The influence of HIV-1 genomic target region selection and sequence length on the accuracy of inferred phylogenies and clustering outcomes

by

Zandile Sibisi

Submitted in fulfillment for the degree in Masters of Medical Science

In the College of Health Sciences
Department Of Medical Virology
School of Laboratory Medicine and Medical Sciences
University of KwaZulu-Natal

Supervisor: Professor Tulio de Oliveira

Date: April 2017
AUTHOR’S DECLARATION

The experimental work described in this dissertation was carried out in the School of Laboratory Medicine and Medical Sciences, in the faculty of Health Sciences, at the University of KwaZulu-Natal, from June 2014 to April 2017, under the supervision of Professor Tulio de Oliveira.

This dissertation submitted for the degree of Masters in Medical Science, is the candidate’s original work and has not been submitted, in part or in whole for a degree or diploma to any other university. Where use has been made of the work of others, it has been accordingly acknowledged in the text. My contribution to the project was from protocol design, DNA sequencing, programming and formulating code, bioinformatical analysis, and thesis writing. The contributions of others to the project included the assistance in data generation, data analysis, and evaluation of the present dissertation.

........................................................................................................

Zandile Sibisi                  Date: 18 April 2017

I certify that the above statement is correct.
PLAGIARISM DECLARATION

I, ……Zandile Sibisi…………………………………………………………., declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons’ data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   a. Their words have been re-written but the general information attributed to them has been referenced
   b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed …………………………………………………………………………..
Date ……18 April 2017……………………………………………………
ETHICS

The ethics for the study was approved by Biomedical Research Ethics committee of the University of KwaZulu-Natal (ref. BF052/10).
ACKNOWLEDGEMENTS

To my supervisor Prof. Tulio de Oliveira and analytical advisor Dr. Tiago Graf, a very earnest and sincere thank you – for your patience, your time and your dedication to this project. Tulio, the success of this project was possible only because of your tremendous generosity, wisdom and guidance. Tiago, I truly appreciate all of the hours of discussions and advice that you offered. I must also give thanks to Dr. Eduan Wilkinson for his guidance and advice with the technical aspects of this project. Thank you for your kindness and patience and for taking the time to help me even with your hectic schedule.

I would like to acknowledge the staff at Africa Centre for Health and Population Studies, particularly Dr. Justen Manasa and Dr. Siva Danaviah, your assistance and imparting of DNA sequencing expertise was invaluable. To the bioinformatics course coordinators at SANBI, the analytical training I received was crucial for conducting my phylogenetic investigations. Thank you also to the Medical Research Council for funding the generation of data for this project.

To my mother, your unwavering support and motivation throughout the conduction of this study is what propelled me to remain dedicated and absorbed in my research, this has resulted in the ultimate delivery of a commendable study.
Siva Danaviah, Justen Manasa, Eduan Wilkinson, Suressnee Pillay, Zandile Sibisi, Sthembiso Msweli, Deenan Pillay, Tulio de Oliveira, Near Full Length HIV-1 Sequencing to Understand HIV Phylodynamics in Africa in Real Time. An abstract of the PANGEA-HIV consortium that was accepted for CROI 2015, which took place in Seattle, USA from the 23rd to the 26th of February.

Training attended

**K-RITH Interactive Biostatistics Course**: Statistical methods used in medical research, taught by Lori Chibnik, PhD, MPH, Assistant Professor at Harvard Medical School and the Harvard T.H. Chan School of Public Health and research scientist at the Broad Institute of Harvard and MIT. November 2014

**Medical Educational Partnership Initiative Biostatistics Workshop**: Biostatistical reasoning in health research, taught by Professor Mary Lou Thompson, Department of Biostatistics, School of Public Health, University of Washington. 26 January – 6 February 2015.

**National Bioinformatics Course**: Bioinformatics Support Platform Introduction to Bioinformatics Course provided by South African National Bioinformatics Institute (SANBI) and the Department of Science and Technology. 16 February – 2 April 2015.

**K-RITH Introduction to R Programming and Intermediate Biostatistics**: Regression techniques applied to various data types and research questions, taught by Lori Chibnik, PhD, MPH, Assistant Professor at Harvard Medical School and the Harvard T.H. Chan School of Public Health and research scientist at the Broad Institute of Harvard and MIT. 10 – 18 March 2016.
# TABLE OF CONTENTS

**AUTHOR’S DECLARATION** ................................................................................................................................. ii

**PLAGIARISM DECLARATION** .......................................................................................................................... iii

**ETHICS** ................................................................................................................................................. iv

**ACKNOWLEDGEMENTS** ................................................................................................................................. v

**THESIS OUTPUT** ........................................................................................................................................... vi

**TABLE OF CONTENTS** .................................................................................................................................... vii

**LIST OF FIGURES** ........................................................................................................................................... x

**LIST OF TABLES** ............................................................................................................................................ xii

**LIST OF ABBREVIATIONS** ............................................................................................................................... xiii

**ABSTRACT** ....................................................................................................................................................... xvii

**CHAPTER ONE: INTRODUCTION** ................................................................................................................... 18

  Justification .................................................................................................................................................. 19

  1. Literature Review ................................................................................................................................. 21

    1.1. History and present epidemiology of the HIV pandemic ............................................................... 21

    1.2. Virion structure and genomic organization of HIV-1 ................................................................. 22

    1.3. Genetic diversity of HIV-1 ............................................................................................................. 24

    1.4. HIV-1 phylogenetic analysis pipeline ............................................................................................ 26

      1.4.1. An overview of phylogenetics ................................................................................................. 27

      1.4.2. Generation of viral sequences and retrieval of homologous genetic information ... 29

      1.4.3. Sequence alignments .................................................................................................................. 30

      1.4.4. Nucleotide substitution models ................................................................................................ 31

      1.4.5. Phylogenetic tree structure and inference methods ................................................................. 32

    1.5. Application of HIV-1 phylogenetic analysis in modern research ........................................... 34

    1.6. HIV-1 transmission cluster analysis and its shortcomings ....................................................... 35

    1.7. Factors that may influence the outcomes of viral clustering .................................................... 39

      1.7.1. Target gene selection and sequence length ............................................................................. 41

      1.7.2. Contrasting the usage of full genome and sub-genomic sequences for HIV-1 cluster analysis .................................................................................................................................................. 42

    1.8. Assessing the accuracy of phylogenetic trees ..................................................................................... 45
CHAPTER TWO: METHODOLOGY ................................................................. 50
2.1. Ethics statement ............................................................................... 50
2.2. Reagents and equipment .................................................................. 51
2.3. HIV-1 full-genome sequences .......................................................... 52
   2.3.1. Full-genome sequences from the LANL HIV Database .................. 52
   2.3.2. Full-genome HIV-1 sequences from a South African cohort ............ 52
      2.3.2.1. Sample collection, transport and processing ............................ 53
      2.3.2.2. RNA extraction .................................................................. 53
      2.3.2.3. Amplicon generation ......................................................... 54
      2.3.2.5. PCR purification ............................................................... 56
      2.3.2.6. DNA quantification ............................................................. 57
      2.3.2.7. Next generation DNA Sequencing ........................................ 57
      2.3.2.8. Assembly and consensus generation .................................... 58
2.4. Multiple sequence alignment ............................................................ 58
2.5. Analyzed sub-genomic regions of the HIV-1 genome ....................... 59
2.6. Sliding window analysis ...................................................................... 59
2.7. Phylogenetic inference ....................................................................... 60
2.8. Assessment of the accuracy of the inferred phylogenies ..................... 61
   2.8.1. Estimation of tree certainty ....................................................... 61
   2.8.2. Estimation of Shimodaira–Hasegawa [SH]-aLRT .......................... 61
   2.8.3. Evaluation of cluster quality .................................................... 61
2.9. Cluster enumeration .......................................................................... 62
   2.9.1. Identification of phylogenetic clusters using PhyloPart .................. 62
   2.9.2. Identification of phylogenetic clusters using PhyloType StandAlone PST07 .... 62

CHAPTER THREE: RESULTS ........................................................................ 64
3.1 Accuracy of inferred phylogenies ........................................................ 64
   3.1.1. Hierarchy of tree certainty ......................................................... 64
   3.1.2. Hierarchy of quality of clustering ............................................... 66
      3.1.2.1. Quantified by subtype diversity ratio ..................................... 67
3.1.2.2. Quantified by subtype diversity variance .......................................................... 67

3.1.3. Hierarchy of Shimodaira-Hasegawa (SH) – like support ....................................... 68

3.2. Extent of HIV clustering across the HIV-1 genome ................................................. 69

3.2.1. Clusters enumerated by PhyloPart in each genomic region phylogeny .................. 69

3.2.2. Clusters enumerated by PhyloType in each genomic region phylogeny ............... 71

3.2.3. Clusters enumerated by PhyloType in each sliding window length phylogeny ..... 81

CHAPTER FOUR: DISCUSSION .......................................................................................... 83

CHAPTER FIVE: APPENDICES ...................................................................................... 87

Appendix 1: Consent form in IsiZulu ............................................................................. 87

Appendix 2: Consent form in English ........................................................................... 88

REFERENCES ................................................................................................................ 89
LIST OF FIGURES

Figure 1: HIV-1 virion structure. ................................................................. 23
Figure 2: A diagrammatical representation of the genome layout of HIV-1 .......... 23
Figure 3: Features associated with HIV-1 variability .................................... 25
Figure 4: Phylogenetic topology of the nine (A, B, C, D, F, G, H, J, K) pure subtypes of HIV-1 group M ................................................................. 26
Figure 5: A breakdown of the basic steps involved in any phylogenetic investigation........ 28
Figure 6: DNA sequences given in the form of a multiple sequence alignment.......... 30
Figure 7: Phylogenetic tree structures .......................................................... 33
Figure 8: Phylogenetic tree showing evolutionary relationships and clustering between a set of viruses ................................................................. 36
Figure 9: Neighbour-joining (NJ) phylogenies constructed using sub-genomic regions a) gag and b) pol ................................................................. 44
Figure 10: Diagrammatical depiction of bootstrap analysis ................................ 47
Figure 11: Subtype diversity ratio and subtype diversity variance ....................... 48
Figure 12: Hlabisa sub-district, northern KwaZulu-Natal, showing position of primary hospital with one on-site clinic and 15 peripheral clinics ............................................. 53
Figure 13: An illustration of the location and amplicon size of each pan primer ........ 55
Figure 14: Gel visualization of the PCR DNA products .................................... 56
Figure 15: Python interface showing script to generate datasets for sliding window analysis... 60
Figure 16: An illustration of a Phylotype file annotation ................................... 63
Figure 17: Graph comparing relative tree certainty amongst phylogenies inferred from different HIV-1 gene regions ......................................................... 65
Figure 18: Subtype diversity ratio (SDR) score for each HIV-1 targeted genomic region phylogeny ................................................................. 67
Figure 19: Subtype diversity variance (SDV) score for each HIV-1 targeted genomic region phylogeny ................................................................. 67
Figure 20: Graph comparing SH-like support values amongst phylogenies inferred from different HIV-1 genomic regions ............................................. 69
Figure 21: Number of clusters in the targeted regions of the HIV-1 genome .......... 70
Figure 22: The extent of HIV clustering in each targeted HIV-1 genomic region ......... 70
Figure 23: Number of phylotypes in the targeted regions of the HIV-1 genome ........ 71
Figure 24: Percentage of strains associated with phylotypes in the targeted regions of the HIV-1 genome ................................................................. 72
Figure 25: A subset of the env phylotype map (ACCTRAN) of global HIV-1 ............ 73
Figure 26: A subset of the full genome phylotypes map (ACCTRAN) of global HIV-1 ........... 74
Figure 27: A subset of the gag phylotype map (ACCTRAN) of global HIV-1 .................... 75
Figure 28: A subset of the product 2 phylotype map (ACCTRAN) of global HIV-1 .......... 76
Figure 29: A subset of the product 4 phylotype map (ACCTRAN) of global HIV-1 ........... 77
Figure 30: A subset of the partial pol phylotype map (ACCTRAN) of global HIV-1 .......... 78
Figure 31: A subset of the partial env phylotype map (ACCTRAN) of global HIV-1 .......... 79
Figure 32: A subset of the pol phylotype map (ACCTRAN) of global HIV-1 .................... 80
Figure 33: Sliding window analysis across the HIV-1 genome depicting number of phylotypes .......................................................................................................... 81
Figure 34: Sliding window analysis across the HIV-1 genome depicting extent of clustering . 82
## LIST OF TABLES

**Table 1:** Summary of the various methods of tree construction ........................................... 27

**Table 2:** Phylogenetic methodologies of HIV-1 clustering analysis ..................................... 37

**Table 3:** Papers exploring parameters that may affect the outcomes of HIV-1 phylogenetic clustering .......................................................................................................................... 40

**Table 4:** Characteristics of target genes for phylogenetic analysis ...................................... 41

**Table 5:** List of chemicals and commercial products used in the study ................................. 51

**Table 6:** Equipment used to perform sample analysis .............................................................. 51

**Table 7:** Software programs and online analytical tools that were used in the analysis of sequence information ............................................................................................................. 52

**Table 8:** Summary of the primers for NGS amplification ......................................................... 54

**Table 9:** Master Mix for DNA amplification ............................................................................. 55

