deficiency Syndrome (AIDS)-United States’. *MMWR – Morbidity and Mortality Weekly Report*. 31:37 507-508.

Mock NB, Duale S, Brown LF, Mathys E, O’Maonaigh HC, Abul-Husn NKL, Elliott S. Conflict and HIV: A framework for risk assessment to prevent in conflict-affected settings in Africa. *Emerging Themes in Epidemiology*. 2004. 1(6): 1-16.

Morris M and Mirjam K. Concurrent partnerships and the spread of HIV. *AIDS*. 1997. 11(5): 641-648.

Morris M, Kurth AE, Hamilton DT, Moody J, Wakefield S, for the Network Modelling Group. Concurrent Partnerships and HIV Prevalence Disparities by Race: Linking Science and Public Health Practice. *American Journal of Public Health*. 2009. 99(6): 1023-1031.

Murphy E, Korber B, Georges-Courbot MC, You B, Pinter A, Cook D, Kieny MP, Georges A, Mathiot C, Barre-Sinoussi F, Girard M. Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the Central African Republic. *AIDS Research and Human Retroviruses*. 1993. 9: 997-1007.

Murtagh F. Complexities of Hierarchic Clustering Algorithms: the state of the art. *Computational Statistics Quarterly*. 1984. 1: 101–113.

Myriam BH, Christophe P, Amine S, Taoufik CB, Zakia A, Jacqueline P, Saida BR, Jacques I. First Molecular Characterization of HIV-1 Tunisian Strains. *Journal of Acquired Immune Deficiency Syndrome*: 2001. 28(1): 94-96.

Naganawa S, Sato S, Nossik D, Takahashi K, Hara T, Tochikubo O, Kitamura K, Honda M, Nakasone T. First Report of CRF03_AB Recombinant HIV Type 1 in Injecting Drug Users in Ukraine. *AIDS Research and Human Retroviruses*. 2002. 18(15): 1145-1149.

Nahmias AJ, Weiss J, Yao X, Lee F, Kodsi R, Schanfield M, Matthews T, Bolognesi D, Durack D, Mothulsky A, Kanki P, Essex M. Evidence for Human Infection with an HTLV III/LAV-like virus in Central Africa, 1959. *The Lancet*. 1986. 1: 1279-1280.

Ndiaye HD, Toure-Kane C, Vidal N, Niama FR, Niang-Diallo PA, Die'ye T, Gaye-Diallo A, Wade AS, Peeters M, Mboup S. Surprisingly High Prevalence of Subtype C and Specific HIV-1 Subtype/CRF Distribution in Men Having Sex With Men in Senegal. *Journal of Acquired Immune Deficiency Syndrome*. 2009. 52: 249 – 252.

Nelson MI, Lemey P, Tan Y, Vincent A, Lam TT-Y, Detmer S, Viboud C, Suchard MA, Rambaut A, Holmes EC, Gramer M. Spatial Dynamics of Human-Origin H1 Influenza A Virus in North American Swine. *PLOS Pathogens*. 2011. 7(6): e1002077.

Neogi U, Bontell I, Shet A, Costa AD, Gupta S, Diwan V, Laishram RS, Eanchu A, Ranga U, Banerjea AC, Sönnnerborg A. Molecular Epidemiology of HIV-1 Subtypes in India: Origin and Evolutionary History of the Predominant Subtype C. *PLoS ONE*. 2012. 7(6): e39819.

Neto MBB, Aguiar CV, Maciel JG, Oliveira BMC, Sevilleja JE, Oriá RB, Brito GAC, Warren CA, Guerrant RL, Lima AAM. Evaluation of HIV protease and nucleoside reverse transcriptase inhibitors on proliferation, necrosis, apoptosis in intestinal epithelial cells and electrolyte and water transport and epithelial barrier function in mice. *BMC Gastroenterology*. 2010. 10: 90-103.

Newman CJ, Fogarty L, Makoae LN, Reavely E. Occupational Segregation, gender essentialism and male primacy as major barriers to equity in HIV care giving: Findings from Lesotho. *International Journal of Equity in Health*. 2011. 10(24): doi:10.1186/1475-9276-10-24.

Nicholson LK, Yamazaki T, Torchia DA, Grzesiek S, Bax A, Stahl SJ, Kaufman JD, Wingfield PT, Lam PYS, Jadhav PK, Hodge CN, Domaille PJ, Chang C-H. Flexibility and function in HIV-1 protease. *Nature Structural Biology*. 1995. 2: 274 – 280.

Nofemela A, Bandawe G, Thebus R, Marais J, Wood N, Hoffmann O, Maboko L, Hoelscher M, Woodman Z, Williamson C. Defining the human immunodeficiency virus type 1 transmission genetic bottleneck in a region with multiple circulating subtypes and recombinant forms. *Virology*. 2011. 415(2) 107-113.

Novitsky V, Wang R, Lagakos S, Essex M. HIV-1 Subtype C Phylodynamics in the Global Epidemic. *Viruses*. 2010. 2: 33 - 54.

Novitsky V, Wang R, Margolin L, Baca J, Kebaabetswe L, Rossenkhan R, Bonney C, Herzig M, Hkwe D, Moyo S, Musonda R, Woldegabriel E, van Widenfelt E, Makhema J, Lagakos S, Essex M. Timing Constraints of in Vivo Gag Mutations during Primary HIV-1 Subtype C Infection. *PLoS ONE*. 2009. 4:11 e7727.

Novitsky V, Wester CW, DeGruttola V, Bussmann H, Gaseitsiwe S, Thomas A, Moyo S, Musonda R, Van Widenfelt E, Marlink RG, Essex M. The Reverse Transcriptase 67N 70R 215Y Genotype Is the Predominant TAM Pathway Associated with Virologic Failure among HIV Type 1C-Infected Adults Treated with ZDV/ddI-Containing HAART in Southern Africa. *AIDS Research and Human Retroviruses*. 2007. 23(7) 868-878.

Novitsky VA, Gaolekwe S, McLane MF, Ndung'u TP, Foley BT, Vannberg F, Marlink R, Essex M. HIV type 1 A/J recombinant with a pronounced pol gene mosaicism. *AIDS Research and Human Retroviruses*. 2000. 16(10) 1015-1020.

Ojo K, and Delaney M. Economic and demographic consequences of AIDS in Namibia: rapid assessment of the costs. *International Journal of Health Planning and Management*. 1997. 12(4): 315-326.

Opgen-Rhein R, Fahrmeir L, Strimmer K. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evolutionary Biology*. 2005. 5: 6.

Palma AC, Araújo F, Duque V, Borges F, Paixão MT, Camacho R, on behalf of the Portuguese SPREAD Network. Molecular epidemiology and prevalence of drug resistance-associated mutations in newly diagnosed HIV-1 patients in Portugal. *Infection, Genetics and Evolution*. 2007. 7: 391 – 398.

Papathanasopoulos MA, Hunt GM, Tiemessen CT. Evolution and diversity of HIV-1 in Africa-a review. *Virus Genes*. 2003. 26(2): 151-163.

Papathanasopoulos MA, Vardas E, Wallis C, Glashoff R, Butto S, Poli, G, Malnati M, Clerici M, Ensoli B. Characterization of HIV Type 1 Genetic Diversity Among South African Participants Enrolled in the AIDS Vaccine Integrated Project (AVIP). *AIDS Research and Human retroviruses* 2010;26:705-709.

Paxton NA. *Plague, War, and Democracy: Political Processes and the spread of HIV/AIDS*. 2010. Harvard university, Department of Government.

Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*. 1996. 271: 1582–1586.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biology*. 2011. 9(3): e1000602.

Phits'ane K. AIDS soars in Lesotho. *South African Politics and Economics Monthly*. 1994. 7(9): 10-11.

Plantier J-C, Dachraoui R, Lemée V, Gueudin M, Borsa-Lebas F, Caron F, Simon F. HIV-1 resistance genotyping on dried serum spots. *AIDS*. 2005. 19(4): 391 - 397.

Plummer FA, Simonsen JN, Cameron DW, et al. Cofactors in male-female sexual transmission of human immunodeficiency virus type 1. *Journal of Infectious Diseases*. 1991. 163: 233-239.

Polanski A, Kimmel M, Chakraborty R. Application of a time- dependent coalescence process for inferring the history of population size changes from DNA sequence data. *Proceedings of the National Academy of Sciences of the United States of America*. 1998. 95: 5456–5461.

Preston BD, Poiesz BJ, Loeb LA. Fidelity of HIV-1 reverse transcriptase. *Science*. 1988. 242: 1169–1171.

Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*. 2000. 155: 1429–1437.

Pybus OG. *Inferring evolutionary and epidemiologic processes from molecular phylogenies*. University of Oxford. Oxford. 2000.

Quinn TC, Overbaugh J. HIV/AIDS in Women: An Expanding Epidemic. *Science*. 2005. 308(5728): 1582-1583.

Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. *Nature Genetics Review*. 2004. 5: 52-61.

Rambaut A, Pybus OG, Nelson M, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. *Nature*. 2008. 453: 615 - 619.

Rannala B and Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*. 1996. 43: 304 – 311.

Ras GJ, Simson IW, Anderson R, Prozeksy OW, Hamersma T: Acquired immunodeficiency syndrome: a report of 2 South African cases. *South African Medical Journal* 1983; 64: 140-142.

Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, Doran ER, Rafalski JA, Whitehorn EA, Baumeister K. Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature*. 1985. 313(6000): 277-284.

Reeve PA. HIV infection in patients admitted to a general hospital in Malawi. *BMJ*. 1989. 298: 1567-1568.

Rhee SY, Kantor R, Katzenstein DA, Camacho R, Morris L, Sirivichayakul S, Jorgensen L, Brigido LF, Schapiro JM, Shafer RW. HIV-1 Pol Mutation Frequency by Subtype and Treatment

Experience: Extension of the HIVseq Program to Seven Non-B Subtypes. *AIDS*. 2006. 20:5 643-651.

Rodrigo AG and Learn GH Jr. *Computational and Evolutionary analysis of HIV molecular sequences*. Kluwer Academic Publishers. Dordrecht, The Netherlands. 2001.

Rodriguez F, Oliver JF, Marin A, and Medina JR. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*. 1990. 142: 485-501

Ronquist F and Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 2003. 19(12): 1572 - 1574.

Rousseau CM, Learn GH, Bhattacharya T, Nickle DC, Heckerman PJR, Walker BD, Brander C, Goulder PJR, Walker BD, Kiepiela P, Korber BT, Mullins JI. Extensive Intrasubtype Recombination in South African Human Immunodeficiency Virus Type 1 Subtype C Infection. *Journal of Virology*. 2007. 81(9): 4492 – 4500.

Rozenbaum W, Coulaud JP, Saimot AG, Klatzmann D, Mayaud C, Carette MF. Multiple opportunistic infections in a male homosexual in France. *The Lancet*. 1982. 1: 572-573.

Rzhetsky A and Nei M. A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution*. 1992. 9: 945-967.

Rzhetsky A and Nei M. Theoretical Foundation of the Minimum-Evolution Method of Phylogenetic Inference. *Molecular Biology and Evolution*. 1993. 10: 1073-1095.

Saad MD, Al-Jaufy A, Graham RR, Nadai Y, Earhart KC, Sanchez JL, Carr JK. HIV Type 1 Strains Common in Europe, Africa, and Asia Cocirculate in Yemen. *AIDS Research and Human Retroviruses*. 2005. 21(7): 644 – 648.

Saitou N, and Nei, M. The Neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*. 1987. 4(4): 406 - 425.

Salemi M and Vandamme A-M. *The Phylogenetic Handbook: A practical approach to DNA and protein phylogeny*. Cambridge University Press. Cambridge, United Kingdom. 2003

Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, Peeters M, and Vandamme A-M. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1

subtypes using a new method to uncover clock-like molecular evolution. *The FASEB Journal*. 2001. 15: 276 - 278.

Santiago ML, Rodenburg CM, Kamenya S, Bibollet-Ruche F, Gao F, Bailes E, Meleth S, Soong SJ, Kilby JM, Moldoveanu Z, et al. SIVcpz in wild chimpanzees. *Science*. 2002. 295: 465.

Santos AF and Soares MA. HIV Genetic Diversity and Drug Resistance: Review. *Viruses*. 2010. 2(2): 503 - 531.

Sarngadharan MG, Popovic M, Bruch L, Schüpbach J, Gallo RC. Antibodies reactive with human T-Lymphotropic retroviruses (HTLV-III) in the serum of patients with AIDS". *Science*. 1984. 224(4648): 506 – 508.

Sawers L and Stillwaggon E. Concurrent sexual partnerships do not explain the HIV epidemics in Africa: a systematic review of the evidence. *Journal of International AIDS Society*. 2010. 13(1): 34.

Schultz AK, Zhang M, Leitner T, Kuiken C, Korber B, Morgenstern B, and Stanke M. A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC Bioinformatics*. 2006. 7: 265 - 280

Schüpbach J, Popovic M, Gilden RV, Gonda MA, Sarngadharan MG, Gallo RC. Serological analysis of a subgroup of human T-Lymphotropic retroviruses (HTLV-III) associated with AIDS. *Science*. 1984. 224(4648): 503 – 505.

Scriba TJ, Treurnicht FK, Zeier M, Engelbrecht S, and van Rensburg EJ. Characterization and Phylogenetic Analysis of South African HIV-1 Subtype C Accessory Genes. *AIDS Research and Human Retroviruses*. 2001. 17(8): 775 - 781.

Serwadda D, Sewankambo NK, Carswell JW, Bayley AC, Tedder RS, Weiss RA, Mugerwa RD, Lwegaba A, Kirya GB, Downing RG, Clayden SA, Dalgleish AG. SLIM DISEASE: A NEW DISEASE IN UGANDA AND ITS ASSOCIATION WITH HTLV-III INFECTION. *The Lancet*. 1985. 2: 849 - 52.

Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Fazadegan H, Gupta P, Rinaldo CR, Learn GH, He XL and Mullins JI. Consistent viral evolutionary dynamics associated with the progression of HIV-1 infection. *Journal of Virology*. 1999. 73: 10489-10502.

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

Sharp PM, Hahn BH. Origins of HIV and the AIDS Pandemic. *Cold Spring Harbour Perspective Medicine*. 2011. 1: 1 - 22.

Sharp PM, Robertson DL, Gao F, Hahn BH. Origins and diversity of human immunodeficiency viruses. 1994. *AIDS* 8: S27 – S42.

Shell R. Halfway to the Holocaust: the Economic, Demographic and Social Implications of the AIDS Pandemic to the Year 2010 in the Southern African Region. Published in *HIV/AIDS: a Threat to the African Renaissance?* 2001. Part of the Konrad Adenauer Foundation.

Shen C, Craigo J, Ding M, Chen Y, Gupta P. Origin and Dynamics of HIV-1 Subtype C Infection in India. *PLoS ONE*. 2011. 6(10): e25956.

Sher R. HIV infection in South Africa, 1982 – 1988 – review. *South African Medical Journal*. 1989. 76: 314 - 318.

Sokal R and Michener C. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*. 1958. 38: 1409 – 1438.

Spang R, Rehmsmeier M, and Stoye J. A Novel Approach to Remote Homology Detection: Jumping Alignments. *Journal of Computational Biology*. 2002. 9: 747 - 760

Spiegel PB, Bennedsen AR, Claass J, Bruns L, Patterson N, Yiweza D, Schilperoord. Prevalence of HIV infection in conflict-affected and displaced people in seven sub-Saharan African countries: a systematic review. *The Lancet*. 2007. 369: 2187 - 2195.

Strimmer K, Pybus OG. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution*. 2001. 18: 2298 – 2305.

Suchard MA, Weiss RE, Sinsheimer JS. Bayesian Selection of Continuous-Time Markov Chain Evolutionary Models. *Molecular Biology and Evolution*. 2001. 18(6): 1001 - 1013.

Swanson P, Devare SG, and Hackett J Jr. Molecular Characterization of 39 HIV-1 Isolates Representing Group M (Subtype A-G) and Group O: Sequence Analysis of gag p24, pol Integrase, and env gp41. *AIDS Research and Human Retroviruses*. 2003. 19: 625 - 629

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol*. 2011. 28(10): 2731-2739.

Tapia N, Franco S, Puig-Basagoiti F, Menendez C, Alonso PL, Mshinda H, Clotet B, Saiz JC, Martinez MA. Influence of human immunodeficiency virus type 1 subtype on mother-to-child transmission. *Journal of General Virology*. 2003. 84(3): 607-613.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG. The CLUSTAL X windows interface: Flexible strategies for multiple-sequence alignment aided by quality analysis tools. *Nucleic Acids Research*. 1997. 25: 4876 - 4882

Thomson MM, de Parga EV, Vazque de Parga E, Vinogradova A, Sierra M, Yakovlev A, Rakhmanova A, Delgado E, Casado E, Munoz G, Carmona M, Vega Y, Perez-Alvarez L, Contreras G, Medrano L, Osmanov S, Najera R. New Insights into the Origin of the HIV Type 1 Subtype A Epidemic in Former Soviet Union's Countries Derived from Sequence Analyses of Preepidemically Transmitted Viruses. *AIDS Research and Human Retroviruses*. 2007. 23(12): 1599 – 1604.

Torbeev VY, Raguraman H, Hamelberg D, Tonelli M, Westler Wm, Perozo E, Kent SBH. Protein conformational dynamics in the mechanism of HIV-1 protease catalysis. *Proceedings of the National Academy of Science*. 2011. 108(52): 20982 - 20987.

Torques K, Bourgeois A, Saragosti S, Vidal N, Mpoudi-Ngolle E, Nzilambi N, Apetrei C, Ekwilanga M, Delaporte E, and Peeters M. High diversity of HIV-1 subtype F strains in Central Africa. *Virology*. 1999. 259: 99 - 109

Travers SAA, Clewley JP, Glynn JR, Fine PEM, Crampin AC, Sibande F, Mulawa D, McInerney JO, McCormack GP. Timing and Reconstruction of the Most Recent Common Ancestor of the Subtype C Clade of Human Immunodeficiency Virus Type 1. *Journal of Virology*. 2004. 78(19): 10501 - 10506.

Tully DC and Wood C. Chronology and evolution of the HIV-1 subtype C epidemic in Ethiopia. *AIDS*. 2010. 19(24): 1577 - 1582.

UNAIDS 2012 - UNAIDS Report on the global AIDS epidemic (2012).

Vallari A, Holzmayer V, Harris B, Yamaguchi J, Ngansop C, Makamche F, Mbanya D, Kaptué L, Ndembe N, Gürtler L, Devare S, Brennan CA. Confirmation of Putative HIV-1 Group P in Cameroon. *Journal of Virology*. 2011. 85(3): 1403 - 1407.

van Harmelen JH, van der Ryst E, Loubser AS, York D, Madurai S, Lyons S, Wood R, Williamson C. A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations. *AIDS Research and Human Retroviruses*. 1999. 15(4): 395 - 398.

van Harmelen JH, Wood R, Lambrick M, Rybicki EP, Williamson AL, and Williamson C. An association between HIV-1 subtypes and mode of transmission in Cape Town, South Africa. *AIDS*. 1997. 11: 81 - 87.

Van Heuverswyn F, Li Y, Bailes E, Neel C, Lafay B, Keele BF, Shaw KS, Takehisa J, Kraus MH, Loul S. Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. *Virology*. 2007. 368: 155 – 171.

Vangroenweghe D. The earliest cases of human immunodeficiency virus type 1 group M in Congo-Kinshasa, Rwanda and Burundi and the origin of acquired immune deficiency syndrome. *Phil Trans of the Royal Society*. 2001. 356: 923 – 925.

Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B, and Delaporte E. Unprecedented Degree of Human Immunodeficiency Virus Type 1 (HIV-1) Group M Genetic Diversity in the Democratic Republic of Congo Suggests that the HIV-1 Pandemic Originated in Central Africa. *Journal of Virology*. 2000. 74: 10498 - 10507

Vilaseca J, Arnau JM, Bacardi R, Mieras C, Serrano A, Navarro C. Kaposi's sarcoma and *Toxoplasma gondii* brain abscess in a Spanish homosexual. *The lancet*. 1982. 1: 572.

Vuylsteke B, Bastos R, Barreto J, Crucitti T, Folgosa E, Mondlane J, Dusauchoit T, Piot P, Laga M. High prevalence of sexually transmitted diseases in a rural area in Mozambique. *Genitourin Medicine*. 1993. 69: 427 - 430.

Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr., Swanstrom R, Burch CL, Weeks KM. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*. 2009. 460(6): 711 - 720.

Weller I, Crawford DH, Iliescu V, MacLennan K, Sutherland S, Tedder RS, and Adler MW. Homosexual men in London: Lymphadenopathy, immune status, and Epstein-Barr virus infection. Edited by Selikoff IJ, Teirstein AS, and Hirschman SZ. *Annals of the New York Academy of Science*. 1984. 437: 248 - 249.

Wensing AMJ, van Maarseveen NM, Nijhuis M. Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Research*. 2010. 85(1): 59 - 74.

Wertheim JO and Worobey M. Dating the Age of the SIV Lineages That Gave Rise to HIV-1 and HIV-2. *PLoS Computational Biology*. 2009. 5(5): 1 - 9.

Wilkinson E and Engelbrecht S. Molecular characterization of non-subtype C and recombinant HIV-1 viruses from Cape Town, South Africa. *Infection, Genetics and Evolution*. 2009. 9(5): 840 - 846.

Williams BG, Gilgen D, Campbell C, Taljaard, MacPhail C. HIV/AIDS in South Africa. A biomedical and social survey in Carletonville. CSIR, Johannesburg, 2000. ISBN 0-620-26235-4.

Williams BG, Lloyd-Smith JO, Gouws E, Hankins C, Getz WM, Hargrove J, de Zooyes I, Dye C, Auvert B. The Potential Impact of Male Circumcision on HIV in sub-Saharan Africa. *PLoS Medicine*. 2006. 3:7. 1032 - 1040.

Williamson C, Engelbrecht S, Lambrick M, Janse van Rensburg E, Wood R, Bredell W, and Williamson A. HIV-1 subtypes in different risk groups in South Africa. *Lancet*. 1995. 346: 782

Wiseman CC. Aliens and AIDS in Southern Africa: The Malawi-South Africa Debate. *African Affairs*. 1998. 97: 53-79.

Wlodawer A, Miller M, Jaskólski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent SB. Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science*. 1989. 245(4918): 616 - 621.

Wollants E, Schoenenberg M, Figueroa C, Shor-Posner G, Klaskala W, Baum MK. Risk factors and patterns of HIV-1 transmission in the El Salvador military during war time. *AIDS*. 1995. 9: 1291-1292.

Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, Muyembe JJ, Kabongo JMM. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*. 2008. 455(7213): 661 – 664.

Worobey M, Santiago ML, Keele BF, Ndjango JB, Joy JB, Labama BL, Dhed AB, Rambaut A, Sharp PM, Shaw GM, et al. Origin of AIDS: Contaminated polio vaccine theory refuted. *Nature*. 2004. 428: 820.

Wyatt R, Kwong PD, Hendrickson WA, Sodroski JG. Structure of the Core of the HIV-1 gp120 Exterior Envelope Glycoprotein. 1998. pp. III-3-9 in *Human Retroviruses and AIDS 1998*. Edited by: Korber B, Kuiken CL, Foley B, Hahn B, McCutchan F, Mellors JW, and Sodroski J. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Yamaguchi K, Groopman JE, Byrn RA. The regulation of HIV by retinoic acid correlates with cellular expression of the retinoic acid receptors. *AIDS*. 1994. 8: 1675 - 1682.

Yang Y. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press. 2008.

Yang Z and Rannala B. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. *Molecular Biology and Evolution*. 1997. 14: 717 - 724.

Yang Z, Goldman N, and Friday A. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Molecular Biology and Evolution*. 1994. 11: 316 – 324.

Zarandia M, Tsertsvadze T, Carr JK, Sanchez JL, Nelson AK. HIV-1 genetic diversity and genotypic drug susceptibility in the Republic of Georgia. *AIDS Research and Human Retroviruses*. 2006. 22(5): 470 – 476.

Zhang M, Schultz AK , Calef C, Kuiken C, Leitner T, Korber B, Morgenstern B, and Stanke M. jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. *Nucleic Acids Research*. 2006. 34: 463 – 465.

Zuckerandl E, and Pauling LB. Evolutionary divergence and convergence in proteins. In Bryson V and Vogel HJ (editors). *Evolving Genes and Proteins*. Academic Press, New York. pp 97 – 116.

Zuckerandl E, and Pauling LB. Molecular disease, evolution, and genetic heterogeneity. 1962. Kasha M and Pullman B (editors). *Horizons in Biochemistry*. Academic Press, New York. pp 189 – 225.

Zwickl DJ and Hillis DM. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*. 2002. 51(4): 588 – 598.

CHAPTER SIX – TABLE OF CONTENT

	Page
APPENDIX A Patient information of Cape Town data sets	223
APPENDIX B Root-to-tip regression analysis of the final <i>gag</i> p24 Cape Town data set	236
APPENDIX C Composition of various Southern Africa data sets for Bayesian inference	237
APPENDIX D Bayes factor model comparisons from Cape Town and Southern African BEAST runs	238
APPENDIX E Convergence in trace files of various BEAST runs	243
APPENDIX F Clustering analyses for putative transmission clusters	245

APPENDIX A

Composition of Cape Town *gag* p24 data setTable 6.1: Composition of the Cape Town *gag* p24 data set

Sample number	Date of Sample	Age	Patients Race	Patients sex	ARV Status	CD4 cell count	Viral Load	Country of Infection	Suspected mode of transmission
R4 368	17/10/1989	No Data	No Data	No Data	Negative	No Data	No Data	South Africa	Heterosexual
R4 369	17/10/1989	No Data	No Data	No Data	Negative	No Data	No Data	South Africa	Heterosexual
R4 370	17/10/1989	No Data	No Data	No Data	Negative	No Data	No Data	South Africa	Heterosexual
R8 597	15/04/1991	27	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R8 864	15/05/1991	25	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R9 684	12/07/1991	23	Mixed Race	Male	Negative	No Data	No Data	South Africa	Homosexual
R11 391	07/10/1991	37	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R11 397	07/10/1991	43	Mixed Race	Male	Negative	No Data	No Data	South Africa	No Data
R11 582	01/11/1991	42	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R11 983	18/11/1991	32	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R11 961	15/11/1991	35	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R11 988	18/11/1991	49	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R12 077	21/11/1991	39	Mixed Race	Female	Negative	No Data	No Data	South Africa	Heterosexual
R13 003	04/02/1992	36	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R14 747	19/06/1992	No Data	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R16 022	31/08/1992	2	Mixed Race	Female	Negative	No Data	No Data	South Africa	Heterosexual
R16 200	23/09/1992	25	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 335	02/10/1992	46	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 469	14/10/1992	31	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 662	29/10/1992	24	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 812	08/11/1992	32	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R16 885	15/11/1992	32	Mixed Race	Female	Negative	No Data	No Data	South Africa	Heterosexual
R16 911	17/11/1992	32	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R17 042	27/11/1992	22	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R17 373	23/12/1992	28	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
DQ866205	13/03/2002	31	African	Female	Negative	382	No Data	South Africa	Heterosexual
DQ866206	12/02/2002	37	African	Female	Negative	208	No Data	South Africa	Heterosexual
DQ866207	13/02/2002	26	African	Male	Negative	239	No Data	South Africa	Heterosexual
DQ866208	13/03/2002	34	Mixed Race	Male	Negative	54	No Data	South Africa	Heterosexual

DQ866209	13/03/2002	45	Indian	Male	Negative	No Data	No Data	Somalia	Heterosexual
DQ866211	13/03/2002	33	African	Female	Negative	509	No Data	South Africa	Heterosexual
DQ866212	13/03/2002	31	African	Female	Negative	49	No Data	Transkei	Heterosexual
DQ866213	13/03/2002	31	African	Male	Negative	14	No Data	Transkei	Heterosexual
DQ866214	13/03/2002	27	African	Female	Negative	57	No Data	Transkei	Heterosexual
DQ866215	13/03/2002	28	African	Female	Negative	315	No Data	South Africa	Heterosexual
DQ866216	13/03/2002	40	African	Female	Positive	463	No Data	South Africa	Heterosexual
DQ866217	13/03/2002	28	African	Female	Negative	265	No Data	South Africa	Heterosexual
DQ866218	14/03/2002	28	African	Female	Positive	104	LDL	South Africa	Heterosexual
DQ866219	18/03/2002	33	African	Female	Negative	596	No Data	South Africa	Heterosexual
DQ866220	18/03/2002	22	African	Female	Negative	292	No Data	South Africa	Heterosexual
DQ866221	19/03/2002	29	Mixed Race	Female	Negative	116	No Data	South Africa	Heterosexual
DQ866222	19/03/2002	29	African	Female	Negative	65	No Data	Transkei	Heterosexual
DQ866223	20/03/2002	56	Mixed Race	Female	Negative	102	No Data	South Africa	Heterosexual
DQ866224	20/03/2002	21	African	Female	Negative	770	No Data	South Africa	Heterosexual
DQ866240	03/04/2002	29	African	Female	Negative	66	No Data	South Africa	Heterosexual
DQ866241	08/04/2002	30	Mixed Race	Female	Negative	183	No Data	South Africa	Heterosexual
DQ866242	08/04/2002	34	African	Female	Negative	389	No Data	South Africa	Heterosexual
DQ866243	08/04/2002	25	African	Female	Negative	9	No Data	South Africa	Heterosexual
DQ866244	09/04/2002	26	African	Male	Negative	53	No Data	South Africa	Heterosexual
DQ866246	08/04/2002	42	Caucasian	Male	Positive	295	No Data	South Africa	Homosexual
DQ866247	03/04/2002	26	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
DQ866248	10/04/2002	33	African	Female	Negative	60	No Data	South Africa	Heterosexual
DQ866249	10/04/2002	31	African	Female	Negative	303	No Data	South Africa	Heterosexual
DQ866250	10/04/2002	38	African	Female	Negative	170	No Data	Transkei	Heterosexual
DQ866251	11/04/2002	28	African	Female	Negative	439	No Data	South Africa	Heterosexual
DQ866252	10/04/2002	22	African	Female	Negative	350	No Data	South Africa	Heterosexual
DQ866253	10/04/2002	40	African	Female	Negative	604	No Data	South Africa	Heterosexual
DQ866254	10/04/2002	49	African	Female	Negative	179	No Data	Transkei	Heterosexual
DQ866255	10/04/2002	45	African	Female	Negative	209	No Data	South Africa	Heterosexual
DQ866256	16/04/2002	19	African	Female	Negative	236	No Data	South Africa	Heterosexual
DQ866257	16/04/2002	44	African	Female	Negative	385	No Data	South Africa	Heterosexual
DQ866258	16/04/2002	26	African	Female	Negative	129	No Data	South Africa	Heterosexual
DQ866259	17/04/2002	40	African	Female	Negative	171	No Data	South Africa	Heterosexual
DQ866260	17/04/2002	47	African	Female	Negative	308	No Data	South Africa	Heterosexual
DQ866261	17/04/2002	27	African	Male	Negative	224	No Data	South Africa	Heterosexual
DQ866262	18/04/2002	29	Mixed Race	Female	Negative	0	No Data	South Africa	Heterosexual
DQ866264	22/04/2002	42	Mixed Race	Male	Negative	459	No Data	South Africa	Homosexual
DQ866265	23/04/2002	24	African	Female	Negative	455	No Data	South Africa	Heterosexual