**Table 10:** Thermal cycler PCR amplification cycling conditions ............................................ 57

**Table 11:** Characteristics and hierarchy of tree certainty amongst genomic regions ............ 65

**Table 12:** Characteristics and hierarchy of SDR and SDV amongst genomic regions .......... 66

**Table 13:** Characteristics and hierarchy of SH support amongst genomic regions ............... 68
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCTRAN</td>
<td>Ancestral character state reconstruction using parsimony</td>
</tr>
<tr>
<td>ACL</td>
<td>Africa Centre Laboratory</td>
</tr>
<tr>
<td>AF</td>
<td>Afghanistan</td>
</tr>
<tr>
<td>AIDS</td>
<td>Acquired Immune Deficiency Syndrome</td>
</tr>
<tr>
<td>aLRT</td>
<td>Approximate Likelihood Ratio Test</td>
</tr>
<tr>
<td>APOBEC</td>
<td>Apolipoprotein B Editing Catalytic Polypeptide</td>
</tr>
<tr>
<td>ART</td>
<td>Antiretroviral Therapy</td>
</tr>
<tr>
<td>ATM</td>
<td>Amplicon Tagment Mix</td>
</tr>
<tr>
<td>BAMBE</td>
<td>Bayesian Analysis in Molecular Biology and Evolution (Software)</td>
</tr>
<tr>
<td>BEAST</td>
<td>Bayesian Evolutionary Analysis Sampling Trees (Software)</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>Bp</td>
<td>Base-pair</td>
</tr>
<tr>
<td>BR</td>
<td>Brazil</td>
</tr>
<tr>
<td>BW</td>
<td>Botswana</td>
</tr>
<tr>
<td>CD</td>
<td>Congo, The Democratic Republic of</td>
</tr>
<tr>
<td>CDC</td>
<td>Centre for Disease Control and Prevention</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>CM</td>
<td>Cameroon</td>
</tr>
<tr>
<td>CMV</td>
<td>Cytomegalovirus</td>
</tr>
<tr>
<td>CN</td>
<td>China</td>
</tr>
<tr>
<td>CRF</td>
<td>Circular Recombination Form</td>
</tr>
<tr>
<td>CU</td>
<td>Cuba</td>
</tr>
<tr>
<td>CY</td>
<td>Cyprus</td>
</tr>
<tr>
<td>DDBJ</td>
<td>DNA Data Bank of Japan</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>EMBL</td>
<td>European Molecular Biology Laboratory</td>
</tr>
<tr>
<td>Env</td>
<td>Envelope</td>
</tr>
<tr>
<td>ES</td>
<td>Spain</td>
</tr>
<tr>
<td>FG</td>
<td>Full Genome</td>
</tr>
<tr>
<td>FR</td>
<td>France</td>
</tr>
<tr>
<td>G</td>
<td>Gamma distribution</td>
</tr>
<tr>
<td>Gag</td>
<td>Group-specific antigen</td>
</tr>
<tr>
<td>GB</td>
<td>United Kingdom</td>
</tr>
</tbody>
</table>
GH  Ghana
GHz  Gigahertz
GP120  Envelope glycoprotein 120
GP41  Glycoprotein 41
GRID  Gay-Related Immune Deficiency
GTR  Generalized Time-Reversal
HAART  Highly Active Antiretroviral Therapy
HIV-1  Human Immunodeficiency Virus type 1
HIV-1C  Human Immunodeficiency Virus type 1 Subtype C
HIV-2  Human Immunodeficiency Virus type 2
IC  Internode Certainty
IN  India
IR  Iran
IVDU  Intravenous Drug User
JC69  Jukes and Cantor
JP  Japan
KE  Kenya
KITSCH  Fitch-Margoliash method assuming a molecular clock
KR  South Korea
KS  Kaposi's Sarcoma
LANL  Los Alamos National Laboratory
LAV  Lymphadenopathy-Associated Virus
LNA  Library Normalization Additives 1
LNB1  Library Normalization Beads 1
LNS1  Library Normalization Storage Buffer 1
LTR  Long Terminal Repeat
LU  Luxembourg
ML  Maximum Likelihood
MMWR  Morbidity and Mortality Weekly Report
MY  Malaysia
NaOH  Sodium Hydroxide
Nef  Negative Regulatory Factor
NG  Niger
NGS  Next-Generation Sequencing
NJ  Neighbour-Joining
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNI</td>
<td>Nearest Neighbour Interchange</td>
</tr>
<tr>
<td>NPN</td>
<td>Nextera PCR Master Mix</td>
</tr>
<tr>
<td>Nt</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>OTU</td>
<td>Operational Taxonomic Unit</td>
</tr>
<tr>
<td>P-Env</td>
<td>Partial Env</td>
</tr>
<tr>
<td>P-Pol</td>
<td>Partial Pol</td>
</tr>
<tr>
<td>PAUP</td>
<td>Phylogenetic Analysis Using Parsimony (Software)</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PE</td>
<td>Peru</td>
</tr>
<tr>
<td>PHYLIP</td>
<td>Phylogeny Inference Package (Software)</td>
</tr>
<tr>
<td>PHYML</td>
<td>Phylogenetic Maximum Likelihood analyses (Software)</td>
</tr>
<tr>
<td>Pol</td>
<td>Polymerase</td>
</tr>
<tr>
<td>Pro-2</td>
<td>Product 2</td>
</tr>
<tr>
<td>Pro-4</td>
<td>Product 4</td>
</tr>
<tr>
<td>Prot</td>
<td>Protease</td>
</tr>
<tr>
<td>RAxML</td>
<td>Randomized Axelerated Maximum Likelihood (Software)</td>
</tr>
<tr>
<td>RELL</td>
<td>Resampling of estimated log likelihoods</td>
</tr>
<tr>
<td>Rev</td>
<td>Transactivating protein</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RSB</td>
<td>Resuspension Buffer</td>
</tr>
<tr>
<td>RT</td>
<td>Reverse Transcriptase</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>Reverse Transcriptase - Polymerase chain reaction</td>
</tr>
<tr>
<td>RU</td>
<td>Russia</td>
</tr>
<tr>
<td>RW</td>
<td>Rwanda</td>
</tr>
<tr>
<td>SDR</td>
<td>Subtype Diversity Ratio</td>
</tr>
<tr>
<td>SDV</td>
<td>Subtype Diversity Variance</td>
</tr>
<tr>
<td>SE</td>
<td>Sweden</td>
</tr>
<tr>
<td>SH</td>
<td>Shimodaira-Hasegawa</td>
</tr>
<tr>
<td>SH-aLRT</td>
<td>Shimodaira-Hasegawa-approximate likelihood ratio test</td>
</tr>
<tr>
<td>SPR</td>
<td>Subtree Pruning and Regrafting</td>
</tr>
<tr>
<td>T cell</td>
<td>Thymus derived lymphocyte cell</td>
</tr>
<tr>
<td>Tat</td>
<td>Transcriptional Transactivator</td>
</tr>
<tr>
<td>TBE</td>
<td>Tris buffer solution</td>
</tr>
<tr>
<td>TC</td>
<td>Tree Certainty</td>
</tr>
<tr>
<td>TD</td>
<td>Tagment DNA Buffer</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Name</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>TH</td>
<td>Thailand</td>
</tr>
<tr>
<td>TRIM5</td>
<td>Human Tripartite Motif-Containing Protein 5</td>
</tr>
<tr>
<td>TZ</td>
<td>Tanzania</td>
</tr>
<tr>
<td>UA</td>
<td>Ukraine</td>
</tr>
<tr>
<td>UG</td>
<td>Uganda</td>
</tr>
<tr>
<td>UPGMA</td>
<td>Unweighted Pair Group Method with Arithmetic Mean</td>
</tr>
<tr>
<td>URF</td>
<td>Unique Recombinant Form</td>
</tr>
<tr>
<td>US</td>
<td>United States of America</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>UZ</td>
<td>Uzbekistan</td>
</tr>
<tr>
<td>Vif</td>
<td>Viral Infectivity Factor</td>
</tr>
<tr>
<td>VN</td>
<td>Vietnam</td>
</tr>
<tr>
<td>Vpr</td>
<td>Viral Protein R</td>
</tr>
<tr>
<td>Vpu</td>
<td>Viral Protein U</td>
</tr>
<tr>
<td>WPGMA</td>
<td>Weighted Pair Group Method with Arithmetic Mean</td>
</tr>
<tr>
<td>ZA</td>
<td>South Africa</td>
</tr>
</tbody>
</table>
ABSTRACT

To improve the methodology of HIV-1 cluster analysis, we addressed how analysis of HIV-1 clustering is associated with parameters that can affect the outcome of viral clustering. The extent of HIV clustering, tree certainty, subtype diversity ratio (SDR), subtype diversity variance (SDV) and Shimodaira-Hasegawa (SH)-like support values were compared between 2881 HIV-1 full genome sequences and sub-genomic regions of which 2567 were retrieved from the LANL HIV Database and 314 were sequenced from blood samples from a cohort in KwaZulu-Natal. Sliding window analysis was based on 99 windows of 1000 bp, 45 windows of 2000 bp and 27 windows of 3000 bp. Clusters were enumerated for each window sequence length, and the optimal sequence length for cluster identification was probed. Potential associations between the extent of HIV clustering and sequence length were also evaluated. The phylogeny based on the full-genome sequences showed the best tree accuracy; it ranked highest with regards to both tree certainty and SH-like support. Product 4, a region associated with env, had the best tree accuracy among the sub-genomic regions. Among the HIV-1 structural genes, env had the best tree certainty, SH-like support, SDR score and the best SDV score overall. The hierarchy of cluster phylotype enumeration mirrored the tree accuracy analysis, with the full genome phylogeny showing the highest extent of clustering, and the product 4 region being second best. Among the structural genes, the highest number of phylotypes was enumerated from the pol phylogeny, followed by env. The extent of HIV-1 clustering was slightly higher for sliding windows of 3 000 bp than 2000 bp and 1000 bp, thus 3000 bp was found to be the optimal length for phylogenetic cluster analysis. We found a moderate association between the length of sequences used and proportion of HIV sequences in clusters; the influence of viral sequence length may have been diminished by the substantial number of taxa. Full-genome sequences could provide the most informative HIV cluster analysis. Selected sub-genomic regions with the best combination of high extent of HIV clustering and high tree accuracy, such as env, could also be considered as a second choice.

Key words: HIV-1, viral cluster analysis, tree accuracy, sequence length, sub-genomic region
CHAPTER ONE: INTRODUCTION

The human immunodeficiency viruses (HIV-1 and HIV-2) are the etiologic agents for AIDS in humans\(^1\)-\(^3\). AIDS is a condition in which progressive failure of the immune system allows life-threatening opportunistic infections and cancers to thrive; it is arguably one of the most devastating infectious diseases to emerge in modern times\(^4\). Despite advancements in HIV prevention over the past 30 years, an estimated 2.1 million persons were newly infected in 2015, bringing the number of people living with HIV worldwide to 36.7 million\(^5\). While cases have been reported in all regions of the world, 95% of new infections occur in individuals living in low- and middle-income countries. The HIV burden continues to be greatest in sub-Saharan Africa; this region accounts for approximately 70% (25.8 million) of worldwide infections\(^6\).

The virus itself belongs to the group of retroviruses known as lentiviruses or “slow viruses”, which are characterized by a long interval between infection and disease development\(^7\). These are spherical in shape, roughly 80 – 100 nm in diameter and possess two usually identical copies of a single stranded RNA genome approximately 10kb in length\(^8\). Following reverse transcription the genome must integrate into the host cell’s DNA in order to replicate. The lifecycle within an individual host is characterized by exceptionally high replication\(^9\), \(^10\), mutation\(^11\) and recombination\(^12\), \(^13\) rates which combined with the positive selection, promoted by the hosts immune response\(^10\), \(^14\), result in the huge amount of diversity observed at both an intra – as well as at an inter – host level\(^15\), \(^16\).

The pandemic of HIV infection continues to pose an enormously difficult public health challenge for a multitude of reasons. HIV-1 continually evolves and migrates through individual hosts, overcoming barriers to transmission, avoiding different immune responses, and resisting various antiretroviral regimens. Immunological host restriction factors in the form of proteins such as TRIM5/22, APOBEC, and Tetherin have been to some extent ineffective in blocking early HIV-1 infection\(^17\)-\(^20\). Highly Active Antiretroviral Therapy (HAART) has been very effective at reducing viral loads within patients and thereby significantly prolonging life expectancy for HIV infected individuals, particularly in those countries where HAART is accessible. However, even when HAART is available, effective control remains elusive due to the number of evolved mechanisms that HIV uses to evade the host immune system\(^21\), \(^22\) the evolution of drug resistance\(^23\), \(^24\), and the isolation of viral reservoirs from drug treatments\(^25\), \(^26\).
Over the past thirty years, HIV-1/AIDS has evolved into an increasingly heterogeneous disease composed of multiple epidemics each influenced by a complex array of biological, behavioral, and cultural factors. The extremely high level of genetic diversity among the 9 recognized global pure subtypes and at least 88 circulating recombinant forms is one of the most daunting aspects of the HIV epidemic. It has been postulated that viral sequence variability may dictate biologic differences that partially explain the different epidemic patterns seen in different regions of the world. Several reports have documented that HIV-1 subtypes may differ with respect to viral load, chemokine co-receptor usage, transcriptional activation levels and antiretroviral drug susceptibility.

While our knowledge of HIV biology is still limited, we have gained significant insights through the application of phylogenetics to HIV diversity. Phylogeny provides a unique framework to capture underlying structures of transmission networks that could not be otherwise identified. Phylogenetics can identify the genetic interrelatedness of viruses in HIV-infected persons. The “clustering” of sequences can infer transmission networks whereby dynamic HIV spread can be assessed on chronological and stage of infection time scales. Phylogenetic cluster analysis can be combined with epidemiological, demographic, and behavioural data to describe the underlying factors contributing to the growth of individual epidemics. Without the immediate prospect of a broadly effective vaccine for HIV-1, the study of HIV transmission networks provides insight into the spread of HIV, and thus into opportunities for intervention.

Justification

A major goal of public health in relation to HIV/AIDS is to prevent new transmissions in communities. A better understanding of the structure and dynamics of HIV transmission through comprehensive HIV cluster analysis could facilitate the achievement of this goal. However, there is confusion surrounding HIV clustering due to differences in sampling, methodological approaches, and interpretation of HIV clustering results across studies. Presently, there has been interest in seeking to establish how the selection of region across the HIV-1 genome and its length affects the extent of HIV clustering. The choice of gene fragment for the reconstruction of phylogenies is crucial as it has been documented that phylogenetic trees constructed from different genes frequently contradict each other, giving rise to incongruence. Several studies examining hundreds of genes in fungi, plants and mammals found that the vast majority of gene trees are not topologically congruent either with each other or with the species phylogeny.
With regards to HIV-1 phylogenetic analysis, the HIV-1 pol gene has been predominantly used for phylogenetic reconstruction of transmission events\textsuperscript{35} and for HIV cluster analysis over the past decade\textsuperscript{33,41-43,45-48}. Although initially the env gene was considered to present the strongest phylogenetic signal, it was argued that some env fragments were too short and/or variable for a robust analysis\textsuperscript{55}. After pol was demonstrated to accurately reconstruct HIV transmission\textsuperscript{35}, its analysis for phylogenetic studies became the standard due to the fact that HIV-1 pol sequences are generated as a part of routine clinical care and thus very large datasets are available for analysis. In the last few years, the increasing availability of HIV whole genome sequences has made possible the analysis of other genetic regions, which has raised discussion about whether full-length genome trees should be used or which viral genes provide the best trees.

Leitner et al. explored this very issue in a study that evaluated the contribution of different regions across the HIV-1 genome to the reconstruction of viral phylogeny. Their findings highlighted the importance of the choice of the HIV-1 gene fragment for reconstruction of true phylogeny, and showed that combining data on gag p17 and env V3 performed better than data on either p17 or V3 evaluated separately\textsuperscript{56}. Results of a study conducted by Harris et al. in Ethiopia also support the notion of the importance of target gene selection in HIV-1 cluster analysis. Phylogenies reconstructed from sub-genomic regions showed a weaker clustering relative to the near full-length genome phylogenies\textsuperscript{57}; however, the study was limited by the small set of viral sequences analyzed. In a study also exploring the parameters that may affect the outcomes of viral clustering, Novitsky et al. hypothesized that the size, complexity, and number of variable or informative sites in the multiple sequence alignment are important factors that impact the extent of HIV clustering\textsuperscript{48}. Their analyses elucidated that the extent of detectable HIV clustering is directly associated with the length of viral sequences used, as well as the number of variable and informative sites.

A common conclusion amongst previous studies is that the combination of more than one gene provides the best estimation of the true tree. However, all were limited to small sample sizes, and, in some cases, short nucleotide sequences. Additionally, demographic and socioeconomic data, as well as stage of HIV infection at the time of sampling, were unavailable for most sequences. There is thus a need to further investigate the parameters that influence the outcomes of viral clustering, examined in a multifaceted manner, using bigger and more comprehensive data. The nature of this influence could help inform the choice of design in studies employing
HIV cluster analysis. It could also help in making choices regarding how subjects are sampled, requirements for laboratory facilities, duration of studies, and budget.

To improve the methodology of HIV cluster analysis, we addressed how analysis of HIV clustering is associated with parameters that can affect the outcome of viral clustering patterns of HIV-1. This is imperative as the extent of viral clustering is one of the key factors in making inferences about epidemiologic processes inferred from viral phylogenies.

1. Literature Review

In the following section the history and current epidemiology of the HIV pandemic, HIV-1 genomic structure, the genetic diversity of HIV-1, phylogenetic methods commonly used in the analysis of HIV, factors that may influence the outcomes of HIV cluster analysis, as well as phylogeny accuracy assessment methods, will be briefly reviewed.

1.1. History and present epidemiology of the HIV pandemic

In 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, oral thrush, high viral loads for cytomegalovirus (CMV), and a 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’s) Morbidity and Mortality Weekly Report (MMWR) weekly newsletter\textsuperscript{59}. 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\textsuperscript{60}, 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 where starting to appear in the heterosexual population\textsuperscript{61}.

The first group in the general heterosexual population to present with these unusual symptoms were intravenous drug users (IVDU’s). The second group was young Haitian immigrants in the USA\textsuperscript{59}. Shortly after that, reports of the disease were documented amongst haemophiliacs who
had been treated with blood and other blood products\textsuperscript{62}. By the end of 1982 reports of the disease in new born babies, who were born to IVDU mothers, were documented\textsuperscript{63}. 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.\textsuperscript{64}. 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\textsuperscript{65-67}. 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\textsuperscript{68}. The findings of the French team were confirmed by research teams in the USA\textsuperscript{69-72}. 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\textsuperscript{73}.

In 2015, 34 years since the first reported cases of AIDS appeared in the USA, HIV-1 global infections had reached an estimated 36.7 million\textsuperscript{5}, showing an increase from 35.3 million in 2012\textsuperscript{74}. The largest burden of HIV-1 is found in Sub-Saharan Africa with more than 65% of adult worldwide infections\textsuperscript{75}. South Africa has the biggest and most high profile HIV epidemic in the world, with an estimated 7 million people living with HIV in 2015\textsuperscript{76}. With the expansion of the ART treatment programme the number of newly infected people globally had decreased from 2.3 million\textsuperscript{74} to 2.1 million\textsuperscript{5} between 2012 and 2015. In addition, global deaths decreased to nearly 1.1 million by 2015\textsuperscript{5}, compared to almost 2.3 million deaths in 2005\textsuperscript{77}, an approximate 50% decrease since the scale-up of ART in the past 10 years. Despite this decrease, HIV is presently one of the world’s most prominent infectious killers\textsuperscript{78}.

1.2. Virion structure and genomic organization of HIV-1

HIV comprises of an outer lipid envelop with glycoproteins, gp120 and gp41, that cover the lipid membrane and matrix protein (p17) underneath\textsuperscript{79}. The conical shape of the HIV capsid (p24) encloses two copies of the single-stranded RNA genome and the 3 viral enzymes; reverse transcriptase, integrase and protease (Figure 1). The HIV-1 genome length is about 9.7-kilo base pairs (kbp) and consists of several major genes coding for structural proteins that are found
in all retroviruses as well as several nonstructural ("accessory") genes unique to HIV. The HIV genome contains three major genes, 5'gag-pol-env-3', encoding major structural proteins as well as essential enzymes\textsuperscript{80}. These are synthesized as polyproteins which produce proteins for virion interior, called Gag (group specific antigen); the viral enzyme Pol (polymerase) or the glycoproteins of the virion env (envelope)\textsuperscript{81}. A diagrammatic representation of the genomic layout of HIV-1 is indicated in Figure 2.

![Diagram of HIV virion structure](image1)

**Figure 1**: HIV-1 virion structure\textsuperscript{82}.

![Diagram of genome layout](image2)

**Figure 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\textsuperscript{83}.

The gag gene codes for the capsid of the virus. 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\textsuperscript{84}. 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. The pol gene is a common feature of retroviruses. As with the gag gene, pol is transcribed in a single protein, which is then spliced into the four functional polypeptides: reverse transcriptase, the RNase, the integrase and the protease. The function of the reverse transcriptase gene is to transcribe the viral RNA to double stranded DNA. The protease gene is responsible for the cleaving/splicing of large protein segments of gag, pol, env, and nef into the separate functional units. The integrase fragment of the pol gene is responsible for the integration of the double stranded viral DNA into the host cells genome. 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. 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. The glycoprotein gp41 is noncovalently bound to gp120, and facilitates the second step of viral entry into the target cells. 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.