DQ866266	23/04/2002	31	African	Male	Negative	63	No Data	South Africa	Heterosexual
DQ866267	23/04/2002	37	African	Female	Positive	No Data	No Data	South Africa	Heterosexual
DQ866268	24/04/2002	32	African	Male	Negative	451	No Data	South Africa	Heterosexual
DQ866269	24/04/2002	32	African	Female	Negative	485	No Data	South Africa	Heterosexual
DQ866270	24/04/2002	28	African	Female	Negative	144	No Data	South Africa	Heterosexual
DQ866271	24/04/2002	29	African	Female	Negative	199	No Data	Transkei	Heterosexual
DQ866272	24/04/2002	25	African	Female	Negative	344	No Data	Transkei	Heterosexual
DQ866273	24/04/2002	25	African	Female	Positive	9	1000000	South Africa	Heterosexual
DQ866275	24/04/2002	23	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
DQ866276	26/04/2002	18	Caucasian	Female	Negative	No Data	No Data	South Africa	Heterosexual
DQ866277	06/05/2002	26	African	Female	Negative	282	No Data	South Africa	Heterosexual
DQ866278	07/05/2002	40	African	Male	Negative	287	No Data	South Africa	Heterosexual
DQ866279	07/05/2002	40	African	Male	Negative	80	No Data	South Africa	Heterosexual
DQ866280	08/05/2002	44	Mixed Race	Female	Positive	303	No Data	South Africa	Heterosexual
DQ866281	03/04/2002	30	Mixed Race	Female	Negative	286	No Data	South Africa	Heterosexual
DQ866282	08/05/2002	30	Mixed Race	Female	Negative	101	No Data	South Africa	Heterosexual
DQ866283	08/05/2002	24	African	Female	Negative	530	No Data	South Africa	Heterosexual
DQ866284	13/05/2002	27	African	Female	Negative	233	No Data	South Africa	Heterosexual
DQ866285	14/05/2002	33	African	Male	Negative	584	No Data	South Africa	Heterosexual
DQ866286	14/05/2002	38	Mixed Race	Male	Negative	516	No Data	South Africa	Heterosexual
DQ866287	13/05/2002	29	African	Female	Negative	558	No Data	South Africa	Heterosexual
DQ866288	13/05/2002	25	African	Female	Negative	565	No Data	South Africa	Heterosexual
DQ866289	14/05/2002	26	African	Male	Negative	53	No Data	South Africa	Heterosexual
DQ866290	14/05/2002	35	Mixed Race	Male	Negative	55	1500000	South Africa	Heterosexual
DQ866291	14/05/2002	35	Mixed Race	Male	Negative	795	9600	South Africa	Homosexual
DQ866292	15/05/2002	56	Caucasian	Male	Positive	243	101853	United States of America	Homosexual
DQ866293	15/05/2002	37	Caucasian	Male	Negative	327	38000	South Africa	Homosexual
DQ866294	15/05/2002	40	African	Female	Negative	210	No Data	South Africa	Heterosexual
DQ866295	15/05/2002	36	African	Female	Negative	184	No Data	South Africa	Heterosexual
DQ866296	15/05/2002	40	African	Male	Negative	588	No Data	Transkei	Heterosexual
DQ866297	15/05/2002	46	African	Female	Negative	30	No Data	Ciskei	Heterosexual
DQ866298	15/05/2002	38	African	Female	Negative	376	No Data	Transkei	Heterosexual
DQ866299	16/05/2002	40	Mixed Race	Female	Negative	628	No Data	South Africa	Heterosexual
DQ866300	16/05/2002	30	Caucasian	Male	Negative	188	60000	South Africa	Homosexual
DQ866301	16/05/2002	37	Caucasian	Male	Negative	317	140000	South Africa	Homosexual
DQ866302	16/05/2002	28	African	Female	Negative	474	No Data	South Africa	Heterosexual
DQ866303	16/05/2002	27	African	Female	Negative	594	No Data	South Africa	Heterosexual
DQ866304	16/05/2002	25	African	Female	Negative	521	No Data	South Africa	Heterosexual
DQ866305	16/05/2002	26	African	Female	Negative	178	No Data	South Africa	Heterosexual

DQ866306	16/05/2002	37	African	Female	Negative	438	No Data	South Africa	Heterosexual
DQ866308	16/05/2002	38	African	Female	Negative	184	No Data	South Africa	Heterosexual
DQ866309	16/05/2002	33	African	Female	Negative	575	No Data	Ciskei	Heterosexual
DQ866310	16/05/2002	28	African	Female	Negative	322	No Data	South Africa	Heterosexual
DQ866312	21/05/2002	25	Mixed Race	Female	Negative	383	No Data	South Africa	Heterosexual
DQ866313	21/05/2002	32	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
DQ866314	30/05/2002	28	African	Female	Negative	481	No Data	Ciskei	Heterosexual
JO152	05/02/2009	30	African	Female	Negative	185	No Data	South Africa	Heterosexual
JO161	18/12/2008	32	African	Female	Negative	307	No Data	South Africa	Heterosexual
JO162	12/01/2008	26	African	Female	Negative	361	No Data	South Africa	Heterosexual
JO166	12/01/2008	33	African	Female	Negative	75	No Data	South Africa	Heterosexual
JO167	15/01/2008	35	African	Female	Negative	409	No Data	South Africa	Heterosexual
JO168	19/01/2009	30	African	Female	Negative	335	No Data	South Africa	Heterosexual
JO173	16/02/2009	34	African	Female	Negative	398	No Data	South Africa	Heterosexual
JO176	02/02/2009	26	African	Female	Negative	282	No Data	South Africa	Heterosexual
JO179	09/02/2009	25	African	Female	Negative	143	No Data	South Africa	Heterosexual
JO180	09/03/2009	32	African	Female	Negative	184	No Data	South Africa	Heterosexual
JO183	23/02/2009	33	African	Female	Negative	201	No Data	South Africa	Heterosexual
JO185	26/02/2009	25	African	Female	Negative	318	No Data	South Africa	Heterosexual
JO186	26/02/2009	31	African	Female	Negative	152	No Data	South Africa	Heterosexual
JO189	02/03/2009	32	African	Male	Negative	194	No Data	South Africa	Heterosexual
JO191	09/02/2009	29	African	Male	Negative	112	No Data	South Africa	Heterosexual
JO192	09/03/2009	35	African	Female	Negative	172	No Data	South Africa	Heterosexual
JO193	19/03/2009	26	African	Male	Negative	172	No Data	South Africa	Heterosexual
JO195	02/03/2009	35	African	Male	Negative	146	No Data	South Africa	Heterosexual
JO198	02/03/2009	26	African	Female	Negative	140	No Data	South Africa	Heterosexual
JO200	05/03/2009	34	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
JO201	19/04/2009	30	Caucasian	Male	Negative	No Data	No Data	South Africa	Heterosexual
JO202	09/03/2009	28	African	Female	Negative	107	No Data	South Africa	Heterosexual
JO203	12/03/2009	29	African	Female	Negative	180	No Data	South Africa	Heterosexual
JO204	12/03/2009	34	African	Female	Negative	199	No Data	South Africa	Heterosexual
JO207	19/03/2009	30	African	Female	Negative	192	No Data	South Africa	Heterosexual
JO209	19/03/2009	33	African	Female	Negative	118	No Data	South Africa	Heterosexual
JO210	16/03/2009	33	African	Female	Negative	120	No Data	South Africa	Heterosexual
JO213	23/03/2009	27	African	Male	Negative	164	No Data	South Africa	Heterosexual
JO214	23/03/2009	28	Caucasian	Female	Negative	165	No Data	South Africa	Heterosexual
JO216	30/03/2009	35	African	Male	Negative	107	No Data	South Africa	Heterosexual
JO218	26/03/2009	25	African	Female	Negative	191	No Data	South Africa	Heterosexual
JO219	26/03/2009	36	African	Female	Negative	185	No Data	South Africa	Heterosexual

JO220	26/03/2009	26	African	Male	Negative	196	No Data	South Africa	Heterosexual
JO223	06/04/2009	31	African	Female	Negative	190	No Data	South Africa	Heterosexual
JO224	30/03/2009	35	African	Female	Negative	182	No Data	South Africa	Heterosexual
JO225	02/04/2009	34	African	Female	Negative	196	No Data	South Africa	Heterosexual
JO229	10/04/2009	29	African	Female	Negative	180	No Data	South Africa	Heterosexual
JO229	10/04/2009	32	African	Female	Negative	100	No Data	South Africa	Heterosexual
JO230	10/04/2009	27	African	Female	Negative	190	No Data	South Africa	Heterosexual
JO231	10/04/2009	32	African	Male	Negative	135	No Data	South Africa	Heterosexual
JO232	01/06/2009	30	African	Female	Negative	107	No Data	South Africa	Heterosexual
JO233	10/04/2009	30	Male	Female	Negative	81	No Data	South Africa	Heterosexual
JO235	01/06/2009	30	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
JO236	19/04/2009	32	African	Female	Negative	139	No Data	South Africa	Heterosexual
JO237	23/04/2009	31	African	Female	Negative	184	No Data	South Africa	Heterosexual
JO238	14/05/2009	32	African	Male	Negative	71	No Data	South Africa	Heterosexual
JO239	30/04/2009	32	African	Female	Negative	191	No Data	South Africa	Heterosexual
JO240	11/05/2009	31	African	Female	Negative	177	No Data	South Africa	Heterosexual
JO241	07/05/2009	31	African	Male	Negative	52	No Data	South Africa	Heterosexual
JO244	07/05/2009	28	African	Female	Negative	66	No Data	South Africa	Heterosexual
JO247	18/05/2009	33	African	Female	Negative	184	No Data	South Africa	Heterosexual
JO248	18/05/2009	25	African	Male	Negative	118	No Data	South Africa	Heterosexual
JO252	21/05/2009	25	African	Female	Negative	186	No Data	South Africa	Heterosexual
JO259	01/06/2009	21	African	Male	Negative	182	No Data	South Africa	Heterosexual
AQ123	23/03/2010	29	African	Female	Positive	48	190000	South Africa	Heterosexual
AZ111	22/01/2010	26	African	Female	Negative	490	8000	South Africa	Heterosexual
BD112	21/01/2010	36	African	Female	Negative	508	20000	South Africa	Heterosexual
BL115	27/01/2010	25	African	Female	Negative	586	72000	South Africa	Heterosexual
BS139	08/06/2010	33	African	Female	Negative	231	29000	South Africa	Heterosexual
CM103	11/11/2009	43	African	Female	Negative	218	100000	South Africa	Heterosexual
LN135	11/05/2010	28	African	Female	Positive	46	LDL	South Africa	Heterosexual
MN091	09/09/2009	25	African	Female	Negative	478	22000	South Africa	Heterosexual
MT100	21/10/2009	41	African	Female	Negative	1529	12000	South Africa	Heterosexual
NM082	29/07/2009	28	African	Female	Negative	377	37000	South Africa	Heterosexual
NM090	09/09/2009	33	African	Female	Negative	572	45000	South Africa	Heterosexual
NM114	25/01/2010	48	African	Female	Negative	572	45000	South Africa	Heterosexual
NN087	18/08/2009	27	African	Female	Negative	163	160000	South Africa	Heterosexual
NN117	03/02/2010	29	African	Female	Negative	411	22000	South Africa	Heterosexual
NN138	04/06/2010	42	African	Female	No Data	No Data	No Data	South Africa	Heterosexual
NN140	09/06/2010	28	African	Female	Negative	124	100000	South Africa	Heterosexual
NS092	10/09/2009	26	African	Female	Positive	111	210000	South Africa	Heterosexual

NS118	12/02/2010	25	African	Female	Positive	703	330000	South Africa	Heterosexual
NS121	11/03/2010	34	African	Female	Positive	111	210000	South Africa	Heterosexual
PK116	28/01/2010	35	African	Female	Negative	347	230000	South Africa	Heterosexual
RG084	31/07/2009	35	African	Female	Negative	491	570	South Africa	Heterosexual
TB089	02/09/2009	45	African	Female	Positive	374	1600	South Africa	Heterosexual
TG088	24/08/2009	34	African	Female	Positive	810	270	South Africa	Heterosexual
TM098	01/10/2009	26	African	Female	Negative	690	61000	South Africa	Heterosexual
VN124	24/03/2010	28	African	Female	Positive	228	2400	South Africa	Heterosexual
VN127	13/04/2010	39	African	Female	Positive	46	4800	South Africa	Heterosexual
ZM126	08/04/2010	26	African	Female	Positive	60	3200000	South Africa	Heterosexual
ZN119	25/02/2010	34	African	Female	Positive	333	11000	South Africa	Heterosexual
ZN122	11/03/2010	25	African	Female	Positive	668	350	South Africa	Heterosexual

Composition of Cape Town *gag-pol* concatenated data set**Table 6.2:** Composition of the Cape Town *gag-pol* concatenated data set

Sample number	Date of Sample	Age	Patients Race	Patients sex	ARV status	CD4 cell count	Viral Load	Country of Infection	Suspected mode of transmission
R4 368	17/10/1989	No Data	No Data	No Data	No	No Data	No Data	South Africa	Heterosexual
R4 369	17/10/1989	No Data	No Data	No Data	No	No Data	No Data	South Africa	Heterosexual
R4 370	17/10/1989	No Data	No Data	No Data	No	No Data	No Data	South Africa	Heterosexual
R8 597	15/04/1991	27	African	Female	No	No Data	No Data	South Africa	Heterosexual
R8 864	15/05/1991	25	African	Female	No	No Data	No Data	South Africa	Heterosexual
R9 684	12/07/1991	23	Mixed Race	Male	No	No Data	No Data	South Africa	Homosexual
R11 391	07/10/1991	37	African	Female	No	No Data	No Data	South Africa	Heterosexual
R11 397	07/10/1991	43	Mixed Race	Male	No	No Data	No Data	South Africa	No Data
R11 582	01/11/1991	42	African	Male	No	No Data	No Data	South Africa	Heterosexual
R11 983	18/11/1991	32	African	Male	No	No Data	No Data	South Africa	Heterosexual
R11 961	15/11/1991	35	African	Male	No	No Data	No Data	South Africa	Heterosexual
R11 988	18/11/1991	49	African	Male	No	No Data	No Data	South Africa	Heterosexual
R12 077	21/11/1991	39	Mixed Race	Female	No	No Data	No Data	South Africa	Heterosexual
R13 003	04/02/1992	36	African	Female	No	No Data	No Data	South Africa	Heterosexual
R14 747	19/06/1992	No Data	African	Female	No	No Data	No Data	South Africa	Heterosexual
R16 022	31/08/1992	2	Mixed Race	Female	No	No Data	No Data	South Africa	Heterosexual
R16 335	02/10/1992	46	African	Male	No	No Data	No Data	South Africa	Heterosexual
R16 469	14/10/1992	31	African	Male	No	No Data	No Data	South Africa	Heterosexual
R16 662	29/10/1992	24	African	Male	No	No Data	No Data	South Africa	Heterosexual
R16 885	15/11/1992	32	Mixed Race	Female	No	No Data	No Data	South Africa	Heterosexual
R16 911	17/11/1992	32	African	Male	No	No Data	No Data	South Africa	Heterosexual
R17 042	27/11/1992	22	African	Female	No	No Data	No Data	South Africa	Heterosexual
R17 373	23/12/1992	28	African	Female	No	No Data	No Data	South Africa	Heterosexual
JO191	09/02/2009	29	African	Male	Negative	112	No Data	South Africa	Heterosexual
JO192	09/03/2009	35	African	Female	Negative	172	No Data	South Africa	Heterosexual
JO193	19/03/2009	26	African	Male	Negative	172	No Data	South Africa	Heterosexual
JO195	02/03/2009	35	African	Male	Negative	146	No Data	South Africa	Heterosexual
JO198	02/03/2009	26	African	Female	Negative	140	No Data	South Africa	Heterosexual
JO200	05/03/2009	34	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
JO204	12/03/2009	34	African	Female	Negative	199	No Data	South Africa	Heterosexual
JO210	16/03/2009	33	African	Female	Negative	120	No Data	South Africa	Heterosexual

JO213	23/03/2009	27	African	Male	Negative	164	No Data	South Africa	Heterosexual
JO214	23/03/2009	28	Caucasian	Female	Negative	165	No Data	South Africa	Heterosexual
JO219	26/03/2009	36	African	Female	Negative	185	No Data	South Africa	Heterosexual
JO220	26/03/2009	26	African	Male	Negative	196	No Data	South Africa	Heterosexual
JO223	06/04/2009	31	African	Female	Negative	190	No Data	South Africa	Heterosexual
JO225	02/04/2009	34	African	Female	Negative	196	No Data	South Africa	Heterosexual
JO230	10/04/2009	27	African	Female	Negative	190	No Data	South Africa	Heterosexual
JO231	10/04/2009	32	African	Male	Negative	135	No Data	South Africa	Heterosexual
JO232	01/06/2009	30	African	Female	Negative	107	No Data	South Africa	Heterosexual
JO233	10/04/2009	30	African	Female	Negative	81	No Data	South Africa	Heterosexual
JO235	01/06/2009	30	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
JO236	19/04/2009	32	African	Female	Negative	139	No Data	South Africa	Heterosexual
JO237	23/04/2009	31	African	Female	Negative	184	No Data	South Africa	Heterosexual
JO238	14/05/2009	32	African	Male	Negative	71	No Data	South Africa	Heterosexual
JO239	30/04/2009	32	African	Female	Negative	191	No Data	South Africa	Heterosexual
JO240	11/05/2009	31	African	Female	Negative	177	No Data	South Africa	Heterosexual
JO241	07/05/2009	31	African	Male	Negative	52	No Data	South Africa	Heterosexual
JO244	07/05/2009	28	African	Female	Negative	66	No Data	South Africa	Heterosexual
JO247	18/05/2009	33	African	Female	Negative	184	No Data	South Africa	Heterosexual
JO248	18/05/2009	25	African	Male	Negative	118	No Data	South Africa	Heterosexual
JO252	21/05/2009	25	African	Female	Negative	186	No Data	South Africa	Heterosexual

Composition of Cape Town *pol* data set**Table 6.3:** Composition of the Cape Town *pol* data set

Sample number	Date of Sample	Age	Patients Race	Patients sex	ARV status	CD4 cell count	Viral Load	Country of Infection	Suspected mode of transmission
R4 368	17/10/1989	No Data	No Data	No Data	Negative	No Data	No Data	South Africa	Heterosexual
R4 369	17/10/1989	No Data	No Data	No Data	Negative	No Data	No Data	South Africa	Heterosexual
R4 370	17/10/1989	No Data	No Data	No Data	Negative	No Data	No Data	South Africa	Heterosexual
R4 714	01/02/1990	38	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R4 794	14/02/1990	50	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R4 846	27/02/1990	38	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R5 785	04/09/1990	50	Mixed Race	Male	Negative	No Data	No Data	South Africa	Heterosexual
R6 191	23/10/1990	35	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R6 201	25/10/1990	32	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R6 742	14/12/1990	55	Mixed Race	Female	Negative	No Data	No Data	South Africa	Heterosexual
R6 978	11/01/1991	32	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R7 148	27/01/1991	16	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R7 663	20/02/1991	23	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R7 788	27/02/1991	61	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R8 597	15/04/1991	27	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R8 864	15/05/1991	25	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R9 684	12/07/1991	23	Mixed Race	Male	Negative	No Data	No Data	South Africa	Homosexual
R11 391	07/10/1991	37	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R11 397	07/10/1991	43	Mixed Race	Male	Negative	No Data	No Data	South Africa	No Data
R11 582	01/11/1991	42	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R11 983	18/11/1991	32	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R11 961	15/11/1991	35	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R11 988	18/11/1991	49	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R12 077	21/11/1991	59	Mixed Race	Female	Negative	No Data	No Data	South Africa	Heterosexual
R13 003	04/02/1992	36	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R13 800	31/03/1992	38	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R14 747	19/06/1992	No Data	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R15 124	15/07/1992	37	Mixed Race	Male	Negative	No Data	No Data	South Africa	Heterosexual
R15 682	21/01/1992	18	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R15 791	28/08/1992	30	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 022	31/08/1992	18	Mixed Race	Female	Negative	No Data	No Data	South Africa	Heterosexual
R16 087	16/09/1992	52	African	Female	Negative	No Data	No Data	South Africa	Heterosexual

R16 166	21/09/1992	39	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 200	23/09/1992	25	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 314	28/09/1992	36	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 335	02/10/1992	46	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 469	14/10/1992	31	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 510	16/10/1992	33	Mixed Race	Female	Negative	No Data	No Data	South Africa	Heterosexual
R16 575	22/10/1992	28	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R16 662	29/10/1992	24	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R16 812	08/11/1992	32	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R16 885	15/11/1992	32	Mixed Race	Female	Negative	No Data	No Data	South Africa	Heterosexual
R16 911	17/11/1992	32	African	Male	Negative	No Data	No Data	South Africa	Heterosexual
R17 042	27/11/1992	22	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
R17 373	23/12/1992	28	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
TV21	10/02/2000	23	African	Female	Negative	57	No Data	South Africa	Heterosexual
TV22	15/02/2000	31	African	Female	Negative	198	No Data	South Africa	Heterosexual
TV27	22/02/2000	40	Mixed Race	Female	Negative	299	No Data	South Africa	Heterosexual
TV28	17/02/2000	47	Mixed Race	Male	Negative	92	No Data	South Africa	Heterosexual
TV30	03/03/2000	46	Mixed Race	Female	Negative	13	No Data	South Africa	Heterosexual
TV40	18/03/2000	25	African	Female	Negative	251	No Data	South Africa	Heterosexual
TV45	03/07/2000	27	African	Female	Negative	388	No Data	South Africa	Heterosexual
TV47	08/07/2000	29	Mixed Race	Male	Negative	265	29800	South Africa	Heterosexual
TV50	21/07/2000	46	Mixed Race	Male	Negative	348	No Data	South Africa	Heterosexual
TV51	21/07/2000	32	African	Male	Negative	170	No Data	South Africa	Heterosexual
TV54	21/07/2000	29	African	Male	Negative	276	No Data	South Africa	Heterosexual
TV55	21/07/2000	26	African	Female	Negative	66	No Data	South Africa	Heterosexual
TV57	21/07/2000	27	African	Female	Negative	519	No Data	South Africa	Heterosexual
TV58	21/07/2000	25	African	Male	Negative	275	No Data	South Africa	Heterosexual
TV59	22/07/2000	24	Mixed Race	Female	Negative	No Data	No Data	South Africa	Unknown
TV64	22/07/2000	22	Mixed Race	Female	Negative	500	No Data	South Africa	Unknown
TV66	22/07/2000	36	Mixed Race	Female	Negative	No Data	No Data	South Africa	Heterosexual
TV67	25/07/2000	37	African	Female	Negative	329	No Data	South Africa	Heterosexual
TV69	25/07/2000	52	African	Male	Negative	156	No Data	South Africa	Heterosexual
TV71	25/08/2000	17	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
TV72	25/08/2000	24	African	Female	Negative	416	No Data	South Africa	Heterosexual
TV76	25/08/2000	30	African	Female	Negative	300	No Data	South Africa	Heterosexual
TV77	28/08/2000	43	Mixed Race	Male	Negative	81	No Data	South Africa	Heterosexual
TV86	31/08/2000	35	African	Male	Negative	207	No Data	Zimbabwe/Botswana	Heterosexual
TV87	31/08/2000	43	African	Female	Negative	156	No Data	South Africa	Heterosexual
TV88	31/08/2000	42	Caucasian	Male	Negative	143	No Data	South Africa	Heterosexual

TV89	31/08/2000	34	African	Female	Negative	160	No Data	South Africa	Heterosexual
TV155	02/10/2000	31	African	Male	Negative	400	No Data	South Africa	Heterosexual
TV161	04/10/2000	No Data	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
TV1748	23/02/2004	28	African	Female	Negative	359	No Data	South Africa	Heterosexual
TV1753	23/02/2004	39	African	Female	Negative	338	No Data	South Africa	Heterosexual
TV1755	25/02/2004	22	African	Female	Negative	507	No Data	South Africa	Heterosexual
TV1757	25/02/2004	40	African	Female	Negative	337	No Data	South Africa	Heterosexual
TV1758	25/02/2004	28	African	Female	Negative	614	No Data	South Africa	Heterosexual
TV1761	25/02/2004	41	Mixed Race	Male	Negative	328	No Data	South Africa	Heterosexual
TV1762	25/02/2004	35	African	Female	Negative	272	No Data	South Africa	Heterosexual
TV1763	25/02/2004	29	African	Female	Negative	618	No Data	South Africa	Heterosexual
TV1767	08/03/2004	30	African	Female	Negative	181	No Data	South Africa	Heterosexual
TV1769	08/03/2004	28	African	Female	Negative	258	No Data	South Africa	Heterosexual
TV1770	08/03/2004	23	African	Male	Negative	559	No Data	South Africa	Heterosexual
TV1771	08/03/2004	31	African	Female	Negative	409	No Data	South Africa	Heterosexual
TV1774	10/03/2004	41	African	Male	Negative	137	No Data	South Africa	Heterosexual
TV1775	10/03/2004	40	African	Female	Negative	65	No Data	South Africa	Heterosexual
TV1777	10/03/2004	42	African	Male	Negative	142	No Data	South Africa	Heterosexual
TV1778	15/03/2004	25	Mixed Race	Male	Negative	458	No Data	South Africa	Homosexual
TV1779	15/03/2004	24	African	Female	Negative	130	No Data	South Africa	Heterosexual
TV1781	17/03/2004	32	African	Male	Negative	899	No Data	South Africa	Heterosexual
TV1785	17/03/2004	38	African	Male	Negative	501	No Data	South Africa	Heterosexual
TV1786	17/03/2004	24	African	Female	Negative	173	No Data	South Africa	Heterosexual
TV1787	17/03/2004	19	African	Female	Negative	280	No Data	South Africa	Heterosexual
TV1788	17/03/2004	25	African	Female	Negative	355	No Data	South Africa	Heterosexual
TV1790	31/03/2004	23	African	Female	Negative	101	No Data	South Africa	Heterosexual
TV1792	31/03/2004	43	African	Female	Negative	6	No Data	South Africa	Heterosexual
TV1793	31/03/2004	42	Mixed Race	Female	Negative	124	No Data	South Africa	Heterosexual
TV1795	05/04/2004	32	African	Male	Negative	178	No Data	South Africa	Heterosexual
TV1796	05/04/2004	25	Mixed Race	Female	Negative	361	No Data	South Africa	Heterosexual
TV1798	05/04/2004	31	African	Female	Negative	358	No Data	South Africa	Heterosexual
TV1799	05/04/2004	38	Caucasian	Male	Negative	109	No Data	South Africa	Heterosexual
TV1804	21/04/2004	25	African	Female	Negative	4	No Data	South Africa	Heterosexual
TV1805	21/04/2004	30	African	Female	Negative	363	No Data	South Africa	Heterosexual
TV1806	21/04/2004	39	African	Female	Negative	133	No Data	South Africa	Heterosexual
JO152	05/02/2009	30	African	Female	Negative	185	No Data	South Africa	Heterosexual
JO166	12/01/2008	33	African	Female	Negative	75	No Data	South Africa	Heterosexual
JO168	19/01/2009	30	African	Female	Negative	335	No Data	South Africa	Heterosexual
JO173	16/02/2009	34	African	Female	Negative	398	No Data	South Africa	Heterosexual

JO176	02/02/2009	26	African	Female	Negative	282	No Data	South Africa	Heterosexual
JO179	09/02/2009	25	African	Female	Negative	143	No Data	South Africa	Heterosexual
JO180	09/03/2009	32	African	Female	Negative	184	No Data	South Africa	Heterosexual
JO183	23/02/2009	33	African	Female	Negative	201	No Data	South Africa	Heterosexual
JO189	02/03/2009	32	African	Male	Negative	194	No Data	South Africa	Heterosexual
JO191	09/02/2009	29	African	Male	Negative	112	No Data	South Africa	Heterosexual
JO192	09/03/2009	35	African	Female	Negative	172	No Data	South Africa	Heterosexual
JO193	19/03/2009	26	African	Male	Negative	172	No Data	South Africa	Heterosexual
JO195	02/03/2009	35	African	Male	Negative	146	No Data	South Africa	Heterosexual
JO198	02/03/2009	26	African	Female	Negative	140	No Data	South Africa	Heterosexual
JO202	09/03/2009	28	African	Female	Negative	107	No Data	South Africa	Heterosexual
JO204	12/03/2009	34	African	Female	Negative	199	No Data	South Africa	Heterosexual
JO210	16/03/2009	33	African	Female	Negative	120	No Data	South Africa	Heterosexual
JO213	23/03/2009	27	African	Male	Negative	164	No Data	South Africa	Heterosexual
JO214	23/03/2009	28	Caucasian	Female	Negative	165	No Data	South Africa	Heterosexual
JO219	26/03/2009	36	African	Female	Negative	185	No Data	South Africa	Heterosexual
JO220	26/03/2009	26	African	Male	Negative	196	No Data	South Africa	Heterosexual
JO223	06/04/2009	31	African	Female	Negative	190	No Data	South Africa	Heterosexual
JO225	02/04/2009	34	African	Female	Negative	196	No Data	South Africa	Heterosexual
JO230	10/04/2009	27	African	Female	Negative	190	No Data	South Africa	Heterosexual
JO231	10/04/2009	32	African	Male	Negative	135	No Data	South Africa	Heterosexual
JO232	01/06/2009	30	African	Female	Negative	107	No Data	South Africa	Heterosexual
JO233	10/04/2009	30	African	Female	Negative	81	No Data	South Africa	Heterosexual
JO234	30/04/2009	30	African	Male	Negative	81	No Data	South Africa	Heterosexual
JO235	01/06/2009	30	African	Female	Negative	No Data	No Data	South Africa	Heterosexual
JO236	19/04/2009	32	African	Female	Negative	139	No Data	South Africa	Heterosexual
JO237	23/04/2009	31	African	Female	Negative	184	No Data	South Africa	Heterosexual
JO238	14/05/2009	32	African	Male	Negative	71	No Data	South Africa	Heterosexual
JO239	30/04/2009	32	African	Female	Negative	191	No Data	South Africa	Heterosexual
JO240	11/05/2009	31	African	Female	Negative	177	No Data	South Africa	Heterosexual
JO241	07/05/2009	31	African	Male	Negative	52	No Data	South Africa	Heterosexual
JO244	07/05/2009	28	African	Female	Negative	66	No Data	South Africa	Heterosexual
JO247	18/05/2009	33	African	Female	Negative	184	No Data	South Africa	Heterosexual
JO248	18/05/2009	25	African	Male	Negative	118	No Data	South Africa	Heterosexual
JO252	21/05/2009	25	African	Female	Negative	186	No Data	South Africa	Heterosexual
JO259	01/06/2009	21	African	Male	Negative	182	No Data	South Africa	Heterosexual
CD007	2008/09/02	35	African	Female	No	No Data	6200	South Africa	Heterosexual
CS006	2008/08/26	33	African	Female	No	165	1800000	South Africa	Heterosexual
EF031	2008/12/03	34	Mixed Race	Female	No	268	1800	South Africa	Heterosexual

MD045	2009/03/05	27	African	Female	Yes	845	7400	South Africa	Heterosexual
MH020	2008/10/29	23	African	Female	No Data	257	98000	South Africa	Heterosexual
MN019	2008/10/10	24	African	Female	No	456	6500	South Africa	Heterosexual
ND021	2008/10/30	27	African	Female	No	360	1900	South Africa	Heterosexual
NJ039	2009/02/05	25	African	Female	Yes	416	2500	South Africa	Heterosexual
PM004	2008/07/23	27	African	Female	No	No Data	2800	South Africa	Heterosexual
STO18	2008/10/09	21	African	Female	No	628	7900	South Africa	Heterosexual
TB037	2009/01/30	29	African	Female	Yes	374	1600	South Africa	Heterosexual
TG005	2008/07/05	23	African	Female	No	810	4000	South Africa	Heterosexual
TM040	2009/02/17	34	African	Female	Yes	519	270	South Africa	Heterosexual
WJ027	2008/11/21	30	African	Female	No	320	52000	South Africa	Heterosexual

APPENDIX B

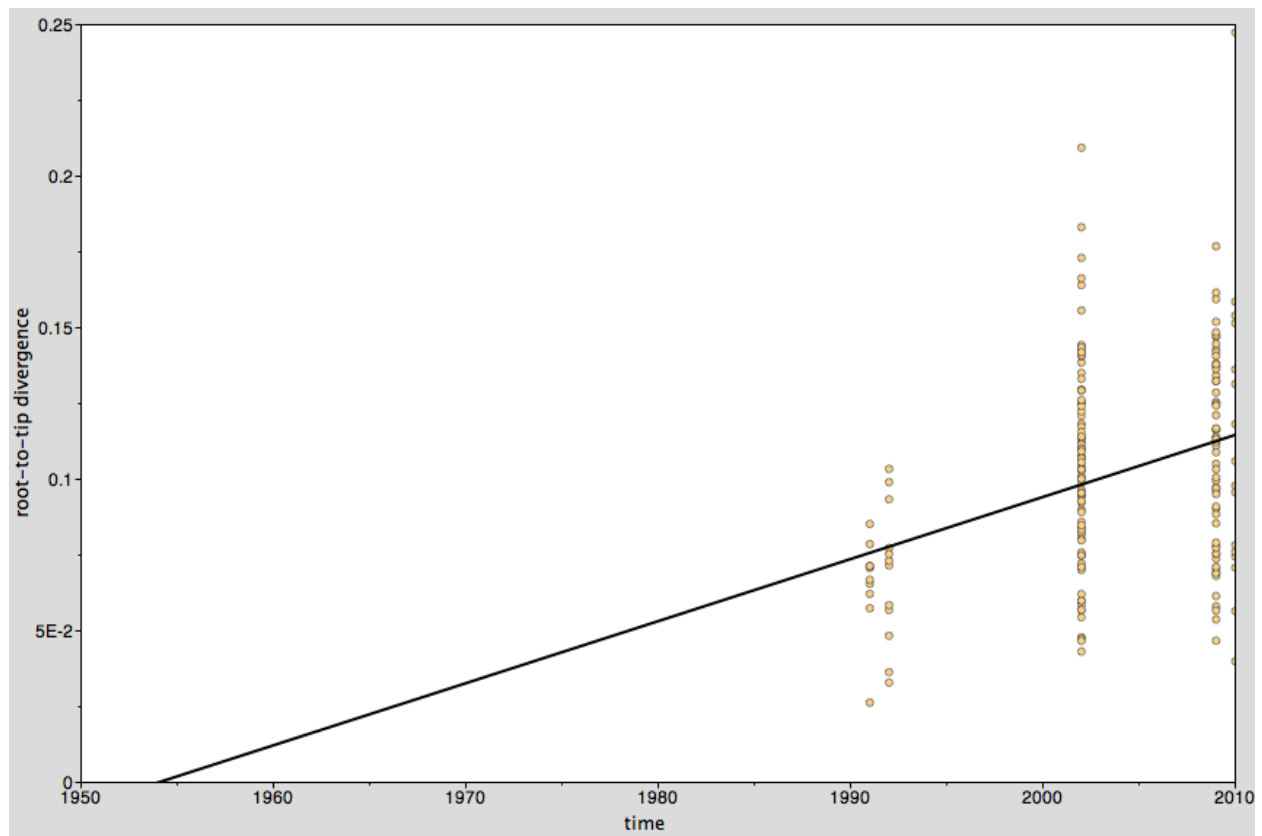


Figure 6.1: A diagrammatical representation of a molecular clock analysis that was performed in Path-O-Gen on the final *gag* p24 Cape Town data set. On the vertical axis the degree of genetic diversity is represented, while time is represented on the horizontal axis. Please note the increasing degree of genetic diversity over time. This tree contains a total of 193 *gag* p24 sequences from the final Cape Town data set.

APPENDIX C

Table 6.4: Total number of taxa that were used in each of the various data sets for the Bayesian inference of the Southern African HIV-1 subtype C epidemic.

Country or region	<i>gag</i> p24 data sets		concatenated <i>gag-pol</i> data sets		<i>pol</i> data sets	
	Southern Africa (with Cape Town)	Southern Africa (no Cape Town)	Southern Africa (with Cape Town)	Southern Africa (no Cape Town)	Southern Africa (with Cape Town)	Southern Africa (no Cape Town)
Botswana	83	83	49	49	65	65
Cape Town	192	-	52	-	165	-
Malawi	11	11	-	-	-	-
South Africa	94	94	61	61	75	75
Swaziland	-	-	-	-	21	21
Zambia	118	118	23	23	27	27
Zimbabwe	9	9	9	9	31	31
Total	507	315	194	142	384	219

APPENDIX D

Table 6.5: Results of Bayes factor comparison for the Cape Town *gag* p24 data set.

BEAST run under various parameters	Log score
Const.relax.est.2	-10106,57
Const.relax.est.1	-10106,58
Const.strict.est.1	-10112,66
BSP.relax.est.1	-10113,86
Const.strict.est.2	-10118,42
BSP.relax.est.2	-10119,28
BSP.strict.est.1	-10132,79
BSP.strict.est.2	-10140,81
BSP.relax.fix.2	-10162,42
BSP.relax.fix.1	-10162,97
BSP.strict.fix.2	-10169,40
BSP.strict.fix.1	-10174,14

BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

Table 6.6: Results of Bayes factor comparison for the Cape Town concatenated *gag-pol* data set.

BEAST run under various parameters	Log score
BSP.relax.est.1	-11000,31
BSP.relax.est.2	-11001,42
Const.relax.fix.2	-11002,94
Const.relax.est.1	-11003,43
Const.relax.est.2	-11003,44
Const.relax.fix.1	-11003,45
BSP.strict.est.2	-11004,31
BSP.strict.est.1	-11004,76
BSP.relax.fix.2	-11004,88
BSP.relax.fix.1	-11005,06
BSP.strict.fix.2	-11030,33
BSP.strict.fix.1	-11030,44

BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

Table 6.7: Results of Bayes factor comparison for the Cape Town *pol* data set.

BEAST run under various parameters	Log score
Const.relax.est.1	-17899,39
Const.relax.fix.2	-17900,34
Const.relax.fix.1	-17902,92
Const.relax.est.2	-17906,52
BSP.relax.fix.2	-17908,80
BSP.relax.fix.1	-17910,40
BSP.strict.est.2	-17982,04
BSP.strict.est.1	-17983,04
BSP.strict.fix.2	-17994,45
Const.strict.est.2	-17995,79
BSP.strict.fix.1	-17995,96
Const.strict.est.1	-17996,37
Const.strict.fix.2	-18005,56
Const.strict.fix.1	-18007,67

BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

Table 6.8: Results of Bayes factor comparison for the Southern African *gag* p24 data set, excluding sequence data from the original Cape Town data set.

BEAST run under various parameters	Log score
BSP.relax.est.2	-15867,59
BSP.relax.est.1	-15893,83
BSP.strict.est.2	-15914,43
Const.relax.est.2	-15917,64
BSP.strict.fix.1	-15961,20
Const.strict.est.2	-15972,24
Const.relax.est.1	-15980,43
Const.strict.est.1	-15986,26
BSP.strict.fix.2	-15995,92
BSP.strict.est.1	-16002,48
Const.strict.fix.1	-16014,73
Const.strict.fix.2	-16036,77
Const.relax.fix.1	-16264,59
Const.relax.fix.2	-16660,77

BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

Table 6.9: Results of Bayes factor comparison for the concatenated Southern African *gag-pol* data set, excluding sequence data from the original Cape Town data set.

BEAST run under various parameters	Log score
Const.relax.est.2	-82001,81
Const.relax.fix.1	-82048,03
Const.relax.est.1	-82081,04
BSP.BSP.fix.2	-82090,52
BSP.strict.est.2	-82106,71
BSP.relax.est.2	-82118,85
BSP.relax.est.1	-82118,95
BSP.BSP.fix.1	-82186,84
BSP.strict.fix.1	-82198,58
BSP.strict.est.1	-82254,85
BSP.strict.fix.2	-82272,28
Const.relax.fix.2	-87378,01

BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

Table 6.10: Results of Bayes factor comparison for the Southern African *pol* data set, excluding sequence data from the original Cape Town data set.

BEAST run under various parameters	Log score
BSP.relax.est.1	-16156,83
Const.strict.est.2	-16159,41
BSP.strict.est.2	-16160,10
BSP.relax.est.2	-16161,34
Const.relax.est.2	-16210,24
Const.relax.est.1	-16210,63
BSP.strict.est.1	-16237,70
Const.strict.est.1	-16235,10

BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

Table 6.11: Results of Bayes factor comparison for the Southern African *gag* p24 data set, including sequence data from the original Cape Town data set.

BEAST run under various parameters	Log score
BSP.relax.est.1	-24695,23
BSP.relax.fix.2	-24739,66
Const.strict.est.2	-24767,78
Const.relax.est.2	-25773,08
BSP.relax.est.2	-24792,08
Const.strict.est.1	-24798,48
BSP.strict.est.2	-24839,59
BSP.relax.fix.1	-24842,01
BSP.strict.est.1	-24857,88
Const.relax.est.1	-24902,98

BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

Table 6.12: Results of Bayes factor comparison for the concatenated Southern African *gag-pol* data set, including sequence data from the original Cape Town data set.

BEAST run under various parameters	Log score
BSP.relax.fix.1	-75728,94
BSP.strict.est.2	-75736,75
BSP.relax.fix.2	-75770,62
BSP.relax.est.2	-75795,03
Const.relax.est.2	-76018,20
BSP.relax.est.1	-76025,80
Const.relax.fix.2	-76128,81
Const.relax.fix.1	-76681,66
Const.relax.est.1	-77404,68
BSP.strict.est.1	-77899,02

BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

Table 6.13: Results of Bayes factor comparison for the Southern African *pol* data set, including sequence data from the original Cape Town data set.

BEAST run under various parameters	Log score
Const.relax.est.1	-25808,06
BSP.relax.est.2	-25823,75
BSP.relax.est.1	-25849,14
Const.relax.est.2	-25850,92
BSP.relax.fix.1	-25858,88
BSP.relax.fix.2	-25902,23
BSP.strict.est.2	-25915,47
BSP.strict.fix.2	-25923,64
BSP.strict.est.1	-25932,28
BSP.strict.fix.1	-25935,53
Const.strict.est.1	-25975,99
Const.strict.est.2	-25999,67

BSP – Bayesian Skyline Plot tree prior, relax – Relaxed Molecular Clock assumption; fix – Fixed mutation rate; est – Estimated mutation rate; strict – Strict Molecular Clock assumption; Const – Constant Population Size tree prior

APPENDIX E

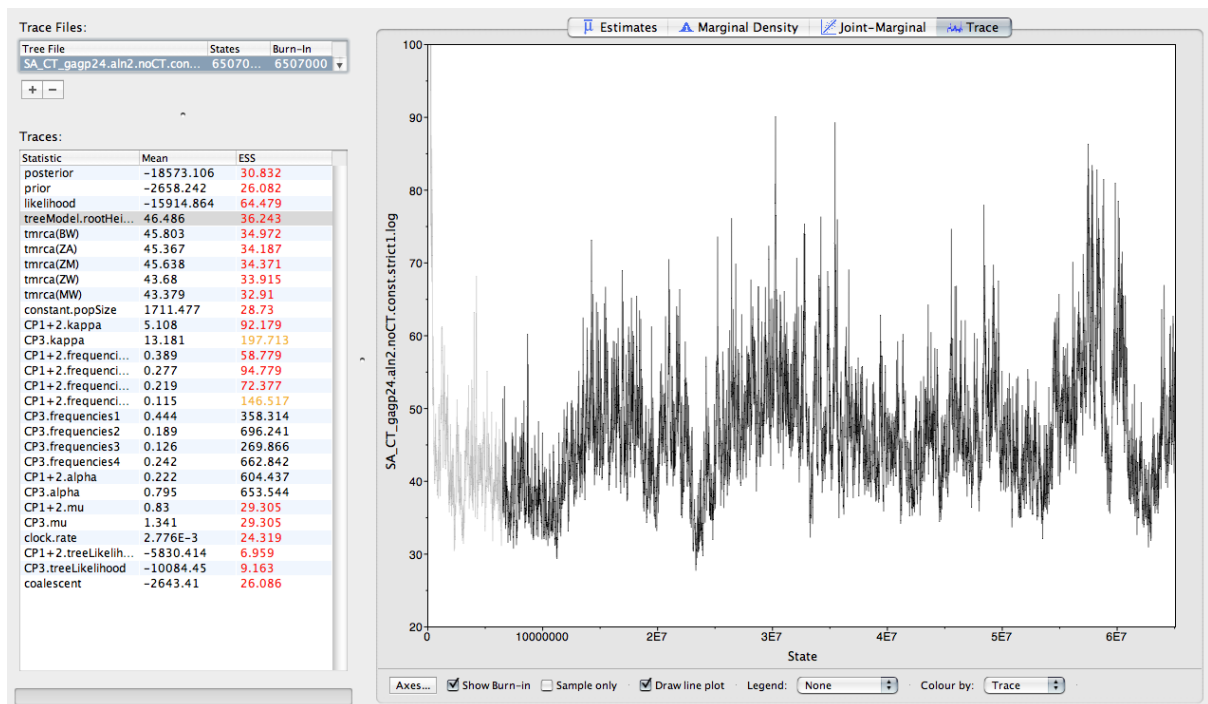


Figure 6.2: Convergence in the trace file for the run **Const.strict.est.2** in the *gag* p24 Southern African (excluding Cape Town isolates) data set. The ESS for the root height was only 36.243, however good convergence in the trace file was observed. The Bayesian MCMC chain was only run for 65 million steps in the chain, as can be seen in the X-axis of the graph, before it were stopped.

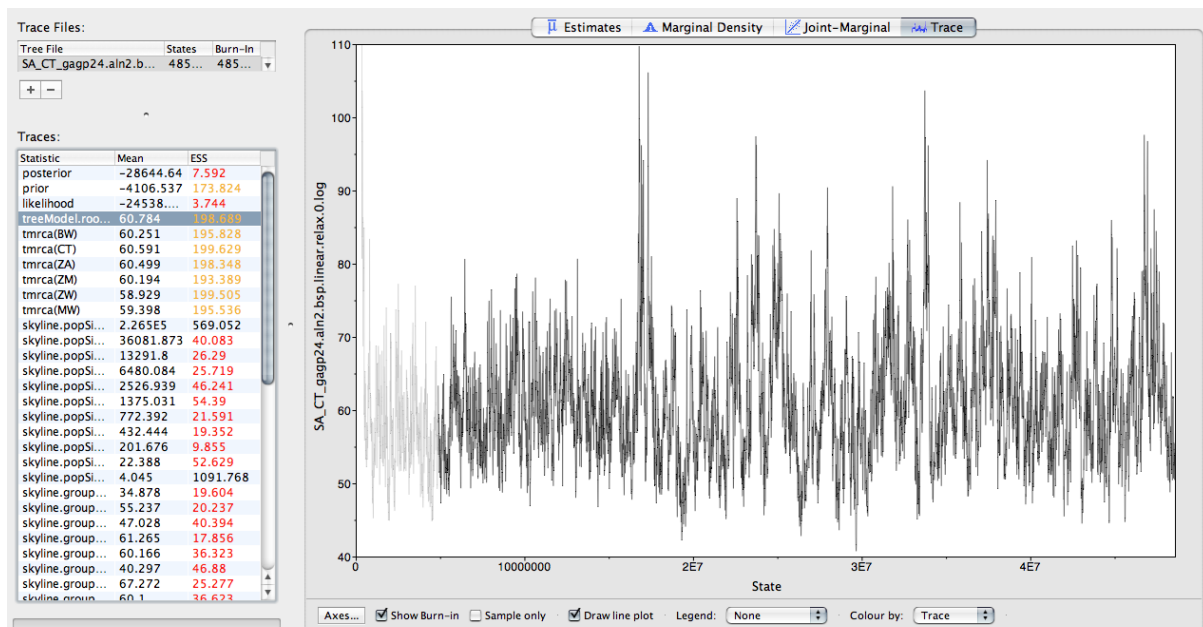


Figure 6.3: Convergence in the trace file for the run **BSP.relax.est.1** in the *gag* p24 Southern African (including Cape Town isolates) data set. The ESS for the root height was only 198.689, however good convergence in the trace file was observed. The Bayesian MCMC chain was only run for 80 million chains, as can be seen in the X-axis of the graph, before it were stopped.

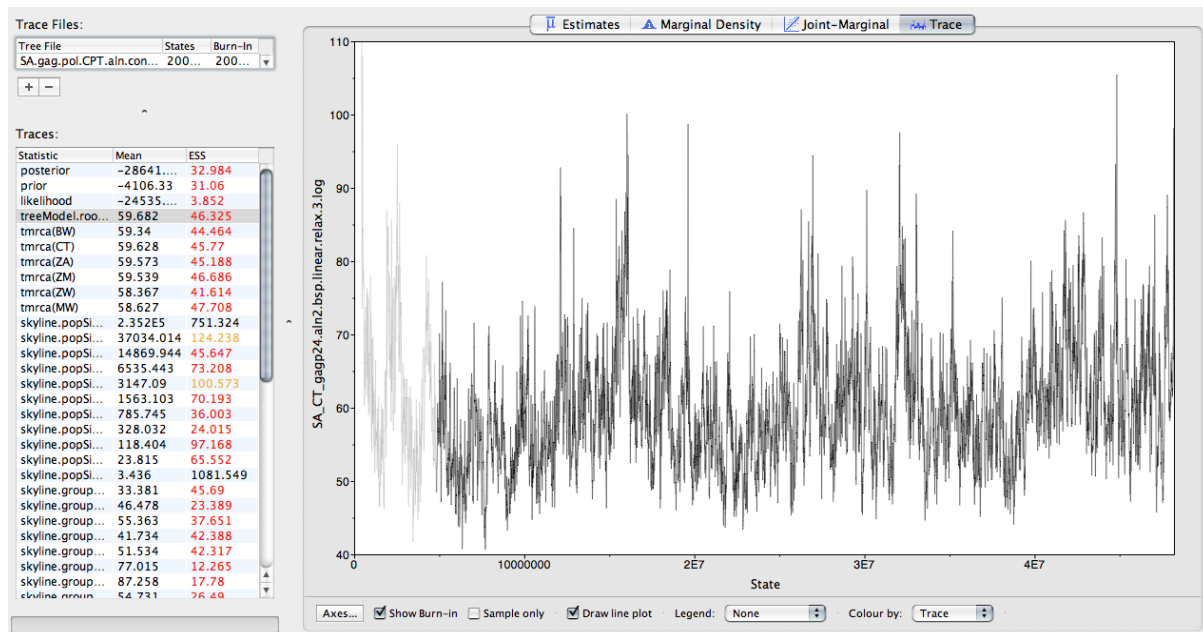


Figure 6.4: Convergence in the trace file for the run **Const.relaxed.fix.1** in the concatenated *gag-pol* Southern African (including Cape Town isolates) data set. The ESS for the root height was only 46.325, however fairly good convergence in the trace file was observed. The Bayesian MCMC chain was only run for 80 million chains, as can be seen in the X-axis of the graph, before it was stopped.

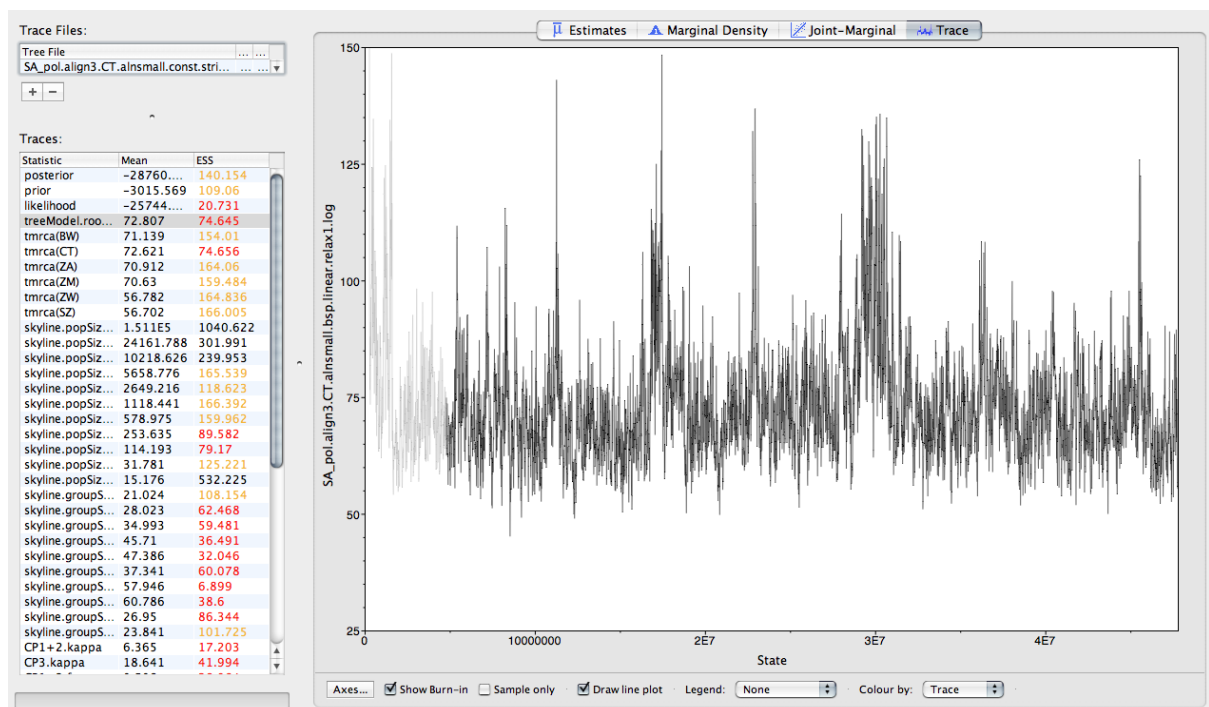


Figure 6.5: Convergence in the trace file for the **BSP.relax.est.1** model parameter run in the *pol* Southern African (including Cape Town isolates) data set. The ESS for the root height was only 74.645, however fairly good convergence in the trace file was observed. The Bayesian MCMC chain was only run for 50 million chains, as can be seen in the X-axis of the graph, before it was stopped.

APPENDIX F

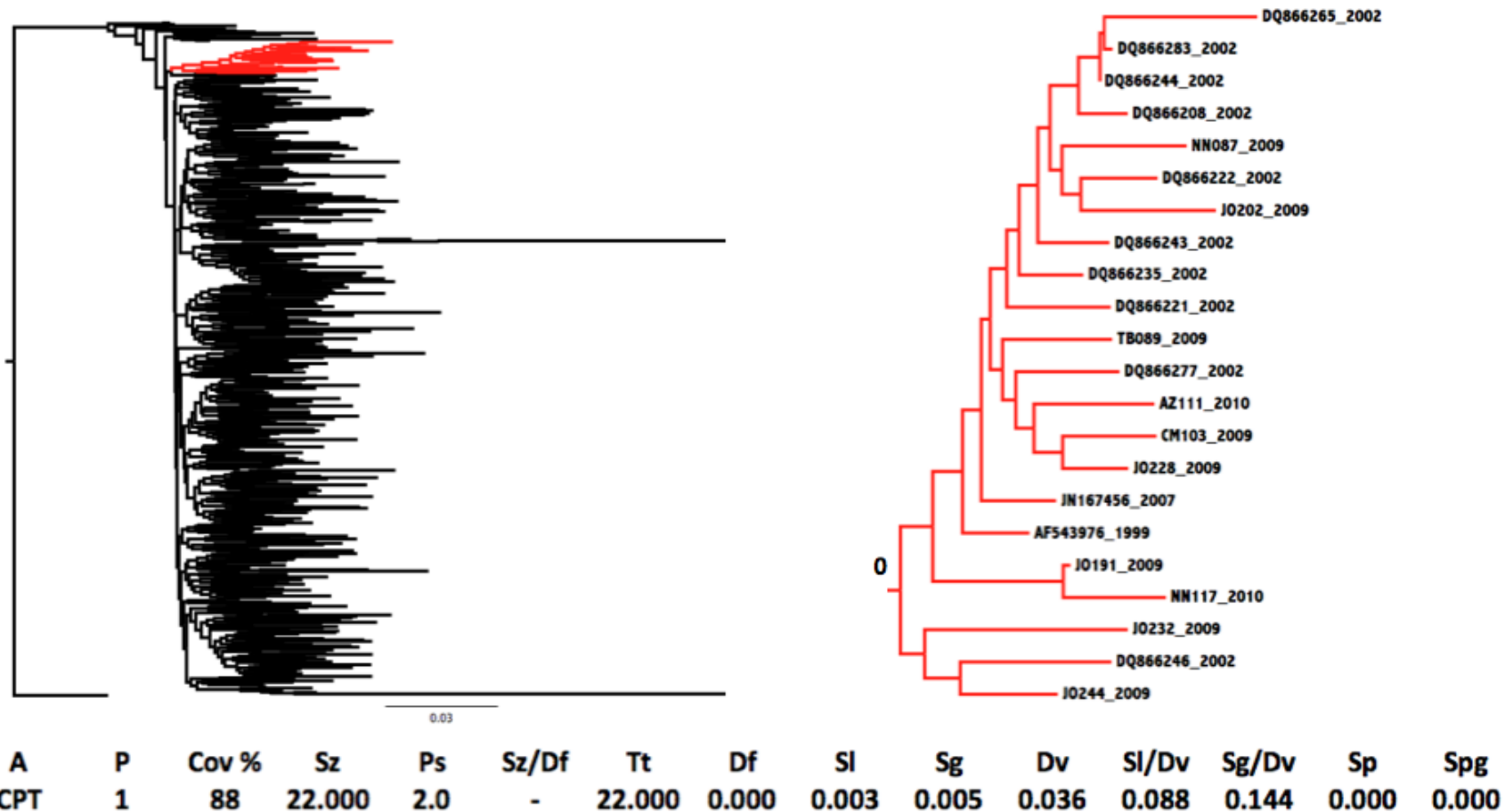


Figure 6.6: NJ-tree topology of the *gag*.cluster.1 data set with bootstrap resampling. Only 22 out of the 25 Cape Town isolates in this data set clustered together in a monophyletic clade. This tree was inferred from an alignment totalling 441 nucleotide base pairs in MEGA v 5.0 with the K2P model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

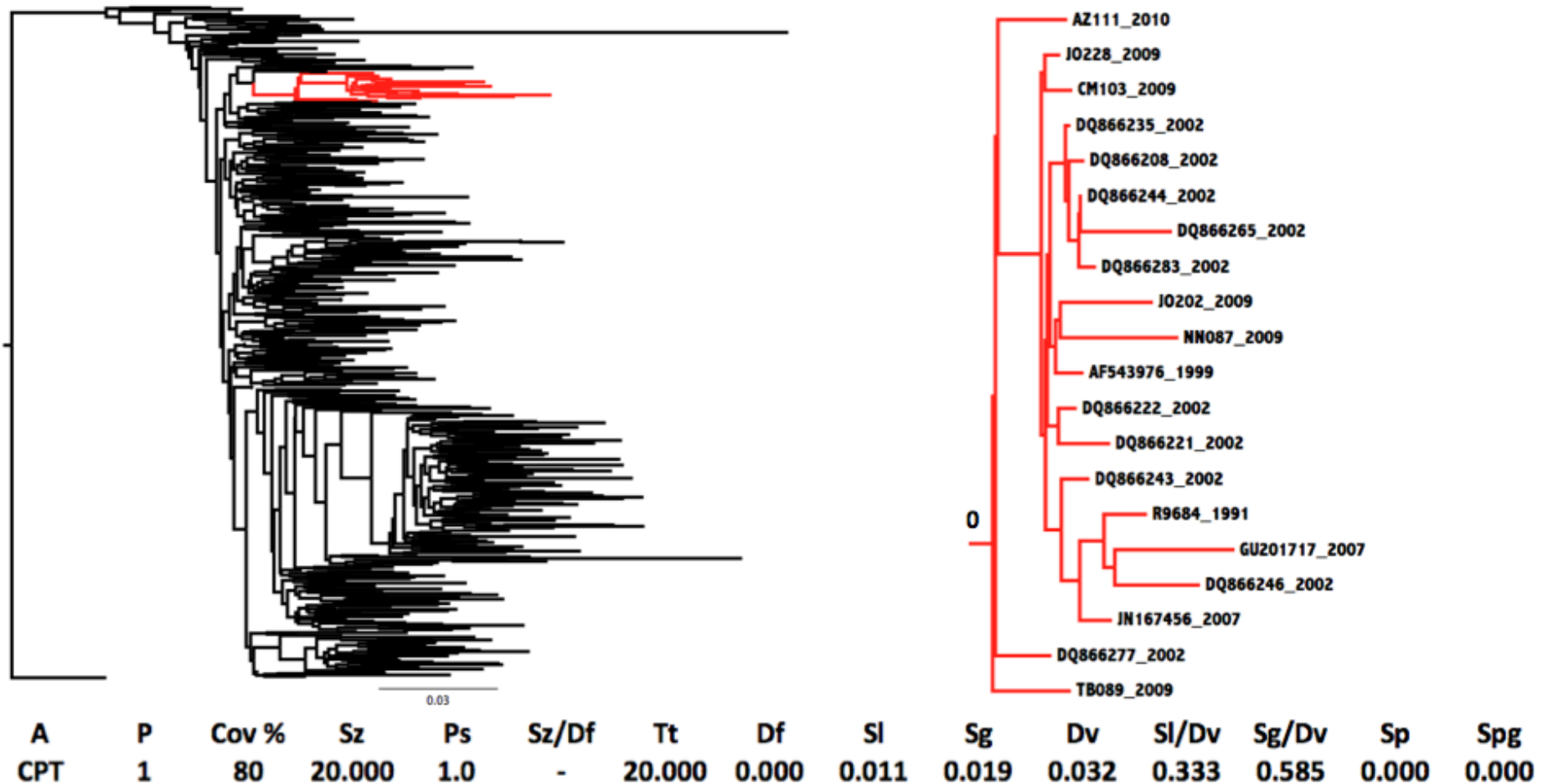


Figure 6.7: ME-tree topology of the *gag*.cluster.1 data set with bootstrap resampling. Only 20 out of the 25 Cape Town taxa clustered in a monophyletic cluster. This tree was inferred from an alignment totalling 441 nucleotide base pairs in fastME with the K2P model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