In addition to gag, pol and env, HIV encodes for proteins which have certain regulatory and auxiliary functions as well. HIV-1 has two important regulatory elements: Tat and Rev and few important accessory proteins such as Nef, Vpr, Vif and Vpu which are not essential for replication in certain tissues. The gag gene provides the basic physical infrastructure of the virus, and pol provides the basic mechanism by which retroviruses reproduce, while the others help HIV to enter the host cell and enhance its reproduction. Though they may be altered by mutation, all of these genes except rev exist in all known variants of HIV.

1.3. Genetic diversity of HIV-1

HIV variability is a consequence of at least three features peculiar to the virus (Figure 3). First, viral replication is rapid, generating a large number of virions per day (estimated at approximately $10^{10}$ virions per day in an infected individual); second, two or more variants of HIV can undergo recombination within the same infected individual and third, HIV RT is highly error prone, introducing on average one substitution per genome per replication round. 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. As with other RNA viruses, HIV forms complex distributions of closely related but non-identical genomes that are subject to a continuous process of genetic variation, competition, and
selection. These viral quasi-species are highly mutable entities that can quickly adapt to new environments and ecological challenges.

**Figure 3:** Features associated with HIV-1 variability. a) The viral reverse transcriptase is highly error prone, resulting in each new virion encoding approximately one mutation. b) Viral recombination in CD4+ T cells can also generate HIV-1 genetic variation. When two HIV-1 virions with different genetic sequences enter the same cell, they can both integrate and produce viral RNA. Homologous recombination or packaging of RNA from parent viruses leads to the creation of entirely new HIV-1 genomes.

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 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. Four groups of HIV-1 occur, including M, N, O and P. Group M is found worldwide and is the major cause of the HIV epidemic, whereas clusters N, O and P remain mainly in Central West-Africa.
The diversity present within the group M is extensive when compared to other rapidly evolving viral genomes such as influenza\textsuperscript{16}. When represented on a phylogenetic tree, strains within group M form well defined clusters. Nine of these clusters are currently termed subtypes. These are labeled A to D, F to H, J and K plus many circular recombination forms (CRFs), such as CRF01\_AE and CRF02\_AG\textsuperscript{100, 101}. These subtypes, supported by a low subtype diversity ratio\textsuperscript{102} (Figure 4) as well as high bootstrap values\textsuperscript{103}, are roughly equidistant to each other when represented on a phylogenetic tree. They also have long characteristic evolutionary branch lengths stretching into them. As a result of these combined characterizes HIV-1 group M’s phylogeny is often described as being “double starburst” like in nature. Three group M subtypes predominate in the world: subtype A in Central Africa, Western Europe and North America has subtype B and, sub-Saharan Africa plus India has subtype C\textsuperscript{99}. HIV-1 subtype C contributes approximately 50% of the worldwide HIV infections\textsuperscript{104}.

\textbf{Figure 4}: Phylogenetic topology of the nine (A, B, C, D, F, G, H, J, K) pure subtypes of HIV-1 group M. * indicates bootstrap support of 100%\textsuperscript{16}.

\textbf{1.4. HIV-1 phylogenetic analysis pipeline}

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 variety of different phylogenetic methods are used throughout the
course of this study it is of importance to briefly introduce some of the basic concepts of modern phylogenetic practices.

1.4.1. An overview of phylogenetics

Broadly, modern molecular phylogeny is the science of estimating evolutionary histories using DNA and amino acid sequences. 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- 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\textsuperscript{105}. 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\textsuperscript{106, 107}, the Maximum Likelihood method\textsuperscript{108}, the Fitch-Margoliash method\textsuperscript{109}, the Neighbor-Joining method\textsuperscript{110}, the Minimum Evolution method\textsuperscript{111}, and lastly the Bayesian method\textsuperscript{112} of tree inference. These various techniques can broadly be divided into two main categories (\textbf{Table 1}) based on the kind of data they use to infer tree topologies: distance based methods and character based methods\textsuperscript{113, 114}.

\textbf{Table 1: Summary of the various methods of tree construction}

<table>
<thead>
<tr>
<th>Method</th>
<th>Optaimality search criterion</th>
<th>Clustering algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance based</td>
<td>Fitch-Margoliash</td>
<td>UPGMA</td>
</tr>
<tr>
<td></td>
<td>Minimum Evolution</td>
<td>Neighbour-joining</td>
</tr>
<tr>
<td>Character based</td>
<td>Maximum Parsimony</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum Likelihood</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayesian Inference</td>
<td></td>
</tr>
</tbody>
</table>
In HIV-1 phylogenetics, epidemics are characterized based on the genetic interrelatedness of HIV-1 viral sequences, capturing the underlying structure of transmission networks within a given population. HIV-1 phylogenetic analysis initially requires the generation of viral sequences, which in most cases, are derived from virions isolated from the peripheral blood of infected persons. Once HIV has been isolated from the blood or other body fluids and then amplified through a single or series of polymerase chain reactions, a viral sequence can be obtained from the resulting viral amplicon. The next step involves the aligning of the different sequences with one another in order to obtain position homology. Subsequently, 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. A schematic breakdown of the basic steps involved in any phylogenetic investigation is presented in Figure 5.

![Schematic breakdown of the basic steps involved in any phylogenetic investigation](image)

**Figure 5**: 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.
1.4.2. Generation of viral sequences and retrieval of homologous genetic information

HIV sequencing methodologies include direct Sanger sequencing\textsuperscript{117}, single genome amplification or cloning (SGA/cloning)\textsuperscript{118}, and next–generation sequencing (NGS)\textsuperscript{119}. 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\textsuperscript{120}. 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\textsuperscript{120}.

The easiest method to obtain homologous sequence information is to use the Basic Local Alignment Search Tool or BLAST method\textsuperscript{121}. 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\textsuperscript{121}.

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)\textsuperscript{122}. Over the past two decades, HIV data have accumulated rapidly in public and specialized databases thereby creating one of the richest datasets we have for a single entity in terms of sequence tallies and epidemiological information. For instance, the number of available sequences in the Los Alamos database has exploded to 339,306 sequences, a 45% increase over the preceding year, with 2576 complete genomes\textsuperscript{123}. Increasing amounts of HIV-1 nucleic acid sequence data of have led to an ever-growing number of methods to organize and analyze data.
1.4.3. Sequence alignments

In the phylogenetic analysis pipeline, these HIV-1 nucleic acid sequences are aligned with each other in a multiple sequence alignment. 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. Aligned sequences of nucleotides or amino acid residues are typically represented as rows within a matrix (Figure 6). 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. Most sequences alignments require the alignment of large numbers of lengthy, and sometime highly variable, sequences that cannot be aligned solely by human effort, thus in the modern digital age, algorithms are used for the construction of sequence alignments.

\[
A : \text{AACCCCTT} - - - - \\
B : \text{AACCC-CTT} - - - - \\
C : \text{AACCT-C-T--G} \\
D : \text{AACCT-C---G} \\
E : \text{CCTTTT--TTT} \\
F : \text{CCTCTCC-T-CTT} \\
G : \text{ACG---------------}
\]

**Figure 6:** DNA sequences given in the form of a multiple sequence alignment

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. They are however extremely easy to calculate and are therefore often used for methods that do not require extreme precision.

A multiple sequence alignment is an extension of pairwise alignment to accommodate more than two sequences at a time. 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\textsuperscript{114}. 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. 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\textsuperscript{114}.

1.4.4. 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\textsuperscript{114}. To study the dynamics of these changes in sequences, one needs to use mathematical algorithms that take into account different rates of nucleotide substitution (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\textsuperscript{125}. This model operates under the assumption that the equilibrium base frequencies of the four nucleotides are 25% for each nucleotide ($\pi_1 = \pi_2 = \pi_3 = \pi_4 = \frac{1}{4}$). 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 $\mu$. 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.

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 (+G). 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.

1.4.5. Phylogenetic tree structure and inference methods

The resultant phylogenetic tree is a mathematical structure that can be used to graphically depict the relationship among sequences within the alignment. The tree itself consists of internal nodes, external nodes and edges (Figure 7, A and B). From a biological point of view the external nodes (green) are the input sequences. These can also be called leaf nodes. The internal nodes (blue) represent the ancestral relationships between these sequences. These eventually converge to the root of the tree – the estimated most recent common ancestor. Often a closely related sequence, that is not part of the sampled dataset, will be added to the alignment so that it is easier to determine where the true root lies within the group. This is referred to as an outgroup. The choice of outgroup is essential for understanding the evolution of traits along a phylogeny. The chosen outgroup is hypothesized to be closely related to the other groups but less closely related than any single one of the other groups is to each other. Edge lengths connecting the nodes represent the amount of change that occurs between each node. In summary, HIV phylogenies are evolutionary trees in which the leaves of the tree are the sampled sequences or taxa, branches are the genetic distance between taxa, and the nodes denote estimated speciation events. There are a number of programs for inferring phylogenies including, but not limited to, PAUP*, BAMBE, BEAST, PHYLIP, RAxML and MrBayes.
Figure 7: Phylogenetic tree structures. (A) A tree inferred from 13 HIV-1 sequences labelled A – M. (B) Identical tree but represented using a different trigonometric pattern.

Phylogenetic trees are generally inferred based on an optimality criterion. Most commonly utilized methods are distance based methods and evolutionary methods. Distance based methods, such as Neighbor-Joining and UPGMA, explicitly rely on a measure of genetic distance between operational taxonomic units (OTUs = [neighbours]) based on their sequence differences. Distance measures are derived from pairwise comparisons of the sequences. Whereas the distance based methods represent sequence divergence by a single number, the evolutionary methods attempt to infer the phylogeny by fitting individual characters (nucleotides or amino acids) to the tree. Most popular approaches for evolutionary methods are maximum likelihood, maximum parsimony.

The parsimony score is the minimum number of character changes implied by a tree given a multiple sequence alignment. Thus the principle of parsimony as applied to phylogenetics, states that the topology that requires the fewest evolutionary changes is the one that should be assumed as correct. A smaller parsimony score indicates a better tree. Parsimony-based methods count the total number of substitutions in the tree by summing the substitutions between
sequences of every pair of adjacent nodes. Sequences for internal nodes may be reconstructed using algorithms such as the Sankoff algorithm. The maximum likelihood method is broadly similar to the maximum-parsimony method, but maximum likelihood allows additional statistical flexibility by permitting varying rates of evolution across both lineages and sites. In fact, the method requires that evolution at different sites and along different lineages must be statistically independent. Maximum likelihood is thus well suited to the analysis of distantly related sequences, but it is believed to be computationally intractable to compute due to its NP-hardness.

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 and retesting to establish if the new tree is a better fit. 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.

**1.5. Application of HIV-1 phylogenetic analysis in modern research**

With the advent of rapid and inexpensive DNA sequencing and the development of bioinformatics tools for comparing the genetic sequences from different organisms, we are now in the era of molecular phylogeny. Recent studies utilizing new sequencing technologies and genomics tools have begun to reveal a vast amount of data regarding the genetics of both HIV and the human immune system. Today, phylogenetic analysis has become a common practice of many HIV/AIDS research programs, due mainly to the many insights these analyses
can provide and the novel questions they can address over a variety of topics related to HIV biology. Accurate phylogenetic inference is integral to properly understanding how evolutionary processes have shaped living organisms and their genomes.

In the context of infectious diseases, a phylogenetic perspective can also be particularly helpful for drawing important medical, epidemiological, and forensic conclusions. For example, phylogenetic analysis of HIV-1 has been used to determine the impact of antiretroviral therapy (ART) on viral evolution\(^\text{141, 142}\), to infer migration patterns across broad geographic ranges and time-scales\(^\text{143, 144}\), and to understand transmission dynamics in populations with narrow geographic and temporal bounds\(^\text{56, 145-147}\). The application of phylogenetic methods to identify individuals who are probable sources of infection within small transmission clusters has been found to “meet the judicial standards of evidence admissibility”\(^\text{119, 148}\) and the results have been presented as supporting evidence in courts of law\(^\text{119, 148}\). Due to the uncertainty in infection time, evolutionary rate and potential contacts, it is generally not possible to reconstruct the exact transmission network from a phylogenetic tree alone. However patients sharing similar viruses are potentially epidemiologically linked, so local outbreaks within the larger epidemic can be identified by finding transmission clusters.

1.6. HIV-1 transmission cluster analysis and its shortcomings

Clusters in epidemiology are broadly described as an unusual aggregation of infection, perceived to be greater than that expected by chance. In transmission networks, clusters are quantitatively defined as a group of nodes having a local clustering coefficient significantly greater than that of a random graph with the same number of vertices and the same mean shortest path\(^\text{149}\). In a phylogenetic tree, clusters contain sequences from different patients which share a recent common ancestor. These clusters are manifest as groupings in the phylogenetic tree which we have high confidence and which are likely to reflect recent or ongoing transmission. However, defining and detecting meaningful transmission clusters from a population sample in a phylogenetic tree is not straightforward, and various strategies have been proposed and used in the literature.

As shown in Figure 8, clusters are often defined based on high support (bootstrap or posterior probability) and/or low within cluster genetic distance, but the thresholds for both vary. The numbers next to each node, in pink, present a measure of support for the node. These are generally numbers between 0 and 1, where 1 represents maximal support. For HIV, bootstraps ranging from 70% and up to 99% have been used\(^\text{35, 47, 150-153}\), in combination with within-cluster genetic distances from 1% to 4.5% substitutions per site\(^\text{39, 47, 152-155}\). Thus a drawback of this
procedure is that there is an onus on the user to determine the appropriate support/distance thresholds and rationale for threshold selection is rarely provided; however, these decisions may affect study inferences. Robinson et al. enumerated this in a study where they simulated HIV spread and pathogen phylogenies on two different network topologies. They showed that threshold choices affect the size and distribution of phylogenetic clusters obtained.$^{156}$

![Figure 8: Phylogenetic tree showing evolutionary relationships and clustering between a set of viruses.](image)

The blue circles represent ancestors, which in this context, mean an infected host at sometime in the past that in turn infected 2 or more new hosts producing chains of infections that lead to the sampled viruses (green circles). The branches then represent this chain of infections.$^{157}$

The method for calculating within cluster genetic distance also varies: the mean of the pairwise genetic distances of clustered sequences has been employed$^{155}$, as well as their median$^{158}$. In addition to being data and user-specific, threshold values can also be affected by the statistical approach used to measure support. Namely, in a formal investigation conducted by Alfaro et al.$^{159}$, the result of the phylogenetic tree branch support assessment was dependent on the method chosen, posterior probabilities were higher than their corresponding bootstrap values on average. Furthermore, reconstructing phylogenetic trees can be computationally intensive, especially when a large number of sequences are being considered. Another alternative is “single linkage”, where a sequence is included in a cluster if its distance to just one other sequence in the cluster is below the threshold.$^{160, 161}$ If time resolved trees are used (which require knowledge or inference of a molecular clock), clusters can be defined based on time to most recent common ancestor.$^{33}$ These most resemble clusters generated using maximum genetic distance in a non-time resolved distance-based tree. Table 2 shows the lack of standardization of phylogenetic methods amongst HIV-1 transmission clustering analysis studies. There appears to be no consensus with regards to target gene selection, sequence length, sampling density and HIV-1 transmission cluster definition.
Table 2: Convenient sample of HIV-1 clustering analysis studies

<table>
<thead>
<tr>
<th>Authors</th>
<th>Location</th>
<th>Sample Size</th>
<th>Sequence used</th>
<th>Method of phylogeny reconstruction</th>
<th>HIV-1 cluster definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonia dou et al.162</td>
<td>Greece</td>
<td>98</td>
<td>Pol</td>
<td>NJ</td>
<td>≥85% bootstrap, ≤0.015 mean intra-cluster genetic distance</td>
</tr>
<tr>
<td>Bezeme r et al.163</td>
<td>Kenya</td>
<td>674</td>
<td>Pol</td>
<td>ML, Bayesian</td>
<td>≥70% bootstrap, ≤0.015 mean intra-cluster genetic distance</td>
</tr>
<tr>
<td>Ruelle et al.164</td>
<td>Belgium</td>
<td>55</td>
<td>RT</td>
<td>ML, Bayesian</td>
<td>≥0.89 posterior probability</td>
</tr>
<tr>
<td>Frentz et al.165</td>
<td>Europe, Israel</td>
<td>4260</td>
<td>Pol</td>
<td>ML</td>
<td>≥98% bootstrap, ≤0.030 mean intra-cluster genetic distance</td>
</tr>
<tr>
<td>Dennis et al.166</td>
<td>El Salvador</td>
<td>119</td>
<td>Pol</td>
<td>Bayesian</td>
<td>Posterior probability=1, ≤0.015 mean intra-cluster genetic distance</td>
</tr>
<tr>
<td>Li et al.167</td>
<td>China</td>
<td>253</td>
<td>Gag and Pol</td>
<td>ML</td>
<td>≥70% bootstrap</td>
</tr>
<tr>
<td>Yebra et al.168</td>
<td>Spain</td>
<td>1293</td>
<td>Pol</td>
<td>Bayesian</td>
<td>≥90 posterior probability, cluster depth cutoff</td>
</tr>
<tr>
<td>Ng et al.169</td>
<td>Malaysia</td>
<td>496</td>
<td>Pol and RT</td>
<td>ML, Bayesian</td>
<td>≥2 individuals from same geographic location, &gt;90% bootstrap, posterior probability=1</td>
</tr>
<tr>
<td>Feng et al.138</td>
<td>China</td>
<td>75</td>
<td>Near full length genome</td>
<td>ML</td>
<td>≥90% bootstrap</td>
</tr>
<tr>
<td>Siljic et al.170</td>
<td>Serbia</td>
<td>221</td>
<td>Pol</td>
<td>ML, Bayesian</td>
<td>≥90% bootstrap, ≤0.015 mean intra-cluster genetic distance, ≥0.9 posterior probability</td>
</tr>
<tr>
<td>Yebra et al.171</td>
<td>Spain</td>
<td>278</td>
<td>Pol</td>
<td>ML, Bayesian</td>
<td>≥95% bootstrap, ≥0.95 posterior probability</td>
</tr>
<tr>
<td>Murrillo et al.172</td>
<td>Central America</td>
<td>625</td>
<td>Pol</td>
<td>ML</td>
<td>Shimodaira-Hasegawa test (p-value &lt;0.01), patristic distance threshold (25th percentile)</td>
</tr>
<tr>
<td>Temer anca et al.173</td>
<td>Romania</td>
<td>61</td>
<td>Pol</td>
<td>ML</td>
<td>≤0.01 maximum intra-cluster genetic distance</td>
</tr>
<tr>
<td>Audelin et al.174</td>
<td>Denmark</td>
<td>1515</td>
<td>Partial pol</td>
<td>NJ, Bayesian</td>
<td>≥90 % bootstrap, ≤0.025/ ≤0.050 mean/maximum intra-cluster genetic distance, posterior probability=1</td>
</tr>
<tr>
<td>Chen et al.178</td>
<td>China</td>
<td>308</td>
<td>Partial gag and env</td>
<td>Neighbor-joining</td>
<td>≥70% bootstrap</td>
</tr>
<tr>
<td>Han et al.176</td>
<td>China</td>
<td>583</td>
<td>1.0-kb prot-RT</td>
<td>Neighbor-joining</td>
<td>≥70% bootstrap</td>
</tr>
<tr>
<td>Ivanov et al.177</td>
<td>Bulgaria</td>
<td>125</td>
<td>Pol</td>
<td>ML, Bayesian</td>
<td>≥96 % bootstrap, ≤0.10 maximum intra-cluster genetic distance, ≥0.97 posterior probability</td>
</tr>
<tr>
<td>Avidor et al.178</td>
<td>Israel</td>
<td>318</td>
<td>Prot and RT</td>
<td>Bayesian</td>
<td>≥0.95 posterior probability</td>
</tr>
<tr>
<td>Ndziye et al.179</td>
<td>Senegal</td>
<td>109</td>
<td>Pol, gag and env</td>
<td>ML</td>
<td>≥98% bootstrap</td>
</tr>
<tr>
<td>Tramuto et al.180</td>
<td>Sicily</td>
<td>155</td>
<td>Pol</td>
<td>ML</td>
<td>≥75% bootstrap</td>
</tr>
</tbody>
</table>

NJ=Neighbor-joining, ML=Maximum Likelihood, RT=Reverse Transcriptase, Prot=Protease, Pol=Polymerase, Env=Envelope, Gag=Group-specific antigen. Adapted from181.
Phylogenetic clusters have been used to provide crucial insights about the spread and transmission of the disease. In the case of HIV, analyses of phylogenetic clusters have been used to identify correlates of transmission including risk group, stage of infection, cluster size, the presence or absence of co-infections, including other sexually transmitted infections as well as drug treatment and compliance. A recent study used a phylogenetic approach to determine the relative contribution of each of these variables to the risk of onward transmission, finding that antiretroviral treatment decreased HIV transmission risk.

The origin and geographic expansion of HIV-1 have been well characterized using phylogenetic approaches, but it has been argued that these methods are suboptimal for describing recent HIV transmission. Wertheim et al. contended that phylogenies are well suited for differentiating distinct viral lineages but not for identifying transmission partners. Several authors have concluded that phylogenetic analysis is most powerful at excluding potential transmission partners, rather than establishing linkage. Although there are a number of programs available for clustering nucleotide sequences (e.g., BLASTClust, UPGMA and WPGMA, neighbor-joining (NJ), and phyclust), phylogenetic approaches have been ubiquitous in the literature involving HIV-1 transmission clusters.

As studies investigating HIV transmission clusters typically begin by inferring a phylogeny and then identifying those clades (sub-trees) that have appropriate statistical support, Wertheim et al. argued that this identification alone is insufficient for epidemiological purposes, because such an analysis lacks the concept of recency. Additionally, high statistical support (e.g., bootstrap) for any specific clade indicates that there is no close relative to the clade in question, not that the members of the clade itself are necessarily closely related to each other. Another documented problem with using phylogenetic methods to infer transmission clusters is that often only a single geographic region is considered, and the data must be subsampled for computational tractability. Both of these simplifications can seriously bias the interpretation due to the limited scope of the analysis, as closely related or relevant sequences may be inadvertently excluded during sequence selection. Also, most transmissions from one individual to another involve transmission of a minor variant from the circulation of the donor; this makes inference of transmission pairs via phylogeny difficult to impossible.

Despite these shortcomings, the best method of identifying and establishing transmission events of HIV between individuals or within a community is through the use of high-resolution phylogenetic methods of HIV sequence data. Phylogenetic analysis has greatly improved our understanding of the epidemic, and remains at the forefront of cluster analysis on HIV.
sequences, however there is great space for improvement. Thus the importance of the detection
and evaluation of the parameters that effect viral clustering as this will inform the design for
better methodologies for HIV phylogenetic cluster analysis.

1.7. Factors that may influence the outcomes of viral clustering

Phylogenetic analysis can be very informative, but the accuracy of phylogenetic conclusions is
highly dependent on the method chosen and sampling strategy. Besides the inherent issues
about model selection and phylogenetic inference, data availability also plays a major role.
Some of the concerns are related to the direction of transmission or who infected whom,
availability of all involved sexual contacts, and interpretation of the phylogeny given that
certain individuals could be infected with more than one strain. Similarly, criminal cases of HIV
transmission that rely solely on phylogenetic evidence are precarious. Issues of convergent
evolution can also erroneously link individuals in the absence of any other independent source
of evidence^{201}.

Table 3 shows research that has probed the parameters that may affect the outcome of viral
clustering; these include certain biological processes^{202}, sampling density^{203}, sequence
properties^{58}, sequence length^{58} and target gene selection^{35, 57}. A common limitation amongst
previous research conducted is sample size and the predominant utilization of HIV-1
polymerase datasets. Another commonality is the conclusion that the use of HIV-1 full-length
genome sequences results in the best phylogeny for cluster enumeration; however, none have
fully probed the optimal sequence length when full-length genome sequences aren’t feasible.
Table 3: Papers exploring parameters that may affect the outcomes of HIV-1 phylogenetic clustering

<table>
<thead>
<tr>
<th>Authors</th>
<th>Parameter</th>
<th>Results</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doyle et al.</td>
<td>Biological processes</td>
<td>Convergent evolution and high rates of insertions and deletions (causing alignment uncertainty) lead to spurious phylogenetic signal with forensic relevance.</td>
<td>Full-genome sequencing of HIV-1, combined with careful phylogenetic analyses based on biologically realistic models of sequence evolution, will greatly increase the information available for inference of transmission histories while avoiding many of the biases inherent to individual genes.</td>
</tr>
<tr>
<td>Novitsky et al.</td>
<td>Sampling density</td>
<td>HIV clustering increased linearly at sampling density &gt; 10%, and was accompanied by narrowing confidence intervals. HIV clustering increased linearly at sampling density &gt; 10%, and was accompanied by narrowing confidence intervals.</td>
<td>The detectability of HIV clusters is substantially affected by sampling density. A minimal genotyping density of 10% and sampling density of 50–70% are suggested for HIV-1 V1C5 cluster analysis.</td>
</tr>
<tr>
<td>Novitsky et al.</td>
<td>Sequence properties</td>
<td>Found a moderate association between the number of variable and informative sites and the proportion of HIV sequences in clusters</td>
<td>Use sequences with adequate number of variable and informative sites.</td>
</tr>
<tr>
<td>Harris et al.</td>
<td>Target gene selection</td>
<td>A cluster of HIV-1 sequences from Ethiopia, observed in full genome analysis, is not sustained in sub-genomic regions.</td>
<td>Results elucidate the advantages of the usage of near full-length genome sequences for cluster analysis.</td>
</tr>
<tr>
<td>Novitsky et al.</td>
<td>Sequence length</td>
<td>The near full-length genome HIV sequences showed the highest extent of HIV clustering and the highest tree certainty. Found a strong association between the sequence length and proportion of HIV sequences in clusters</td>
<td>Near full-length genome sequences could provide the most informative HIV cluster analysis; selected subgenomic regions with a high extent of HIV clustering and high tree certainty could also be considered as a second choice.</td>
</tr>
<tr>
<td>Hue et al.</td>
<td>Target gene selection</td>
<td>The topology of the pol tree was consistent after exclusion of the drug resistance associated codons. Identical topologies were obtained in tress implemented from gag and env gene alignments.</td>
<td>Despite its genetic conservation, the HIV-1 pol gene holds sufficient variability to permit the phylogenetic reconstruction of transmissions, when compared to the env and gag genes.</td>
</tr>
<tr>
<td>Gifford et al.</td>
<td>Target gene selection</td>
<td>Due to the presence of recombinant strains, the internal topologies of phylogenies differ depending on which sub-genomic region is analyzed.</td>
<td>Results advocate the usage of full-length HIV-1 genome sequences for cluster analysis.</td>
</tr>
<tr>
<td>Robinson et al.</td>
<td>Transmission cluster definition/threshold selection</td>
<td>Threshold choices affect the size and distribution of phylogenetic clusters obtained.</td>
<td>Rationale for threshold selection should be provided.</td>
</tr>
</tbody>
</table>
1.7.1. Target gene selection and sequence length

Genetic marker choice is important when seeking to capture transmission and other desired signals. Extensive debate exists concerning the gene(s) choice in HIV phylogenetics as they are a number of factors to be considered (Table 4). The pol gene has been suggested as a candidate due to the large dataset of HIV-1 pol sequences available, however some researchers have been reluctant to use it given the number of drug resistance mutations associated with this region. The env gene is sometimes preferable relative to the pol gene on the basis of high genetic variability. This was highlighted by a study done by Frange et al., 2008, who analyzed 15 samples whose subtype/CRF could not be identified using RT sequences. By using env sequences, 6 were found to be divergent A, 2 were distantly related to E or D, 2 C, 1 B; the variability of env was paramount for this investigation. However, indications of convergent evolution on the env region can preclude its use since it violates the unique evolutionary history assumption made by phylogenetic methods.

Table 4: Characteristics of target genes for phylogenetic analysis

<table>
<thead>
<tr>
<th>Target gene</th>
<th>Region</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>env</td>
<td>Codes for the external glycoproteins gp120 and gp41; contains high variable regions including v3 loop</td>
<td>High number of sequences available in genetic databases</td>
<td>High variability, Can lead to convergent evolution whereby two sequences are similar because of homoplasy, not homology</td>
</tr>
<tr>
<td>gag</td>
<td>Codes for internal virion proteins</td>
<td>Less variable target than env</td>
<td>Fewer database sequences than env and pol</td>
</tr>
<tr>
<td>pol</td>
<td>Codes for the enzymes reverse transcriptase, protease and integrase</td>
<td>Least variable, Highest number of sequences in genetic databases</td>
<td>May be argued its too conserved to contain useful phylogenetic information, Sight of drug resistance mutations</td>
</tr>
</tbody>
</table>

With the conundrum of gene target selection in mind, Lemey et al., 2005 explored across the length of full-genome sequences for phylogenetic support of HIV-1 transmission events. They investigated three known and distinct transmission cases for which full-genome sequence data was available. To evaluate which genome regions are the most informative for transmission chain reconstruction, they performed a sliding window analysis using the same maximum likelihood method for different window sizes.
For a window size of 400 nucleotides, only the \textit{vif} gene region provided considerable bootstrap support (\textgreater{} 90\%) for all three transmission clusters. More importantly, extensive variation in gene-specific bootstrap support was observed among the three transmission chains. Increasing the window size up to 800 nucleotides resulted, on average, in an increase in transmission cluster support and more distinct patterns of gene-specific support. The 3’ part of the \textit{pol} gene up to the \textit{env} gene appeared to provide relatively good support for all three transmission clusters. However, there was still considerable variability in transmission cluster bootstrap support for \textit{gag} and \textit{env}. A window size of 1200 nucleotides resulted in a further average increase in bootstrap support but still showed the \textit{gag} and \textit{env} differences. They concluded that transmission chain support across the genome can be case specific and it does not appear to be largely moderated by functional constraints across the genome. Harris et al. also argued that phylogenetic relationships determined with partial genomes may not be reproduced when other regions of the genome are considered\textsuperscript{57}.

Controversy has been expressed on the adequacy of single gene phylogenies to establish confidently true relationships in transmission cases\textsuperscript{207,208}. The initial analysis by Sturmer et al.\textsuperscript{209} of the \textit{pol} and \textit{env} regions in strains derived from the two case patients and several controls, both local and from the databases, led the authors to conclude that the \textit{pol} sequence on its own did not provide enough information to clarify the relationship between the two patients. Hue et al.\textsuperscript{35} came to the conclusion that ‘there is no such thing as an ultimate gene for evolutionary analyses of HIV-1 and, ideally, full-length sequences should be used’.

\textbf{1.7.2. Contrasting the usage of full genome and sub-genomic sequences for HIV-1 cluster analysis}

To improve the methodology of HIV cluster analysis, Novitsky et al. investigated the influence of the usage full genome versus sub-genomic sequences. Their analyses indicated that near full-length genome HIV sequences showed the highest extent of HIV clustering and the highest tree certainty. They concluded that near full-length genome sequences could provide the most informative HIV cluster analysis\textsuperscript{58}. To highlight the shortfalls of using sub-genomic data, Harris et al. showed that a cluster of HIV-1C sequences from Ethiopia, observed in full genome analysis, was not sustained in sub-genomic regions\textsuperscript{57}.

One of the key advantages of full HIV-1 genome analysis is thorough subtype classification, which, after taking appropriate consideration for the exclusion of recombination events taking place during amplification, can then allow the examination of complex recombination events.
occurring in vivo. Such recombination events can be observed within sequences from different compartments within a single individual, from different individuals with the same subtype, or from different subtypes\textsuperscript{210,211}. Full HIV-1 genome analysis provides robust information for investigations of HIV-1 recombination. Yamaguchi et al.\textsuperscript{212} noted that HIV-1 strain classification based on partial genome fragments has limitations and recombinant strains may go unrecognized in the absence of complete genome sequences. With the use of full genome sequencing, they were able to identify one circulating recombinant form (CRF) and six unique recombinant forms (URF) in the Saudi Arabian population. The benefits of complete versus targeted sequencing are numerous and include characterization of the complete genome allowing for identification of accessory / compensatory mutation and detection of rare genetic variants\textsuperscript{213-215}.

Analysis done by Gifford et al.\textsuperscript{204} also shows how phylogenies inferred from sub-genomic regions can be spurious. They conducted a strain-level classification of HIV-1 genetic diversity using 976 complete-genome sequences (Figure 9). The phylogenetic trees depicted here illustrate that due to the presence of recombinant strains, the internal topologies of phylogenies (shown as dotted lines) differ depending on which sub-genomic region is analyzed. Thus, the CRF03 strain illustrated, highlighted in trees by shaded circles, can be seen to group with subtype A in trees constructed using \textit{gag} (A) and with subtype B in trees constructed using \textit{pol} (B).
**Figure 9**: Neighbour-joining (NJ) phylogenies constructed using sub-genomic regions a) **gag** and b) **pol**. CRF03 AB groups with subtype A in the **gag** phylogeny and subtype B in the **pol** phylogeny\textsuperscript{204}.

For transmission cluster analysis in the present study, we utilized PhyloPart, a relatively novel software tool for large-scale phylogeny partition\textsuperscript{158}. This method is based on a depth-first search algorithm and conjugates the evaluation of node reliability, tree topology and patristic distance analysis. Other available methods have used different cluster selection schemes by performing nested phylogenetic analyses, and/or adding criteria for geographical consistency\textsuperscript{39, 47, 154, 184, 186, 194} but in most cases the assessment of transmission clusters is still subject to a visual tree inspection. The definition of a transmission cluster proposed in this method is general and can be tuned to accommodate any of the previous definitions, making it an ideal tool to explore parameters that may influence viral clustering outcomes.

We also utilized a novel phylotype-based exploratory tool that has the ability to use phylogenies constructed with any of the most popular methods, while providing fast inference of ancestral traits and enabling hypothesis testing and visual data interpretation of evolutionary scenarios. The method combines ancestral trait reconstruction using parsimony, with combinatorial and numerical criteria measuring tree shape characteristics and the diversity and separation of the potential phylotypes\textsuperscript{216}. 

\textsuperscript{204} CRF03 AB groups with subtype A in the **gag** phylogeny and subtype B in the **pol** phylogeny.

\textsuperscript{158} PhyloPart, a relatively novel software tool for large-scale phylogeny partition.

\textsuperscript{39, 47, 154, 184, 186, 194} Other available methods have used different cluster selection schemes by performing nested phylogenetic analyses, and/or adding criteria for geographical consistency.