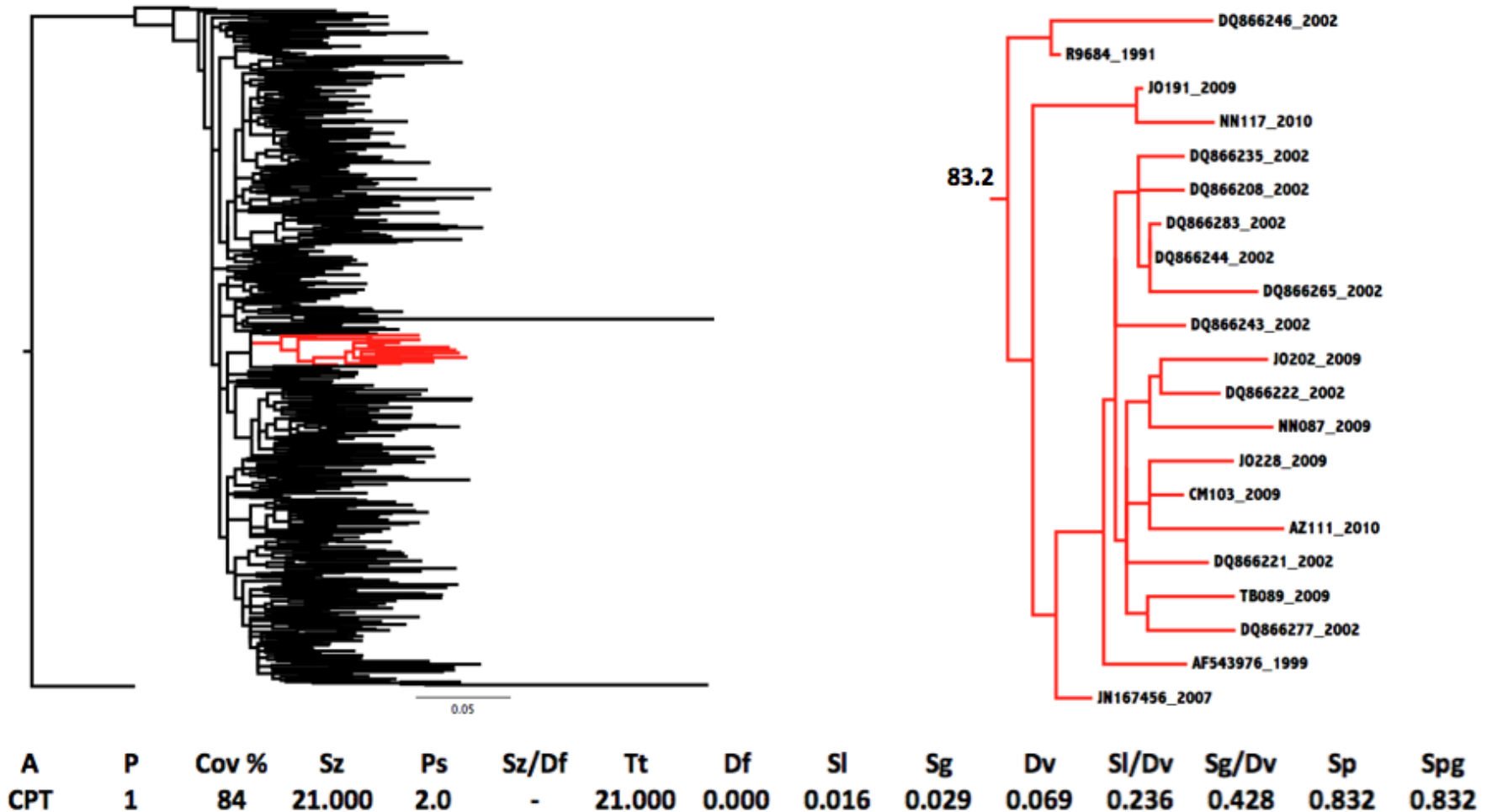


Figure 6.8: ML-tree topology of the *gag.cluster.1* data set with aLRT. Only 21 out of the 25 Cape Town taxa clustered in a monophyletic cluster. This tree was inferred from an alignment totalling 441 nucleotide base pairs in phyML with the HKY85 model of nucleotide substitution and aLRT. The aLRT support for the internal branch of the large monophyletic cluster of Cape Town sequences is 83,2%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

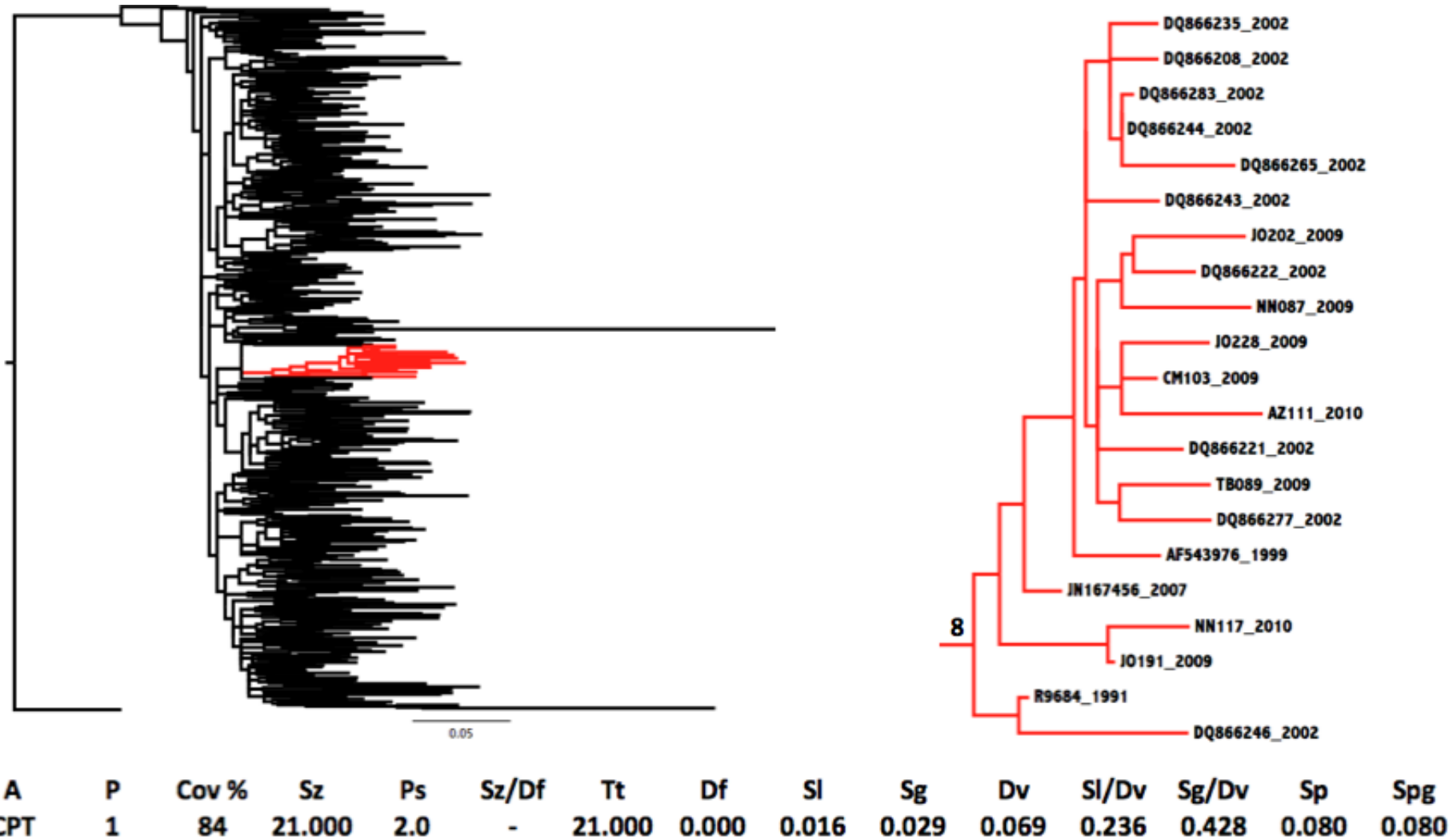


Figure 6.9: ML-tree topology of the *gag.cluster.1* data set with bootstrap resampling. Only 21 out of the 25 Cape Town taxa clustered in a monophyletic cluster. This tree was inferred from an alignment totalling 441 nucleotide base pairs in phyML with the HKY85 model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 8,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

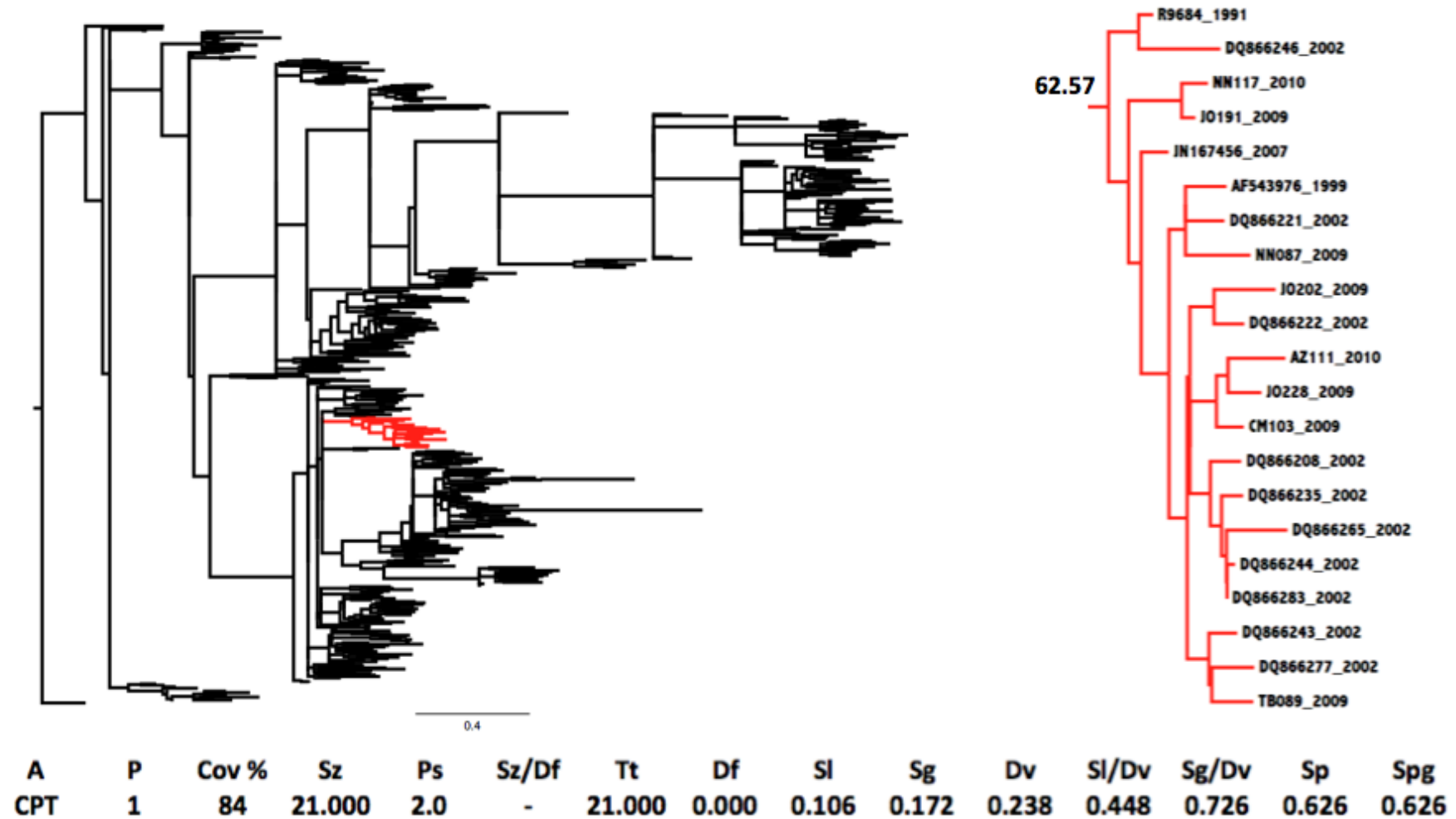


Figure 6.10: Bayesian tree topology of the *gag*.cluster.1 data set. Only 21 out of the 25 Cape Town taxa clustered in a monophyletic cluster. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in MrBayes with the GTR model of nucleotide substitution. The posterior support for the internal branch of the large monophyletic cluster of Cape Town sequences is 62,57%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

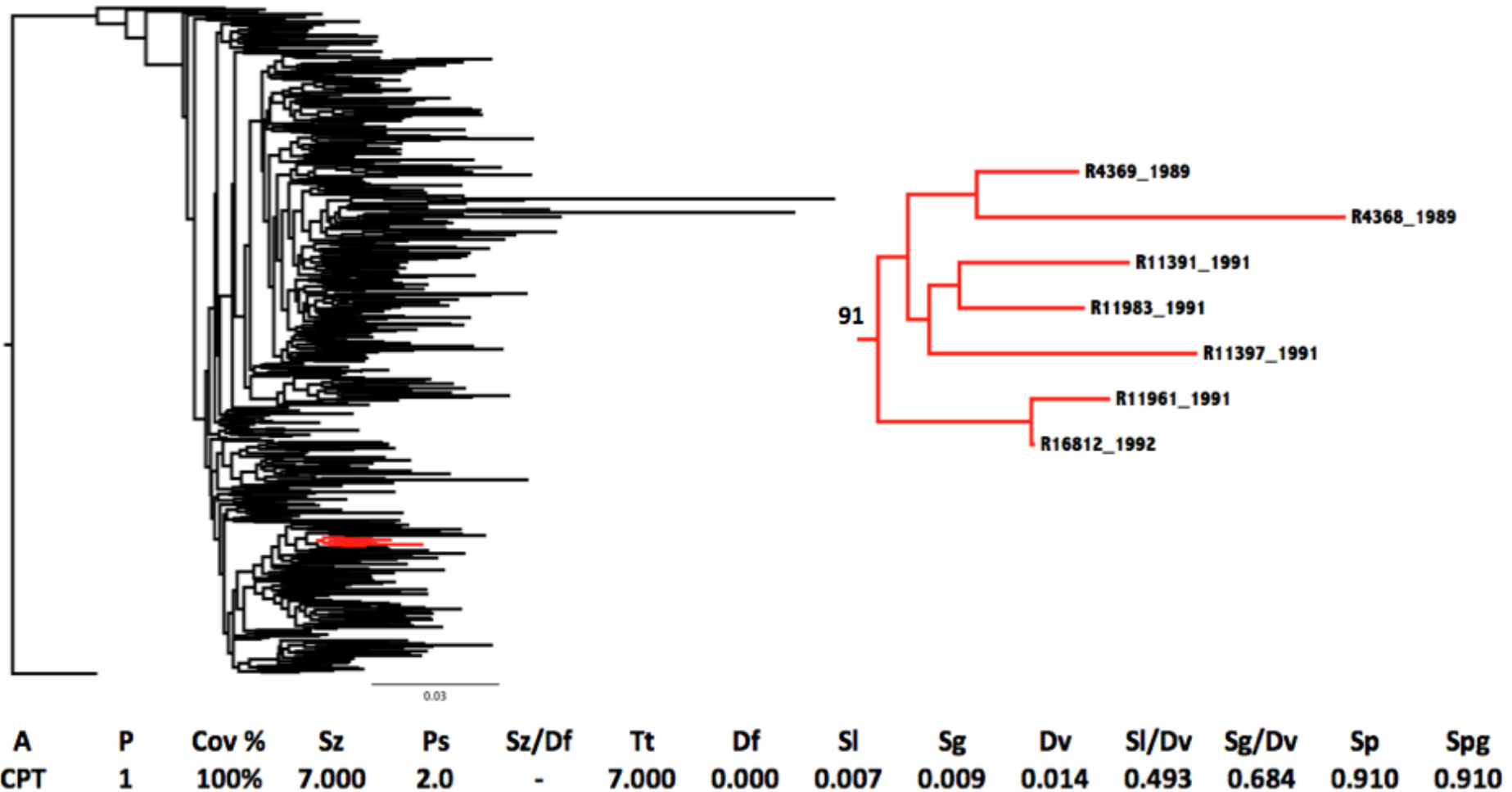


Figure 6.11: NJ-tree topology of the *gag.cluster.2* data set with bootstrap resampling. All 7 Cape Town taxa clustered in a monophyletic cluster. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in MEGA v 5.0 with the K2P model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 91,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

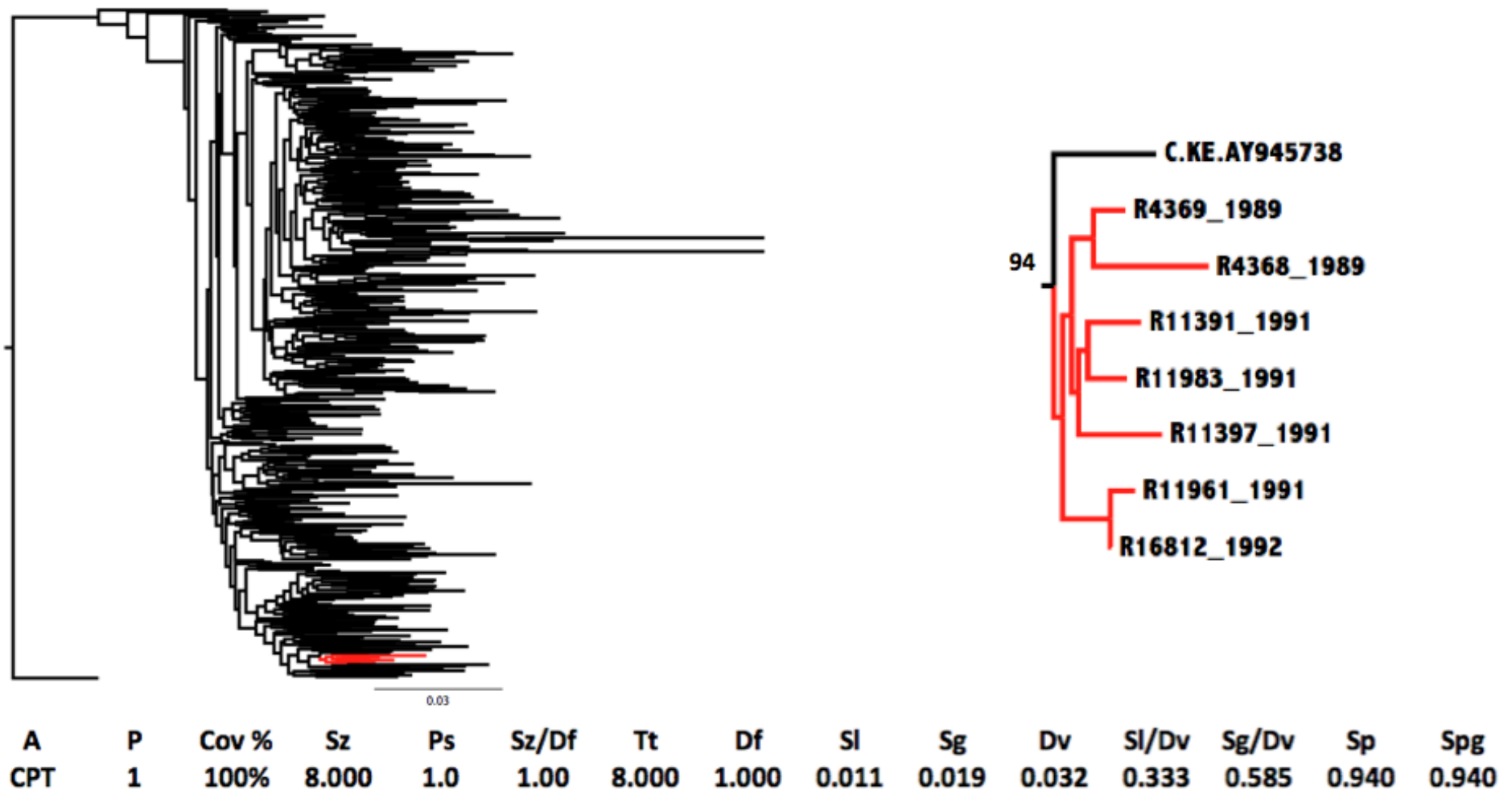


Figure 6.12: ME-tree topology of the *gag*.cluster.2 data set with bootstrap resampling. All 7 of the Cape Town taxa clustered in a monophyletic cluster broken once by an isolate from Kenya. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in fastME with the K2P model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 94,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

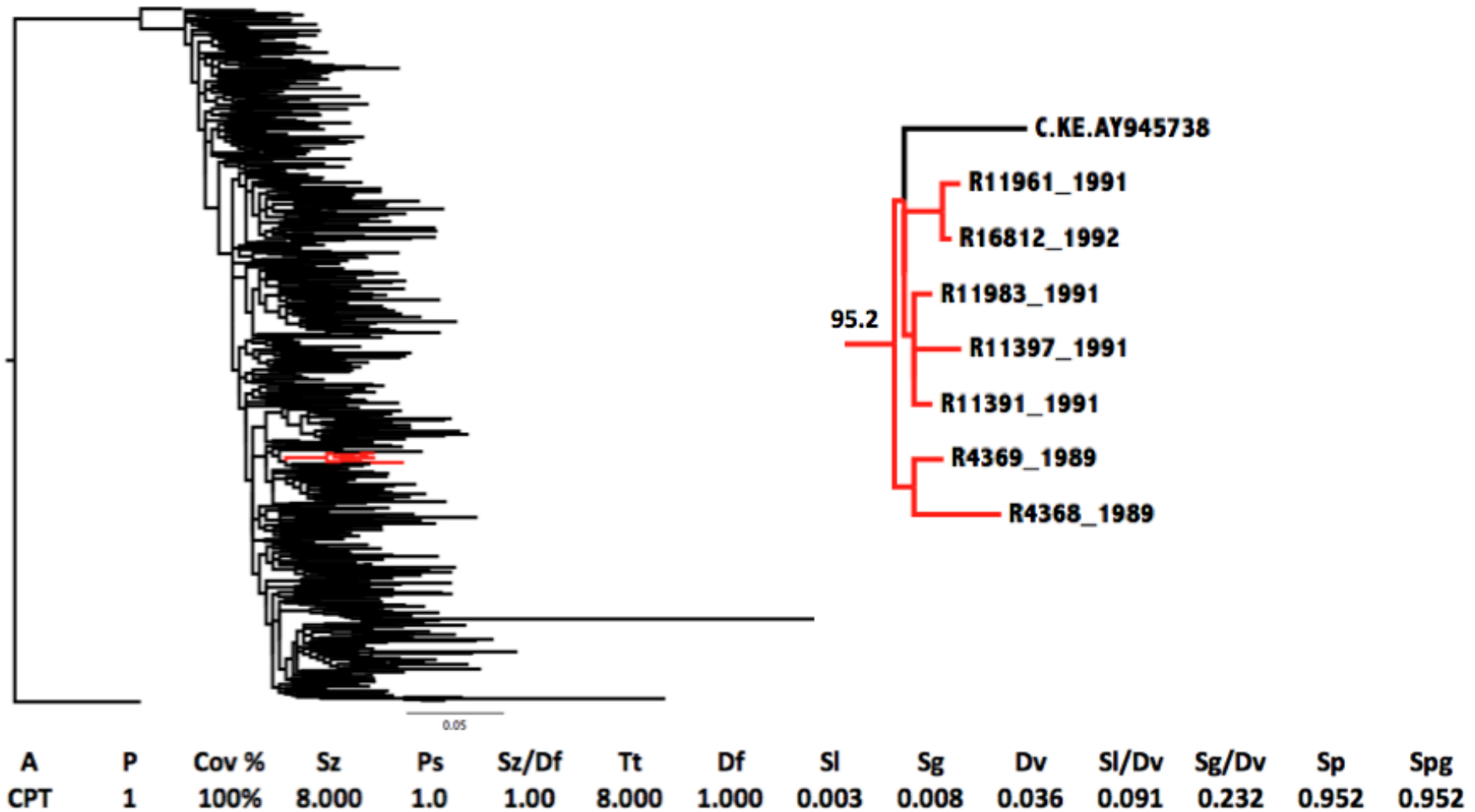


Figure 6.13: ML-tree topology of the *gag*.cluster.2 data set with aLRT. All of the 7 Cape Town taxa clustered in a monophyletic cluster broken once by an isolate from Kenya. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in phyML with aLRT support and the use of the GTR model of nucleotide substitution. The aLRT support for the internal branch of the large monophyletic cluster of Cape Town sequences is 95,2%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

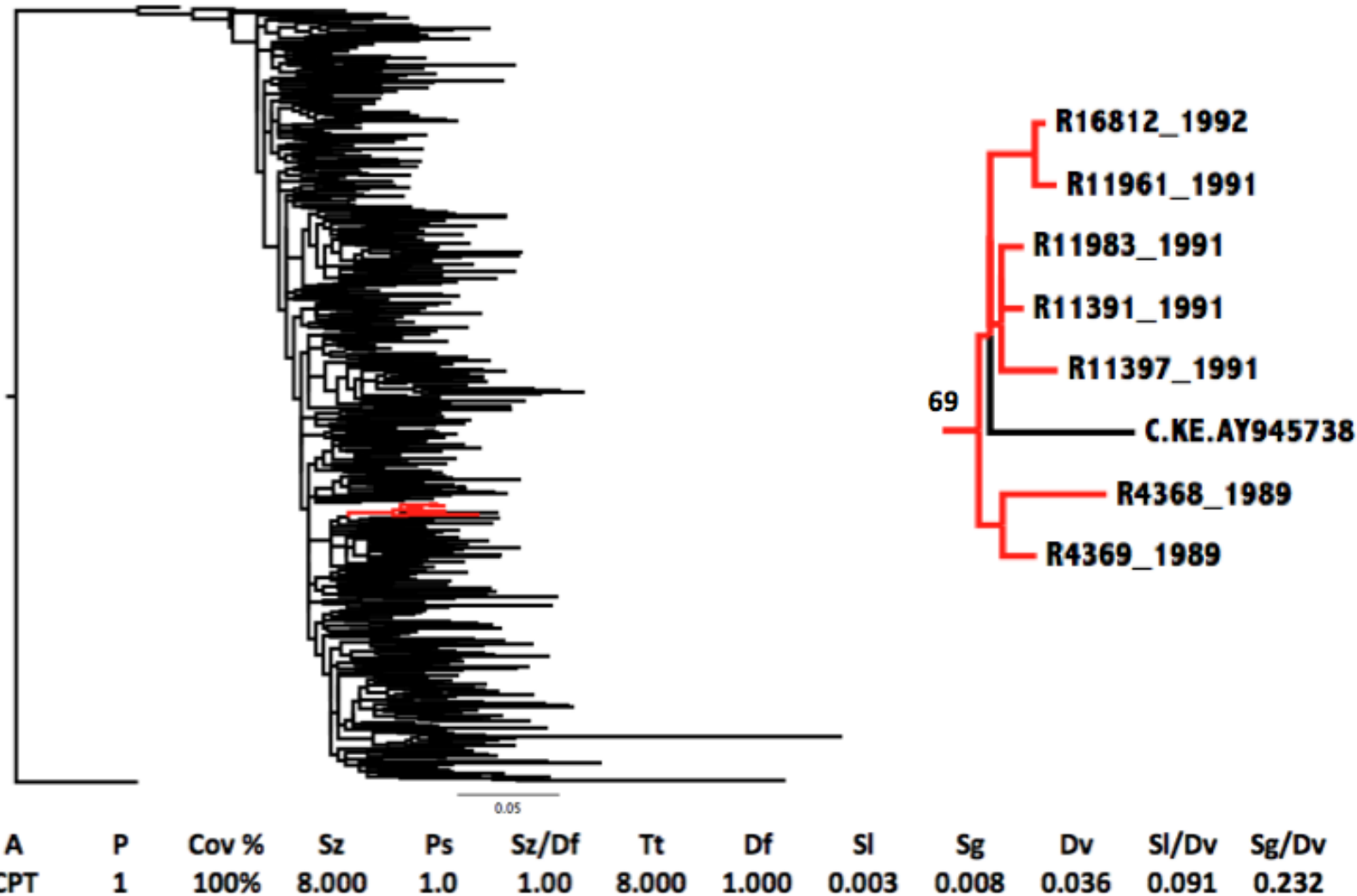


Figure 6.14: ML-tree topology of the *gag*.cluster.2 data set with bootstrap resampling. All 7 of the Cape Town taxa clustered in a monophyletic cluster broken once by a Kenyan isolate. This tree was inferred from an alignment 441 bp long in phyML v 3.0 with the use of the HKY85 model of nucleotide substitution and a total of 100 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 69,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

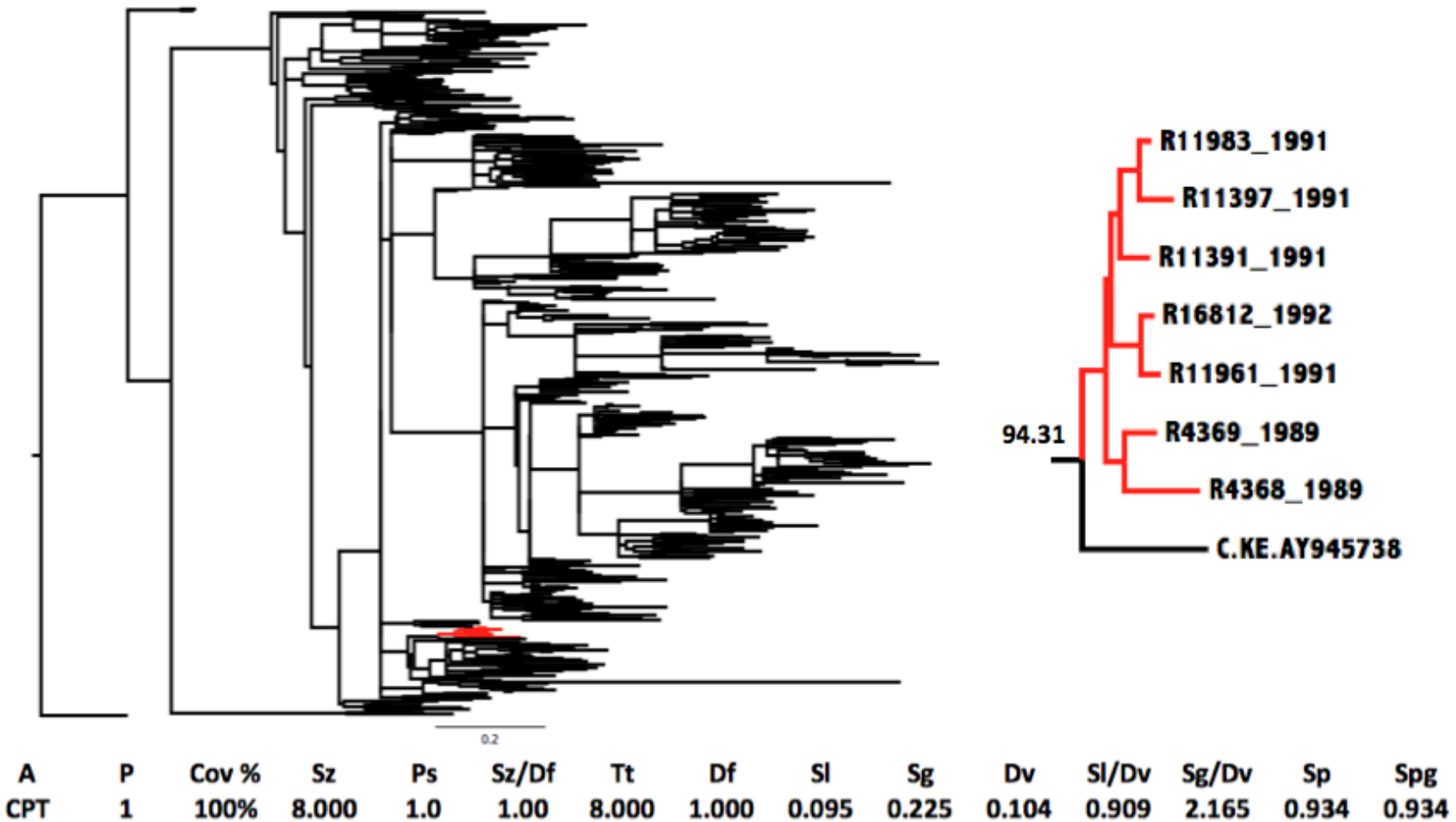
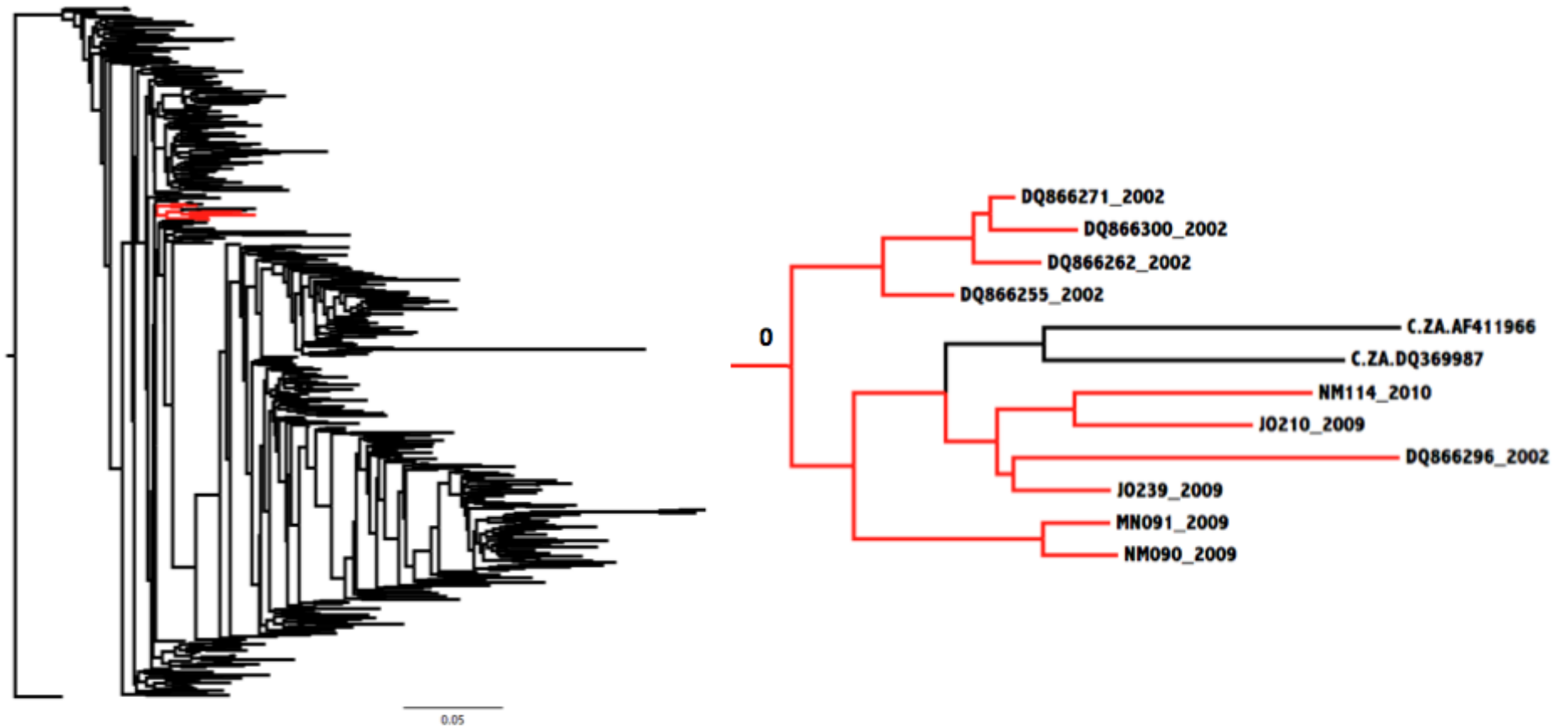


Figure 6.15: Bayesian tree topology of the *gag*.cluster.2 data set. All 7 of the Cape Town taxa clustered in a monophyletic cluster broken once by an isolate from Kenya. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in MrBayes with the GTR model of nucleotide substitution. The posterior support for the internal branch of the large monophyletic cluster of Cape Town sequences is 94,31%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.



A	P	Cov %	Sz	Ps	Sz/Df	Tt	Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
CPT	1	100%	10.000	2.0	10.000	12.000	1.000	0.001	0.003	0.029	0.021	0.090	0.000	0.000

Figure 6.16: NJ-tree topology of the *gag*.cluster.3 data set with bootstrap resampling. All 10 of the Cape Town taxa clustered in a monophyletic cluster broken once by two isolates from South Africa. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in MEGA v 5.0 with the use of the K2P model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

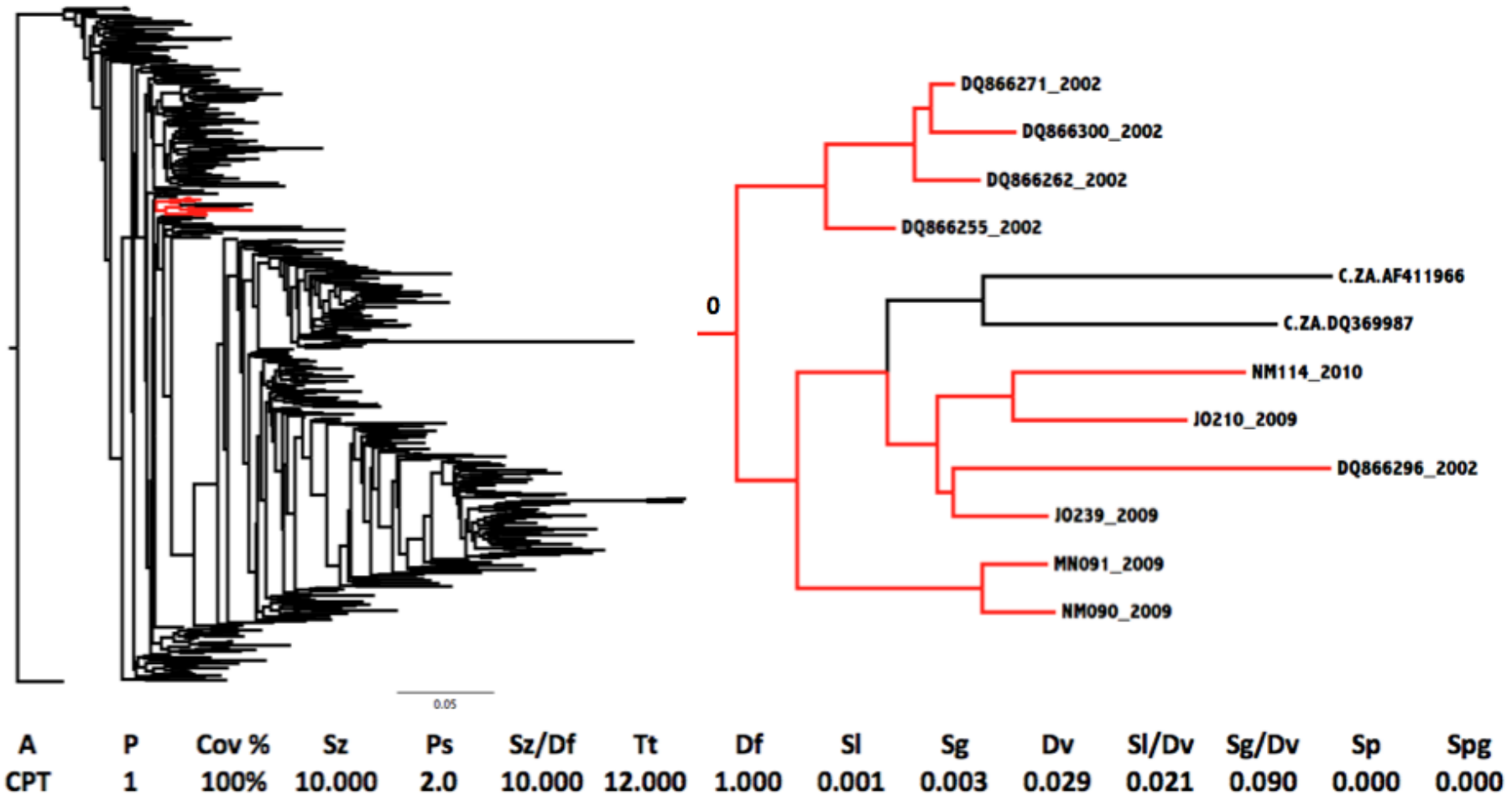


Figure 6.17: ME-tree topology of the *gag*.cluster.3 data set with bootstrap resampling. All of 10 of the Cape Town taxa clustered in a monophyletic cluster broken once by two isolates from South Africa. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in fastME with the use of the K2P model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

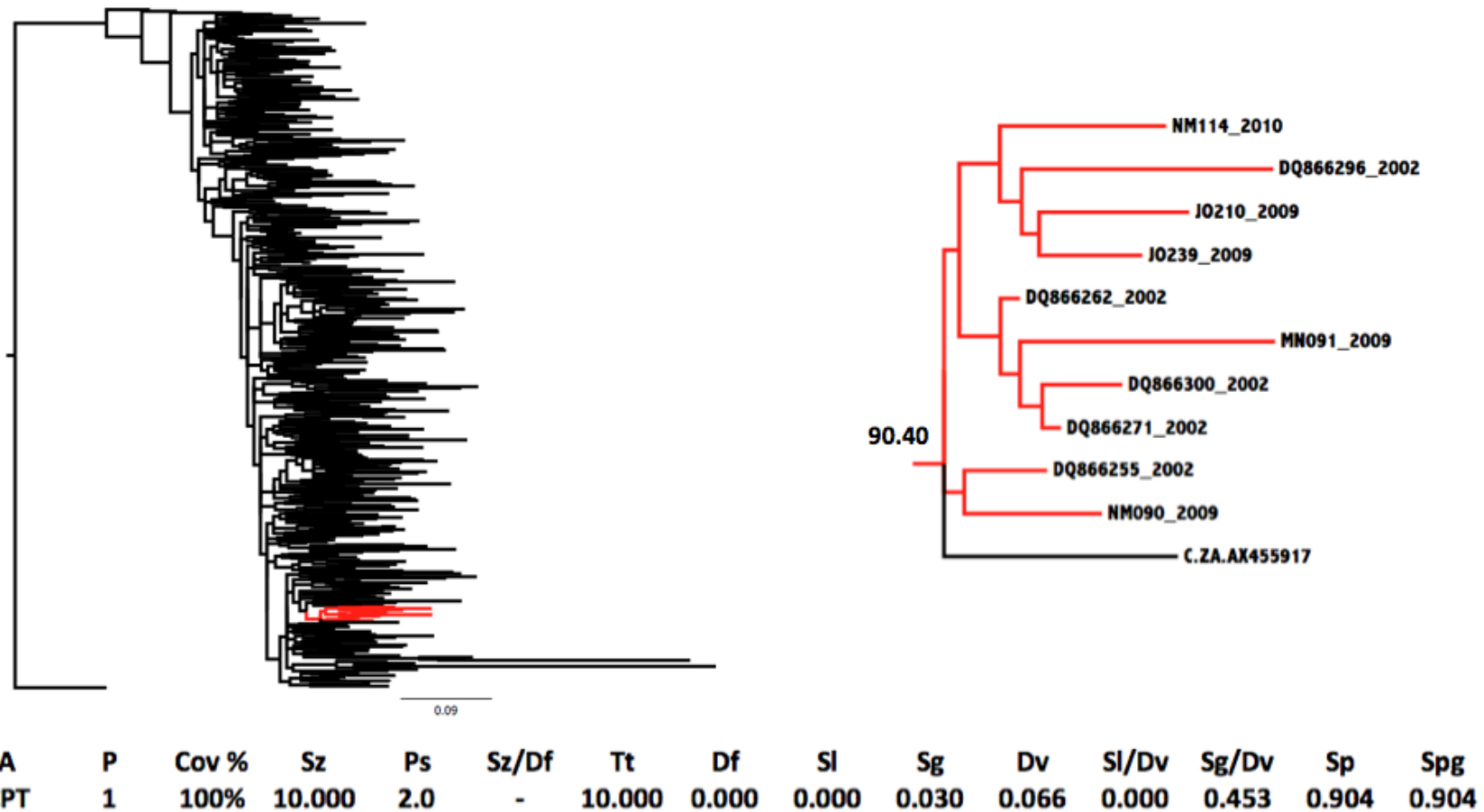


Figure 6.18: ML-tree topology of the *gag.cluster.3* data set with aLRT. All of the 10 Cape Town taxa clustered in a monophyletic cluster. Manual inspection revealed that the cluster was broken once by an isolate from South Africa. However, the phylotype analyses of the tree topology suggest that this South African isolate cluster outside of the monophyletic clade. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in phyML with the HKY85 model of nucleotide substitution and aLRT. The aLRT support for the internal branch of the large monophyletic cluster of Cape Town sequences is 90,4%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

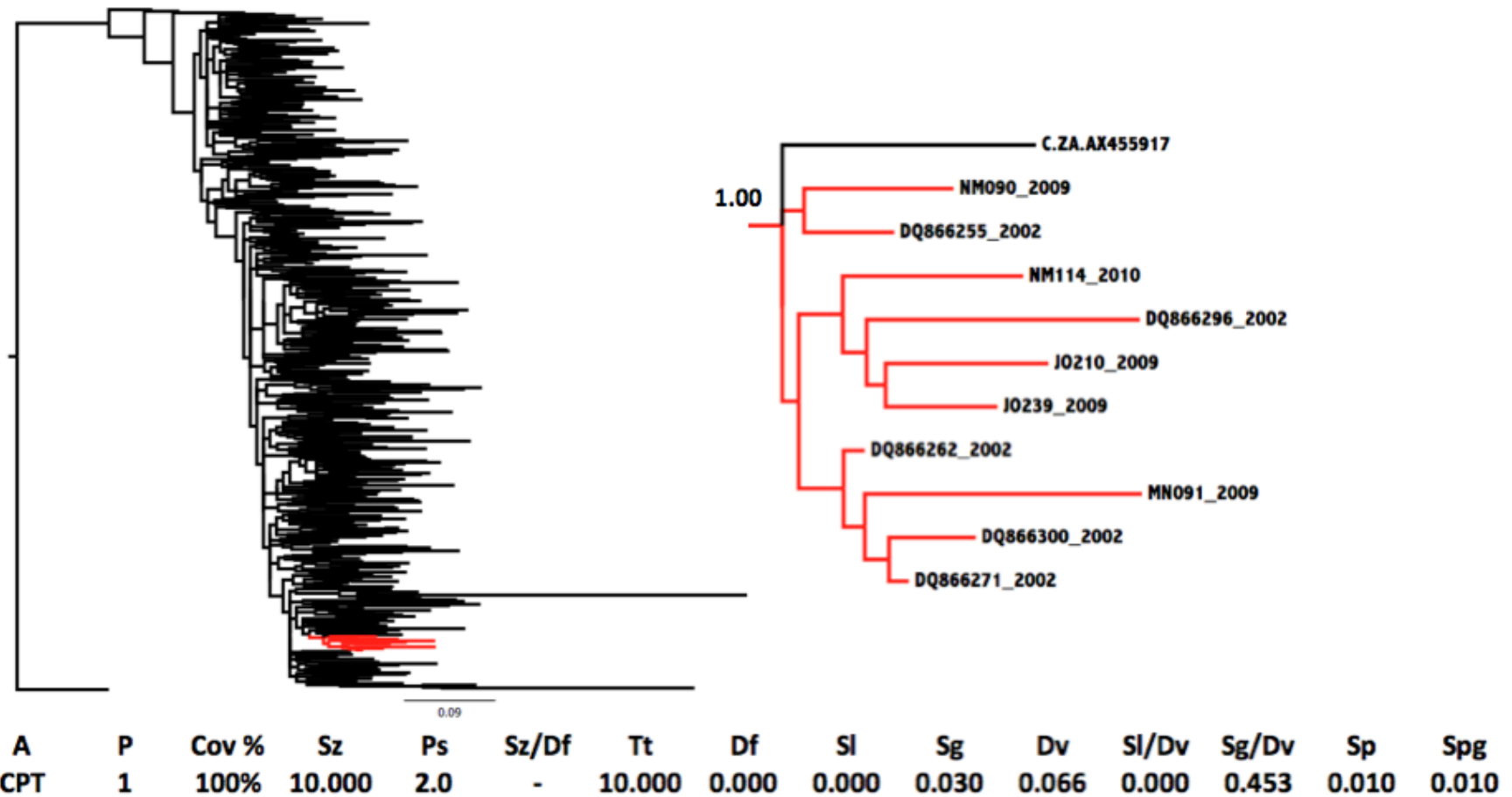


Figure 6.19: ML-tree topology of the *gag*.cluster.3 data set with bootstrap resampling. All 10 of the Cape Town isolates clustered in a monophyletic cluster. Manual inspection revealed that the cluster was broken once by an isolate from South Africa. However, the phylotype analyses of the tree topology suggest that this South African isolate cluster outside of the monophyletic clade. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in phyML with the HKY85 model of nucleotide substitution with a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 1,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

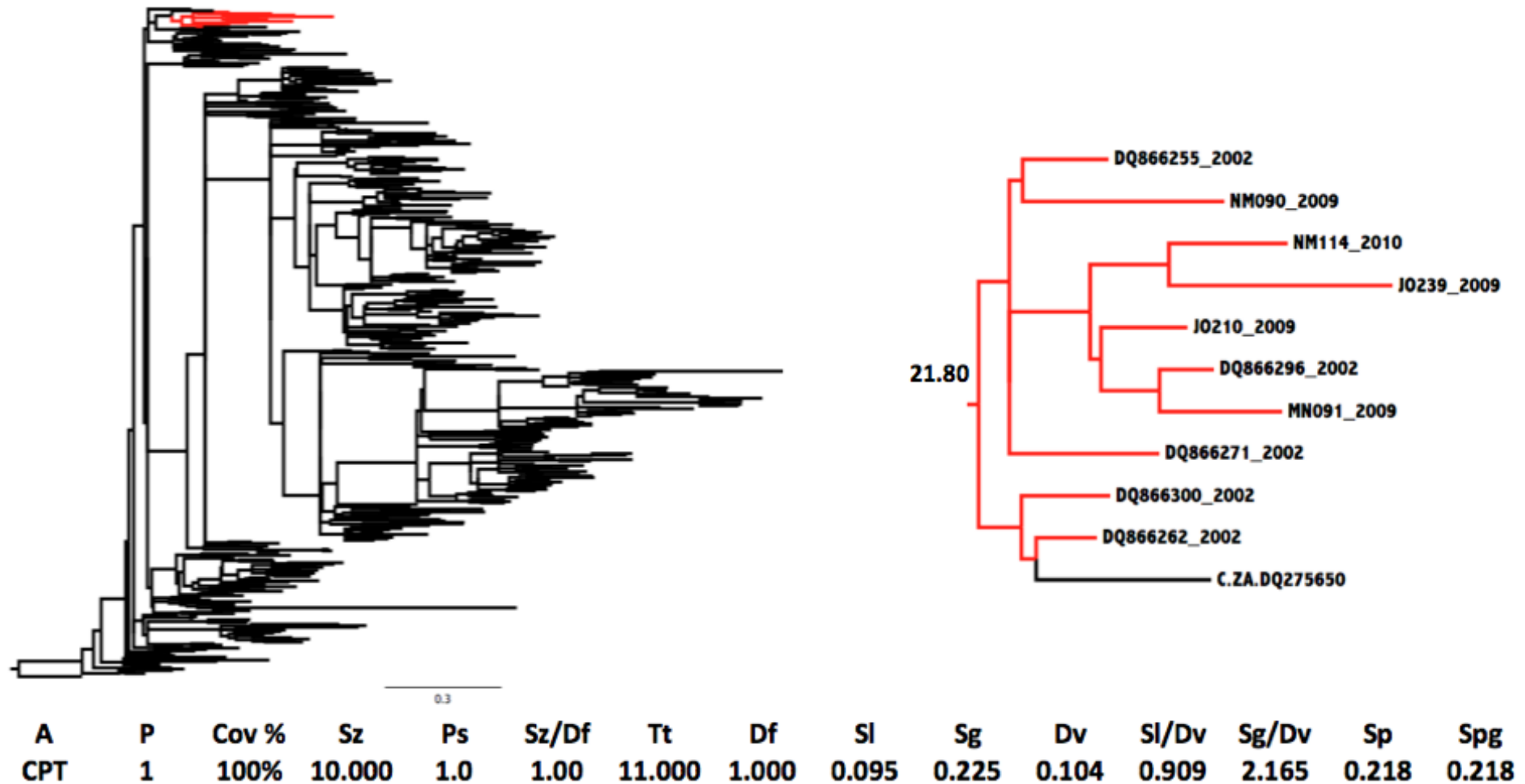


Figure 6.20: Bayesian tree topology of the *gag*.cluster.3 data set. All of the 10 Cape Town taxa clustered in a monophyletic cluster, which were broken once by an isolate from South Africa. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in MrBayes with the GTR model of nucleotide substitution. The posterior support for the internal branch of the large monophyletic cluster of Cape Town sequences is 21,80%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

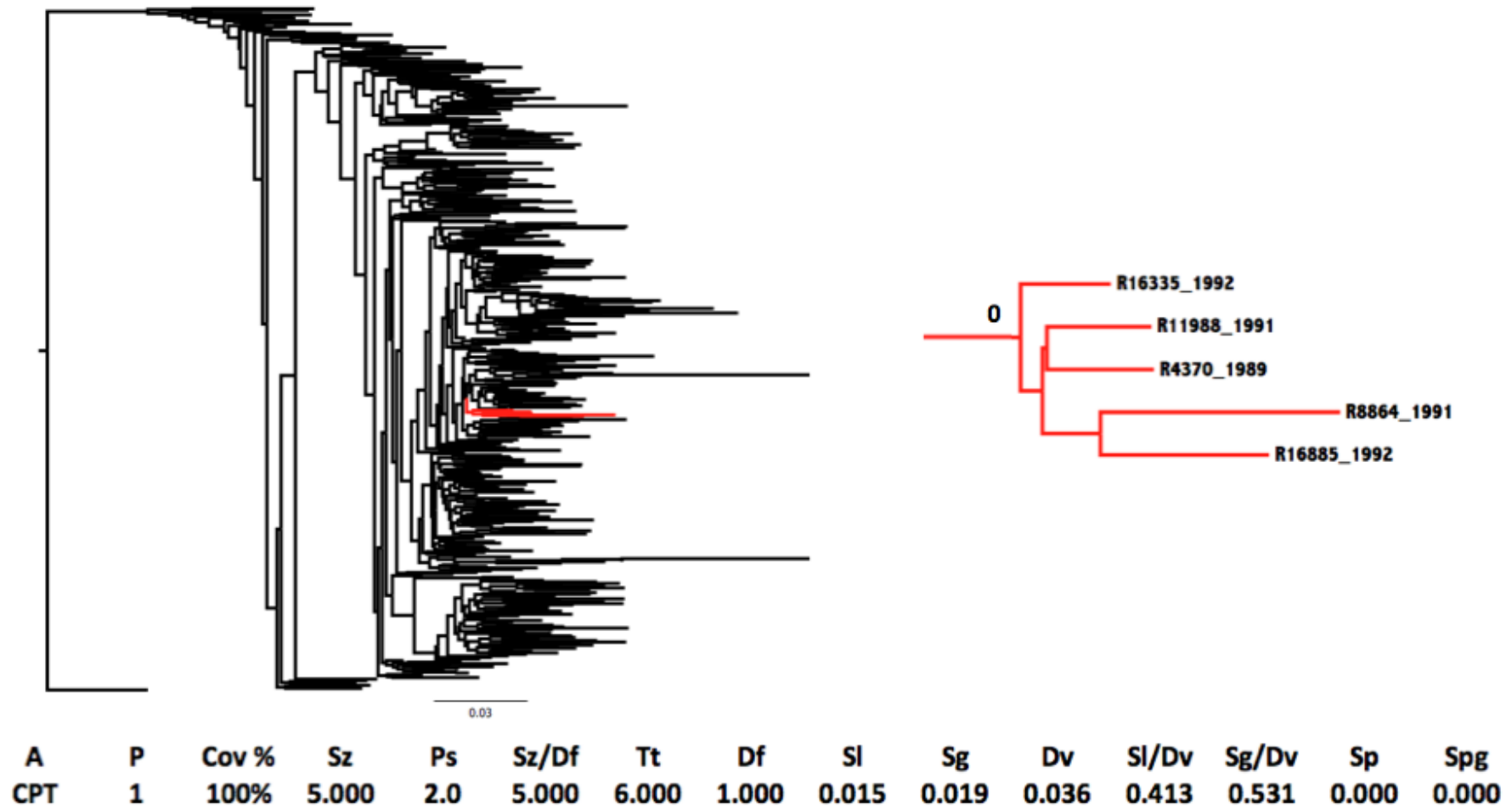


Figure 6.21: NJ-tree topology of the *gag*.cluster.4 data set with bootstrap resampling. All 5 isolates clustered in a single monophyletic clade. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in MEGA v 5.0 with the K2P model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