\textsuperscript{216} The method combines ancestral trait reconstruction using parsimony, with combinatorial and numerical criteria measuring tree shape characteristics and the diversity and separation of the potential phylotypes.
1.8. Assessing the accuracy of phylogenetic trees

In assessing which gene target or sequence length is optimal for phylogenetic cluster analysis, an important step is contrasting the accuracy of phylogenies inferred from each gene fragment and sequence length. Four principal methods have been used for assessing phylogenetic tree accuracy in HIV-1 clustering analysis studies: simulation, known phylogenies, statistical analyses, and congruence. Simulations are useful for studying accuracy of methods under idealized conditions and can be used to make general predictions about the behaviour of methods if the limitations of the models are taken into account. Yebra et al. recently used this method to determine which gene(s) provide(s) the best approximation to the real phylogeny by sub-sampling a simulated dataset of 4662 sequences. Although the sample size was significant, many biological systematists dismiss simulation results because there is usually a complete fit between the evolutionary model used to simulate the sequence data and the model used for analysing it.

Studies of known phylogenies can also be used to test predictions from simulation studies, thus providing a check on the robustness of the models. This was a methodology utilized by Leitner et al. where they constructed a true phylogenetic tree based on he knowledge about when the transmissions had occurred and when the samples were obtained. This complex, known HIV-1 transmission history was then compared with reconstructed molecular trees, which were calculated from the DNA sequences by several commonly used phylogenetic inference methods [Fitch-Margoliash, neighbor-joining, minimum-evolution, maximum-likelihood, maximum-parsimony, unweighted pair group method using arithmetic averages (UPGMA), and a Fitch-Margoliash method assuming a molecular clock (KITSCH)]. A drawback of the utilization of known phylogenies is the data required for the complex reconstruction of the true tree. An alternative method for phylogeny accuracy assessment is statistical analysis. Statistical analyses allow general predictions to be applied to specific results, facilitate assessments as to whether or not sufficient data have been collected to formulate a robust conclusion, and indicate whether a given data set is any more structured than random noise. Finally, congruence analyses of multiple data sets can be used to assess the degree to which independent results agree and thus the minimum proportion of that can be attributed to an underlying phylogeny.

1.8.1. Tree branch support assessment

1.8.1.1. Tree certainty

Computing and evaluating tree branch supports (measures of confidence in given branches) are indispensable parts of phylogenetic inference. In particular, support measures are crucial to
validating or refuting biological hypotheses on the basis of trees. Parallel to the development of phylogenetic inference methods, various measures of branch support have been proposed. In the statistical paradigm, the perhaps three most desirable properties of a branch support measure are high accuracy, power, and robustness. High accuracy implies that under the true model, incorrectly inferred branches should not be statistically supported. High power implies that correctly inferred branches should have high statistical support. As for high robustness, it conveys the notion that modeling inadequacies, which are unavoidable when dealing with real biological data, do not strongly affect the accuracy of the measure.

In the present study, the accuracy of each inferred phylogeny was assessed by estimating phylogenetic tree certainty. This calculation is based on a set of novel measures proposed by Salichos and Rokas for quantifying the confidence for bipartitions in a phylogenetic tree. These measures are the so-called Internode Certainty (IC) and Tree Certainty (TC), which are calculated for a specific reference tree given a collection of other trees with the exact same taxon set. The underlying idea of IC is to assess the degree of conflict of each internal branch (a branch connecting two internal nodes) of a phylogenetic reference tree by calculating Shannon’s Measure of Entropy. This score is evaluated for each bipartition in the reference tree independently. The basis for the calculations is the frequency of occurrence of this bipartition and the frequencies of occurrences of a set of conflicting bipartitions from the collection of trees. In contrast to classical scoring schemes for the branches, such as simple bipartition support or posterior probabilities, the IC score also reflects to which degree the most favoured bipartition is contested.

1.8.1.2. Bootstrapping and approximate likelihood ratio tests

To probe phylogenetic tree accuracy, bootstrapping can be done to infer the reliability of branch order. Bootstrap resampling as a statistical tool was invented in the late 1970’s by Bradley Efron and was introduced into the field of molecular phylogenetics by Joseph Felsenstein in the mid 1980’s. Briefly, bootstrapping in modern molecular phylogenetics entails continuous resampling of taxa, over a user specified number of iterations (Figure 10). 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.

**Figure 10:** Diagrammatical depiction of bootstrap analysis. A neighbour joining tree is (red box) is inferred from the input alignment (purple box). Columns on the input alignment are then randomly sampled (green boxes) and a tree is inferred. This sampling-inference process is repeated – usually 1000 times. Branches on the correct tree are compared to branches on the trees from the random samples (circle)\(^\text{16}\).

Despite the popularity of nonparametric bootstrap frequencies and Bayesian posterior probabilities, the interpretation of these measures of tree branch support remains a source of discussion\(^\text{223}\). Furthermore, both methods are computationally expensive and become prohibitive for large data sets. Recent fast approximate likelihood-based measures of branch supports (approximate likelihood ratio test [aLRT] and Shimodaira–Hasegawa [SH]-aLRT) provide a compelling alternative to these slower conventional methods, offering not only speed advantages but also excellent levels of accuracy and power\(^\text{223}\). SH-aLRT was developed and implemented in the PHYML phylogenetic inference software\(^\text{223}\). It is derived from the SH
multiple tree comparison procedure\textsuperscript{224} and is fast due to the RELL technique based on the resampling of estimated log likelihoods\textsuperscript{225}.

\subsection*{1.9.2. Tree metrics}

In this study, we evaluated the cluster quality in inferred phylogenies by using phylogenetic tree metrics that quantitatively measure the extent of strain clustering. The tree metrics include two statistics that are used to quantify the degree of clustering present on the tree topology. The subtype diversity ratio (\textbf{Figure 11, Panel A}), SDR, is defined as the ratio of the mean intra-cluster pairwise distance to the mean inter-cluster pairwise distance\textsuperscript{226}. The SDR is therefore a quantitative measure of the extent of clustering found within a tree. Low intra-cluster pairwise distances relative to inter-cluster pairwise distances implies more defined clustering in the tree. Thus trees with lower SDR values are characterized by well defined clusters. An SDR approaching one would indicate a lack of clustering is present in the tree. As the SDR does not take into account the variability that can occur between individual clusters the subtype diversity variance (\textbf{Figure 11, Panel B}), SDV, was devised. The SDV statistic is a measure of the variation within the ratio of the mean intra-cluster pairwise distance to the mean inter-cluster pairwise distance calculated for each cluster on the tree. The lower the SDV value the more symmetrical, or equidistant, the clusters in a tree are relative to each other.

\textbf{Figure 11:} Subtype diversity ratio and subtype diversity variance. (A) The relationship between the SDR and the quality of clustering on a phylogenetic tree (B) The relationship between the SDR and the variability of the quality of clustering\textsuperscript{16}.
In the present study, not only were the outcomes of HIV-1 clustering probed, but also the accuracy of the inferred phylogenies from which clustering analysis was done. This is significant as the accuracy of the outcomes of viral clustering, that is, the proportion of sequences in clusters enumerated, is dependent on the quality of the reconstructed phylogenetic tree.

**Aims**

1. To determine if the usage of full-genome sequence data as opposed to sub-genomic data for phylogenetic tree construction results in a more accurate phylogenetic tree, measurable by tree certainty; subtype diversity ratio; subtype diversity variance; and Shimodaira-Hasegawa-like support values.
2. To elucidate the optimal window size (sequence length) and genomic region to accurately identify transmission clusters.

**Objectives**

1. To sequence the full-length of subtype C HIV-1 genomes using next generation sequencing platform (Illumina Miseq).
2. To reconstruct HIV phylogenies using different sub-genomic sequences and sequence lengths.
3. To evaluate and contrast the outcomes of viral clustering and phylogeny accuracy per parameter investigated.
CHAPTER TWO: METHODOLOGY

The methodology that was used during the course of this study will be discussed in this chapter. This project involved five main steps: (1) the retrieval of HIV-1 full genome sequences from the LANL Database and the generation of full genome sequences from patient samples, (2) the stratification of data per parameter investigated, (3) the inferring of phylogenies from different sequence lengths and sub-genomic regions, (4) the phylogeny accuracy assessment and clustering analysis for each phylogeny, and (5) the interpretation of the analyzed data.

Briefly, the general aim was to improve the methodology of HIV cluster analysis by addressing how analysis of HIV clustering is associated with parameters that can affect the outcome of viral clustering. The extent of HIV clustering, tree certainty, subtype diversity ratio, subtype diversity variance and Shimodaira-Hasegawa-like support values were compared between 2881 HIV-1 full genome sequences and sub-genomic regions of which 2567 were retrieved from the LANL HIV Database and 314 were sequenced from blood samples from a cohort in KwaZulu-Natal. Sliding window analysis was based on 99 windows of 1000 bp, 45 windows of 2000 bp and 27 windows of 3000bp. Clusters were enumerated for each window sequence length, and the optimal sequence length for cluster identification was probed. Potential associations between the extent of HIV clustering and sequence length were also evaluated.

2.1. Ethics statement

This study was approved by the Biomedical Research Ethics committee of the University of KwaZulu Natal (ref. BF052/10), the Health Research Ethics committee of the KwaZulu Natal Department of Health (ref. HRKM 176/10) and all study participants provided written informed consent for the collection of samples and subsequent analyses. The investigations also comply with the South Africa National Health Act No 612003 and abide by the ethical norms and principles for research as established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Department of Health Guidelines.
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) 5 - 7. All chemical and biological agents or commercial kits that were used in this study are summarized in Table 5.

Table 5: List of chemicals and commercial products used in the study

<table>
<thead>
<tr>
<th>Chemical or Commercial products and kits used</th>
<th>Suppplying Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>QIAamp Viral RNA Mini kit</td>
<td>QIAGEN</td>
</tr>
<tr>
<td>SuperScriptIII One-Step RT-PCR system</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>Taq DNA Hgh Fidelity Polymerase</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>dNTP's</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>Nuclease free water</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>Ethidium Bromide</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>Agarose</td>
<td>Whitehead Scientific (Pty) Ltd</td>
</tr>
<tr>
<td>Novel Juice 6x</td>
<td>Whitehead Scientific (Pty) Ltd</td>
</tr>
<tr>
<td>QIAquick PCR Purification Kit</td>
<td>QIAGEN</td>
</tr>
<tr>
<td>Nextera XT DNA Sample Preparation Kit</td>
<td>Illumina</td>
</tr>
<tr>
<td>Nextera PCR Master Mix (NPM)</td>
<td>Illumina</td>
</tr>
<tr>
<td>AMPure XP beads</td>
<td>Illumina</td>
</tr>
<tr>
<td>Resuspension buffer (RSB)</td>
<td>Illumina</td>
</tr>
</tbody>
</table>

A brief summary of all the equipment that was used during the course of this study is listed in Table 6.

Table 6: Equipment used to perform sample analysis

<table>
<thead>
<tr>
<th>Piece of Equipment</th>
<th>Supplying Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>QIAcube nucleic acid isolation system</td>
<td>QIAGEN</td>
</tr>
<tr>
<td>GeneAmp PCR System 9700 thermal cycler</td>
<td>Applied BioSystems</td>
</tr>
<tr>
<td>Qubit fluorometer</td>
<td>Qubit</td>
</tr>
<tr>
<td>Illumina Miseq Sequencer</td>
<td>Illumina</td>
</tr>
</tbody>
</table>

The various software applications, and/or, online analytical tool that were used during the phylogenetic analysis of the sequence data are listed in Table 7.
Table 7: Software programs and online analytical tools that were used in the analysis of sequence information

<table>
<thead>
<tr>
<th>Software package</th>
<th>Reference and/or licensed companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClustalW</td>
<td>Thompson et al&lt;sup&gt;227&lt;/sup&gt;</td>
</tr>
<tr>
<td>FigTree ver 1.3.1</td>
<td>Rambaut (<a href="http://tree.bio.ed.ac.uk">http://tree.bio.ed.ac.uk</a>)</td>
</tr>
<tr>
<td>FastTree v2.1.4</td>
<td>Price et al&lt;sup&gt;228&lt;/sup&gt;</td>
</tr>
<tr>
<td>eTree</td>
<td>Archer et al&lt;sup&gt;229&lt;/sup&gt;</td>
</tr>
<tr>
<td>RAxML ver. 8.2.9</td>
<td>Stamatakis&lt;sup&gt;131&lt;/sup&gt;</td>
</tr>
<tr>
<td>PhyloPart ver. 2.1</td>
<td>Ragonnet-Cronin et al&lt;sup&gt;230&lt;/sup&gt;</td>
</tr>
<tr>
<td>PhyloPart StandAlone PST07</td>
<td>Chevenet et al&lt;sup&gt;231&lt;/sup&gt;</td>
</tr>
<tr>
<td>Geneious ver 8</td>
<td>Biomatters Ltd</td>
</tr>
<tr>
<td>Anaconda Python ver 2.7</td>
<td>Python&lt;sup&gt;232&lt;/sup&gt;</td>
</tr>
<tr>
<td>Se-al ver. 2.0</td>
<td>Rambaut&lt;sup&gt;233&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

2.3. HIV-1 full-genome sequences

2.3.1. Full-genome sequences from the LANL HIV Database

A set of 2567 HIV-1 full-genome sequences was retrieved from the LANL HIV Database (www.hiv.lanl.gov/). The set of 2567 HIV-1 full-genome sequences included 7 HIV-1 subtypes (A to D, F to H), sub-subtypes and circulating recombinant forms.

2.3.2. Full-genome HIV-1 sequences from a South African cohort

To supplement the data retrieved from the LANL HIV database, 314 full-genome sequences were sequenced from blood samples obtained from residents of Hlabisa, a predominantly rural sub-district within the uMkhanyakhude District of northern KwaZulu-Natal (Figure 12), with a population of 228 000. This area formed a cohort that included HIV-infected people enrolled at 17 primary health care facilities served by a single district hospital. The local district hospital (Hlabisa) has 296 beds. Six of the Department of Health clinics, and 40% of patients, fall within the Africa Centre Demographic Surveillance Area (DSA), which has a population of 85 000 people in a 438-km<sup>2</sup> area. The DSA population is well characterized<sup>234</sup>. Information on these individuals is collected within the Africa Centre Demographic Information System (ACDIS) and a random 12.5% are tracked each year for the HIV surveillance, allowing a more complete understanding of the determinants of HIV infection<sup>235</sup>.
**Figure 12:** Hlabisa sub-district, northern KwaZulu-Natal, showing position of primary hospital with one on-site clinic and 15 peripheral clinics. Location of the Africa Centre is also shown. The Hlabisa sub-district extends to the area at the bottom-right of the map to encompass Zwelisha and Mtubatuba clinics.

2.3.2. Sample collection, transport and processing

At all the 17 primary health clinics, blood specimens were collected from study participants using 5ml EDTA tubes. On the same day of collection, the blood samples were transported from the Africa Centre head office in Mtubatuba to the Africa Centre Laboratory (ACL) in Durban (200km away); samples were kept on ice during transportation. Samples were received at the ACL and subsequently recorded in the laboratory information management system. The blood plasma was isolated, aliquoted and stored at -80°C.

2.3.2.2. RNA extraction

HIV-1 RNA was extracted using a QIAamp Viral RNA Mini kit (Qiagen NV, Venlo, Netherlands). The protocol as per the manufacturer’s guidelines was modified to extract RNA from 200μl of plasma instead of from 140μl of plasma; this was done to provide adequate viral RNA needed for full genome sequencing. The stored plasma was retrieved, thawed and centrifuged at a speed of 13 000 x g for 1 hour to concentrate the viral RNA. 200μl of plasma
was added to 800µl of lysis buffer, and was incubated for 10 minutes. 800µl of absolute ethanol was added to precipitate the RNA. The mixture was then loaded into the spin column, and centrifuged at 13 000 x g for 1 minute, to allow the extracted viral RNA to bind to the silica membrane. This step was followed by two washes with Wash Buffer 1 and Wash Buffer 2 with centrifugation in between the washes. After the second wash, a second dry spin was done to remove any excess of the washing buffer. Finally, 60µl of elution buffer was added to the column and RNA was eluted into a clean 1.5ml collecting tube. Eluted viral RNA was then stored at -80°C.

2.3.2.3. Amplicon generation

The extracted RNA was used for PCR amplification. We adapted the published protocol from Gall and colleagues\textsuperscript{238} for amplification, sequencing and assembly of full-length HIV-1 genomes; the original primers pairs (Pan1, Pan2, Pan3 and Pan4) were redesigned to be HIV-1 subtype C specific (Table 8) and were reviewed using Quick Align tool\textsuperscript{239}. Four overlapping amplicons, covering 9.7kb of HIV-1 genome, were generated (1.9 kbp, 3.6 kbp, 3 kbp, and 3.5 kbp respectively) as shown in Figure 13. This was achieved by using SuperScriptIII One-Step RT-PCR system with Platinum Taq DNA High Fidelity polymerase (Invitrogen). For all samples, the reactions of the PCR assays contained 4.5µl of water, 12.5µl of 2x Reaction Buffer, 1µl of each primer at 10µM, 1µl of SuperScriptIII One-Step RT/ Platinum Taq High fidelity mix and 5µl of the RNA template. The following cycling conditions were used (Table 9): one cycle of denaturation at 94°C for 4 minutes, followed by 40 cycles of denaturing at 94°C for 15 seconds, primer annealing at 60°C for 30 seconds, primer extension at 68°C for 4 minutes 30 seconds, one final step of primer extension at 68°C for 10 minutes, after which samples were cooled down and stored at 4°C.