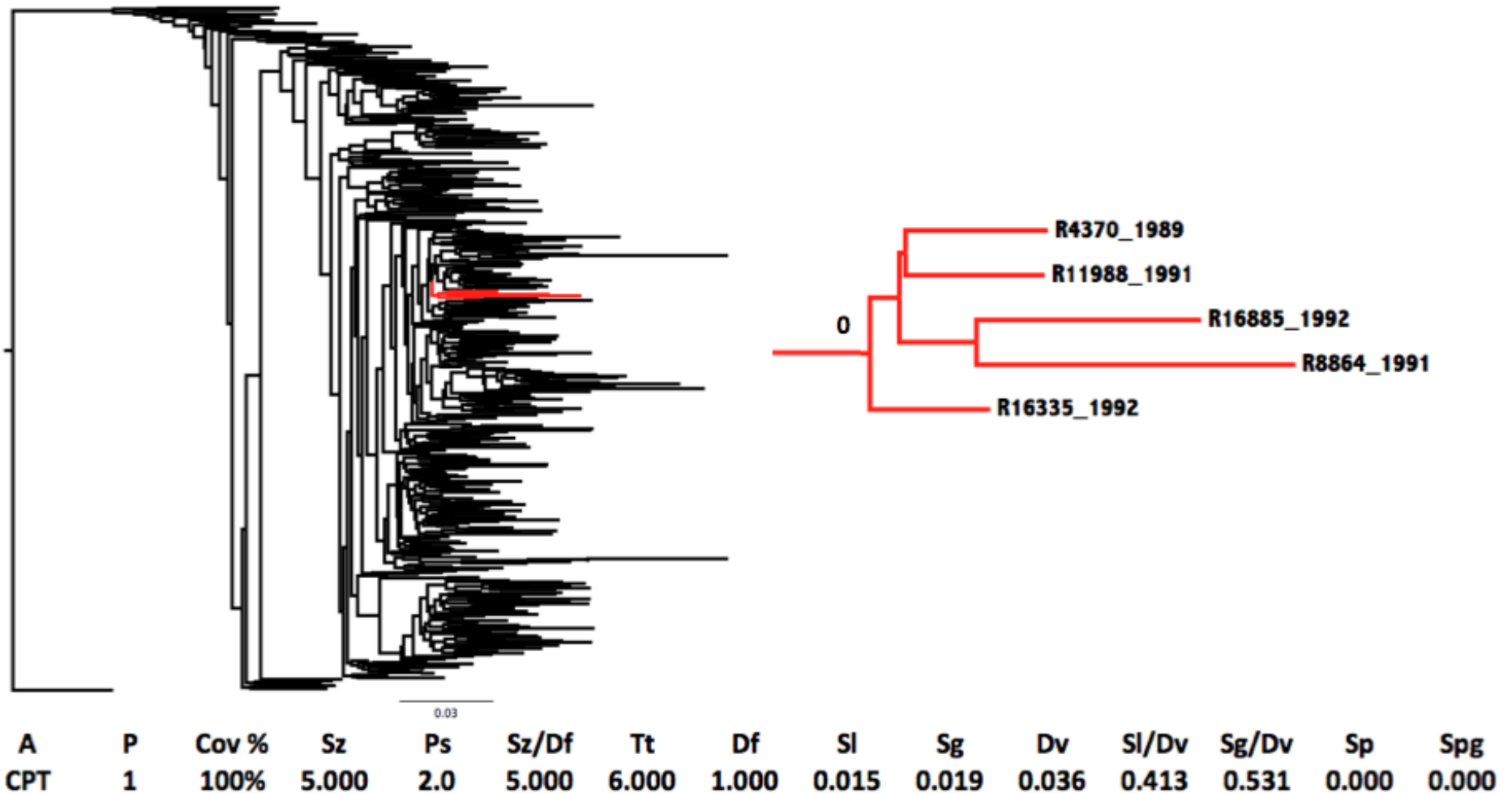
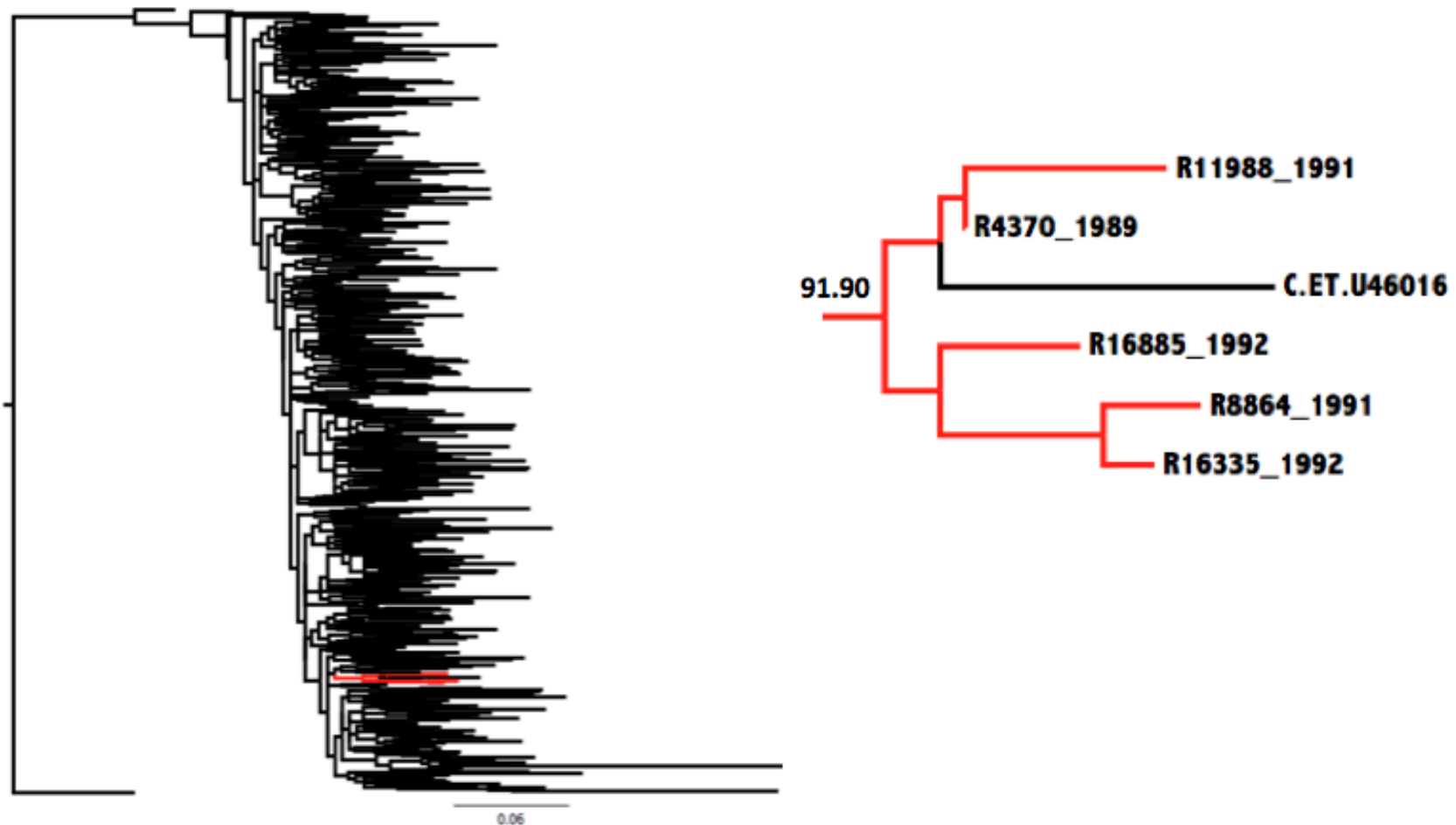
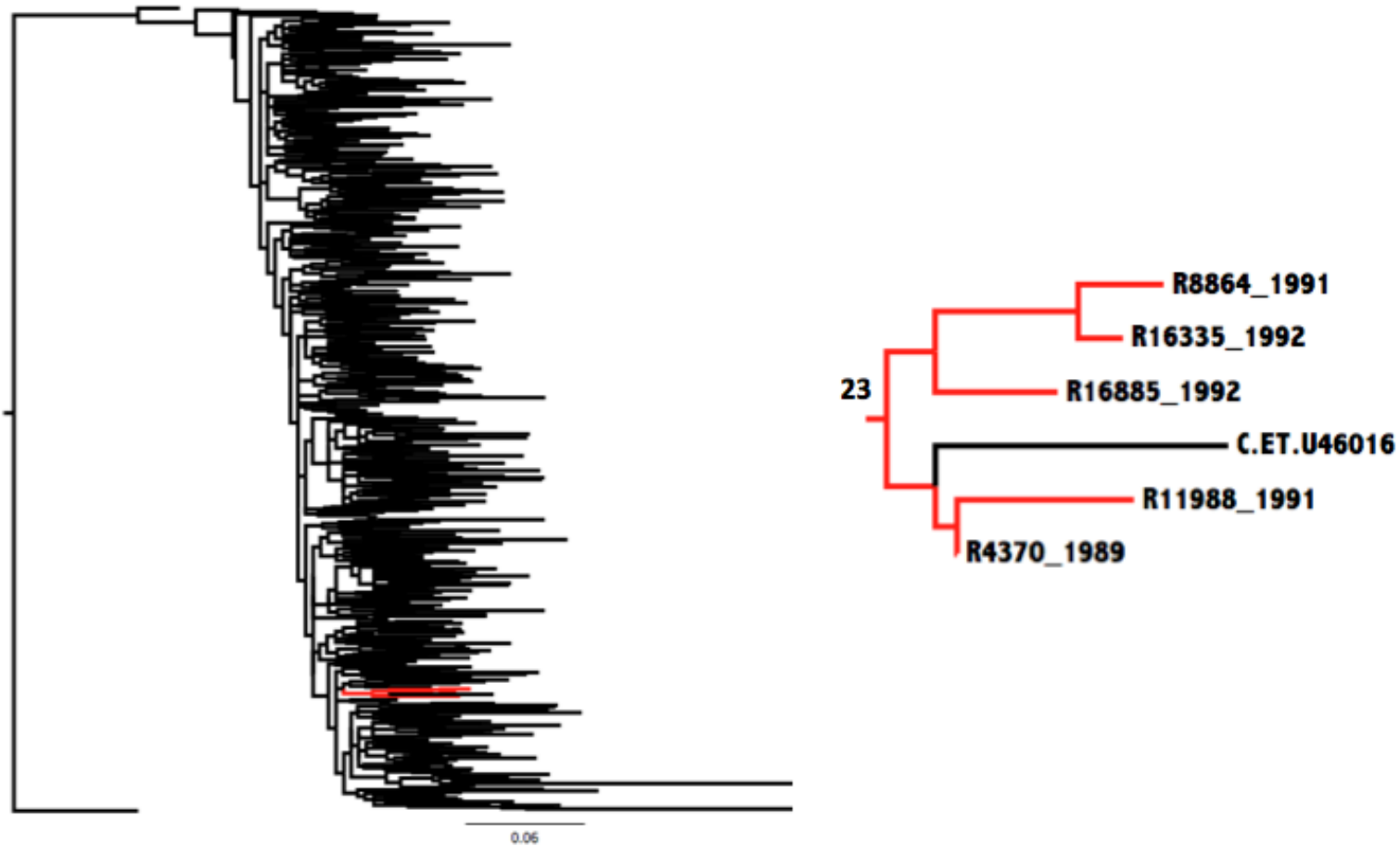


Figure 6.22: ME-tree topology of the *gag*.cluster.4 data set with bootstrap resampling. All 5 isolates clustered in a single monophyletic clade. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in fastME with the K2P model of nucleotide substitutions and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.



A	P	Cov %	Sz	Ps	Sz/Df	Tt	Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
CPT	1	100%	5.000	2.0	5.000	6.000	1.000	0.015	0.019	0.036	0.413	0.531	0.919	0.919

Figure 6.23: ML-tree topology of the *gag*.cluster.4 data set with aLRT. All 5 isolates clustered in a single monophyletic clade. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in phyML v 3.0 with the use of the HKY85 model of nucleotide substitution and aLRT. The aLRT support for the internal branch of the large monophyletic cluster of Cape Town sequences is 91,9%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.



A	P	Cov %	Sz	Ps	Sz/Df	Tt	Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
CPT	1	100%	5.000	2.0	5.000	6.000	1.000	0.015	0.019	0.036	0.413	0.531	0.230	0.230

Figure 6.24: ML-tree topology of the *gag.cluster.4* data set with bootstrap resampling. All 5 of the Cape Town isolates clustered in a monophyletic cluster broken once by an isolate from Ethiopia. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in phyML v 3.0 with the HKY85 model of nucleotide substitution with 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 23,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

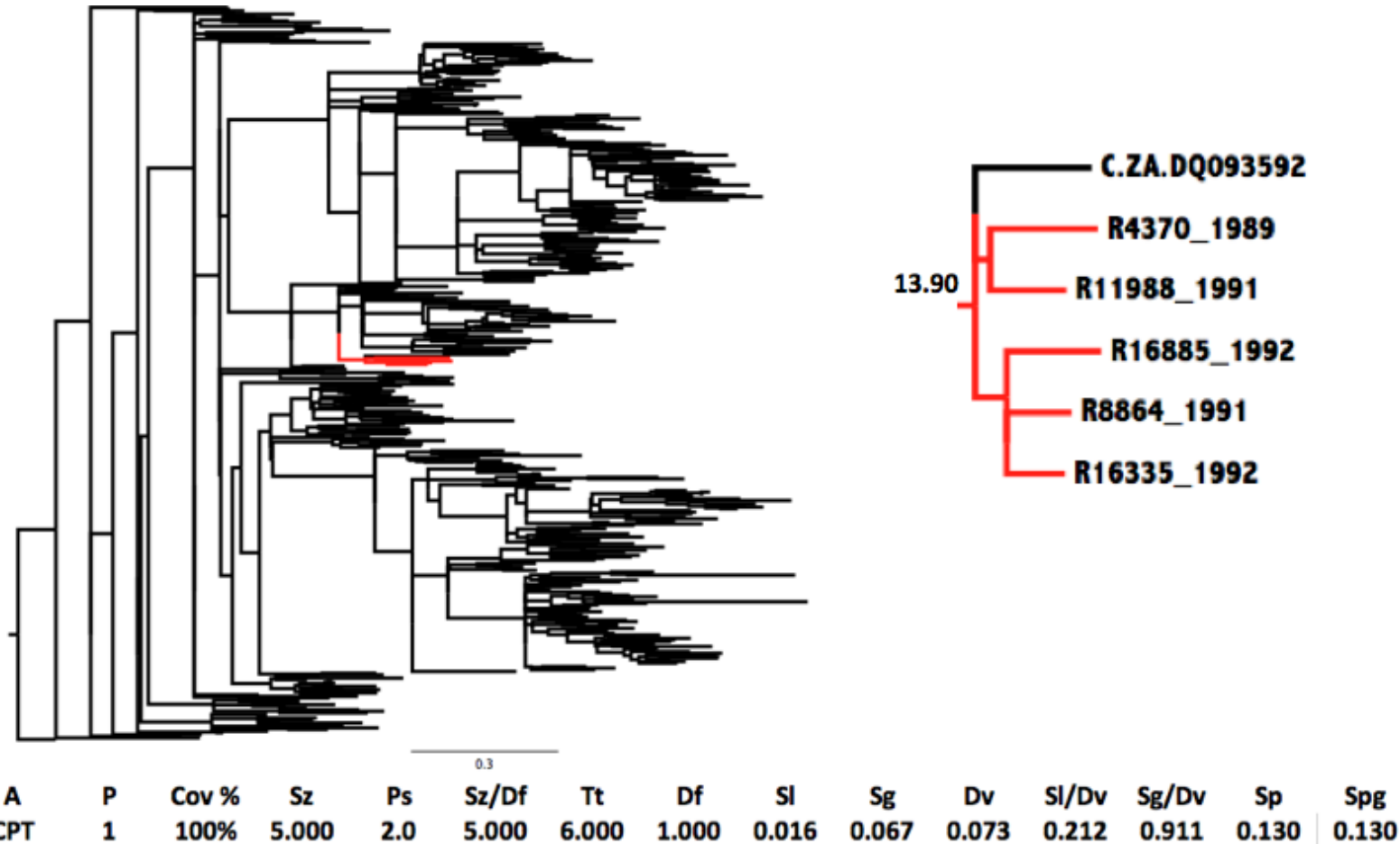


Figure 6.25: Bayesian tree topology of the *gag*.cluster.4 data set. All 5 of the Cape Town taxa clustered in a monophyletic cluster, which were broken once by a South African isolate. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in MrBayes with the GTR model of nucleotide substitution. Cape Town isolates are marked in red, Indian in blue and Brazilian isolates in green. The results of the PhyloType analyses of the tree topology can be seen in the table at the bottom. The posterior support for the internal branch of the large monophyletic cluster of Cape Town sequences is 13,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

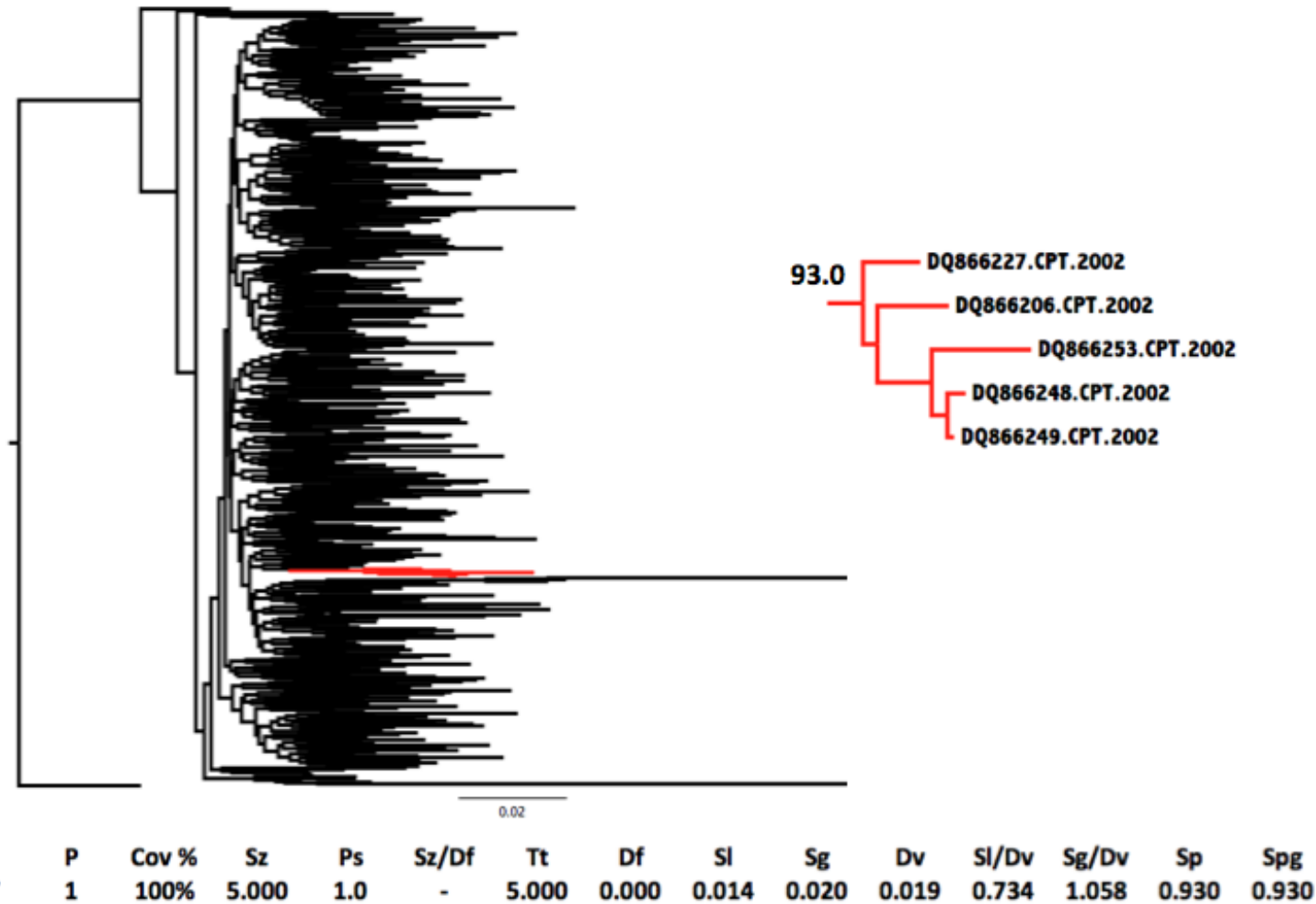


Figure 6.26: NJ-tree topology of the *gag*.cluster.5 data set with bootstrap resampling. All 5 isolates clustered in a single monophyletic clade. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in MEGA v 5.0 with the use of the K2P model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 93,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

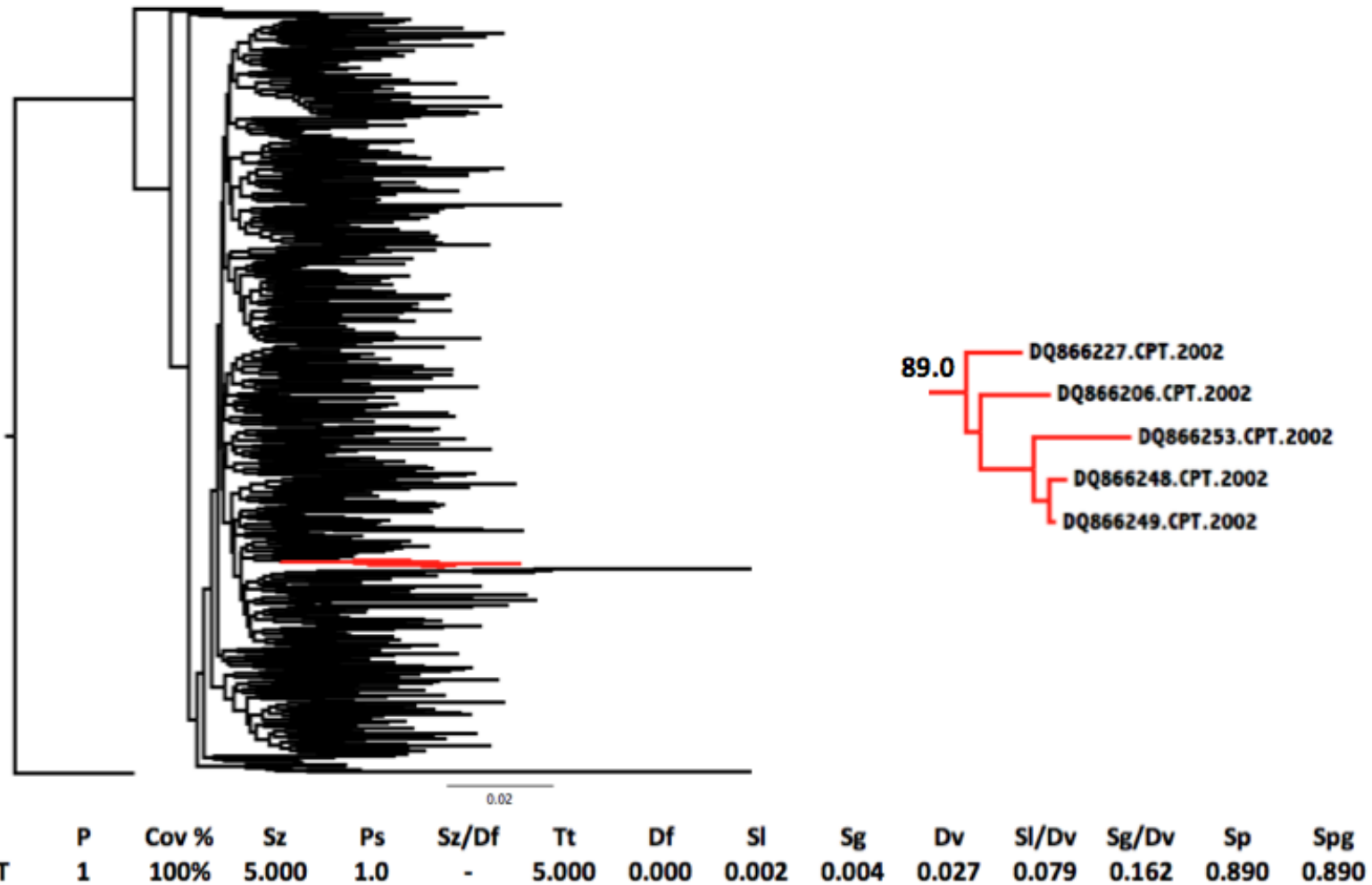


Figure 6.27: ME-tree topology of the *gag*.cluster.5 data set with bootstrap resampling. All 5 isolates clustered in a single monophyletic clade. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in phyML with the use of the K2P model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 89,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

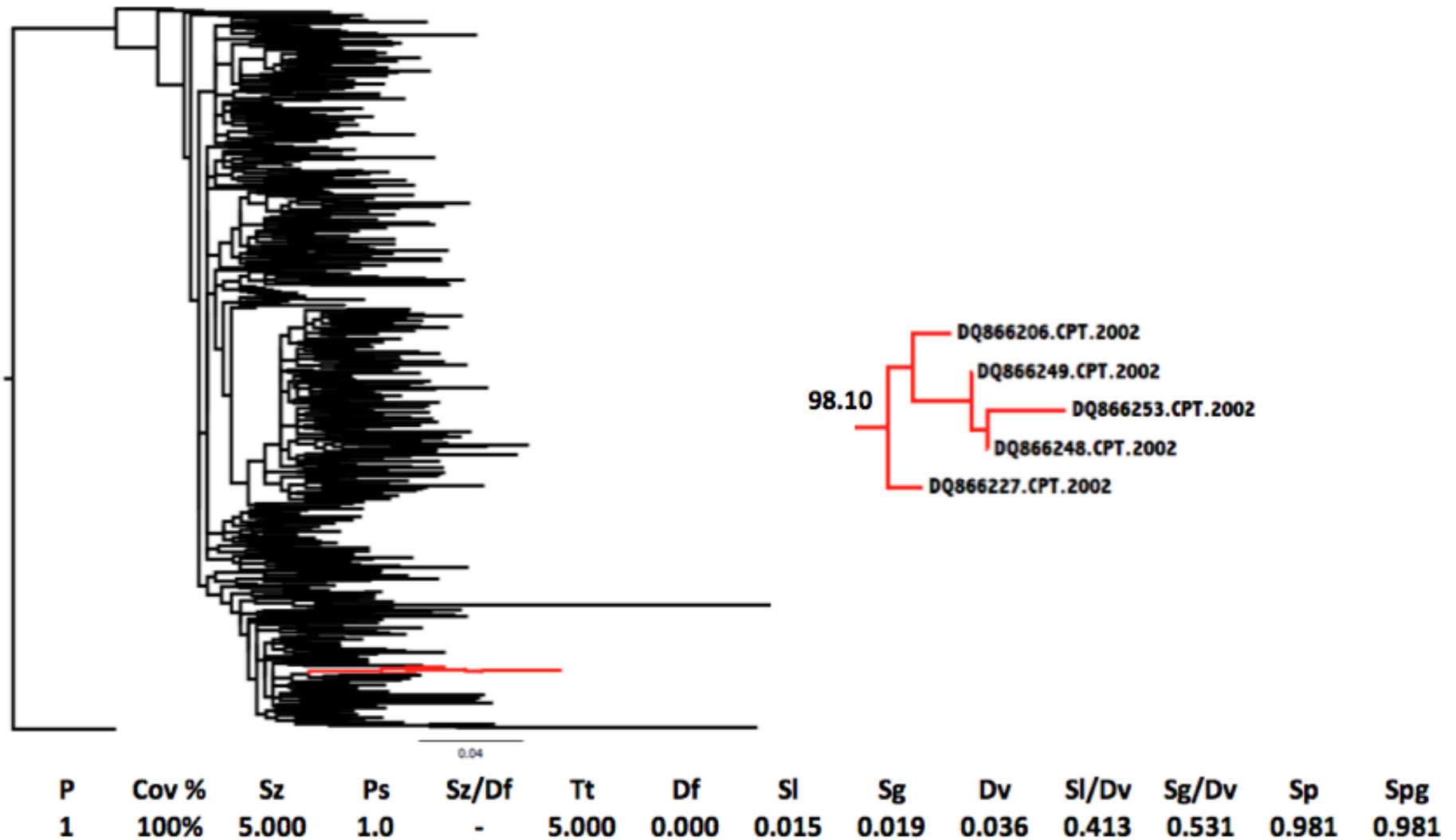


Figure 6.28: ML-tree topology of the *gag*.cluster.5 data set with aLRT. All 5 isolates clustered in a single monophyletic clade. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in phyML with the use of the HKY85 model of nucleotide substitution and aLRT. The aLRT support for the internal branch of the large monophyletic cluster of Cape Town sequences is 98,1%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

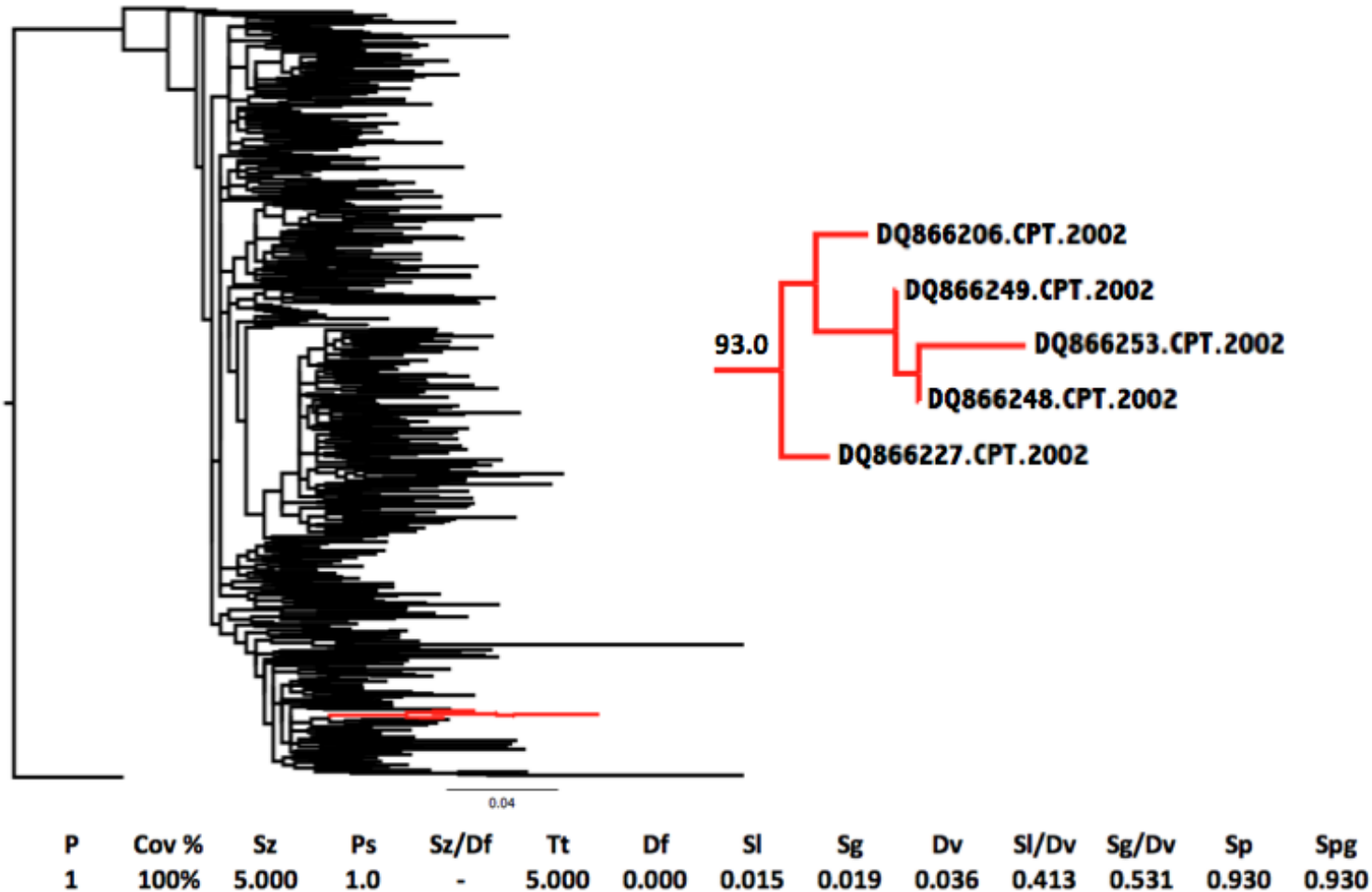


Figure 6.29: ML-tree topology of the *gag.cluster.5* data set with bootstrap resampling. All 5 of the Cape Town isolates clustered in a monophyletic cluster. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in phyML with the use of the HKY85 model of nucleotide substitution and a total of 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 93,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

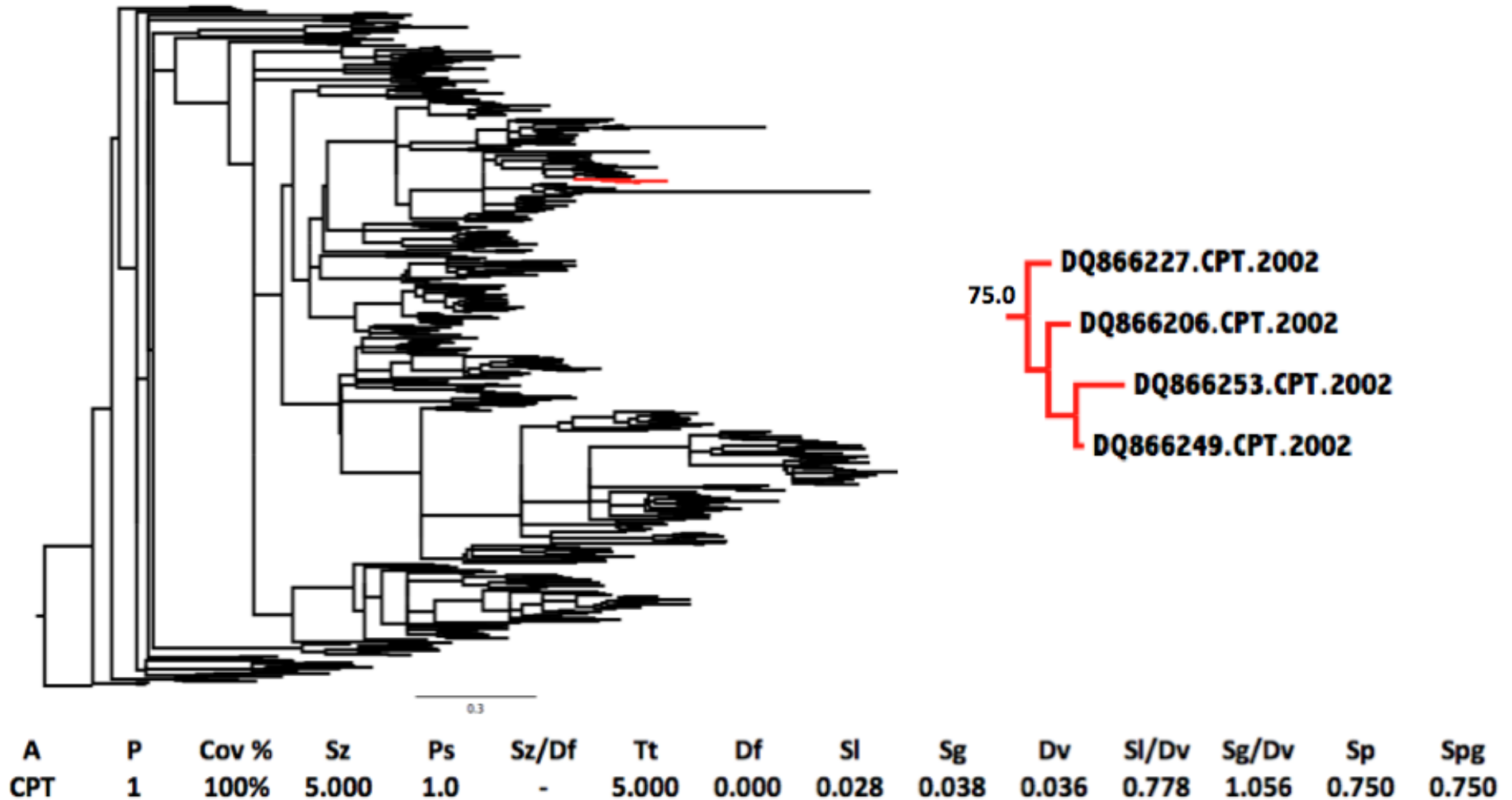


Figure 6.30: Bayesian tree topology of the *gag*.cluster.5 data set. All of the 5 Cape Town taxa clustered in a monophyletic cluster. This tree was inferred from an alignment totalling 441 nucleotide base pairs in length, stretching from position 1258 to 1698 in the HIV-1 genome relative to HXB2. The tree was inferred in MrBayes with the use of the GTR model of nucleotide substitution. The posterior support for the internal branch of the large monophyletic cluster of Cape Town sequences is 75,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

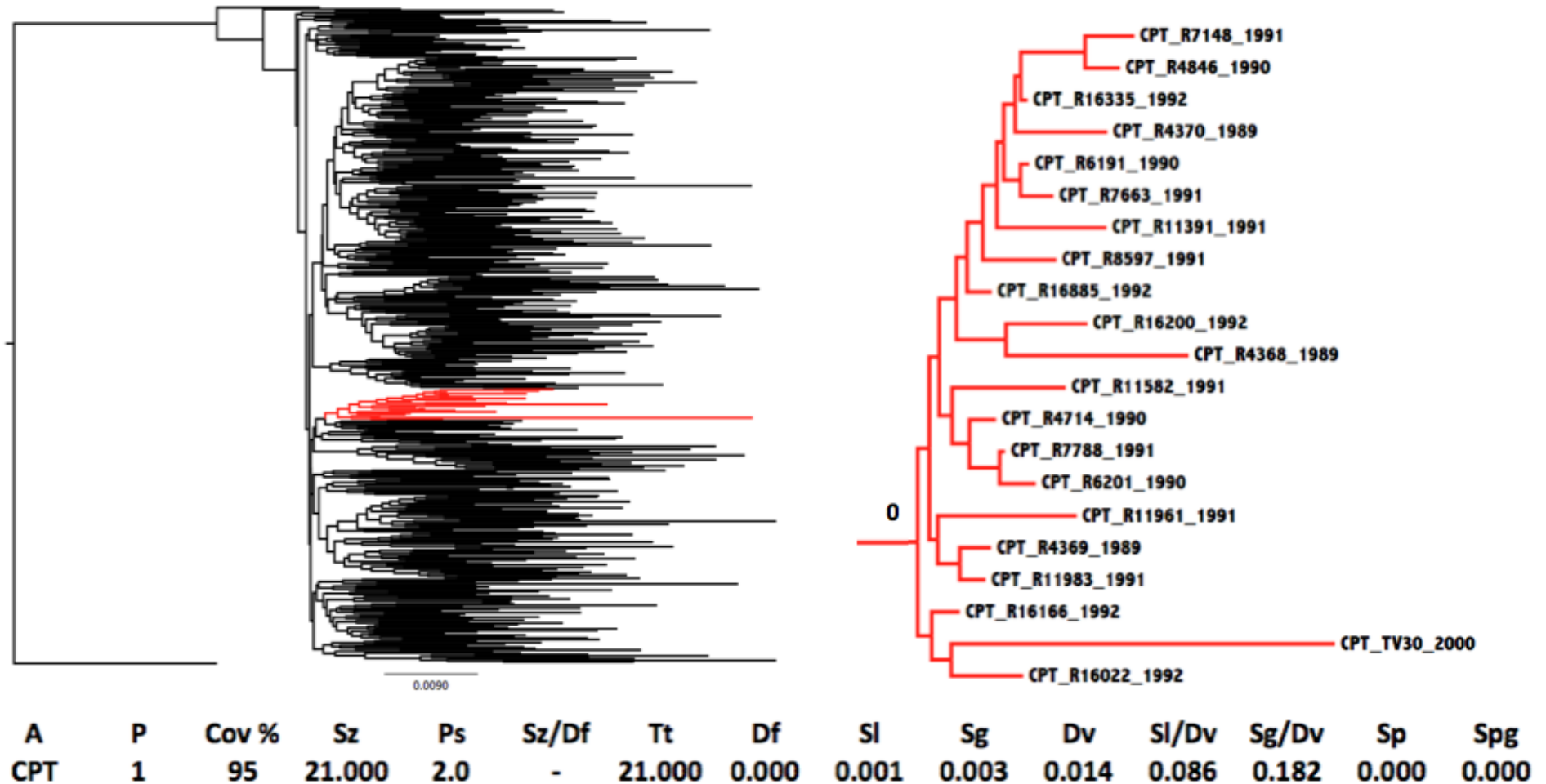
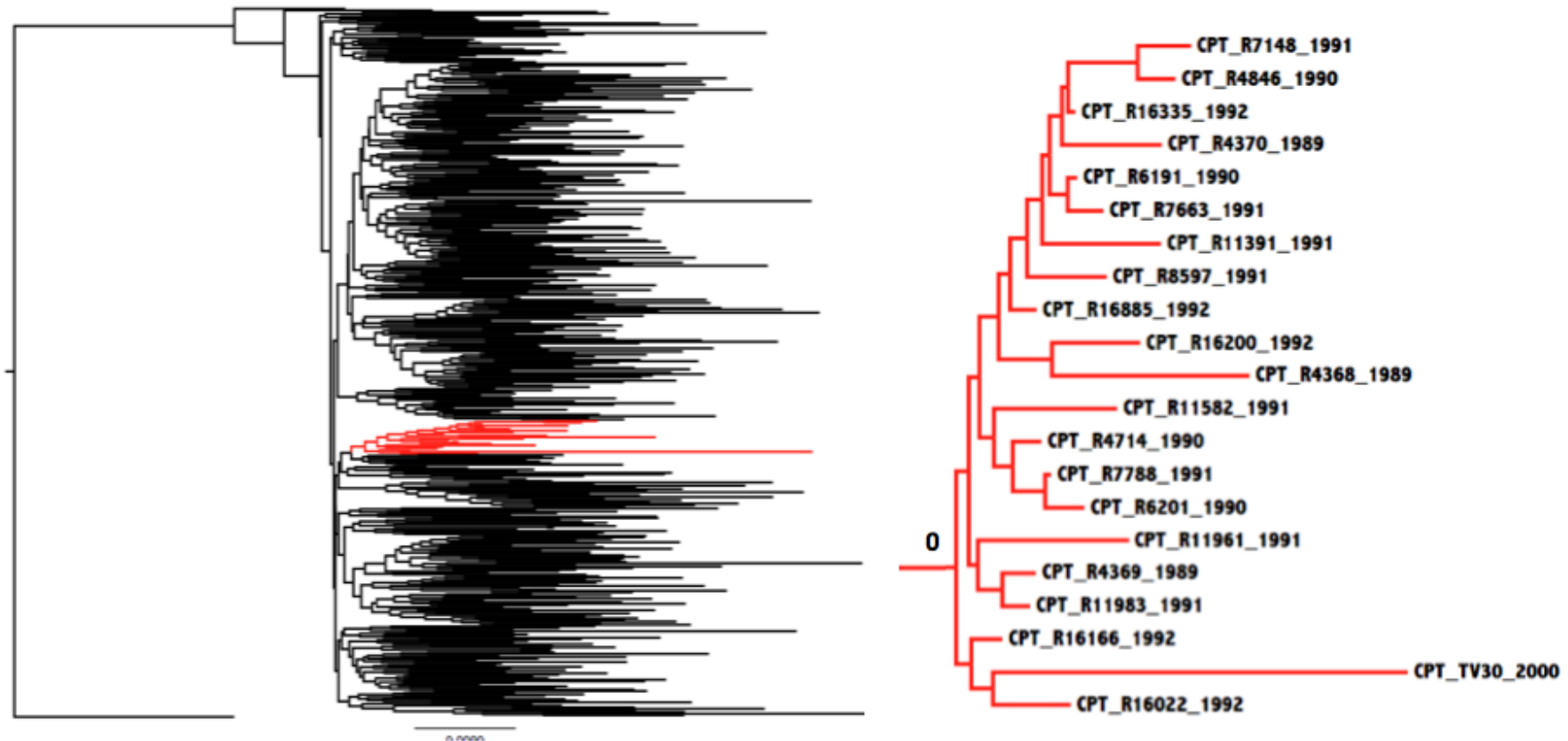


Figure 6.31: NJ-tree topology of the *pol.cluster.1* data set with bootstrap resampling. Only 21 out of the 22 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in MEGA v 5.0 with the K2P model on nucleotide substitution. Bootstrap resampling was performed on the tree topology totalling 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.



A	P	Cov %	Sz	Ps	Sz/Df	Tt	Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
CPT	1	95	21.000	2.0	-	21.000	0.000	0.001	0.003	0.014	0.086	0.182	0.000	0.000

Figure 6.32: ME-tree topology of the *pol.cluster.1* data set with bootstrap resampling. Only 21 of the 22 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in fastME with the K2P model on nucleotide substitution. Bootstrap resampling was performed on the tree topology totalling 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

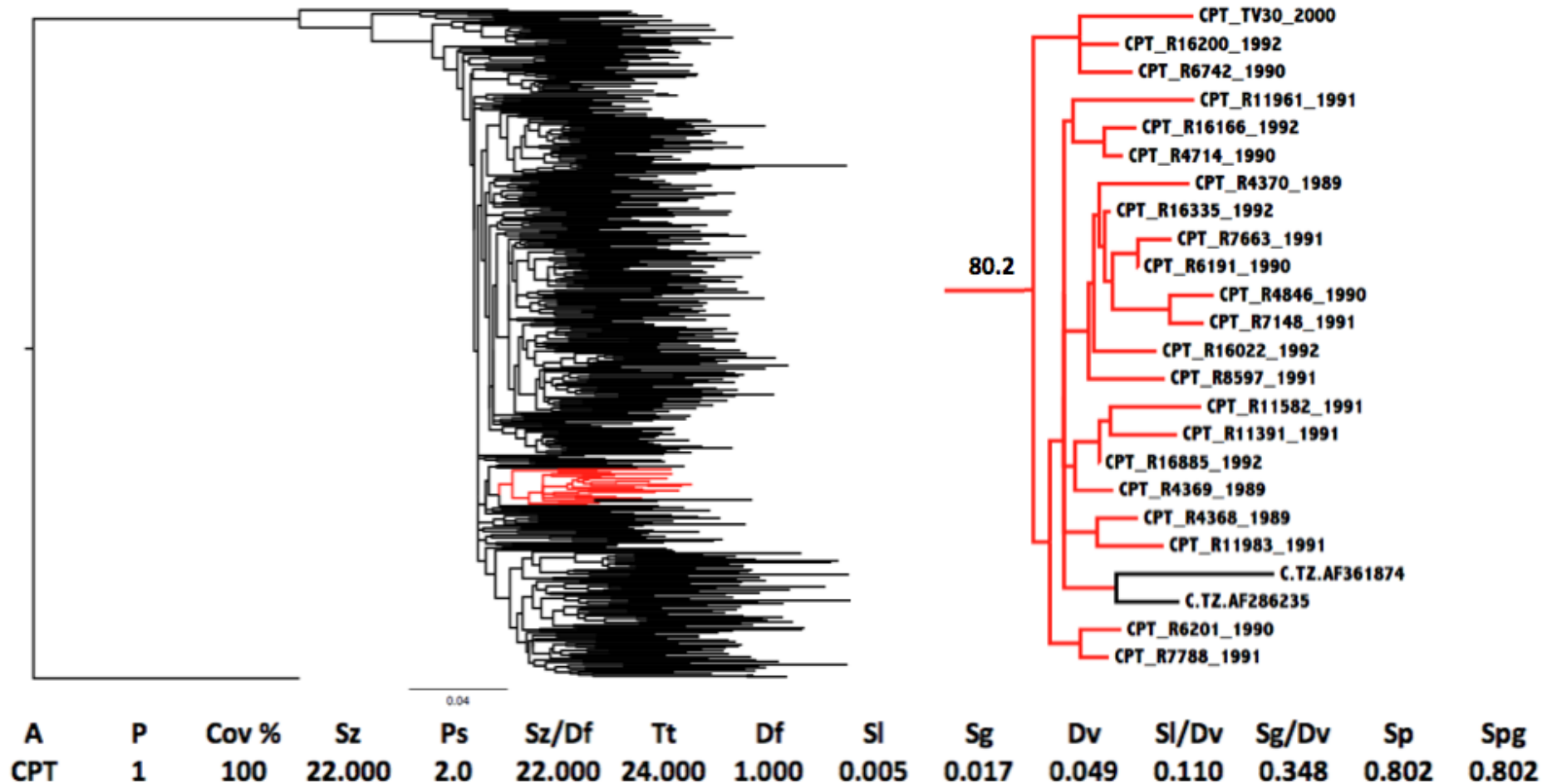


Figure 6.33: ML-tree topology of the *pol.cluster.1* data set with aLRT. All 22 of the Cape Town isolates in the data set clustered in a single monophyletic clade broken once by two isolates from Tanzania. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in phyML with the HKY85 model on nucleotide substitution. An aLRT was performed to assess confidence for each of the internal nodes. The aLRT support for the internal branch of the large monophyletic cluster of Cape Town sequences is 80,2%. Cape Town sequences are marked in red. The results of the Phylotype analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

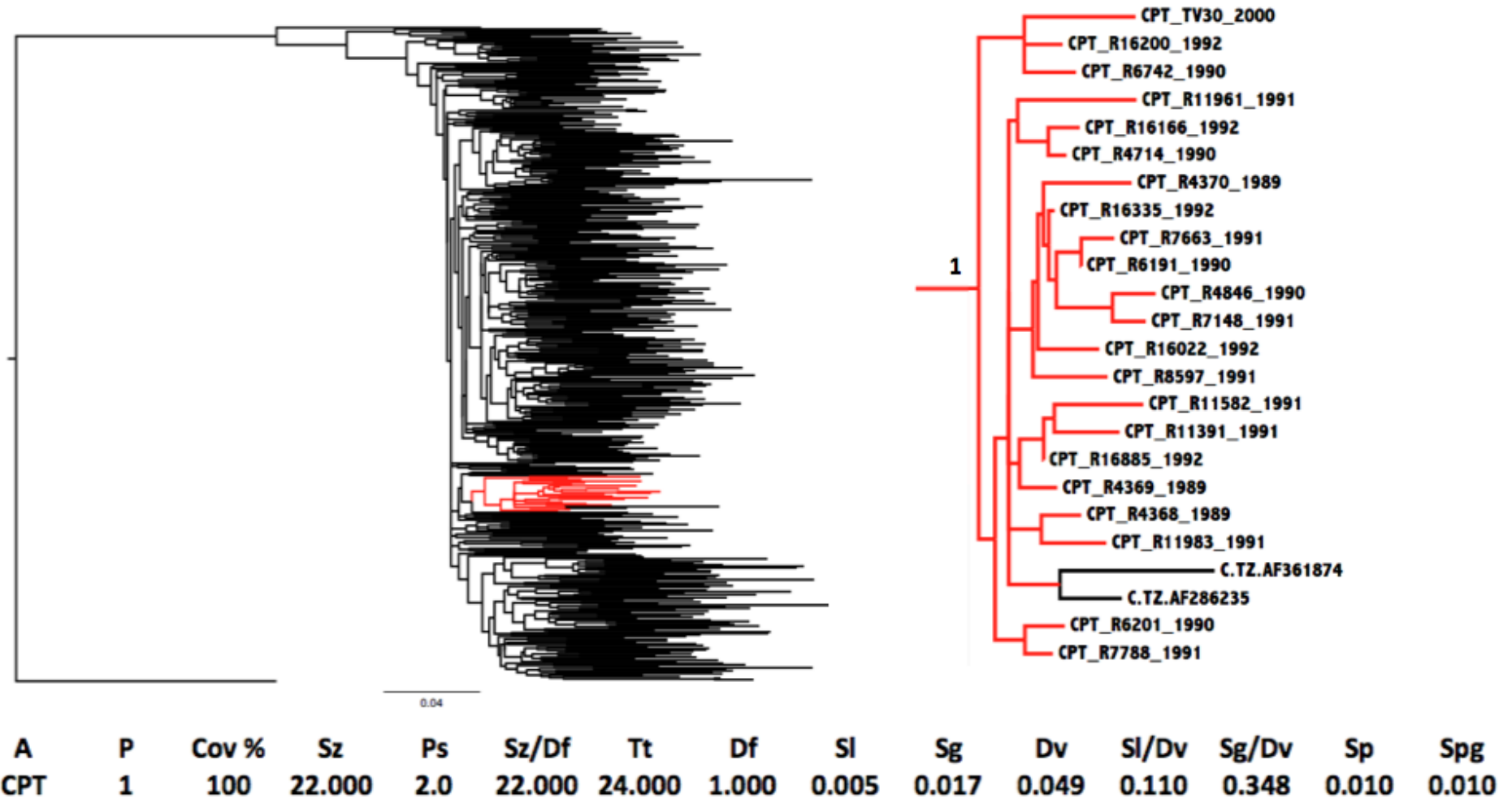


Figure 6.34: ML-tree topology of the *pol.cluster.1* data set with bootstrap resampling. All 22 of the Cape Town isolates in the data set clustered in a single monophyletic clade broken once by two isolates from Tanzania. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in phyML with the HKY85 model on nucleotide substitution. Bootstrap resampling was performed on the tree topology totalling 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 1,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

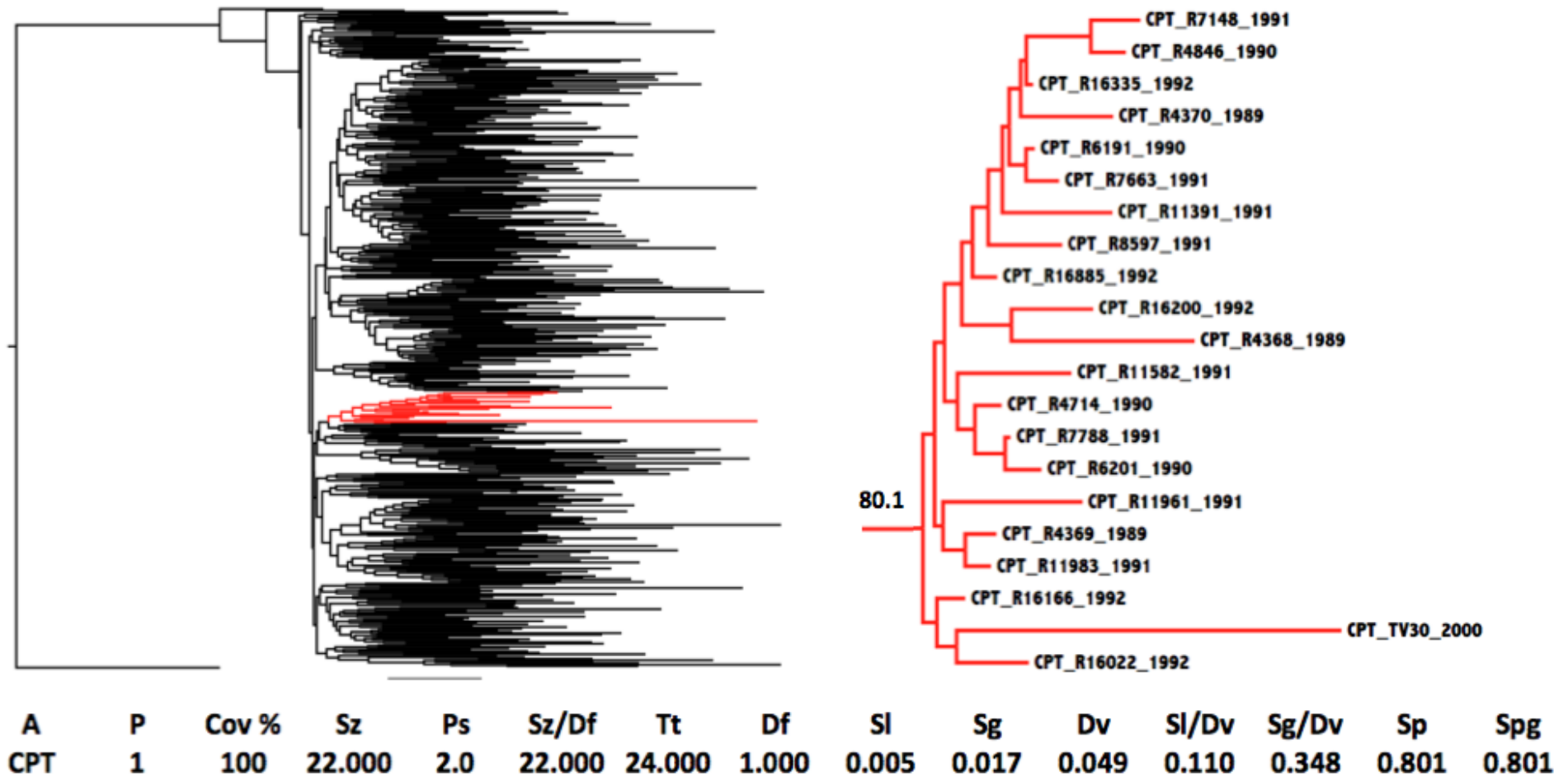


Figure 6.35: Bayesian tree topology of the *pol.cluster.1* data set. All of the 22 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in MrBayes with the GTR model on nucleotide substitution. Posterior support values are indicated as fractions for the internal node of the Cape Town cluster. The posterior support for the internal branch of the large monophyletic cluster of Cape Town sequences is 80,1%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

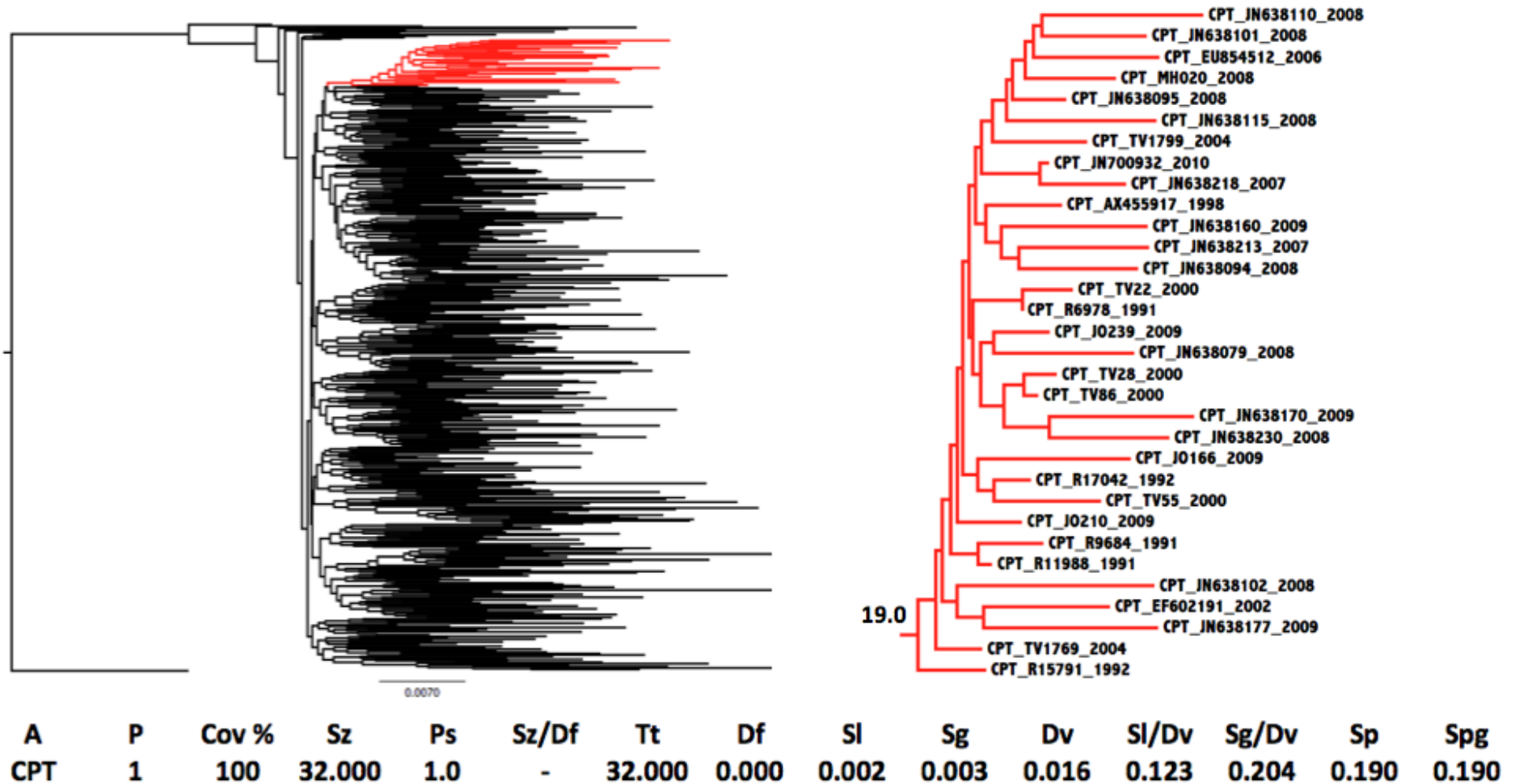


Figure 6.36: NJ-tree topology of the *pol.cluster.2* data set with bootstrap resampling. All of the 32 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in MEGA v 5.0 with the K2P model on nucleotide substitution. Bootstrap resampling was performed on the tree topology totalling 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 19,0%. Cape Town sequences are marked in red. The results of the Phylotype analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

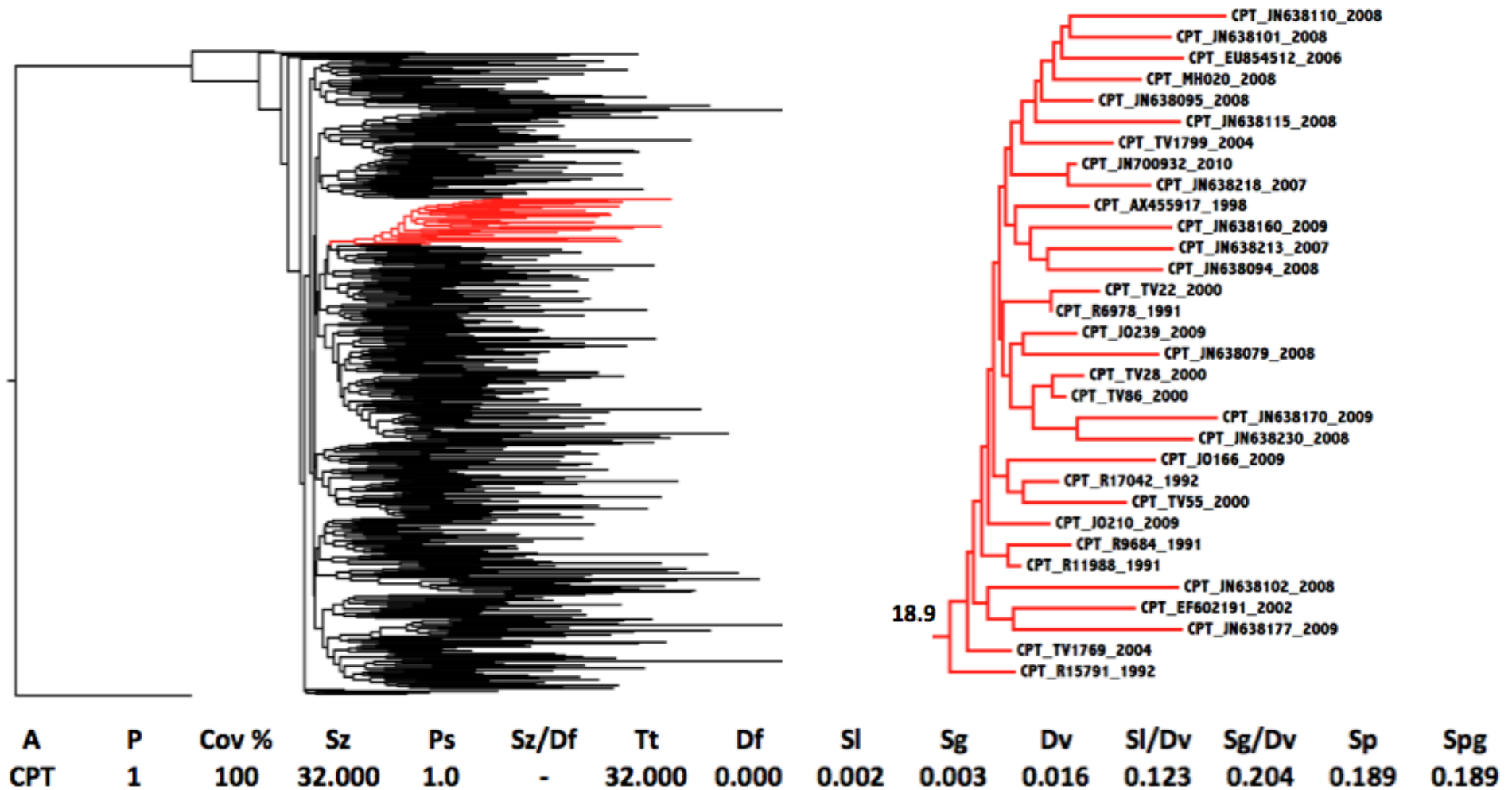


Figure 6.37: ME-tree topology of the *pol.cluster.2* data set with bootstrap resampling. All of 32 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in fastME with the K2P model on nucleotide substitution. Bootstrap resampling was performed on the tree topology totalling 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 18,9%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