Table 8: Summary of the primers for NGS amplification\textsuperscript{238}

<table>
<thead>
<tr>
<th>Set and primer</th>
<th>Sequence (5'-3')</th>
<th>Position (nucleotide)</th>
<th>Product size (base pair)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pan-HIV-1_1F</td>
<td>AGC CYG GGA GCT CTG TG</td>
<td>26-42</td>
<td>1928</td>
</tr>
<tr>
<td>Pan-HIV-1_1R</td>
<td>CCT CCA ATT CCY ATC ATT TT</td>
<td>1953-1931</td>
<td></td>
</tr>
<tr>
<td>Pan-HIV-1_2F</td>
<td>CGG AAG TGA YAT AGC WGG AAC</td>
<td>1031-1051</td>
<td>3574</td>
</tr>
<tr>
<td>Pan-HIV-1_2R</td>
<td>CTG CCA TCT GTT TTC CAT ARTC</td>
<td>4604-4583</td>
<td></td>
</tr>
<tr>
<td>Pan-HIV-1_3F</td>
<td>TTA AAA GAA AGG GGG GGA TTG GG</td>
<td>4329-4351</td>
<td>3066</td>
</tr>
<tr>
<td>Pan-HIV-1_3R</td>
<td>TGG CYT GTA CCG TCA GCG</td>
<td>7394-7377</td>
<td></td>
</tr>
<tr>
<td>Pan-HIV-1_4F</td>
<td>CCT ARG GCA GGA AGA AGC G</td>
<td>5513-5531</td>
<td>3551</td>
</tr>
<tr>
<td>Pan-HIV-1_4R</td>
<td>CTT WTA TGC AGC WTC TGA GGG</td>
<td>9063-9043</td>
<td></td>
</tr>
</tbody>
</table>
**Figure 13:** An illustration of the location and amplicon size of each pan primer. The overlapped primers span almost the full HIV genome from the 5’ to the 3’ LTR.

For samples that failed amplification using One-Step RT-PCR, the RT step was redone with specific reverse primers in the SuperScript III First Strand Synthesis kit to generate cDNA. Following that, the PlatinumTaq High Fidelity DNA Polymerase kit was used for amplification. Briefly, the reaction used 5µl of RNA in a 12.5µl reaction volume and resulted in a final primer concentration of 1.6µM. The cycling conditions for the reverse transcription step were: 65°C for 5 minutes, a hold for 4 minutes at 4°C during the addition of the SuperScript III first strand for the synthesis of cDNA, followed by 30 cycles of 95°C for 30 seconds, 58°C for 20 seconds, 72°C for 2 minutes, and a final extension at 72°C for 10 minutes. Amplicons were generated from 2.5µl of the cDNA in a total reaction volume of 25µl. The cycling conditions for the generation of amplicons were almost identical to those shown in Table 9: one cycle of denaturation at 94°C for 2 minutes, followed by 40 cycles of denaturing at 94°C for 15 seconds, primer annealing at 60°C for 30 seconds, primer extension at 68°C for 4 minutes 30 seconds, one final step of primer extension at 68°C for 10 minutes, after which samples were cooled down and stored at 4°C. Additionally, all PCR assays were run with a positive HIV-1 control sample that amplified well under the same conditions.

**Table 9: Master Mix for DNA amplification and PCR cycling conditions**

<table>
<thead>
<tr>
<th>Reagents</th>
<th>Volume/sample</th>
<th>Vol in MM (µl)</th>
<th>Final Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>4.5</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>2x Reaction Buffer</td>
<td>12.5</td>
<td>200</td>
<td>1x</td>
</tr>
<tr>
<td>Primer mix (10µM each)</td>
<td>2.0</td>
<td>32</td>
<td>0.4µM</td>
</tr>
<tr>
<td>SSIII/Platinum Taq polymerase (5U/l)</td>
<td>1.0</td>
<td>16</td>
<td>5U</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>320</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volume/sample</td>
<td>20.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA Sample</td>
<td>5.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Cycling Conditions**

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Duration</th>
<th>Number of cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>94</td>
<td>2 minutes</td>
<td>1</td>
</tr>
<tr>
<td>94</td>
<td>15 seconds</td>
<td>40</td>
</tr>
<tr>
<td>60</td>
<td>30 seconds</td>
<td>40</td>
</tr>
<tr>
<td>68</td>
<td>4 minutes 30 seconds</td>
<td>40</td>
</tr>
<tr>
<td>68</td>
<td>10 minutes</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>∞</td>
<td>Hold</td>
</tr>
</tbody>
</table>
2.3.2.4. Gel visualization

PCR products were separated on 0.8% ethidium bromide stained agarose gels (10 cm long) for visualization of the PCR DNA products (Figure 14). The gel was prepared by adding 0.5g of agarose tablet to 50ml of TBE buffer, which was followed by heating the mixture to boiling point. The gel was poured into a gel-casting tray containing a comb and was set for 20 to 30 minutes. 1µl of non-mutagenic fluorescent reagent Novel Juice 6x was mixed with 4µl of PCR product. Mixes were then loaded into agarose gel and run at 100 Volts for 60 minutes, the PCR DNA products were then assessed. The DNA bands were confirmed from Pan 1 to Pan 4.

Figure 14: Gel visualization of the PCR DNA products. Gel visualization of the PCR DNA products. Each amplicon runs at different lengths, for example 0.70 - 1 = Pan1, 2 = Pan2, 3 = Pan3 and 4 = Pan4. The scale of the ladder is shown on the left.

2.3.2.5. PCR purification

All successful generated amplicons were cleaned up using a QIAQuick PCR Purification kit (Qiagen)\(^{240}\). This began by adding 100µl of binding buffer to 20µl of each sample PCR product. The mixture was then transferred to the silica-column and centrifuged for 1 minute at 13 000 x g. 650µl of wash buffer PE was added to wash and purify the PCR product and the column was centrifuged for 1 minute at 13 000 x g. A dry centrifugation was then performed to remove possible residual of wash buffer PE. Finally the purified PCR product was eluted in 50µl of elution buffer after 1 minute of centrifugation at 13 000 x g.
2.3.2.6. DNA quantification

The DNA quantification was done using Qubit fluorometer\textsuperscript{241}. This device defines concentrations of particles via fluorescent dyes that only bind to a particular kind of substance such as DNA, RNA or proteins. These dyes have a very low fluorescence before they bind to their target DNA, however, they fluoresce strongly after binding. Therefore, the known DNA standards are used to convert the fluorescence signal into DNA concentration. Samples were diluted to 2ng/µl and pooled in a 1:3:3:3 ratio of Pan-1 to Pan-4 respectively for a final volume of 10µl; the pooling of PCR products was done into equimolar amounts to create the library.

2.3.2.7. Next generation DNA Sequencing

The pooled library was then prepared and sequenced according to the Miseq system user guide instructions. Briefly, 96 samples were prepared using Nextera XT DNA Sample Preparation Kit; this included 2 controls (negative and intra control) and 1 sample repeat. A 96 well Nextera XT Tagment plate was prepared and 10µl Tagment DNA Buffer (TD) buffer was added to each well. This was followed by the addition of 5µl of 2ng/µl input DNA and 5µl of Amplicon Tagment Mix (ATM) to each well using a multichannel pipette. This allows DNA input to be tagmented with adapter sequences added to the end. 15µl of Nextera PCR Master Mix (NPM) was then added to each well. This was followed by the addition of 5µl of index 1(i7) and 5µl of index 2 (i5) primers to the tagmented DNA; the purpose of this is cluster formation. This was then amplified through a limited PCR program on a thermal cycler; Table 10 shows the PCR amplification cycling conditions.

<table>
<thead>
<tr>
<th>Cycling Conditions</th>
<th>Temperature</th>
<th>Duration</th>
<th>Number of cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>72</td>
<td>3 minutes</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>30 seconds</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>10 seconds</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>30 seconds</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>30 seconds</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>5 minutes</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>$\infty$</td>
<td>Hold</td>
</tr>
</tbody>
</table>

PCR amplification was followed by purification of the library DNA via the addition of 30µl AMPure XP beads to each well. The beads were then washed via the addition of 200µl of 80\%
ethanol to each well. This provides a size selection step to remove very short fragments from the library. 52.5µl of Resuspension Buffer (RSB) was then added to each well. We then transferred 50µl of the supernatant to a clean 90 well plate. 45µl of the combined Library Normalization Additives 1/ Library Normalization Beads1 (LNA1/LNB1) was added to each well containing the DNA libraries. This was followed by adding 2 x 45µl of LNW1 to each well to wash the beads. 30µl of 0.1 N NaOH and 30µl of Library Normalization Storage Buffer 1 (LNS1) were added after elution. Following that, 30µl of supernatant was transferred to each well. Lastly, the pooled library was diluted in hybridization buffer, then loaded into a thawed MiSeq reagent cartridge and placed into Illumina MiSeq machine for sequencing.

2.3.2.8. Assembly and consensus generation

FastQ files generated by sequencing were imported into Geneious software version 8.0 (Biomatters Ltd, Auckland, New Zealand), an integrated and extendable software platform for the organization and analysis of genomic and sequence data. Primer sequences were removed, and quality control (removing reads of <200 bp and trimming low-quality bases from the 3’ end of the reads until the median quality of the read was <30) was performed. The remaining reads were trimmed (10bp from 5’ and 30bp from 3’ ends), and then assembled; this helps to decrease the probability of ambiguous read mapping, which occurs when shorter reads of lower accuracy are included in assemblies. Full-genome contigs were generated by assembly of the trimmed reads against a subtype C reference sequence from SA (Accession no AY228557), to derive a consensus sequence.

2.4. Multiple sequence alignment

A data set of 2881 full-genome HIV-1 sequences was obtained (314 sequenced from blood samples combined with the 2567 sequences retrieved from the LANL HIV Database). Multiple alignments of the 2881 sequences, along with 7 HXB2 sub-genomic reference sequences, were constructed in ClustalW (http://www.clustal.org/clustalw2/). In order to increase the speed of the alignment, a quick tree was employed to guide the alignments. Alignment was exported into Se-Al (http://www.tree.bio.ed.ac.uk/software/seal/) and was manually aligned. Gaps were excluded from the alignment if the gaps were not present in more than 20% of the taxa in each of the alignments. Sequences were manually aligned until a perfect codon alignment was achieved.
2.5. Analyzed sub-genomic regions of the HIV-1 genome

The extent of HIV clustering using near full genome sequences was compared with the outcomes of HIV clustering using sub-genomic sequences. To achieve this, sub-genomic regions were extracted from the multiple sequence alignment using the Geneious software platform\textsuperscript{242}. These included regions spanning the three structural HIV-1 genes, \textit{gag}, \textit{pol}, and \textit{env}, and four alternative sub-genomic regions that have been used or proposed for HIV cluster analysis, bringing the total to seven sub-genomic regions. The four alternative sub-genomic regions included (1) a partial \textit{pol} sequence spanning the region encoding HIV-1 protease and the first 335 amino acids of reverse transcriptase, which corresponds to HXB2 nucleotide (nt) positions 2,253 - 3,554\textsuperscript{245-248}, (2) partial \textit{env} sequences spanning the region encoding the gp120 V1C5 region\textsuperscript{249-251}, nt positions 6,570 - 7,757; (3) “product 2” spanning the 3’ -end of \textit{gag} and almost the entire \textit{pol}\textsuperscript{252}, nt positions 1,486 - 5,058; and (4) “product 4” spanning vpu, \textit{env}, nef, and TATA-box in the U3 region of 3’ -LTR, nt positions 5,967- 9,517.

2.6. Sliding window analysis

Sliding window analysis is a commonly used method for studying the properties of molecular sequences\textsuperscript{253}. To estimate the extent of clustering across the HIV-1 genome, a sliding window analysis with windows advancing incrementally across the multiple sequence alignment (a window of a certain length slid along the sequence alignment) was employed. The data sets of multiple sequence alignments of different window lengths were generated using Python programming language, version 2.7\textsuperscript{254}. Figure 15 depicts the script that was specifically formulated for this analysis. Three sizes of sliding windows were used, 1000-bp, 2000-bp and 3000-bp. Sliding steps were equal to 1/10 of the window size; 100 bp for the 1000-bp window, 200 bp for the 2000-bp window and 300 bp for the 3000-bp window; and produced multiple sets of overlapping multiple sequence alignments. A total of 99 alignment sets of 1000 bp each, 45 alignment sets of 2000 bp each and 27 alignment sets of 3000 bp were generated, resulting in 171 sets.
2.7. Phylogenetic inference

The alignment sets (171 of various window lengths, 7 of HIV-1 sub-genomic regions and the full genome alignment) were used to infer phylogenies with the use of the Maximum likelihood (ML) method as implemented in FastTree v2. 1. 4228. For each data set, phylogenies were inferred under the generalized time-reversal (GTR) model, an estimated gamma shape parameter, and the subtree pruning and regrafting (SPR) method of tree rearrangement. For the ML tree topologies, a total of 100 bootstrap replicates were performed for each data set. Inferred tree topologies were visually inspected in FigTree v. 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

The maximum likelihood tree inference was also implemented in RAxML131 under the GAMMA model of rate heterogeneity for the full genome and 7 sub-genomic alignment sets. The statistical support for each node was assessed by bootstrap analysis from 1000 bootstrap replicates performed with the rapid bootstrap algorithm implemented in RAxML. The RAxML runs were performed using RAxML v. 8.2.9. The average time of each of the runs varied depending on the sequence length of each data set. The smallest data sets took on average
around 36 hours, while the largest took up to 10 days. Each of the runs were executed on a Mac OS X 10.10.5 (Yosemite) with a 2.5 GHz Intel Core i5 processor.

2.8. Assessment of the accuracy of the inferred phylogenies

2.8.1. Estimation of tree certainty

Tree certainty quantifies the degree of conflict or incongruence in a set of phylogenetic trees\textsuperscript{219}. The quantification of incongruence is based on Shannon’s entropy\textsuperscript{220}. The internode certainty was measured by quantifying the degree of certainty for each individual internode by considering the two most prevalent conflicting bipartitions and calculating the log magnitude of their difference. An internode certainty close to 1 indicates high certainty of the targeted tree node and a lack of conflict in the data, while values of internode certainty close to 0 show a high degree of incongruence. Tree certainty quantifies the degree of conflict for the whole tree, and is the sum of internode certainty over all internodes in a phylogeny. Tree certainty scores were calculated in RAxML ver. 8.2.9 as described by Salichos and colleagues\textsuperscript{255}. Extended majority-rule consensus trees were computed using bootstrapped trees generated by RAxML for the full genome alignment and for each sub-genomic alignment set.

2.8.2. Estimation of Shimodaira–Hasegawa [SH]-aLRT

Shimodaira-Hasegawa (SH)-like support value computation was implemented in RAxML ver. 8.2.9\textsuperscript{255} as described by Guindon and colleagues\textsuperscript{256}. SH-aLRT is derived from the SH multiple tree comparison procedure\textsuperscript{224} and is fast due to the RELL technique based on the resampling of estimated log likelihoods\textsuperscript{225}. The input for each run was the best-known ML phylogeny found by RAxML analysis. Prior to the application of the SH-like test, the requirement is that each tree has to be NNI (Nearest Neighbour Interchange) optimal. Thus, for the full genome and 7 sub-genomic phylogenies, RAxML initially applied NNI moves to further improve the trees and then computed the SH test for all the inner branches of the trees.

2.8.3. Evaluation of cluster quality

Cluster quality was evaluated in CTree, via an executable jar file from: http://www.manchester.ac.uk/bioinformatics/ctree. CTree was designed for the quantification of clusters within viral phylogenetic tree topologies. Clusters are stored as individual data structures from which statistical data, such as the Subtype Diversity Ratio (SDR), Subtype Diversity Variance (SDV) and pairwise distances can be extracted. The SDR, is defined as the ratio of the mean intra cluster pairwise distance to the mean inter cluster pairwise distance\textsuperscript{257}. 
Low intra pairwise distances relative to inter pairwise distances imply the presence of more defined clusters. The SDV is a measure of the variation within the ratio of the mean intra-cluster pairwise distance to the mean inter-cluster pairwise distance calculated for each cluster on the tree. The lower the SDV the more symmetrical the clusters present. Together these two statistics quantify the presence of clustering within tree topologies. The full genome and sub-genomic phylogenies were uploaded as tree strings in Newick format into CTree, clusters were manually populated and then the tree metrics SDR and SDV were calculated.

2.9. Cluster enumeration

2.9.1. Identification of phylogenetic clusters using PhyloPart

The HIV-1 sequences in clusters in the 7 sub-genomic phylogenies and in the full genome phylogeny were enumerated with PhyloPart ver. 2.1. We defined the HIV cluster as a viral lineage that gives rise to a monophyletic sub-tree of the overall phylogeny with strong statistical support. We employed bootstrap re-sampling with 100 replicates to construct a consensus phylogenetic tree. We use the bootstrapped maximum likelihood method to determine the statistical support of clusters. We evaluated for transmission clusters among sequences that clustered around common proximal nodes with ≥ 90% bootstrap support. Of clusters containing > 2 sequences that met this criterion, we identified final transmission clusters as those in which each sequence had at least one neighbor within a patristic distance of ≤ 4.5% substitutions per site as measured via the length between branch tips on the originally generated phylogenetic tree. Clusters were identified using a depth-first algorithm, a method for traversing or searching tree or graph data structures starting from the root. This approach allowed us to avoid double counting of viral sequences and clusters in any cases in which clusters had internal structure with strong support.