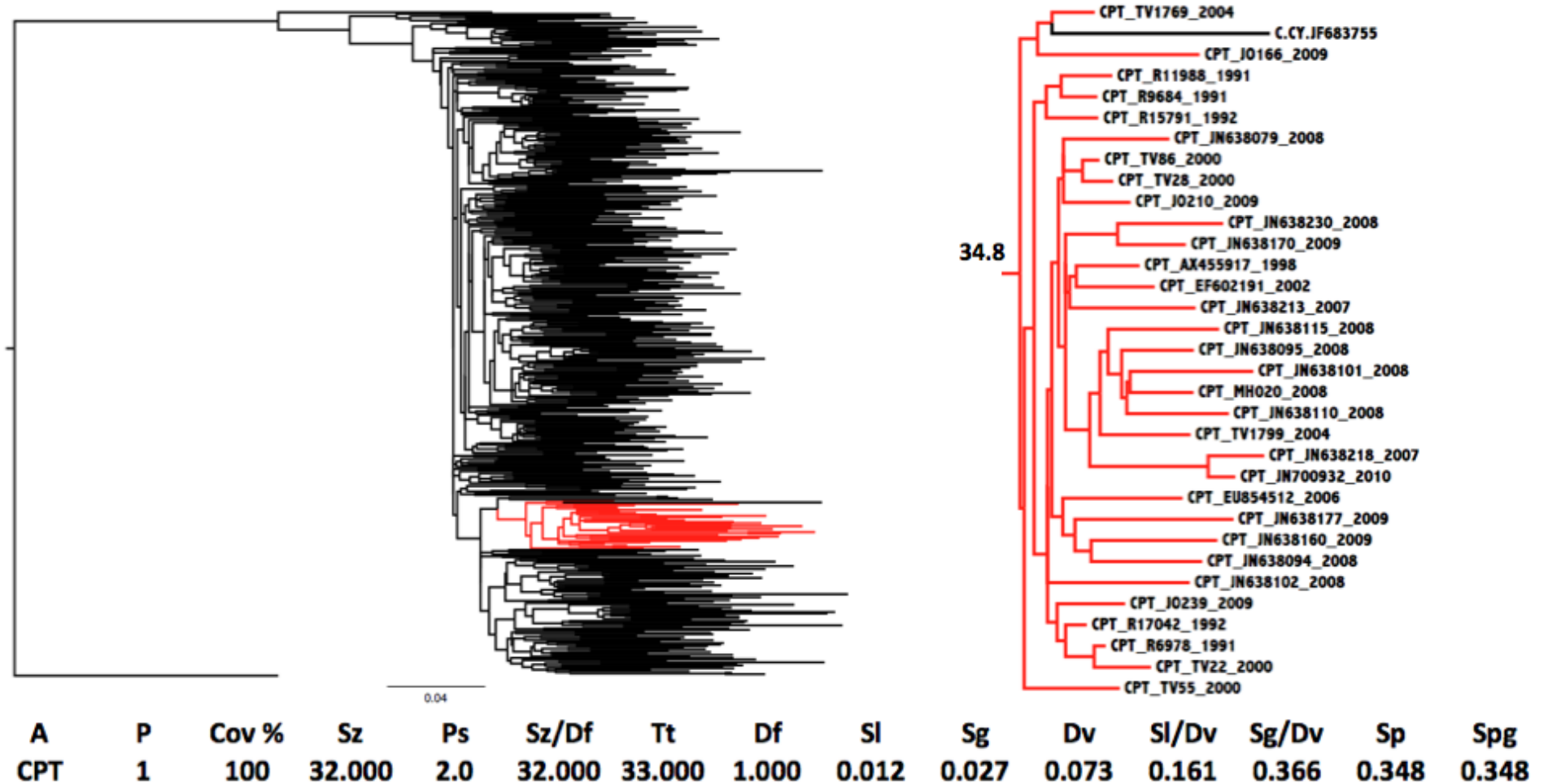


Figure 6.38: ML-tree topology of the *pol.cluster.2* data set with aLRT. All of the 32 Cape Town isolates in the data set clustered in a single monophyletic clade broken once by an isolate from Cyprus. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in phyML with the HKY85 model on nucleotide substitution. An aLRT was performed to assess confidence for each of the internal nodes. The aLRT support for the internal branch of the large monophyletic cluster of Cape Town sequences is 34,8%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

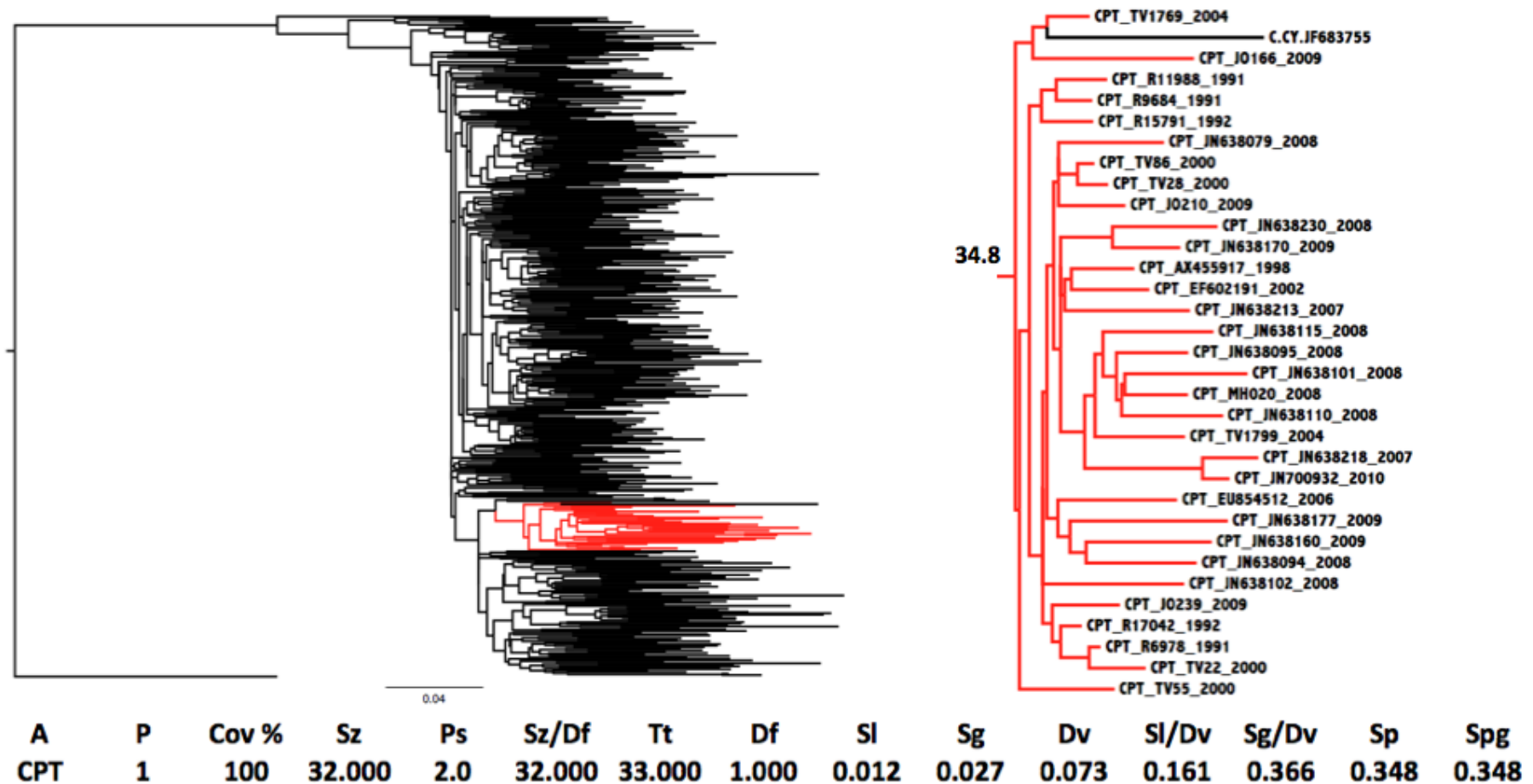


Figure 6.39: ML-tree topology of the *pol.cluster.2* data set with bootstrap resampling. All of the 32 Cape Town isolates in the data set clustered in a single monophyletic clade broken once by an isolate from Cyprus. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in phyML with the HKY85 model on nucleotide substitution. Bootstrap resampling was performed on the tree topology totalling 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 34,8%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

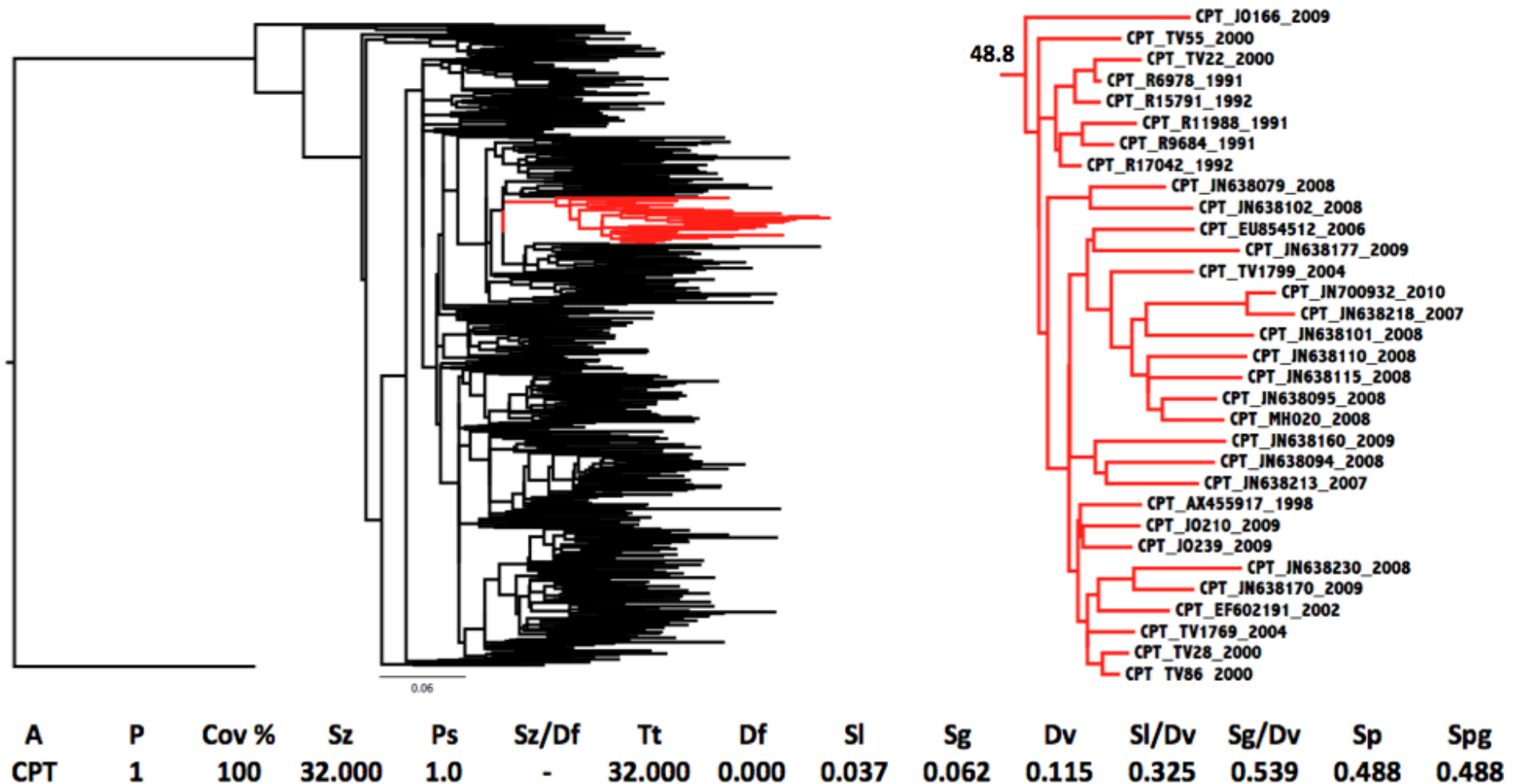


Figure 6.40: Bayesian tree topology of the *pol.cluster.2* data set. All of the 32 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in MrBayes with the GTR model on nucleotide substitution. Posterior support values are indicated as fractions for the internal node of the Cape Town cluster. The posterior support for the internal branch of the large monophyletic cluster of Cape Town sequences is 48,8%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

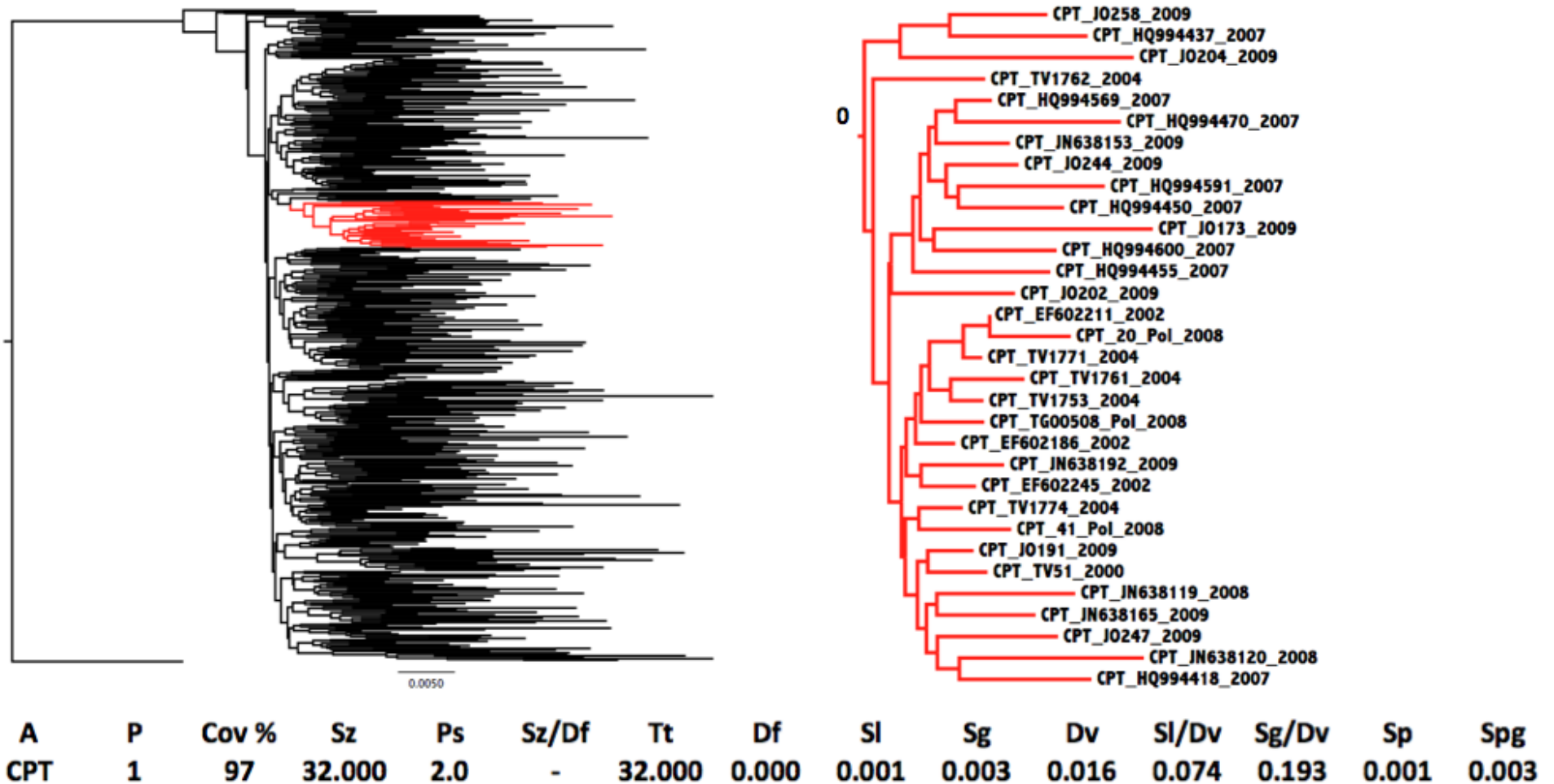


Figure 6.41: NJ-tree topology of the *pol.cluster.3* data set with bootstrap resampling. Only 32 out of the 33 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in MEGA v 5.0 with the K2P model on nucleotide substitution. Bootstrap resampling was performed on the tree topology totalling 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the Phylotype analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

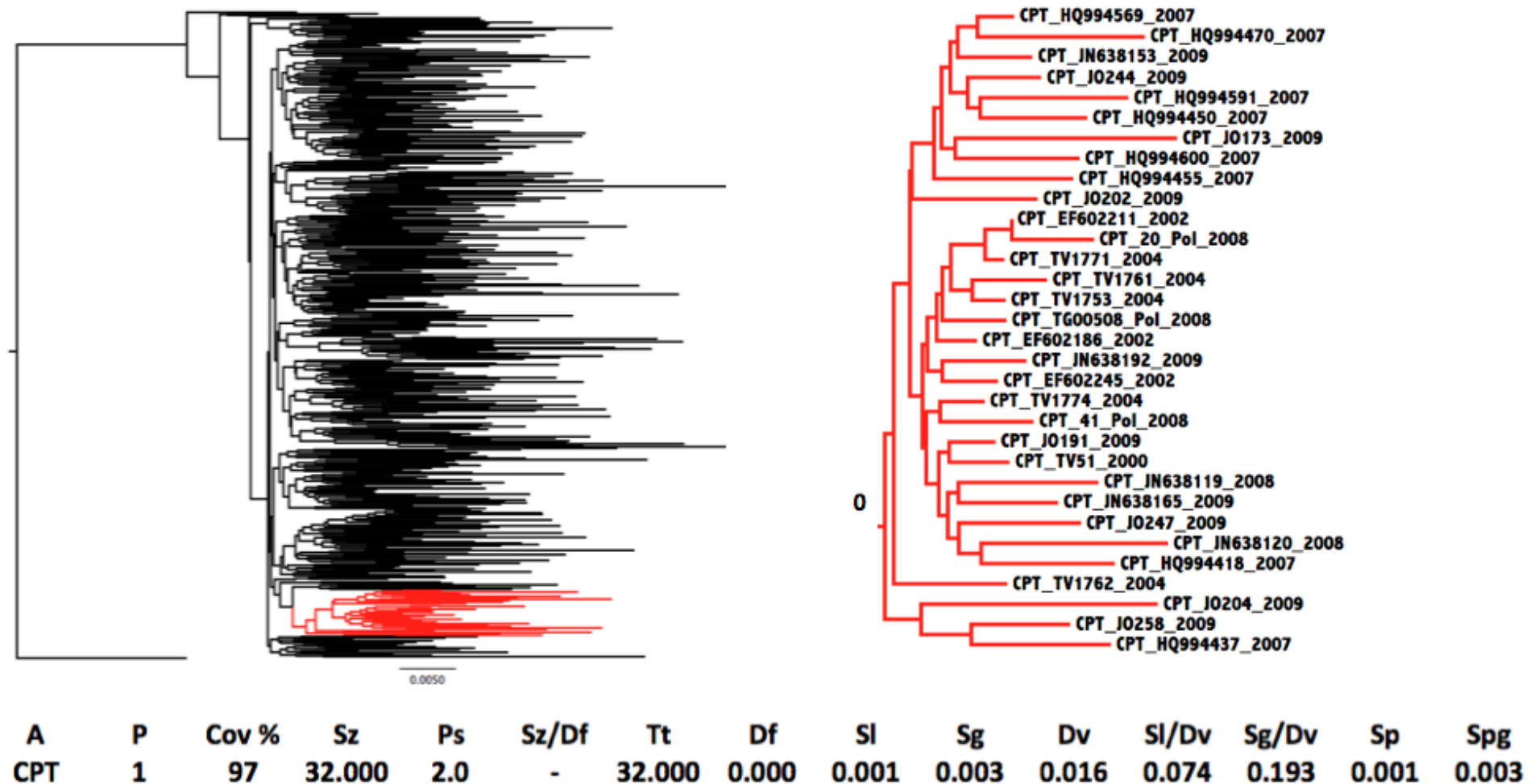


Figure 6.42: ME-tree topology of the *pol.cluster.3* data set with bootstrap resampling. Only 32 out of the 33 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in fastME with the K2P model on nucleotide substitution. Bootstrap resampling was performed on the tree topology totalling 1000 bootstrap replicates. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

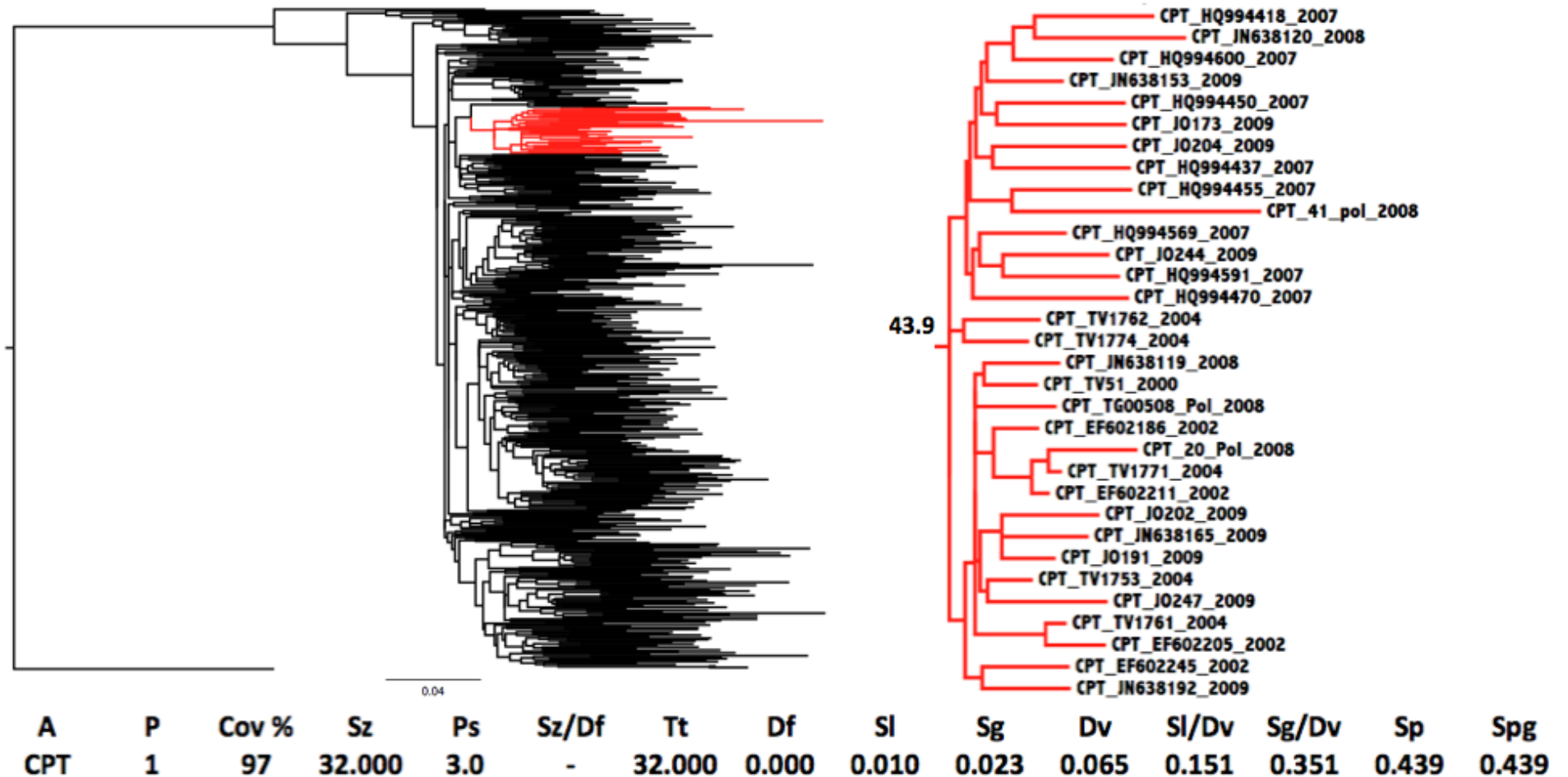


Figure 6.43: ML-tree topology of the *pol.cluster.3* data set with aLRT. Only 32 of the 33 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in phyML with the HKY85 model on nucleotide substitution. An aLRT was performed to assess confidence for each of the internal nodes. The aLRT support for the internal branch of the large monophyletic cluster of Cape Town sequences is 43,9%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

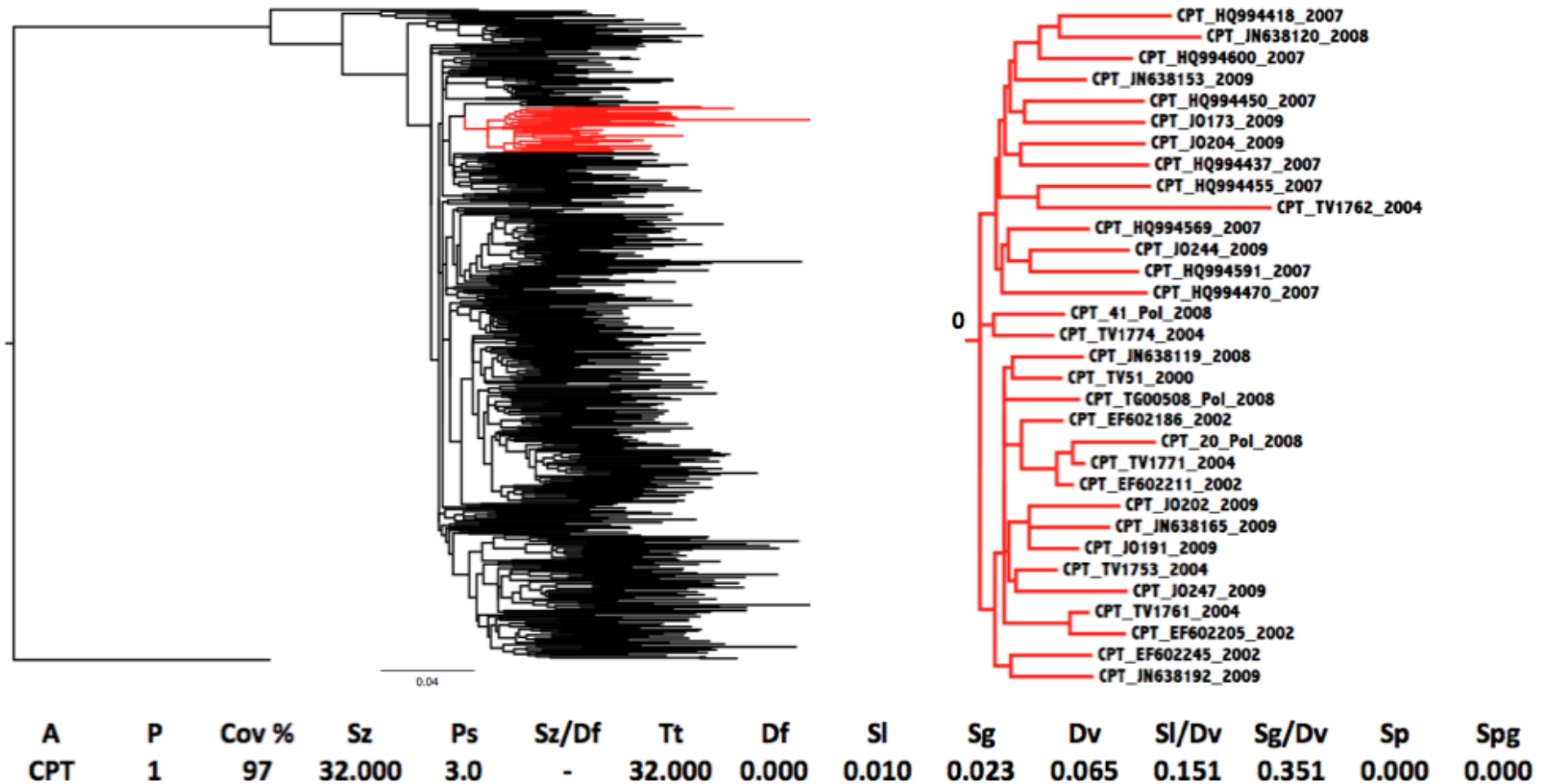


Figure 6.44: ML-tree topology of the *pol.cluster.3* data set with bootstrap resampling. All 32 of the 33 Cape Town isolates in the data set clustered in a single monophyletic clade. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in phyML with the HKY85 model on nucleotide substitution. Bootstrap resampling, totalling 1000 bootstrap replicates, was performed. The bootstrap support for the internal branch of the large monophyletic cluster of Cape Town sequences is 0,0%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.

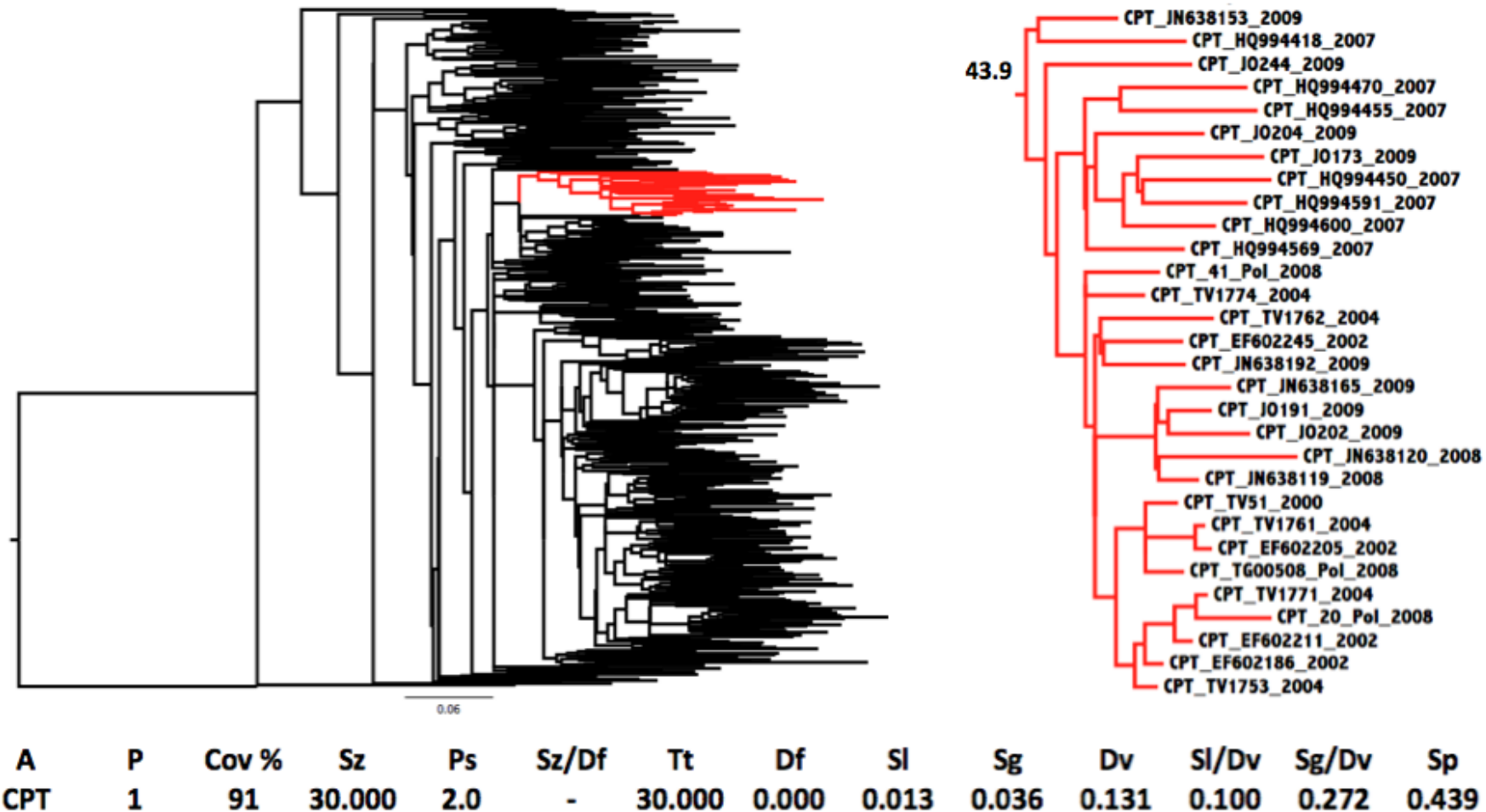


Figure 6.45: Bayesian tree topology of the *pol.cluster.3* data set. Only 30 of the 33 Cape Town isolates in the data set clustered in a single monophyletic clade, broken once by an isolate from Botswana. The alignment that was used for the inference of this tree topology was 963 bp long, stretching from position 2265 to 3227 of the HIV-1 genome, relative to HXB2. This tree topology was inferred in MrBayes with the GTR model on nucleotide substitution. The posterior support for the internal branch of the large monophyletic cluster of Cape Town sequences is 43,9%. Cape Town sequences are marked in red. The results of the PhyloType analysis of the Cape Town cluster have been tabulated and can be seen at the bottom of the figure.