2.9.2. Identification of phylogenetic clusters using PhyloType Stand Alone PST07

The HIV-1 full genome phylogeny, all 7 sub-genomic phylogenies and 171 sliding window length phylogenies were examined with the use of the PhyloType stand-alone application. The PhyloType application is a published application, 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. PhyloType is a tool that inspects phylogenies and combines them with extrinsic traits (e.g. geographic location, risk group, presence of a given resistance mutation), seeking to extract strain groups of specific interest or requiring surveillance. The primary annotations in this data set were the strain subtypes; these were grouped into the
countries in which the studied sequences were collected. An illustration of a PhyloType file annotation is presented in Figure 16. 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 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. The criteria chosen for the PhyloType analysis were Size ≥ 5, Persistence ≥ 1, Size/Different ≥ 1 and Support ≥ 0.7. The ACCTRAN parsimony method was selected and 1000 shuffles were performed to test phylotype significance. Only those phylotypes whose P-value for Size is ≤ 1% were retained. These options, selection criteria and thresholds correspond to PhyloType’s default parameter settings.

<table>
<thead>
<tr>
<th>SequenceID, Subtype, Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>'01_AE.AF.GQ477441.2007', '01_AE', 'AF'</td>
</tr>
<tr>
<td>'01_AE.CF.AF197340.1990', '01_AE', 'CF'</td>
</tr>
<tr>
<td>'01_AE.CF.AF197341.1990', '01_AE', 'CF'</td>
</tr>
<tr>
<td>'01_AE.CF.U51188.1990', '01_AE', 'CF'</td>
</tr>
<tr>
<td>'01_AE.CM.KP718939.2011', '01_AE', 'CM'</td>
</tr>
<tr>
<td>'01_AE.CN.AY008714.1997', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.AY008718.1997', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.DQ859178.2005', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.DQ859179.2005', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.DQ859180.2006', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.EF038527.2005', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.EF038528.2005', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.EF038529.2005', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.EF038530.2005', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.EF038531.2006', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.EF038532.2006', '01_AE', 'CN'</td>
</tr>
<tr>
<td>'01_AE.CN.EF038533.2006', '01_AE', 'CN'</td>
</tr>
</tbody>
</table>

Figure 16: An illustration of a Phylotype file annotation. This particular file annotation contains the strain (sequence ID) of each isolate, the subtype and the country code. This is the query file annotation that was uploaded into PhyloType to search the tree topology. This particular file annotation was generated in TextWrangler v 4.0.1.
CHAPTER THREE: RESULTS

The following chapter contains the results of the study, the results are organized in two parts: (1) Results on the outcomes of the accuracy of phylogenies inferred from each genomic region. This part (section 3.1) will address the first aim of the study, that is, to determine if the usage of full-genome sequence data as opposed to sub-genomic data for phylogenetic tree construction results in a more accurate phylogenetic tree, measurable by tree certainty; subtype diversity ratio; subtype diversity variance; and Shimodaira-Hasegawa-like support values. (2) Results on the outcomes of clustering in phylogenies inferred from different sub-genomic regions and various window lengths. This part (section 3.2) will address the second aim of the study, that is, to elucidate the optimal window size (sequence length) and genomic region to accurately identify transmission clusters. The 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 Accuracy of inferred phylogenies

3.1.1. Hierarchy of tree certainty

The degree of conflict or incongruence in the inferred trees was quantified by measuring tree certainty\(^{255}\). Table 11 displays the characteristics and hierarchy of tree certainty levels of the sub-genomic regions in descending order and comparative tree certainty is graphically presented in Figure 17. The tree based on full genome sequences showed the highest tree certainty. With a shorter sequence length than pol, the env tree had the highest certainty amongst the three structural HIV-1 genes, while the gag-based tree had the lowest certainty. The amount of variation that we find in env (length = 2750 nt) would be equivalent to an approximately 5 Kb-long gag-pol sequence. This could explain that why env outperforms pol (length = 3012 nt). The partial env tree showed relatively low certainty at levels comparable with the gag tree certainty. The partial pol tree certainty was the lowest amongst all sub-genomic regions. The region that had the closest tree certainty levels relative to the full genome was the product 4 region.
Table 11: Characteristics and hierarchy of tree certainty amongst genomic regions

<table>
<thead>
<tr>
<th>Hierarchy of target region</th>
<th>Coding region</th>
<th>Length (Bp)</th>
<th>Nucleotide positions (ref HXB2)</th>
<th>Tree Certainty</th>
<th>Relative Tree Certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Full genome</td>
<td>Entire HIV-1 genome</td>
<td>Entire Length</td>
<td>Entire length</td>
<td>921.7</td>
<td>32.02</td>
</tr>
<tr>
<td>3. Env</td>
<td>External glycoproteins gp120 and gp41</td>
<td>2750</td>
<td>6045 - 8795</td>
<td>638.39</td>
<td>22.18</td>
</tr>
<tr>
<td>4. Product 2</td>
<td>3’-end of <em>gag</em> and almost the entire <em>pol</em></td>
<td>3573</td>
<td>1486 - 5058</td>
<td>614.16</td>
<td>21.34</td>
</tr>
<tr>
<td>5. Pol</td>
<td>Reverse transcriptase, protease and integrase</td>
<td>3012</td>
<td>2085 - 5096</td>
<td>583.02</td>
<td>20.26</td>
</tr>
<tr>
<td>6. Partial env</td>
<td>Gp120 V1C5 region</td>
<td>1188</td>
<td>6570 - 7757</td>
<td>510.27</td>
<td>17.73</td>
</tr>
<tr>
<td>7. Gag</td>
<td>Codes for internal virion proteins</td>
<td>1502</td>
<td>790 - 2292</td>
<td>506.37</td>
<td>17.59</td>
</tr>
<tr>
<td>8. Partial pol</td>
<td>Protease and the first 335 amino acids of reverse transcriptase</td>
<td>1302</td>
<td>2253 - 3554</td>
<td>386.94</td>
<td>13.44</td>
</tr>
</tbody>
</table>

Figure 17: Graph comparing relative tree certainty amongst phylogenies inferred from different HIV-1 gene regions. Internode certainty was quantified by considering the two most prevalent conflicting bipartitions and calculating the log magnitude of their difference. Tree certainty was quantified as the sum of the internode certainty over all internodes in a phylogeny. TC = Tree Certainty, FG = Full genome, P-Env = Partial Env, P-Pol = Partial Pol, Pro-2 = Product 2, Pro-4 = Product 4.
3.1.2. Hierarchy of quality of clustering

Table 12 displays the characteristics and hierarchy of subtype diversity ratio (SDR) and subtype diversity variance (SDV) scores of the genomic regions in descending order of SDR scores. SDR and SDV scores are also graphically presented in Figure 18 and Figure 19 respectively. The SDR is a quantitative measure of the extent of clustering found within a tree, trees with lower SDR values are characterized by well defined clusters. With regards to SDV, the lower the value the more symmetrical, or equidistant, the clusters in a tree are relative to each other.

The tree based on the partial env sequences had the best (lowest) SDR score, and the env tree had the best SDV score. Again, the product 4 region phylogeny had a SDR score comparable to the full genome phylogeny.

Table 12: Characteristics and hierarchy of SDR and SDV amongst genomic regions

<table>
<thead>
<tr>
<th>Hierarchy of target region</th>
<th>Coding region</th>
<th>Length (Bp)</th>
<th>Nucleotide positions (ref HXB2)</th>
<th>Subtype diversity ratio score</th>
<th>Subtype diversity variance score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Partial env</td>
<td>Gp120 V1C5 region</td>
<td>1188</td>
<td>6570 - 7757</td>
<td>0.7626</td>
<td>0.0272</td>
</tr>
<tr>
<td>2. Env</td>
<td>External glycoproteins gp120 and gp41</td>
<td>2750</td>
<td>6045 - 8795</td>
<td>0.7821</td>
<td>0.0063</td>
</tr>
<tr>
<td>3. Product 4</td>
<td>Spanning vpu, env, nef, and TATA-box in the U3 region of 3′ -LTR</td>
<td>3551</td>
<td>5967 - 9517</td>
<td>0.8143</td>
<td>0.0087</td>
</tr>
<tr>
<td>4. Full genome</td>
<td>Entire HIV-1 genome</td>
<td></td>
<td>Entire length</td>
<td>0.8303</td>
<td>0.0727</td>
</tr>
<tr>
<td>5. Gag</td>
<td>Codes for internal virion proteins</td>
<td>1502</td>
<td>790 - 2292</td>
<td>0.8508</td>
<td>0.0150</td>
</tr>
<tr>
<td>6. Pol</td>
<td>Reverse transcriptase, protease and integrase</td>
<td>3012</td>
<td>2085 - 5096</td>
<td>0.9009</td>
<td>0.0110</td>
</tr>
<tr>
<td>7. Product 2</td>
<td>3′ -end of gag and almost the entire pol</td>
<td>3573</td>
<td>1486 - 5058</td>
<td>0.9043</td>
<td>0.0167</td>
</tr>
<tr>
<td>8. Partial pol</td>
<td>Protease and the first 335 amino acids of reverse transcriptase</td>
<td>1302</td>
<td>2253 - 3554</td>
<td>0.9117</td>
<td>0.0888</td>
</tr>
</tbody>
</table>
3.1.2.1. Quantified by subtype diversity ratio

**Figure 18:** Subtype diversity ratio (SDR) score for each HIV-1 targeted genomic region phylogeny. The calculation of SDR scores was implemented in cTree\textsuperscript{29}. Axis-y shows the magnitude of the SDR scores, Axis-x shows targeted regions across the HIV-1 genome. FG = Full genome, P-Env = Partial Env, P-Pol = Partial Pol, Pro-2 = Product 2, Pro-4 = Product 4.

3.1.2.2. Quantified by subtype diversity variance

**Figure 19:** Subtype diversity variance (SDV) score for each HIV-1 targeted genomic region phylogeny. Subtype diversity variance (SDV) score for each HIV-1 targeted genomic region phylogeny. The calculation of SDV scores was implemented in cTree\textsuperscript{29}. Axis-y shows the magnitude of the SDV scores, Axis-x shows targeted regions across the HIV-1 genome. FG = Full genome, P-Env = Partial Env, P-Pol = Partial Pol, Pro-2 = Product 2, Pro-4 = Product 4.
3.1.3. Hierarchy of Shimodaira-Hasegawa (SH) – like support

Table 13 displays the characteristics and hierarchy of SH-like support values of the sub-genomic regions in descending order and comparative SH-like support is graphically presented in Figure 20. Superiority of SH-like support values is graded in descending order, that is, the lower the value, the better the SH-like support. The tree based on full genome sequences had the best SH-like support value. As with the tree certainty outcomes, with a shorter sequence length than pol, the env tree had the best SH-like value amongst the three structural HIV-1 genes, while the gag-based tree had the worst score. Mirroring the hierarchy of tree certainty results, the region that had the closest tree SH-like support levels relative to the full genome was the product 4 region.

Table 13: Characteristics and hierarchy of SH support amongst genomic regions

<table>
<thead>
<tr>
<th>Hierarchy of target region</th>
<th>Coding region</th>
<th>Length (Bp)</th>
<th>Nucleotide positions (ref HXB2)</th>
<th>SH-like support value (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Full genome</td>
<td>Entire HIV-1 genome</td>
<td>Entire length</td>
<td>Entire length</td>
<td>-4539102.4215</td>
</tr>
<tr>
<td>2. Product 4</td>
<td>Spanning vpu, env, nef, and TATA-box in the U3 region of 3’ – LTR</td>
<td>3551</td>
<td>5967 - 9517</td>
<td>-2486652.0891</td>
</tr>
<tr>
<td>3. Env</td>
<td>External glycoproteins gp120 and gp41</td>
<td>2750</td>
<td>6045 - 8795</td>
<td>-2050291.3771</td>
</tr>
<tr>
<td>4. Product 2</td>
<td>3’-end of gag and almost the entire pol</td>
<td>3573</td>
<td>1486 - 5058</td>
<td>-1228841.2981</td>
</tr>
<tr>
<td>5. Partial env</td>
<td>Gp120 V1C5 region</td>
<td>1188</td>
<td>6570 - 7757</td>
<td>-1144029.9217</td>
</tr>
<tr>
<td>6. Pol</td>
<td>Reverse transcriptase, protease and integrase</td>
<td>3012</td>
<td>2085 - 5096</td>
<td>-1020499.4236</td>
</tr>
<tr>
<td>7. Gag</td>
<td>Codes for internal virion proteins</td>
<td>1502</td>
<td>790 - 2292</td>
<td>-884837.7213</td>
</tr>
<tr>
<td>8. Partial pol</td>
<td>Protease and the first 335 amino acids of reverse transcriptase</td>
<td>1302</td>
<td>2253 - 3554</td>
<td>-416874.9752</td>
</tr>
</tbody>
</table>
Figure 20: Graph comparing SH-like support values amongst phylogenies inferred from different HIV-1 genomic regions. The calculation of SH-like support scores was implemented in RAxML. Axis-y shows the magnitude of the SH-like support scores, Axis-x shows targeted regions across the HIV-1 genome. FG = Full genome, P-Env = Partial Env, P-Pol = Partial Pol, Pro-2 = Product 2, Pro-4 = Product 4.

3.2. Extent of HIV clustering across the HIV-1 genome

3.2.1. Clusters enumerated by PhyloPart in each genomic region phylogeny

We addressed whether the extent of HIV clustering is associated with any particular HIV-1 gene or gene sub-region, this analysis was implemented in PhyloPart. This method is based on a depth-first search (an algorithm for traversing a tree where one starts at the root and explores as far as possible along each branch before backtracking) and conjugates the evaluation of node reliability, tree topology and patristic distance analysis. The proportion of clustered sequences was compared between full-genome HIV-1 sequences and sub-genomic regions. Three structural HIV-1 genes, gag, pol, and env, and four regions commonly used in HIV cluster analysis (partial pol, partial env, product 2 and product 4), were targeted. All sets of sequences included the same 2881 HIV-1 sequences. Clusters were enumerated at the bootstrap threshold of 0.9 for cluster definition, with a within-cluster genetic distance of 4.5% substitutions per site, under maximum likelihood inference.

As shown in Figure 21, the highest number of clusters enumerated was observed for the product 2 phylogeny (402); a region associated with pol. Among the three structural HIV-1 genes, the more conserved sub-genomic regions had the higher number of clusters enumerated as well as more sequences in clusters (Figure 22). The highest number of clusters enumerated and the highest proportion of HIV-1 sequences in clusters was found in pol (401 and 1377 respectively) followed by gag (399 and 1345 respectively) and then env (226 and 645
respectively). The full genome phylogeny had the 5th highest number of clusters and sequences in clusters as enumerated by PhyloPart analysis, a method that factors in patristic distance in its analysis.

**Figure 21:** Number of clusters in the targeted regions of the HIV-1 genome. The number of HIV-1 sequences in clusters was estimated in PhyloPart\textsuperscript{20}. Axis-y shows the number of HIV-1 sequences in clusters, Axis-x shows targeted regions across the HIV-1 genome. FG = Full genome, P-\textit{Env} = Partial Env, P-\textit{Pol} = Partial Pol, Pro-2 = Product 2, Pro-4 = Product 4.

**Figure 22:** The extent of HIV clustering in each targeted HIV-1 genomic region. The proportion of HIV-1 sequences in clusters was estimated in PhyloPart\textsuperscript{20}. Axis-y shows the proportion of HIV-1 sequences in clusters, Axis-x shows targeted regions across the HIV-1 genome. FG = Full genome, P-\textit{Env} = Partial Env, P-\textit{Pol} = Partial Pol, Pro-2 = Product 2, Pro-4 = Product 4.
3.2.2. Clusters enumerated by PhyloType in each genomic region phylogeny

Results are provided in Figure 23 and Figure 24 which show the number of phylotypes enumerated in each genomic phylogeny and the percentage of strains associated with phylotypes per genomic phylogeny respectively. The tree based on full genome sequences had the highest number of phylotypes enumerated (69), as well as the highest percent of strains associated with phylotypes (71%). This means that 71% of the 2881 sequences were in clusters. Overall, the profile of phylotype cluster enumeration data resembled the hierarchy of tree certainty and SH-like support values with the tree based on product 4 region sequences having the second highest percent of strains associated with phylotypes (70.3%), slightly lower than that of the full genome phylogeny. Among the three structural HIV-1 genes, the highest proportion of HIV sequences in clusters was found in pol, followed by gag and then env. Pol also had the highest number of phylotypes enumerated, env had the second highest and gag had the lowest. Thus a brief contrast of env and gag is that clustering outcomes that result from the usage of the env phylogeny are that of a relatively greater number of phylotypes with a smaller number of strains per phylotype.

![Figure 23](image)

**Figure 23**: Number of phylotypes in the targeted regions of the HIV-1 genome. The number of HIV-1 sequences in clusters was estimated in PhyloType\textsuperscript{211}. Axis-y shows the number of HIV-1 phylotypes, Axis-x shows targeted regions across the HIV-1 genome. FG = Full genome, P-Env = Partial Env, P-Pol = Partial Pol, Pro-2 = Product 2, Pro-4 = Product 4.
Figure 24: Percentage of strains associated with phylotypes in the targeted regions of the HIV-1 genome. The percentage of strains was estimated in PhyloType\textsuperscript{231}. Axis-\textit{y} shows the proportion of HIV-1 strains in phylotypes, Axis-\textit{x} shows targeted regions across the HIV-1 genome. FG = Full genome, P-\textit{Env} = Partial Env, P-Pol = Partial Pol, Pro-2 = Product 2, Pro-4 = Product 4.

Figure 25 – Figure 32 show the phylotype maps that indicate the succession of founder and migratory events reconstructed from each genomic phylogeny. According to the full genome phylotype map (Figure 26), the virus spread from 7 South African phylotypes and 1 phylotype from Botswana. The Botswanan phylotype spread directly into India, and 1 large South African phylotype (429:ZA) spread into Brazil, Botswana and Cameroon. The phylotype maps reconstructed from the sub-genomic regions show significantly different migratory events and inferior phylotype identification. For example, the \textit{env} and \textit{gag} phylotype maps show that only 3 and 2 founder South African phylotypes were enumerated respectively. The subsequent sequence of migratory events is also varied when contrasting the different phylotype maps, as well as the number and identity of phylotypes of indirect origin (coloured in red). The \textit{env} and product 4 phylotype maps enumerated the least number of phylotypes of indirect origin and the partial \textit{pol} phylotype map had the most.
**Figure 25:** A subset of the *env* phylotype map (ACCTRAN) of global HIV-1. The map summarizes the information contained in the *env* phylogenetic tree; circle surface is proportional to the size value (number of members) of the phylotype. ZA = South Africa, BW = Botswana, CM = Cameroon, TH = Thailand, IR = Iran, RU = Russia, CN = China, UZ = Uzbekistan, AF = Afghanistan, CY = Cyprus, IN = India, JP = Japan, BR = Brazil, US = United States of America. Some of the phylotypes (coloured in red) have indirect origin; for example, 1815:BR and 2646:UZ.
Figure 26: A subset of the full genome phylotypes map (ACCTRAN) of global HIV-1. The map summarizes the information contained in the full genome phylogenetic tree; circle surface is proportional to the size value (number of members) of the phylotype. ZA = South Africa, BW = Botswana, CM = Cameroon, TH = Thailand, IR = Iran, RU = Russia, CN = China, UZ = Uzbekistan, AF = Afghanistan, UG = Uganda, IN = India, JP = Japan, BR = Brazil, US = United States of America, ES = Spain, NG = Niger, CU = Cuba, GB = United Kingdom, PE = Peru, KR = South Korea, LU = Luxembourg, FR = France. Some of the phylotypes (coloured in red) have indirect origin; for example, 4624:BR and 2438:US.
**Figure 27:** A subset of the *gag* phylotype map (ACCTRAN) of global HIV-1. The map summarizes the information contained in the *gag* phylogenetic tree; circle surface is proportional to the size value (number of members) of the phylotype. ZA = South Africa, BW = Botswana, CM = Cameroon, IR = Iran, RU = Russia, UZ = Uzbekistan, AF = Afghanistan, CY = Cyprus, IN = India, JP = Japan, BR = Brazil, NG = Niger, TZ = Tanzania, UG = Uganda, LU = Luxembourg, UA = Ukraine, KR = South Korea. Some of the phylotypes (coloured in red) have indirect origin; for example, 158:IN and 800:CM.
Figure 28: A subset of the product 2 phylotype map (ACCTRAN) of global HIV-1. The map summarizes the information contained in the product 2 phylogenetic tree; circle surface is proportional to the size value (number of members) of the phylotype. ZA = South Africa, BW = Botswana, CM = Cameroon, TH = Thailand, CN = China, IN = India, BR = Brazil, US = United States of America, VN = Vietnam, SE = Sweden, UG = Uganda. Some of the phylotypes (coloured in red) have indirect origin; for example, 21:ZA and 1582:US.
**Figure 29:** A subset of the product 4 phylotype map (ACCTRAN) of global HIV-1. The map summarizes the information contained in the product 4 phylogenetic tree; circle surface is proportional to the size value (number of members) of the phylotype. ZA = South Africa, BW = Botswana, CM = Cameroon, IR = Iran, RU = Russia, CN = China, IN = India, BR = Brazil, US = United States of America, NG = Niger, KE = Kenya, CD = Congo, The Democratic Republic of, MY = Malaysia, ES = Spain, UA = Ukraine, RW = Rwanda, GB = United Kingdom, CY = Cyprus. Some of the phylotypes (coloured in red) have indirect origin; for example, 184:ZA and 2093:BR.
Figure 30: A subset of the partial pol phylotype map (ACCTRAN) of global HIV-1. The map summarizes the information contained in the partial pol phylogenetic tree; circle surface is proportional to the size value (number of members) of the phylotype. ZA = South Africa, BW = Botswana, CM = Cameroon, TH = Thailand, SE = Sweden, RU = Russia, CN = China, VN = Vietnam, JP = Japan, BR = Brazil, US = United States of America, UG = Uganda. Some of the phylotypes (coloured in red) have indirect origin; for example, 287:BR and 4077:ZA.
Figure 31: A subset of the partial env phyotype map (ACCTRAN) of global HIV-1. The map summarizes the information contained in the partial env phylogenetic tree; circle surface is proportional to the size value (number of members) of the phyotype. ZA = South Africa, CM = Cameroon, TH = Thailand, IR = Iran, RU = Russia, CN = China, UZ = Uzbekistan, CY = Cyprus, BR = Brazil, US = United States of America, UG = Uganda, MY = Malaysia, GB = United Kingdom, VN = Vietnam, NG = Niger, UA = Ukraine. Some of the phyotypes (coloured in red) have indirect origin; for example, 3635:ZA and 2499:UA.
Figure 32: A subset of the pol phytype map (ACCTRAN) of global HIV-1. The map summarizes the information contained in the pol phylogenetic tree; circle surface is proportional to the size value (number of members) of the phytype. ZA = South Africa, BW = Botswana, CM = Cameroon, RU = Russia, CN = China, AF = Afghanistan, IN = India, BR = Brazil, US = United States of America, SE = Sweden, UG = Uganda, KE = Kenya, NG = Niger, ES = Spain, GH = Ghana, UA = Ukraine, CD = Congo, The Democratic Republic of. Some of the phylotypes (coloured in red) have indirect origin; for example, 4120:BR and 340:SE.
3.2.3. Clusters enumerated by PhyloType in each sliding window length phylogeny

To assess the extent of HIV clustering across the HIV-1 genome, sliding window analysis was performed with window sizes of 1000-bp; 2000-bp and 3000-bp, and sliding steps of 100 bp; 200 bp; and 300 bp respectively. This analysis allowed us to investigate how patterns of HIV clustering change across the HIV-1 genome.

**Figure 33:** Sliding window analysis across the HIV-1 genome depicting number of phylotypes. (A) HIV-1 genome structure. The map is depicted as a reference to the gene structures associated with the various nucleotide positions across the HIV-1 genome. (B) The extent of HIV-1 clustering for each sliding window length.
The profile of HIV clustering across the HIV-1 genome was “wave shaped” (Figure 33 and Figure 34) suggesting a differential contribution of regions across the HIV genome to clustering. The highest extent of HIV clustering was associated with the region encoding env. HIV-1 gag and the region from 4100nt to 4200nt (p15 RNase) showed the lowest extent of HIV clustering. The size of the sliding window has a moderate effect on the extent of HIV clustering. Longer viral sequences with window size 3000-bp were associated with slightly higher extents of HIV clustering than sequences with window sizes of 2000-bp and 1000-bp across the entire HIV genome (Figure 33). The ups and downs in the profiles of HIV clustering were similar between longer and shorter HIV windows. A deeper look into sliding windows across the viral genome reveals substantial heterogeneity in HIV clustering based on the sub-genomic region and sampling. Analysis of potential reasons for such a differential clustering across the viral genome, such as searching for specific signatures associated with clustering, warrants dedicated future studies and should be taken in the context of sampling.

**Figure 34**: Sliding window analysis across the HIV-1 genome depicting extent of clustering. The percentage of strains associated with phylotypes for each sliding window length is shown in the y-axis and HIV-1 nucleotide positions are shown in the x-axis.
CHAPTER FOUR: DISCUSSION

The dynamics of HIV-1 transmission networks can be investigated through comprehensive HIV cluster analysis. HIV-1 transmission cluster analysis can provide insights into the dynamics of HIV-1 spread, and the results of HIV-1 cluster analysis can help inform public health prevention interventions, such as an optimal balance of Treatment-as-Prevention and Pre-Exposure Prophylaxis strategies. The higher the extent of HIV clustering, the more informative HIV cluster analysis could be. In phylogenetics, there are a wide range of factors that may influence the results or outcomes of any investigation, and the inference of HIV-1 transmission clustering analysis 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: the effect of viral genetic diversity between strains or subtypes, the effect of viral recombination, the number of the taxa and fragment size of the data set, the specific model parameters that was used, and the effect of mutation rates.

In this study we investigated whether the extent of HIV clustering is associated with the length of targeted HIV sequences, or with a particular sub-genomic region across the HIV-1 genome. The extent of HIV clustering was compared between the full genome and sub-genomic regions, and also among different sliding window lengths. The accuracy of each inferred phylogeny (reconstructed from full genome, sub-genomic and various sliding-window length sequences) was also probed through the computation of various clade confidence metrics.

**Tree accuracy and clustering outcomes**

Although previous phylogenetic studies have shown that large input taxa is positively associated with inferred tree accuracy\(^{258}\), we saw great differences between the tree accuracy outcomes of the full genome and sub-genomic phylogenies in spite of the considerable number of input sequences. In phylogenetic studies, the total size of the number of taxa included in transmission cluster analysis has been a concern as too many isolates would unnecessarily slow down analysis, while too few isolates would leave out too much of the genetic information that is needed to infer epidemic histories\(^{259}\). In an attempt to exclude data size as a confounding factor in inferred tree accuracy and clustering outcomes, we utilized a considerable size of input taxa.

In our analysis, the phylogeny based on the full genome sequences showed the best tree accuracy; it ranked highest with regards to both tree certainty and SH-like support. Product 4, a
region spanning vpu, \textit{env}, nef, and TATA-box in the U3 region of 3'–LTR, had the best tree accuracy among the sub-genomic regions. Among the HIV-1 structural genes, \textit{env} had the best tree certainty, SH-like support, SDR score and the best SDV score overall. The full genome HIV-1 sequences were associated with the highest extent of HIV clustering in the phylotype analysis. Among HIV-1 structural genes, \textit{pol} showed the highest extent of clustering, followed by \textit{env}. Combined with the extent of HIV-1 clustering, the tree accuracy estimates provide additional evidence that full genome HIV-1 sequences are the most informative choice for HIV cluster analysis. \textit{Env} appears to be the best choice among the structural genes, as it fared best in the tree accuracy metrics and product 4 (a region that is associated with \textit{env}), exhibited the second highest slightly tree accuracy and clustering outcomes. These results mirror those by Yebra et al\textsuperscript{217} who found that there was increased reliability of phylogeny reconstruction in simulated data when using \textit{env} trees. The \textit{env} portion of the HIV-1 genome is highly variable and is under more selective pressure when compared to the \textit{gag} or the \textit{pol} regions of the HIV-1. The fact that \textit{env} trees can outperform the \textit{pol} trees, suggests that, in principle, the higher evolutionary rate in \textit{env} can improve reconstruction.

\textbf{The influence of sequence length on the outcomes of HIV-1 phylogenetic cluster analysis}

The total size of the nucleic fragments, as with any phylogenetic investigation, has been elucidated to play an important role in HIV-1 phylogenetic cluster analysis; however, this role hasn’t been adequately quantified. 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-1 carries enough genetic information for reasonable phylogenetic inference\textsuperscript{260}.

To explore the optimal sequence length for phylogenetic cluster analysis, sliding window analysis was performed with window sizes of 1000-bp; 2000-bp and 3000-bp, and sliding steps of 100 bp; 200 bp; and 300 bp respectively. The sequence size, or length, used in HIV cluster analysis appeared to have a moderate effect on the extent of HIV clustering. This was evident from the comparison of HIV clustering between three sliding windows, 1 000 bp, 2000 bp and 3000 bp long, which were run across the entire HIV-1 genome with 100 bp, 200 bp and 300bp steps, respectively. The sliding window analysis also allowed us to identify regions across the HIV-1 genome with higher propensities for HIV clustering. Despite fluctuations across the HIV-1 genome, the extent of HIV clustering was moderately higher for larger sliding windows spanning similar regions in the HIV-1 genome.
With a sample size of 401, research conducted by Novitsky et al.\textsuperscript{58} elucidated dramatically higher HIV clustering for larger sliding windows; the influence of sequence length on viral clustering in this study may have been mitigated by the large number of taxa. Another factor that may have contributed to dissimilar results is the exclusive utilization of HIV-1C sequences in the previous study, whereas the dataset in this study was heterogeneous in HIV-1 subtypes. Correlation between sequence length and tree accuracy has also been probed before in a simulation study\textsuperscript{217}. The results showed that the proportion of correct trees increased in almost direct proportion to the length of the sequences used. Thus a consolidation of previous and current findings is that viral sequence length has a positive association with both tree accuracy and extent of clustering, but the magnitude of the influence is negatively affected by the taxa size.

**Limitations and conclusion**

The 2881 sequences used in this study included recombinants and varied in subtype classification. This may be a limitation as the specifics and nature of subtype recombination could either complicate or assist in the analysis of HIV clustering. For example, the analysis could be complicated due to incorrect estimation of evolutionary rates and a skewed molecular clock\textsuperscript{261,262}. HIV-1 subtypes have a large effect on the analysis of HIV-1 transmission clusters. Since the zoonosis of HIV from non-human primates to humans a large degree of genetic variation has accumulated amongst HIV-1 isolates\textsuperscript{263}. These genetic variations have led to the rise of distinct HIV-1 strains or subtypes\textsuperscript{104}. The reconstruction of transmission 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\textsuperscript{264,265}. The inclusion of isolates from multiple subtypes of HIV-1 will therefore inherently have an effect on transmission cluster analysis\textsuperscript{266}, and this effect may have been a limitation to our analysis.

In summary, the results of this study provide evidence that the extent of HIV clustering is associated with the length of viral sequences used in cluster analysis. The use of longer genetic regions (such as concatenated \textit{gag}, \textit{pol} and \textit{env} or \textit{gag-pol}) will allow for a more reliable reconstruction of transmission events and better cluster enumeration. The traditional short \textit{pol} sequences generated for resistance testing that are used in most molecular epidemiology studies are substantially less reliable. Full genome sequences could be considered the top choice for the
most informative HIV cluster analysis. An alternative approach to HIV cluster analysis could be based on selected sub-genomic regions with an elevated extent of HIV clustering and high tree accuracy such as \textit{env}. An effort to generate highly sampled datasets is also needed to increase our ability to reconstruct real HIV epidemics.
CHAPTER FIVE: APPENDICES

Appendix 1: Consent form in IsiZulu

<table>
<thead>
<tr>
<th>HIV drug resistance study</th>
<th>Consent form</th>
</tr>
</thead>
</table>

Mina…………………………ngiyavuma ukuba umntwana wami abe yingxenye yoezwangiso lokuhlola ukungazweli kwemishangaza yesandulela ngeulazi. Sengilehazeliwe ngocwanningo ngaliquqondisiwa nephapha lolwazi.

Ngiyaiqonda imithelela yokungencela komntwana wami kulolu ewaningi nokuthi kunokwenza kuekela oluanye ulwazi mayelana nempilo kanye nokwelushwa kwakhe ngesikhathi socwaringo.

Ngiyathunzayo abasebenzi booezingo ukuba babheke efayelini kanye nasekhadini lakhe nokuthi ulwazi olutholakala kulolu ewaningi lungahlanganiswa noicolwazi oselukhona kwisicilindoloza lwazi sase-Africa Centre. Ngiyaiqonda nokuthi kazoathwa elinye isampula legazi kuluncwanningo.

Ngiyaiqonda ukuthi ngiyolihola ithuba lokubonisana ngemiphumela yomntwana wami nomhlengikazi noma nozokotela.

Ngiyaiqonda ukuthi umntwana wami angushiya noma nini oewaningweni futhi ngeke abanduluwe ngokwenzenjalo. Siyoqhubeka nokusebenzisa imitholampilo ye-ART futhi ngithole ukunakekelwa ngokuwayelelelile.

Isishicielelo sobambe izhaza
Usuku……./……./

Isishicielelo sikafakazi
Usuku……./……./
Appendix 2: Consent form in English

I agree to be part of the HIV drug resistance study. The study has been explained to me and I fully understand the information in the study information sheet.

I understand the implications of me / my child/ward joining the study and that I / my child/ward may be asked additional information regarding my / his/her health and my / his/her treatment during the study visit.

I give permission to the research staff to look at my / my child’s/ward’s clinic file and clinic card and that information from this study may be linked to information already held on the clinical and demographic databases at the Africa Centre. I understand that an extra blood sample will be taken as part of this study.

I understand that I will have the opportunity to discuss the results with a nurse or doctor.

I understand that I / my child/ward may leave the study at any time and I / he/she will not be discriminated for doing so. I will continue to use the ART clinic and be given appropriate care as usual.

Signature of the study participant: __________________________ date: __________/________/_________

Witness signature: _______________________________date: __________/________/_________
REFERENCES

6. UNAIDS. Global Statistics. 2015.


76. UNAIDS. HIV and AIDS estimates. 2016.


